Q

QSAR Modeling and QSAR Based Virtual Screening, Complexity and Challenges of Modern

Alexander Tropsha
School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, USA

Article Outline

Glossary

QSAR – quantitative structure activity relationships a method to predict biological activity from chemical structure

Combi-QSAR – combinatorial QSAR implies concurrent generation of QSAR models using all possible binary combinations of different descriptor types and model optimization techniques

QSAR modeling workflow a hierarchy of QSAR model development and validation protocols that should be followed to establish validated and externally predictive model

kNN – k nearest neighbors a pattern recognition approach used in deriving non-linear QSAR models

Model validation a set of computational routines used to establish internal and external predictive power of QSAR models

Applicability domain restriction on the chemistry space occupied by compounds for which the prediction of their activity from training set QSAR model is considered reliable.

Definition of the Subject

In the early days of Quantitative Structure Activity Relationship (QSAR) modeling the experimental datasets were relatively small and chemically congeneric and the techniques employed were relatively unsophisticated. Since then, the size and complexity of experimental datasets has increased dramatically, and so had the complexity and challenges of data analytical approaches. This chapter examines the strategy and the output of the modern QSAR modeling approaches especially as applied to complex biomolecular datasets. We discuss a data-analytical modeling workflow developed in our laboratory that incorporates modules for combinatorial QSAR model development (i. e., using all possible binary combinations of available descriptor sets and statistical data modeling techniques), rigorous model validation, and virtual screening of available chemical databases to identify novel biologically active compounds. Our approach places particular emphasis on model validation as well as on the need to define model applicability domains in the chemistry space. We present examples of studies where the application of rigorously validated QSAR models for virtual screening identified computational hits that were confirmed by subsequent experimental investigations. The emerging focus of QSAR modeling on target property forecasting brings it forward as predictive, as opposed to evaluative, modeling approach.

Introduction

QSAR methodology was introduced by Hansch et al. in early 1960th [1]. The original approach stemmed from linear free-energy relationships and is based upon the assumption that the difference in structural properties accounts for the difference in biological activities of compounds. According to this approach, the structural changes that affect the biological activities of a set of congeners are of three major types: electronic, steric, and hydrophobic [2]. These structural properties are often described by Hammett electronic constants [2], Verloop STERIMOL parameters [3], hydrophobic constants [2], etc. The relationship between a biological activity (or chemical property) and the structural parameters is obtained using linear or multiple linear regression (MLR) analysis. The fundamentals and applications of this method in chemistry and biology have been summarized by Hansch and Leo [2].

The original QSAR method was relatively straightforward; the datasets were small, and so was the number of relatively simple physical chemical descriptors used in modeling building. Today, even a person familiar with the basic principles of QSAR could be easily confused by the diversity of methodologies and naming conventions used in modern QSAR studies. Two-dimensional (2D) or three-dimensional (3D) QSAR, variable selection or Artificial Neural Network methods, Comparative Molecular Field Analysis (CoMFA) or binary QSAR present examples of various terms that may appear to describe totally independent approaches, which can not be generalized or even easily compared to each other. Thus, modern QSAR modeling is a very complex and complicated field requiring deep understanding and thorough practicing to develop robust models. Nevertheless, an attempt can be made to provide some unifying concepts that underlie practically any QSAR methodology.

Indeed, any QSAR method can be generally defined as an application of mathematical and statistical methods to the problem of finding empirical relationships (QSAR models) of the form $P_i = \hat{k}(D_1, D_2, \ldots, D_n)$, where P_i are biological activities (or other properties of interest) of molecules, $D_1, D_2, \ldots, D_n$ are calculated (or, sometimes, experimentally measured) structural properties (molecular descriptors) of compounds, and $\hat{k}$ is some empirically established mathematical transformation that should be applied to descriptors to calculate the property values for all molecules. The relationship between values of descriptors D and target properties P can be linear or non-linear. The example of the former relationship is given by multiple linear regression (MLR) common to the Hansch QSAR approach [1], where target property can be predicted directly from the descriptor values. On the contrary, nearest neighbor QSAR methods serve as examples of non-linear techniques where descriptor values are used in characterizing chemical similarities between molecules, which are then used to infer compound activity. The goal of QSAR modeling is to establish a trend in the descriptor values, which parallels the trend in biological activity. In essence, all QSAR approaches imply, directly or indirectly, a simple similarity principle, which for a long time has provided a foundation for the experimental medicinal chemistry: compounds with similar structures are expected to have similar biological activities.

The subsequent sections of this chapter present a brief overview of the modern QSAR modeling field without going into specific details of any particular technique; introduce the predictive QSAR modeling workflow developed in our group; present examples of successful applications of the workflow to several datasets resulting in experimentally confirmed computational predictions of biologically active compounds by the means of virtual screening; address the issue of fruitful collaborations between QSAR modelers in developing and supporting "best practices" in QSAR modeling; and summarize most important challenges that the field of QSAR modeling is facing today.

The Complexity of Modern Datasets and QSAR Methods

Traditionally, QSAR approaches have been applied to modeling datasets tested against a single target, e. g., in specific enzymatic or receptor-binding assays. Recent experimental advances in high-throughput screening and multi-target testing of compound libraries have led to the establishment of datasets of biologically active compounds (often publicly available) that we shall define as complex. A complex dataset could include a library of compounds tested against multiple targets, or have the target property measured in the form of gene or protein expression profiles across many genes (chemical genomics), or could be formed by diverse compounds tested against a complex assay where multiple mechanisms leading to the measured response could be involved (e. g., carcinogenicity or mutagenicity). The examples of complex datasets include Pubchem [4], GPCR ligands [5], NCI [6], US FDA [7], NIEHS [8], and EPA DSS-Tox [9] (see more examples in a recent review [10]). Naturally, the complex datasets call for the development of more sophisticated computational tools and corresponding models.

Modern QSAR approaches are characterized by the use of multiple descriptors of chemical structure com-

bined with the application of both linear and non-linear optimization approaches, and a strong emphasis on rigorous model validation to afford robust and predictive models. The most important recent developments in the field concur with a substantial increase in the size of experimental datasets available for the analysis and an increased application of QSAR models as virtual screening tools to discover biologically active molecules in chemical databases and/or virtual chemical libraries [11]. The latter focus differs substantially from the traditional emphasis on developing so called explanatory QSAR models characterized by high statistical significance but only as applied to training sets of molecules with known chemical structure and biological activity.

The differences in various QSAR methodologies can be understood in terms of the types of *target property* values, *descriptors*, and *optimization* algorithms used to relate descriptors to the target properties and generate statistically significant models. *Target* properties (regarded as dependent variables in statistical data modeling sense) can be generally of three types: *continuous* (i. e., real values covering certain range, e. g., IC_{50} values, or binding constants); *categorical related, or rank-based* (e. g., classes of rank-ordered target properties covering certain range of values, e. g., classes of metabolic stability such as unstable, moderately stable, stable); and *categorical unrelated* (i. e., classes of target properties that do not relate to each other in any continuum, e. g., compounds that belong to different pharmacological classes). As simple as it appears, understanding this classification is actually very important since the choice of descriptor types and modeling techniques as well as model accuracy metrics is often dictated by the type of the target properties. Thus, in general the latter two types require *classification* modeling approaches whereas the former type of the target properties allows the use of (multi)linear regression type modeling. The corresponding methods of data analysis are referred to as either classification or continuous property QSAR.

Many QSAR approaches have been developed during the past few decades (e. g., see recent reviews [12,13]). The major differences between various approaches are due to structural parameters (descriptors) used to characterize molecules and the mathematical approaches used to establish a correlation between descriptor values and biological activity. Most of the modeling techniques assume a linear relationship between molecular descriptors and a target property, which may be an adequate methodology for many datasets. However, the advances in combinatorial chemistry and high throughput screening technologies have resulted in the explosive growth of the amount of structural and biological data making the problem of developing robust QSAR models more challenging. This progress has provided an impetus for the development of fast, nonlinear QSAR methods that can capture structure-activity relationships for large and complex data. New nonlinear methods of multivariate analysis such as different types of Artificial Neural Networks [14,15,16,17], Generalized Linear Models [15,18,19,20], Classification and Regression Trees [18,21,22,23,24], Random Forests [25,26,27], MARS (Multivariate Adaptive Regression Splines) [27,28], Support Vector Machines [29,30,31,32], and some other methods have become routine tools in QSAR studies. Interesting examples of applications have been reported for all types of the above methods. In some cases the comparisons between different techniques as applied to the same dataset have been made but in general there appears to be no universal QSAR approach that produces the best QSAR models for any datasets.

For instance, several types of ANNs have been used in QSAR studies, including feed forward back propagation neural networks (FFBPNN), counter propagation neural networks (CPNN), radial basis function neural networks (RBFNN), Bayesian regularized neural networks (BRNN), etc. In [33], analysis of FFBPNN was carried out based on artificial and real data. It has been shown that for neural nets capable of generating linear models, increasing the number of nodes in the hidden layer can improve statistics for the training sets. However, leave-one-out and leave-some-out cross-validation statistics as well as prediction power for the test sets start decreasing when the number of hidden layer nodes passes some threshold that depends on the dataset. On the whole, according to [33], linear FFBPNNs compare unfavorably with more simple MLR methodology. No conclusions have been made for neural nets including quadratic and indicator variables as well as ANNs with other architectures.

In [34], probabilistic and generalized regression neural networks (PNN and GRNN, respectively), which are variants of RBFNN, have been used for classification and continuous QSAR modeling for a data set of soluble epoxide hydrolase inhibitors and prediction of aqueous solubility of some classes of small organic molecules. Final models appeared to have comparable predictive power to kNN classification, MLR and feed-forward neural networks, but they included significantly lower number of descriptors.

Overall, the advantage of ANNs is that they do not explicitly define the relationship between descriptors and target property making these techniques applicable to datasets of complex chemical nature and/or complex response variable where the relationship between descriptor values and the response variable are likely to be non-linear. On the other hand, for the same reason the interpretation

of ANN QSAR models is very difficult, if not impossible. ANNs can be used as part of the combi-QSAR strategy discussed below and may prove to be particularly useful (as with any other QSAR approach) for specific datasets.

Our laboratory among others has concentrated on the development of automated QSAR approaches based on variable selection and stochastic optimization. The examples of methods developed or implemented in our laboratory include k-Nearest Neighbors (kNN) [31,35,36], Simulated Annealing-Partial Least Squares (SA-PLS) [37], Support Vector Machines (SVM) [38,39,40], and the Automated Lazy Learning QSAR (ALL-QSAR) [41].

Recent Progress in Chemical Descriptor Research: 2D Chirality and Fragment Descriptors

In most cases the accurate formal description of molecules is the key to developing successful QSAR models. Many approaches to generating descriptors have been developed during the long history of QSAR modeling; programs such as Dragon [42] or MolConnZ [43] could calculate 100s or even 1000s of molecular descriptors. The molecular descriptors that are commonly used in QSAR studies can be divided into several groups. Physicochemical descriptors reflect electronegativity, partial charges, hydrogen bond acceptor and donor ability, molecular weight, logP, surface area, etc. Another class of descriptors are three-dimensional (3D) descriptors which are derived from spatial structures of molecules (e. g. CoMFA [44], CoMSIA [45], QSiAR [46] etc.). Most of QSAR studies based on 3D descriptors require exhaustive conformational analysis and spatial alignment of molecules, and thus large computational and human resources (e. g., [47]). Approaches such as GRIND [48] and VolSurf [49] generate descriptors based on 3D interaction maps and do not require alignment, but still require conformational search. On the other hand, descriptors based on molecular graphs (e. g. molecular connectivity indices [50], molecular shape indices [51], E-state indices [52], etc.) are naturally insensitive to problems related to conformation or alignment. Since chemical graphs are planar, these descriptors are often referred to as two-dimensional (2D) descriptors. An important note is that one should not be confused by the perceived low dimensionality of 2D descriptors: a molecule is actually described in multidimensional 2D descriptor space, with the number of dimensions (i. e., descriptors) frequently higher than that for 3D descriptors. Thus any modern "2D" or "3D" QSAR model actually operates on a complex high-dimensional space of multiple chemical descriptors. In this regard, the tasks and challenges of modern QSAR are much more similar to those faced by, e. g., bioinformaticians trying to analyse multidimensional data on protein or gene expression profiles of say patients vs. healthy individuals. 2D descriptors have certain advantages over 3D descriptors: they can be easily calculated even for very large datasets and in most case they seem to capture sufficient information about chemical structures to enable important modeling tasks such as similarity searching or QSAR modeling. For instance, it was shown [53] that 2D descriptors outperformed 3D descriptors in chemical similarity and diversity analyses. Our studies have shown that the predictive power of QSAR models based on 2D descriptors is generally comparable (and sometimes superior) to that of the models based on 3D descriptors [35,54]. Finally, recent comparison of protein structure based vs. ligand based virtual screening results [55] demonstrated that the latter approach helped recover a higher number of confirmed hits than the former one.

Overall, the field of chemical descriptor development has probably reached its saturation (as far as active development of new generalized descriptors is concerned) although there are continuing reports on novel 3D descriptors such as for instance inductive descriptors [56]. Still, there remain challenges in developing specialized descriptors such as topological chirality descriptors and fragment descriptors that have been of special interest to several groups including ours. We briefly review recent advances in such specialized descriptor development research.

Chirality Descriptors

Many biologically-active compounds are in fact enantiospecific, and their chiralities are believed to directly influence their bioactivities because these compounds are recognized differently by their corresponding receptors. In 1999, the worldwide annual sales of chiral drugs exceeded $100 billion that constituted almost one-third of all drug sales, and in 2000, these numbers were $133 billion and 40%, respectively. It was projected that the sales for chiral drugs in 2008 could reach $200 billion [57].

In QSAR studies, taking chirality into account has become possible only after the development of the 3D QSAR methods such as the comparative molecular field analysis (CoMFA) [44]. Subsequently, many CoMFA-like methods have appeared (e. g., [45,46,58]). Evidently, the chirality in such methods is taken into account by default, since 3D molecular field values of chiral isomers are different. At the same time as mentioned above 2D QSAR methods offer a clear advantage over 3D methods since they require no conformational analysis, no alignment, and no 3D pharmacophore hypothesis and, as a result, can be easily fully automated. However, one of the main drawbacks

of QSAR approaches utilizing 2D molecular descriptors such as molecular connectivity indices (calculated with the MolconnZ program [43]) or atom pairs [59] has been their inability to take into account the chirality of atoms since the latter is a true 3D property. Naturally, this deficiency has severely limited the range of applications of 2D QSAR methods, especially as compared to 3D QSAR approaches. We shall discuss two approaches that circumvent this limitation of 2D descriptors.

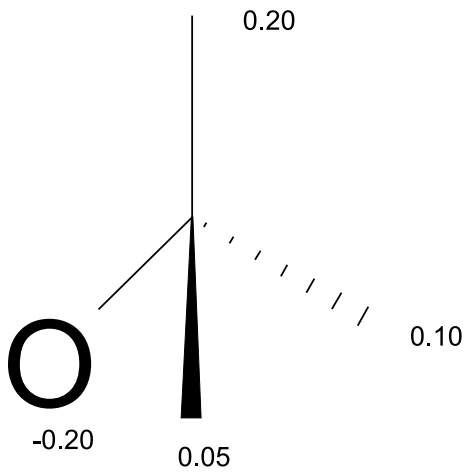
Topological Chirality Indices

The idea of modifying conventional 2D descriptors to make them sensitive to chirality was put forward as early as in 1995 [60]. The authors introduced a so-called chiral factor equal to + 1 or − 1 for an atom in R- or S-configuration, respectively. This factor was used to derive several chirality descriptors which appeared to have different values for linear hexose isomers. The first attempt to consider chirality descriptors in QSAR studies was reported by Ortiz et al. in 1998 [61] when several chirality-sensitive molecular and charge topological indices were introduced. Both chirality descriptors and conventional MolconnZ descriptors have been implemented in QSAR studies of the series of D_2 dopamine and σ receptor ligands. It was shown that the resulting QSAR models had better statistics and predictive power for IC_{50} values of the ligands than those obtained with conventional descriptors only. In addition, using these chirality descriptors, a set of chiral barbiturates were correctly classified as sedatives or stimulants [61].

Several series of chirality molecular topological descriptors have been developed in our laboratory earlier [62]. These descriptors have been developed on the basis of conventional 2D topological descriptors of molecular graphs. They include modified molecular connectivity indices, Zagreb group indices, extended connectivity, overall connectivity, and topological charge indices. These modified descriptors make use of a term called chirality correction, which is added to the vertex degrees of asymmetric atoms in a molecular graph. They have been used to build predictive QSAR models for several datasets [63]. Similar approach was used to develop cis-trans (ZE) isomerism descriptors as well [64].

Chirality-Sensitive Atom Pair Descriptors

Atom pair descriptors were introduced by Carhart et al. [59] and implemented in the GenAP program in our laboratory. The conventional atom pair descriptors can not be applied for a data set of chiral compounds since their numerical values are identical for enantiomeric molecules. Thus, we have recently introduced chiral atom types and thereby augmented the conventional AP descriptors with the novel chiral atom pair (cAP) descriptors [65].



QSAR Modeling and QSAR Based Virtual Screening, Complexity and Challenges of Modern, Figure 1
Example of the chiral atom type definition based on partial charges of the substituents (here, three carbon atoms and one oxygen atom) at the chiral center

A cAP descriptor is defined as a pair of atoms separated by certain chemical graph distance (for 2D representation of a molecule) or a real physical distance (in 3D) where at least one atom of the pair is chiral. Our definition of a chiral atom is similar to the conventional IUPAC nomenclature but different in how we define the seniority of the substituents at the chiral atom. The difference is that the relative priorities of the substituents are not determined by their atomic numbers, but values of a certain property, such as atomic partial charge, van der Waals volume, etc. In the previous studies [65], Gasteiger–Huckel charges as implemented in the SYBYL software were used. In addition, our current implementation does not take into consideration properties of any atoms besides the substituents. Figure 1 shows an example of a chiral atom where all four substituents at a carbon atom have different partial charges.

Since the oxygen atom has the most negative partial charge, we assign the lowest seniority to it, and since the charges on the substituent increase in a counterclockwise order, the central atom is in R_q configuration (index q means that the configuration was defined using partial charges). (Note that the same carbon atom would not be considered chiral based on the standard nomenclature!) Using different atomic properties to assign chirality, we can introduce many different definitions of R and S con-

figurations of chiral atoms; this work is in progress in our laboratory.

Fragment Descriptors

2D descriptors such as molecular connectivity indices have one significant drawback: in most cases models based on 2D descriptors are difficult (if not impossible) to interpret, because most of 2D descriptors have no clear physicochemical meaning. To overcome this drawback, molecular structural keys, molecular fingerprints and molecular holograms are used. Unlike molecular descriptors based on physicochemical properties, 2D and 3D descriptors, structural keys, molecular fingerprints and molecular holograms can provide a mechanistic explanation to the target property of active molecules.

Structural keys and molecular fingerprints were initially introduced by chemical information system companies for querying chemical databases. A structural key of a molecule is defined as a bit string; each bit in state "on" (one) or "off" (zero) represents the presence or absence of a certain atom or group (fragment) in a molecule [66,67]. Alternatively, it is a string of numbers; each number represents how many times an atom or a fragment is represented in a molecule. A predefined library of fragments is used. Usually bit strings are very long and for small molecules almost all bits are zeros.

To reduce the memory necessary to store structural keys, hashing [68] is used to map structural keys on a shorter string of a predefined length. For example, typical Daylight fingerprints have 512 or 1024 bits, but any power of two can be generated. The MACCS public fingerprints include 166 or 320 keys which encapsulate 966 original (private) bits [69]. Barnard Chemical Information Systems (BCI) fingerprints are generated combining both the Daylight and MDL approaches, and molecular fingerprint bit lengths are about 5,000.

Molecular holograms used in HQSAR are based on the structural keys containing the number of times each bit was set. This information is stored in the hologram string which is a string of integers rather than a bit string. As fingerprints, molecular holograms are based on a predefined library of fragments (chemical groups). The development of frequent common subgraph based molecular descriptors represent an important and interesting area of current descriptor research. For instance, recurring substructures in a group of chemicals with similar activity can be identified by finding frequent subgraphs in their related graphical representations. The recurring substructures can implicate chemical features responsible for compounds' biological activities [70].

Critical Importance of Model Validation

It should sound almost axiomatic that validation should be natural part of any model development process. Indeed, what is the (ultimate) purpose of any modeling approach such as QSAR, if not developing models with a significant external predictive power? Unfortunately, as we and others have indicated in many publications (e. g., [46,71,72]), the entire field of QSAR modeling counting nearly 50 years of a rich history has been plagued with insufficient attention paid to the subject of external validation. Indeed, most practitioners have merely presumed that internally cross-validated models built from available training set data should be externally predictive. However, the overwhelming prevalence of QSAR publications exploring small to medium size datasets to produce models with little statistical significance led to the recent editorial published by the *J. Chem. Info. Model.* two years ago [73] that explicitly discouraged researchers from submitting the "introspective" QSAR/QSPR publications and requested that "evidence that any reported QSAR/QSPR model has been properly validated using data not in the training set must be provided". We and others have demonstrated (as we detail below) that the training set statistics using most common internal validation techniques such as leave-one-out or even leave-many-out cross-validation approaches is insufficient and the statistical figures of merit of such models serve as misleading indicators of the external predictive power of QSAR models [72]. We shall refer to studies limited to exploring training sets (and associated model statistical parameters) as *"narcissistic modeling"*.

In our highly cited publication "Beware of q^2!" [71], we have demonstrated the insufficiency of the training set statistics for developing externally predictive QSAR models and formulated the main principles of model validation. At the time of that publication in 2002, the majority of papers on QSAR analysis ignored any model validation except for the cross-validation, performed during model development. Despite earlier observations of several authors [74,75,76] warning that high cross-validated correlation coefficient $R^2(q^2)$ is the necessary, but not the sufficient condition for the model to have high predictive power, many authors continued to consider q^2 as the only parameter characterizing the predictive power of QSAR models. In [71] we have shown that the predictive power of QSAR models can be claimed only if the model was successfully applied for prediction of the external test set compounds, which were not used in the model development. We have demonstrated that the majority of the models with high q^2 values have poor predictive power when applied for prediction of compounds in

the external test set. We believe that paying attention only to the training set statistics equates to "*narcissistic*" modeling in a sense that such models appear "beautiful" only in the eyes of their developers but provide little if any utility to potential users ("viewers") of these models. In another publication [72] the importance of rigorous validation was again emphasized as a crucial, integral component of model development. Several examples of published QSPR models with high fitted accuracy for the training sets, which failed rigorous validation tests, have been considered. We presented a set of simple guidelines for developing validated and predictive QSPR models and discussed several validation strategies such as the randomization of the response variable (*Y*-randomization) external validation using rational division of a dataset into training and test sets. We highlighted the need to establish the domain of model applicability in the chemical space to flag molecules for which predictions may be unreliable, and discussed some algorithms that can be used for this purpose. We advocated the broad use of these guidelines in the development of predictive QSPR models [72,77,78].

Nowadays, most of the QSAR modeling studies include validation of QSAR models. However, some authors still publish QSAR models which lack proper validation. For example, [79] developed a model to predict gastrointestinal absorption of drugs, but did not validate it using a test set. Verma et al. [80] developed QSAR models for predicting cytotoxicity of a group of compounds with antiovarian cancer activity which were validated only using cross-validation and *Y*-randomization.

At the 37th Joint Meeting of Chemicals Committee and Working Party on Chemicals, Pesticides & Biotechnology, held in Paris on 17–19 November 2004, the OECD (Organization for Economic Co-operation and Development) member countries adopted the following five principles that valid (Q)SAR models should follow to allow their use in regulatory assessment of chemical safety. (i) a defined endpoint; (ii) an unambiguous algorithm; (iii) a defined domain of applicability; (iv) appropriate measures of goodness-of-fit, robustness and predictivity; (v) a mechanistic interpretation, if possible. Since then, most of the European authors publishing in QSAR area include a statement that their models fully comply with OECD principles (e. g., see [81,82,83,84]). For instance, two aspects of QSAR modeling outlined in the OECD principles are considered by Estrada and Patlewicz [85]. The first aspect concerns the theoretical approaches used in chemistry in general, and in QSAR in particular, i. e., which method should be selected for theoretical studies: more sophisticated and complex or more simple. The authors criticize the common belief that applying more sophisticated methods should always lead to significantly better results. They considered an example of polycyclic aromatic hydrocarbons (PAHs) the toxicity of which is believed to depend on the energy gap between HOMO and LUMO values. The authors show that a simple Hückel Molecular Orbital theory gives practically the same values of HOMO and LUMO as the sophisticated ab initio methods yet the calculations are 10^{-4} to 10^{-7} times faster. They reach the conclusion that if a more simple method is capable of giving results better or similar to those of more sophisticated method, one should naturally use a more simple method!

The second aspect relates to so called "mechanistic" QSAR. Some authors (e. g., [86]) prefer descriptors which are mechanistically interpretable. On the other hand, Estrada and Patlewicz [85] argue that in many cases a biological response is a result of a multitude of different processes, some of which can be even not known, and its a posteriori mechanistic interpretation is difficult if not impossible. The authors suggest an alternative approach where a biological system is considered as a black box, when considering several possible mechanisms would be more productive. At the same time, some variables included in the model can describe several different mechanisms simultaneously, e. g. $\log P$, so in many cases it makes no sense to suggest that the use of this descriptor in QSAR models affords any mechanistic interpretation (see also [87]). We would add that descriptors which give better models in terms of their predictive power are actually preferable. We consider building predictive models as the main goal of QSAR analysis. Of course, interpretation of the model is also important, and if it is possible, it should be done. However, in many cases it is impossible, even when models with high predictive power have been obtained (e. g., best models were found to be those built using the molecular connectivity indices but these models were disregarded by the authors for the lack of mechanistic interpretability [86]). We believe that mechanistic interpretation of the *externally validated* QSAR model is an important a posteriori exercise that should be done after the model has been internally and externally validated, and descriptors that afford models with the highest predictive power should be always used preferentially.

Validation of QSAR models remains one of the most critical problems of QSAR. Recently, we have extended our requirements for the validation of multiple QSAR models selected by acceptable statistics criteria of prediction of the test set [88]. Additional studies in this critical component of QSAR modeling should establish reliable and commonly accepted "good practices" for model development.

Applicability Domains of QSAR Models

One of the most important problems in QSAR analysis is establishing the models' domain of applicability in the chemistry space. In the absence of the applicability domain, each model can formally predict the activity of any compound, even with a completely different structure from those included in the training set. Thus, the absence of the model applicability domain as a mandatory component of any QSAR model would lead to the unjustified extrapolation of the model in the chemistry space and, as a result, a high likelihood of inaccurate predictions. In our research we have always paid particular attention to this issue [36,41,62,72,89,90,91,92]. The need for establishing the applicability domain for every model adds another critical degree of complexity to the model building process.

The applicability domain problem has been addressed by many researchers. Mandel [93] introduced the so called Effective Prediction Domain which was based on the ranges of descriptors included in the regression equation. Afantitis et al. [94] built a multiple linear regression model for a dataset of apoptotic agents. They defined the applicability domain for each compound as a leverage defined as a corresponding diagonal element of the hat matrix. In fact, it is a method for detecting possible leverage outliers. If for some compound leverage is higher than $3K/N$, where K is the number of descriptors and N is the number of compounds, the compound is an outlier. To use this approach, for each external compound it would be necessary to recalculate the leverage. Netzeva et al. [95] and Saliner et al. [83] defined the applicability domain by ranges of descriptors, i. e., in fact, as a subspace occupied by representative points in the descriptor space. This definition of the applicability domain has a significant drawback, because the representative points could be found only in a small part of the hyper-parallelepiped corresponding to descriptor ranges rather than distributed uniformly. A similar definition of the applicability domain was proposed by Tong et al. [96]. The authors built QSAR models for two datasets of estrogen receptor ligands using the Decision Forest method and studied the dependence of the model predictive power vs. the applicability domain threshold. The prediction accuracy within the domain is defined as a ratio of the number of correct predictions to the total number of compounds in the domain. The accuracy was changing from about 90% for the initial applicability domain to about 50% when the applicability domain increased by 30%. Interestingly, for one of the datasets the prediction accuracy was increasing until the domain was extended by about 20%. Another important aspect of this study was that the authors defined the confidence level of prediction. The probability that a compound belongs to a certain class was defined as the percentage of active compounds in the leaf node that the compound belongs to. The authors found (as expected), that the confidence level correlated with the prediction accuracy.

In [97] a lazy learning kNN-like method was applied for the prediction of rodent carcinogenicity and Salmonella mutagenicity. The applicability domain was defined by a so-called confidence index. A compound was assigned to one of the two classes by a weighted majority vote of its nearest neighbors. The confidence index is the weighted majority quote divided by the number of nearest neighbors. If the absolute value of the confidence index is low (< 0.05) a compound is said to be out of the applicability domain. This definition of the applicability domain captures the areas in the descriptor space where compounds of both classes are close to each other, and possibly mixed. In this area the precise and accurate prediction of a compound's class is impossible. A Tanimoto-like coefficient is used as a similarity measure. Nearest neighbors are defined by the value of this coefficient higher than 0.3, which limits the possibility of over-extrapolation.

In most of our QSAR studies we have defined the applicability domain as the distance cutoff value $D_{\text{cutoff}} = \langle D \rangle + zs$, where Z is a similarity threshold parameter defined by a user, and $\langle D \rangle$ and s are the average and standard deviation of all Euclidian distances in the multidimensional descriptor space between each compound and its nearest neighbors for all compounds in the training set (e. g., see [78]). This definition of the applicability domain has several major drawbacks which we continue to address in our ongoing studies: (i) Currently, applicability domain is direction-independent in the descriptor space. We shall consider the directions in the descriptor space in which the distribution of representative points has smaller spread as less important than those that have higher spread. Thus, the applicability domain will be represented as a multidimensional ellipsoid in the principal component space. (ii) Too strict definition of the applicability domain: if a compound is outside of the model applicability domain, we currently do not predict its activity. Naturally, we shall establish the lower and upper bounds for the applicability domain. (iii) Finally, it seems reasonable to introduce a confidence level of prediction, which will depend on the distance of the compound under prediction from its nearest neighbor of the training set. These considerations provide just a few examples that illustrate the importance of ongoing research in this area of QSAR modeling. Not surprisingly, the model applicability domain was the subject of a special symposium organized at the most recent 235th

meeting of the American Chemical Society in New Orleans, LO.

Combinatorial QSAR and Model Acceptability Criteria

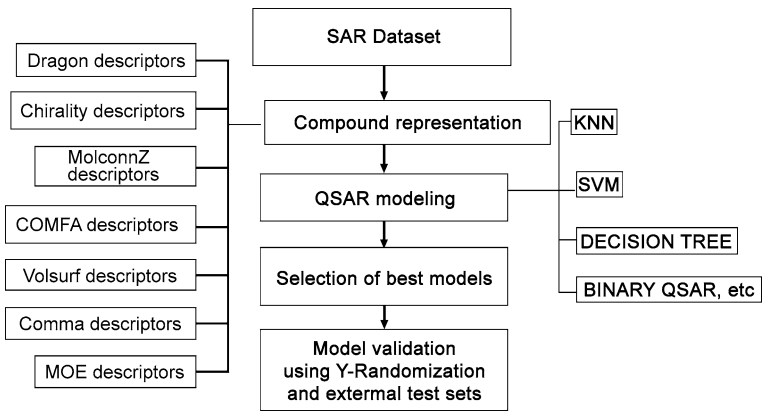
The chief hypothesis of the combi-QSAR approach that we introduced in recent publications [39,98] is that if an implicit structure-activity relationship exists for a given data set, it can be formally manifested via a variety of QSAR models obtained with different descriptors and optimization protocols. Our experience indicates that there is no universal QSAR method that is guaranteed to give the best results for any dataset. Thus we believe that multiple alternative QSAR models should be developed (as opposed to a single model using some favorite QSAR method) for each dataset to identify the most successful technique in the context of the given dataset. Since QSAR modeling is relatively fast, these alternative models could be explored simultaneously when making predictions for external data sets. The consensus predictions of biological activity for novel test set compounds on the basis of several QSAR models, especially when they converge, are more reliable and provide better justification for the experimental exploration of hits.

Our current approach to combi-QSAR modeling is summarized on the workflow diagram (Fig. 2). Our experience suggests that QSAR is a highly experimental area of statistical data modeling where it is impossible to decide a priori as to which particular QSAR modeling method will prove most successful. To achieve QSAR models of the highest internal, and most importantly, *external* accuracy, the combi-QSAR approach explores all possible binary combinations of various descriptor types and optimization methods along with external model validation. Each combination of descriptor sets and optimization techniques is likely to capture certain unique aspects of the structure-activity relationship. Since our ultimate goal is to use the resulting models as reliable activity (property) predictors, application of different combinations of modeling techniques and descriptor sets will increase our chances for success. All types of descriptors and modeling techniques are available within our laboratory and are described in detail in our recent publications on the implementation of the combi-QSAR strategy [39,98].

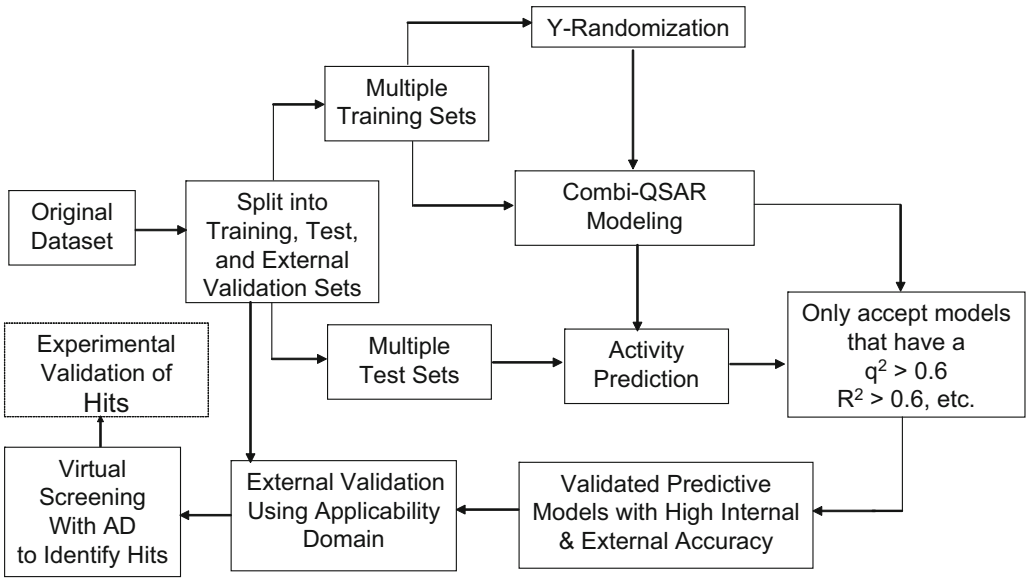
In our critical publications [71,72] we have recommended a set of statistical criteria which must be satisfied by a predictive model. For continuous QSAR, criteria that we will follow in developing activity/property predictors are as follows: (i) correlation coefficient R between the predicted and observed activities; (ii) coefficients of determination [99] (predicted versus observed activities R_0^2, and observed versus predicted activities $R_0'^2$ for regressions through the origin); (iii) slopes k and k' of regression lines through the origin. We consider a QSAR model *predictive*, if the following conditions are satisfied (i) $q^2 > 0.5$; (ii) $R^2 > 0.6$; (iii) $(R^2 - R_0^2)/R^2 < 0.1$ and $0.85 \le k \le 1.15$ or $(R^2 - R_0'^2)/R^2 < 0.1$ and $0.85 \le k' \le 1.15$; (iv) $\left|R_0^2 - R_0'^2\right| < 0.3$ where q^2 is the cross-validated correlation coefficient calculated for the training set, but all other criteria are calculated for the test set.

Predictive QSAR Modeling Workflow and its Application to Virtual Screening

Our experience in QSAR model development and validation has led us to establishing a complex strategy that is



QSAR Modeling and QSAR Based Virtual Screening, Complexity and Challenges of Modern, Figure 2
Flow-chart of the combinatorial QSAR methodology. All descriptor sets and methods currently implemented in our laboratory are listed



QSAR Modeling and QSAR Based Virtual Screening, Complexity and Challenges of Modern, Figure 3
Flowchart of predictive QSAR modeling framework based on the validated combi-QSAR models

summarized in Fig. 3. It describes the predictive QSAR modeling workflow focused on delivering validated models and ultimately, computational hits confirmed for the experimental validation. We start by randomly selecting a fraction of compounds (typically, 10–15%) as an external validation set. The remaining compounds are then divided rationally (using the Sphere Exclusion protocol implemented in our laboratory [78]) into multiple training and test sets that are used for model development and validation, respectively using criteria discussed in more detail below. We employ multiple QSAR techniques based on the combinatorial exploration of all possible pairs of descriptor sets coupled with various statistical data mining techniques (combi-QSAR) and select models characterized by high accuracy in predicting both training and test sets data. Validated models are finally tested using the evaluation set. The critical step of the external validation is the use of applicability domains. If external validation demonstrates the significant predictive power of the models we use all such models for virtual screening of available chemical databases (e. g., ZINC [100]) to identify putative active compounds and work with collaborators who could validate such hits experimentally. The entire approach is described in detail in several recent papers and reviews (e. g., [11,13,72]).

In our recent studies we were fortunate to recruit experimental collaborators who have validated computational hits identified through our modeling of anticonvulsants [36], HIV-1 reverse transcriptase inhibitors [101], D1 antagonists [31], antitumor compounds [102], and beta-lactamase inhibitors [103]. Thus, models resulting from this workflow could be used to prioritize the selection of chemicals for the experimental validation. However, since we can not generally guarantee that every prediction resulting from our modeling effort will be validated experimentally we can not include the experimental validation step as a mandatory part of the workflow on Fig. 3, which is why we used the dotted line for this component. We note that our approach shifts the emphasis on ensuring good (best) statistics for the model that fits known experimental data towards generating testable hypothesis about purported bioactive compounds. Thus, the output of the modeling has exactly same format as the input, i. e., chemical structures and (predicted) activities making model interpretation and utilization completely seamless for medicinal chemists.

The development of truly validated and predictive QSAR models affords their growing application in chemical data mining and combinatorial library design [37,104]. For example, three-dimensional (3D) stereoelectronic pharmacophore based on QSAR modeling was used recently to search the National Cancer Institute Repository of Small Molecules [6] to find new leads for inhibiting HIV type 1 reverse transcriptase at the non-nucleoside binding site [105].

Our studies have shown that QSAR models could be used successfully as virtual screening tools to discover compounds with the desired biological activity in chemical

databases or virtual libraries [11,31,36,102,103,106]. The discovery of novel bioactive chemical entities is the primary goal of computational drug discovery, and the development of validated and predictive QSAR models is critical to achieve this goal.

Computational Chemical Toxicology

Chemical toxicity can be associated with many hazardous biological effects such as gene damage, carcinogenicity, or induction of lethal rodent or human diseases. Toxicity presents an example of complex biological property (cf. the discussion of complex datasets above) where the underlying mechanisms are most frequently unknown and in fact multiple mechanisms are likely to be involved in mediating the end point response. For this reason, toxicity modeling is a very challenging QSAR problem yet alternative approaches such as protein structure based modeling can not be explored in most cases for the same reason of complex or unknown molecular mechanisms.

Although the experimental protocols for toxicity testing have been developed for many years and the cost of compound testing has reduced significantly, computational chemical toxicology continues to be a viable approach to reduce both the amount of efforts and the cost of experimental toxicity assessment [107]. Significant savings could be achieved if accurate predictions of potential toxicity could be used to prioritize compound selection for experimental testing. Many Quantitative Structure Activity Relationship (QSAR) studies have been conducted for different toxicity endpoints to address this challenge, [108,109,110,111]. However, the most critical limitation of many traditional QSAR studies has been their low *external* predictive power, i. e., their ability to predict accurately the underlying end point toxicity for compounds that were not used for model development. We discuss below the results of a recent important study of aquatic toxicity [112] since in our opinion this particular study may serve as a useful example to illustrate the complexity and power of modern QSAR modeling approaches.

The combinational QSAR modeling approach has been applied to a diverse series of organic compounds tested for aquatic toxicity in *Tetrahymena pyriformis* in the same laboratory over nearly a decade [113,114,115, 116,117,118,119]. The unique aspect of this research was that it was conducted in collaboration between six academic groups specializing in cheminformatics and computational toxicology. The common goals for our virtual collaboratory were to explore the relative strengths of various QSAR approaches in their ability to develop robust and externally predictive models of this particular toxicity end point. We have endeavored to develop the most statistically robust, validated, and *externally* predictive QSAR models of aquatic toxicity. The members of our collaboratory included scientists from the University of North Carolina at Chapel Hill in the United States (**UNC**); University of Louis Pasteur (**ULP**) in France; University of Insubria (**UI**) in Italy; University of Kalmar (**UK**) in Sweden; Virtual Computational Chemistry Laboratory (**VCCLAB**) in Germany; and the University of British Columbia (**UBC**) in Canada. Each group relied on its own QSAR modeling approaches to develop toxicity models using the same modeling set, and we agreed to evaluate the realistic model performance using the same external validation set(s) (cf. Table 1 for the summary of approaches).

The *T. pyriformis* toxicity dataset used in this study was compiled from several publications of the Schultz group [113,120,121,122,123] as well as from data available at the Tetratox database website of (http://www.vet.utk.edu/TETRATOX/). After deleting duplicates as well as several compounds with conflicting test results and correcting several chemical structures in the original data sources, our final dataset included 983 unique compounds (the structural information is included in the Appendix). The dataset was randomly divided into two parts: 1) the modeling set of 644 compounds; 2) the validation set including 339 compounds. The former set was used for model development by each participating group and the latter set was used to estimate the external prediction power of each model as a universal metric of model performance. In addition, when this project was already well underway, a new dataset had become available from the most recent publication by the Schultz group [124]. It provided us with an additional *external* set to evaluate the predictive power and reliability of all QSAR models. Among compounds reported in [125] 110 were unique, i. e., not present among the original set of 983 compounds; thus, these 110 compounds formed the second independent validation set for our study.

Universal Statistical Figures of Merit for All Models

Different groups have employed different techniques and (sometimes) different statistical parameters to evaluate the performance of models developed independently for the modeling set (described below). To harmonize the results of this study the same standard parameters were chosen to describe each model's performance as applied to the modeling and external test set predictions. Thus, we have employed Q^2_{abs} (squared leave-one-out cross-validation correlation coefficient) for the modeling set, R^2_{abs} (frequently described as coefficient of determination) for the external

QSAR Modeling and QSAR Based Virtual Screening, Complexity and Challenges of Modern, Table 1
Overview of QSAR modeling approaches employed by six cheminformatic groups involved in this study

Group ID	Modeling Techniques	Descriptor Type	Applicability Domain Definition
UNC	kNN, SVM	MolconnZ, Dragon	Euclidean distance threshold between a test compound and compounds in the modeling set
ULP	MLR, SVM, kNN	Fragments (ISIDA), Molecular (CODESSA-Pro)	Euclidean distance threshold between a compound and compounds in the modeling set; bounding box
UI	MLR/OLS	Dragon	Leverage approach
UK	PLS	Dragon	Residual standard deviation and leverage within the PLSR model
VCCLAB	ASNN	E-state indices	Maximal correlation coefficient of the test molecule to the training set molecules in the space of models
UBC	MLR, ANN, SVM, PLS	IND_I	Undefined

validations sets, and MAE (mean absolute error) for the linear correlation between predicted (Y_{pred}) and experimental (Y_{exp}) data (here, $Y = pIGC_{50}$); these parameters are defined as follows:

$$Q^2_{abs} = 1 - \sum_Y (Y_{exp} - Y_{LOO})^2 \Big/ \sum_Y (Y_{exp} - \langle Y \rangle_{exp})^2 \quad (1)$$

$$R^2_{abs} = 1 - \sum_Y (Y_{exp} - Y_{pred})^2 \Big/ \sum_Y (Y_{exp} - \langle Y \rangle_{exp})^2 \quad (2)$$

$$MAE = \sum_Y |Y - Y_{pred}| \Big/ n \,. \quad (3)$$

Many other statistical characteristics can be used to evaluate model performance; however, we restricted ourselves to these three parameters that provide minimal but sufficient information concerning any model's ability to reproduce both the trends in experimental data for the test sets as well as mean accuracy of predicting all experimental values. The models were considered acceptable if R^2_{abs} exceeded 0.5.

Consensus QSAR Models of Aquatic Toxicity; Comparison Between Methods and Models

The objective of this study from methodological prospective was to explore the suitability of different QSAR modeling tools for the analysis of a dataset with an important toxicological endpoint. Typically, such datasets are analyzed with one (or several) modeling techniques, with a great emphasis on the (high value of) statistical parameters of the training set models. In this study, we went well beyond the modeling studies reported in the original publications [113,126,127,128,129,130] in several respects. First, we have compiled all reported data on chemical toxicity against *T. pyriformis* in a single large dataset and attempted to develop global QSAR models for the entire set. Second, we have employed multiple QSAR modeling techniques thanks to the engagement of six collaborating groups. Third, we have focused on defining model performance criteria not only using training set data but most importantly using external validation sets that were not used in model development in *any* way (unlike any common *cross*-validation procedure) [131]. This focus afforded us the opportunity to evaluate and compare all models using simple and objective universal criteria of *external* predictive accuracy, which in our opinion is the most important single figure of merit for a QSAR model that is of practical significance for experimental toxicologists. Fourth, we have explored the significance of applicability domains and the power of consensus modeling in maximizing the accuracy of external predictivity of our models.

We believe that results of our analysis lend a strong support for our strategy. Indeed, all models performed quite well for the training set (Table 2) with even the lowest Q^2_{abs} among them as high as 0.72. However, there was much greater variation between these models when looking at their (universal and objective) performance criteria as applied to the validation sets I and II (Table 2).

Of 15 QSAR approaches used in this study, nine implemented method-specific applicability domains. Models that did not define the AD showed a reduced predictive accuracy for the validation set II even though they yielded reasonable results for the validation set I. Only CODESSA-MLR (which did not employ any AD) approached in accuracy the lower bound of the models using the AD as measured by $R^2_{abs} = 0.58$ but still had one of the highest MAE of 0.47 (Table 2). On the other hand, among models employing the AD only kNN-MolconnZ had relatively low accuracy of prediction for the validation set II, with R^2_{abs}

QSAR Modeling and QSAR Based Virtual Screening, Complexity and Challenges of Modern, Table 2
Statistical results obtained with all toxicity QSAR models for the modeling and external validation sets

Model	Group ID	Modeling Set ($n = 644$)			Validation Set I ($n = 339$)			Validation Set II ($n = 110$)		
		Q^2_{abs}	MAE	Coverage (%)	R^2_{abs}	MAE	Coverage (%)	R^2_{abs}	MAE	Coverage (%)
kNN-Dragon	UNC	0.92	0.22	100	0.85	0.27	80.2	0.72	0.33	52.7
kNN-MolconnZ	UNC	0.91	0.23	99.8	0.84	0.30	84.3	0.44	0.39	53.6
SVM-Dragon	UNC	0.93	0.21	100	0.81	0.31	80.2	0.83	0.27	52.7
SVM-MolconnZ	UNC	0.89	0.25	100	0.83	0.30	84.3	0.55	0.37	53.6
ISIDA-kNN	ULP	0.77	0.37	100	0.73	0.36	78.5	0.63	0.37	42.7
ISIDA-SVM	ULP	0.95	0.15	100	0.76	0.32	100	0.38	0.50	100
ISIDA-MLR	ULP	0.94	0.20	100	0.81	0.31	95.9	0.65	0.41	51.8
CODESSA-MLR	ULP	0.72	0.42	100	0.71	0.44	100	0.58	0.47	100
OLS	UI	0.86	0.30	92.1	0.77	0.35	97.0	0.59	0.43	98.2
PLS	UK	0.88	0.28	97.7	0.81	0.34	96.1	0.59	0.40	95.5
ASNN	VCCLAB	0.83	0.31	83.9	0.87	0.28	87.4	0.75	0.32	71.8
PLS-IND_I	UBC	0.76	0.39	100	0.74	0.39	99.7	0.45	0.54	100
MLR-IND_I	UBC	0.77	0.39	100	0.75	0.40	99.7	0.46	0.53	100
ANN-IND_I	UBC	0.77	0.39	100	0.76	0.39	99.7	0.46	0.53	100
SVM-IND_I	UBC	0.79	0.31	100	0.79	0.35	99.7	0.53	0.46	100
Consensus Model[a]	–	0.92	0.22	100	0.87	0.27	100	0.70	0.34	100

[a] consensus model: average of the 9 models (kNN-Dragon, kNN-MolconnZ, SVM-Dragon, SVM-MolconnZ, ISIDA-kNN, ISIDA-MLR, OLS, PLS and ASNN) using their individual applicability domains

below 0.5. For all other models the R^2_{abs} ranged between 0.55 and 0.83. On average, the use of applicability domains improved the performance of individual models although the improvement came at the expense of the lower chemistry space coverage (cf. Table 2).

For the most part all models succeeded in achieving reasonable accuracy of external prediction especially when using the AD. It then appeared natural to bring all models together to explore the power of *consensus prediction*. Thus, the *consensus model* was constructed by averaging all available predicted values taking into account the applicability domain of each individual model. In this case we could use only nine of 15 models that had the AD defined. Since each model had its unique way of defining the AD, each external compound could be found within the AD of anywhere between one and nine models so for averaging we only used models covering the compound. The advantage of this data treatment is that the overall coverage of the prediction is still high because it was rare to have an external compound outside of the ADs of all available models. The results (Table 2) showed that the prediction accuracy for both the modeling set (MAE = 0.22) and the validation sets I and II (0.27 and 0.34, respectively) was the best compared to any individual model. The same observation could be made for the correlation coefficient R^2_{abs}. The coverage of this consensus model II was 100% for all three data sets. This observation suggests that consensus models afford both high space coverage and high accuracy of prediction.

In summary, this study presents an example of a fruitful international collaboration between researchers that use different techniques and approaches but share general principles of QSAR model development and validation. Significantly, we did not make any assumptions about the purported mechanisms of aquatic toxicity yet were able to develop statistically significant models for all experimentally tested compounds. In this regard it is relevant to cite an opinion expressed in an earlier publication by Dr. T. Schultz that "models that accurately predict acute toxicity without first identifying toxic mechanisms are highly desirable" [132]. However, the most significant single result of our studies is the demonstrated superior performance of the *consensus modeling* approach when all models are used concurrently and predictions from individual models are averaged. We have shown that both the predictive accuracy and coverage of the final consensus QSAR models were superior as compared to these parameters for individual models. The consensus models appeared robust in terms of being insensitive to both incorporating individual models with low prediction accuracy and the inclusion or exclusion of the AD. Another important result of this study is the power of addressing complex problems in QSAR modeling by forming a virtual collaboratory of independent research groups leading to the formulation and

empirical testing of *best modeling practices*. This latter endeavor is especially critical in light of the growing interest of regulatory agencies to developing most reliable and predictive models for environmental risk assessment [133] and placing such models in the public domain.

Conclusions. Rapid Growth of Publicly Available Databases and Emerging QSAR Research Strategies

With more than 40 years of history behind, QSAR modeling is a well established research field that (as perhaps with any scientific area) has had its ups and downs. There were several recent publications that criticized the current state of the field. Thus, recent editorial published by the leading cheminformatics Journal of Chemical Information and Modeling (JCIM; also reproduced by the Journal of Medicinal Chemistry) introduced severe limitations on the level and quality of QSAR papers to be considered acceptable [73]. Another recent editorial opinion by Dr. Gerry Maggiora [134] outlined limitations and some reasons for failures of QSAR modeling that relate to the so called "activity cliffs". In another recent important paper, Dr. Terry Stouch addressed the question as to why in silico ADME/Tox models fail [135]. These examples naturally lead to an important and perhaps critical question: whether there is any room for further advancement of the field via innovative methodologies and important applications.

Our previous and ongoing research in the area of QSAR suggests that the answer is a resounding 'yes'. We believe strongly that many examples of low impact QSAR research are due to frequent exploration of datasets of limited size with little attention paid to model *external* validation. This limitation leads to models having questionable "mechanistic" explanatory power but perhaps little if any forecasting ability outside of the training sets used for model development. We believe that the latter ability along with the capabilities of QSAR models to explore chemically diverse datasets with complex biological properties should become the chief focus of QSAR studies. This focus requires the re-evaluation of the success criteria for the modeling as well as the development of novel chemical data mining algorithms and model validation approaches. In fact, we think that the most interesting era in QSAR modeling is just beginning with the rapid growth of the experimental SAR data space.

In the past fifteen years, innovative technologies that enable rapid synthesis and high throughput screening of large libraries of compounds have been adopted in almost all major pharmaceutical and biotech companies. As a result, there has been a huge increase in the number of compounds available on a routine basis to quickly screen for novel drug candidates against new targets or pathways. In contrast, such technologies have rarely become available to the academic research community, thus limiting its ability to conduct large scale chemical genetics or chemical genomics research. The NIH Molecular Libraries Roadmap Initiative has changed this situation by forming the national Molecular Library Screening Centers Network (MLSCN) [136] with the results of screening assays made publicly available via PubChem [4]. These efforts have already led to the unprecedented growth of *available* databases of biologically tested compounds (cf. our recent review where we list about 20 available databases of compounds with known bioactivity [10]). This growth creates new challenges for QSAR modeling such as developing novel approaches for the analysis and visualization of large databases of screening data, novel biologically relevant chemical diversity or similarity measures, and novel tools for virtual screening of compound libraries to ensure high expected hit rates. Due to the significant recent increase in publicly available datasets of biologically active compounds and the critical need to improve the hit rate of experimental compound screening there is a strong need in developing widely accessible and reliable computational QSAR modeling techniques and specific endpoint predictors.

Acknowledgments

The studies described in this review were supported in parts by the National Institutes of Health's Cheminformatics Center planning grant P20-RR20751 and the research grants R01GM066940 and R21GM076059.

Bibliography

1. Hansch C, Fujita T (1964) $\rho - \sigma - \pi$ Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. J Amer Chem Soc 86:1616–1626
2. Hansch C, Leo A, Hoekman D (1995) Exploring QSAR: Hydrophobic, Electronic, and Steric Constants. American Chemical Society, Washington, DC
3. Verloop A, Hoogenstraaten W, Tipker J (1976) In: Ariens EJ (ed) Drug Design. Academic Press, New York, pp 165
4. PubChem (2008) http://pubchem.ncbi.nlm.nih.gov/
5. Roth BL, Kroeze WK (2006) Screening the receptorome yields validated molecular targets for drug discovery. Curr Pharm Des 12:1785–1795
6. NCI (2008) http://dtp.nci.nih.gov/docs/3d_database/structural_information/smiles_strings.html
7. FDA (2008) http://www.fda.gov/cder/Offices/OPS_IO/
8. NTP (2008) http://ntp.niehs.nih.gov/ntpweb/
9. DSSTox (2008) http://www.epa.gov/nheerl/dsstox/About.html

10. Oprea T, Tropsha A (2006) Target, Chemical and Bioactivity Databases – Integration is Key. Drug Discov Today 3:357–365
11. Tropsha A (2005) Application of Predictive QSAR Models to Database Mining. In: Oprea T (ed) Cheminformatics in Drug Discovery. Wiley, Darmstadt, pp 437–455
12. Tropsha A (2003) Recent Trends in Quantitative Structure-Activity Relationships. In: Abraham D (ed) Burger's Medicinal Chemistry and Drug Discovery. Wiley, New York, pp 49–77
13. Tropsha A (2006) Predictive QSAR (Quantitative Structure Activity Relationships) Modeling. In: Martin YC (ed) Comprehensive Medicinal Chemistry II. Elsevier, Oxford, pp 113–126
14. Papa E, Villa F, Gramatica P (2005) Statistically validated QSARs, based on theoretical descriptors, for modeling aquatic toxicity of organic chemicals in Pimephales promelas (fathead minnow). J Chem Inf Model 45:1256–1266
15. Tetko IV (2002) Neural network studies 4. Introduction to associative neural networks. J Chem Inf Comput Sci 42:717–728
16. Zupan J, Novic M, Gasteiger J (1995) Neural Networks with Counter-Propagation Learning-Strategy Used for Modeling. Chemom Intell Lab Syst 27:175–187
17. Devillers J (1996) Strengths and Weaknesses of the back propagation neural network in QSAR and QSPR studies. In: Devillers J (ed) Genetic Algorithms in Molecular Modeling. Academic Press, London, pp 1–24
18. Engels MFM, Wouters L, Verbeeck R, Vanhoof G (2002) Outlier mining in high throughput screening experiments. J Biomol Screen 7:341–351
19. Schuurmann G, Aptula AO, Kuhne R, Ebert RU (2003) Stepwise discrimination between four modes of toxic action of phenols in the Tetrahymena pyriformis assay. Chem Res Toxicol 16:974–987
20. Xue Y, Li H, Ung CY, Yap CW, Chen YZ (2006) Classification of a diverse set of Tetrahymena pyriformis toxicity chemical compounds from molecular descriptors by statistical learning methods. Chem Res Toxicol 19:1030–1039
21. Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and Regression Trees. Wadsworth, Co, Belmont
22. Deconinck E, Hancock T, Coomans D, Massart DL, Vander Heyden Y (2005) Classification of drugs in absorption classes using the classification and regression trees (CART) methodology. J Pharm Biomed Anal 39:91–103
23. MOE (2008) http://www.chemcomp.com/fdept/prodinfo.htm#Cheminformatics
24. Put R, Perrin C, Questier F, Coomans D, Massart DL, Vander Heyden YV (2003) Classification and regression tree analysis for molecular descriptor selection and retention prediction in chromatographic quantitative structure-retention relationship studies. J Chromatogr A 988:261–276
25. Breiman L (2001) Random forests. Mach Learn 45:5–32
26. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP (2003) Random forest: A classification and regression tool for compound classification and QSAR modeling. J Chem Inf Comput Sci 43:s1947–1958
27. Put R, Xu QS, Massart DL, Heyden YV (2004) Multivariate adaptive regression splines (MARS) in chromatographic quantitative structure-retention relationship studies. J Chromatogr A 1055:11–19
28. Friedman JH (1991) Multivariate Adaptive Regression Splines. Ann Stat 19:1–67
29. Vapnik VN (1995) The Nature of Statistical Learning Theory. Springer, New York
30. Aires-de-Sousa J, Gasteiger J (2005) Prediction of enantiomeric excess in a combinatorial library of catalytic enantioselective reactions. J Comb Chem 7:298–301
31. Oloff S, Mailman RB, Tropsha A (2005) Application of validated QSAR models of D1 dopaminergic antagonists for database mining. J Med Chem 48:7322–7332
32. Chohan KK, Paine SW, Waters NJ (2006) Quantitative structure activity relationships in drug metabolism. Curr Top Med Chem 6:1569–1578
33. Manallack DT, Ellis DD, Livingstone DJ (1994) Analysis of linear and nonlinear QSAR data using neural networks. J Med Chem 37:3758–3767
34. Mosier PD, Jurs PC (2002) QSAR/QSPR studies using probabilistic neural networks and generalized regression neural networks. J Chem Inf Comput Sci 42:1460–1470
35. Zheng W, Tropsha A (2000) Novel variable selection quantitative structure–property relationship approach based on the k-nearest-neighbor principle. J Chem Inf Comput Sci 40: 185–194
36. Shen M, Beguin C, Golbraikh A, Stables JP, Kohn H, Tropsha A (2004) Application of predictive QSAR models to database mining: identification and experimental validation of novel anticonvulsant compounds. J Med Chem 47:2356–2364
37. Cho SJ, Zheng W, Tropsha A (1998) Rational combinatorial library design 2. Rational design of targeted combinatorial peptide libraries using chemical similarity probe and the inverse QSAR approaches. J Chem Inf Comput Sci 38:259–268
38. Oloff S, Zhang S, Sukumar N, Breneman C, Tropsha A (2006) Chemometric analysis of ligand receptor complementarity: identifying Complementary Ligands Based on Receptor Information (CoLiBRI). J Chem Inf Model 46:844–851
39. Kovatcheva A, Golbraikh A, Oloff S, Xiao YD, Zheng W, Wolschann P, Buchbauer G, Tropsha A (2004) Combinatorial QSAR of ambergris fragrance compounds. J Chem Inf Comput Sci 44:582–595
40. Yao XJ, Panaye A, Doucet JP, Zhang RS, Chen HF, Liu MC, Hu ZD, Fan BT (2004) Comparative study of QSAR/QSPR correlations using support vector machines, radial basis function neural networks, and multiple linear regression. J Chem Inf Comput Sci 44:1257–1266
41. Zhang S, Golbraikh A, Oloff S, Kohn H, Tropsha AA (2006) Novel Automated Lazy Learning QSAR (ALL-QSAR) Approach: Method Development, Applications, and Virtual Screening of Chemical Databases Using Validated ALL-QSAR Models. J Chem Inf Model 46:1984–1995
42. Todeschini R, Consonni V (2000) Handbook of molecular descriptors. Wiley, Weinheim
43. MolconnZ (2008) http://www.edusoft-lc.com/molconn/
44. Cramer RD III, Patterson DE, Bunce JD (1988) Comparative Molecular Field Analysis (CoMFA) 1. Effect of Shape on Binding of Steroids to Carrier Proteins. J Am Chem Soc 110: 5959–5967
45. Klebe G, Abraham U, Mietzner T (1994) Molecular Similarity Indexes in A Comparative-Analysis (Comsia) of Drug Molecules to Correlate and Predict Their Biological-Activity. J Med Chem 37:4130–4146
46. Kubinyi H, Hamprecht FA, Mietzner T (1998) Three-dimensional quantitative similarity-activity relationships (3D QSiAR) from SEAL similarity matrices. J Med Chem 41:2553–2564
47. Waller CL (2004) A comparative QSAR study using CoMFA, HQSAR, and FRED/SKEYS paradigms for estrogen receptor

binding affinities of structurally diverse compounds. J Chem Inf Comput Sci 44:758–765
48. Pastor M, Cruciani G, Mclay I, Pickett S, Clementi S (2000) GRid-INdependent descriptors (GRIND): A novel class of alignment-independent three-dimensional molecular descriptors. J Med Chem 43:3233–3243
49. Cruciani C, Crivori P, Carrupt PA, Testa B (2000) Molecular fields in quantitative structure-permeation relationships: the VolSurf approach. J Mol Struct-Theochem 503:17–30
50. Kier LB, Hall LH (1986) Molecular Connectivity in Structure-Activity Analysis. Wiley, New York
51. Kier LB (1987) Inclusion of symmetry as a shape attribute in kappa-index analysis. Quant Struct-Activit Relatsh 6:8–12
52. Kier LB, Hall LH (1999) Molecular Structure Description: The Electrotopological State. Academic Press, New York
53. Brown RD, Martin YC (1998) An evaluation of structural descriptors and clustering methods for use in diversity selection. SAR QSAR Environ Res 8:23–39
54. Hoffman B, Cho SJ, Zheng W, Wyrick S, Nichols DE, Mailman RB, Tropsha A (1999) Quantitative structure-activity relationship modeling of dopamine D(1) antagonists using comparative molecular field analysis, genetic algorithms-partial least-squares, and K nearest neighbor methods. J Med Chem 42:3217–3226
55. Zhang Q, Muegge I (2006) Scaffold hopping through virtual screening using 2D and 3D similarity descriptors: ranking, voting, and consensus scoring. J Med Chem 49:1536–1548
56. Cherkasov A (2005) 'Inductive' Descriptors. 10 Successful Years in QSAR. Curr Comp Aid Drug Des 1:21–42
57. Stinson SC (2001) Chiral pharmaceuticals. Chem Eng News 79:79
58. Cho SJ, Tropsha A (1995) Cross-validated R2-guided region selection for comparative molecular field analysis: a simple method to achieve consistent results. J Med Chem 38:1060–1066
59. Carhart RE, Smith DH, Venkataraghavan R (1985) Atom pairs as molecular features in structure-activity studies: definition and applications. J Chem Inf Comput Sci 25:64–73
60. Schultz HP, Schultz EB, Schultz TP (1995) Topological Organic-Chemistry 9. Graph-Theory and Molecular Topological Indexes of Stereoisomeric Organic-Compounds. J Chem Inf Comput Sci 35:864–870
61. Julian-Ortiz JV, Alapont CD, Rios-Santamarina I, Garcia-Domenech R, Galvez J (1998) Prediction of properties of chiral compounds by molecular topology. J Mol Graphics Model 16:14–18
62. Golbraikh A, Bonchev D, Tropsha A (2001) Novel chirality descriptors derived from molecular topology. J Chem Inf Comput Sci 41:147–158
63. Golbraikh A, Tropsha A (2003) QSAR modeling using chirality descriptors derived from molecular topology. J Chem Inf Comput Sci 43:144–154
64. Golbraikh A, Bonchev D, Tropsha A (2002) Novel ZE-isomerism descriptors derived from molecular topology and their application to QSAR analysis. J Chem Inf Comput Sci 42:769–787
65. Kovatcheva A, Golbraikh A, Oloff S, Feng J, Zheng W, Tropsha A (2005) QSAR modeling of datasets with enantioselective compounds using chirality sensitive molecular descriptors. SAR QSAR Environ Res 16:93–102
66. Gunner OF, Hughes DW, Dumont LM (1991) An integrated approach to three-dimensional information management with MACCS-3D. J Chem Inf Comput Sci 31:408–414
67. Barnard JM, Downs GM (1995) Fingerprint Descriptor Package, 3.1. Barnard Chemical Information Ltd, Scheffield
68. James CA, Weininger D (1995) Daylight Theory Manual. Daylight Chemical Information Systems, Aliso Viejo
69. Durant JL, Leland BA, Henry DR, Nourse JG (2002) Reoptimization of MDL keys for use in drug discovery. J Chem Inf Comput Sci 42:1273–1280
70. Deshpande M, Kuramochi M, Karypis J (2002) Frequent Sub-Structure-Based Approaches for Classifying Chemical Compounds. Proc of the 8th International Conference on Knowledge Discovery and Data Mining, Edmonton
71. Golbraikh A, Tropsha A (2002) Beware of q2! J Mol Graph Model 20:269–276
72. Tropsha A, Gramatica P, Gombar VK (2003) The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. Quant Struct Act Relat Comb Sci 22:69–77
73. Jorgensen WL, Tirado-Rives J (2006) QSAR/QSPR and Proprietary Data. J Chem Inf Model 46:937
74. Novellino E, Fattorusso C, Greco G (1995) Use of Comparative Molecular Field Analysis and Cluster Analysis in Series Design. Pharm Acta Helv 70:149–154
75. Norinder U (1996) Single and Domain Made Variable Selection in 3D QSAR applications. J Chemomet 10:95–105
76. Tropsha A, Cho SJ (1998) Cross-Validated R2-Guided Region Selection for CoMFA Studies. In: Kubinyi H, Folkers G, Martin YC (eds) 3D QSAR in Drug Design, 3rd edn. Kluwer, Dordrecht, pp 57–69
77. Golbraikh A, Tropsha A (2002) Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. J Comput Aided Mol Des 16:357–369
78. Golbraikh A, Shen M, Xiao Z, Xiao YD, Lee KH, Tropsha A (2003) Rational selection of training and test sets for the development of validated QSAR models. J Comput Aided Mol Des 17:241–253
79. Deconinck E, Coomans D, Vander HY (2007) Exploration of linear modelling techniques and their combination with multivariate adaptive regression splines to predict gastro-intestinal absorption of drugs. J Pharm Biomed Anal 43:119–130
80. Verma RP, Hansch C (2006) Cytotoxicity of organic compounds against ovarian cancer cells: a quantitative structure-activity relationship study. Mol Pharm 3:441–450
81. Pavan M, Netzeva TI, Worth AP (2006) Validation of a QSAR model for acute toxicity. SAR QSAR Environ Res 17:147–171
82. Vracko M, Bandelj V, Barbieri P, Benfenati E, Chaudhry Q, Cronin M, Devillers J, Gallegos A, Gini G, Gramatica P, Helma C, Mazzatorta P, Neagu D, Netzeva T, Pavan M, Patlewicz G, Randic M, Tsakovska I, Worth A (2006) Validation of counter propagation neural network models for predictive toxicology according to the OECD principles: a case study. SAR QSAR Environ Res 17:265–284
83. Saliner AG, Netzeva TI, Worth AP (2006) Prediction of estrogenicity: validation of a classification model. SAR QSAR Environ Res 17:195–223
84. Roberts DW, Aptula AO, Patlewicz G (2006) Mechanistic applicability domains for non-animal based prediction of toxicological endpoints. QSAR analysis of the schiff base appli-

cability domain for skin sensitization. Chem Res Toxicol 19: 1228–1233
85. Estrada E, Patlewicz G (2004) On the usefulness of graph-theoretic descriptors in predicting theoretical parameters. Phototoxicity of polycyclic aromatic hydrocarbons (PAHs). Croat Chem Acta 77:203–211
86. Moss GP, Cronin MTD (2002) Quantitative structure-permeability relationships for percutaneous absorption: re-analysis of steroid data. Int J Pharm 238:105–109
87. Leo AJ, Hansch C (1999) Role of hydrophobic effects in mechanistic QSAR. Perspect Drug Discov Des 17:1–25
88. Zhang S, Golbraikh A, Tropsha A (2006) Development of quantitative structure-binding affinity relationship models based on novel geometrical chemical descriptors of the protein-ligand interfaces. J Med Chem 49:2713–2724
89. Kovatcheva A, Buchbauer G, Golbraikh A, Wolschann P (2003) QSAR modeling of alpha-campholenic derivatives with sandalwood odor. J Chem Inf Comput Sci 43:259–266
90. Shen M, Xiao Y, Golbraikh A, Gombar VK, Tropsha A (2003) Development and validation of k-nearest-neighbor QSPR models of metabolic stability of drug candidates. J Med Chem 46:3013–3020
91. Shen M, LeTiran A, Xiao Y, Golbraikh A, Kohn H, Tropsha A (2002) Quantitative structure-activity relationship analysis of functionalized amino acid anticonvulsant agents using k nearest neighbor and simulated annealing PLS methods. J Med Chem 45:2811–2823
92. Golbraikh A, Shen M, Xiao Z, Xiao YD, Lee KH, Tropsha A (2003) Rational selection of training and test sets for the development of validated QSAR models. J Comput Aided Mol Des 17:241–253
93. Mandel J (1982) Use of the Singular Value Decomposition in Regression-Analysis. Am Stat 36:15–24
94. Afantitis A, Melagraki G, Sarimveis H, Koutentis PA, Markopoulos J, Igglessi-Markopoulou O (2006) A novel QSAR model for predicting induction of apoptosis by 4-aryl-4H-chromenes. Bioorg Med Chem 14:6686–6694
95. Netzeva TI, Gallegos SA, Worth AP (2006) Comparison of the applicability domain of a quantitative structure-activity relationship for estrogenicity with a large chemical inventory. Environ Toxicol Chem 25:1223–1230
96. Tong W, Xie Q, Hong H, Shi L, Fang H, Perkins R (2004) Assessment of prediction confidence and domain extrapolation of two structure-activity relationship models for predicting estrogen receptor binding activity. Environ Health Perspect 112:1249–1254
97. Helma C (2006) Lazy structure-activity relationships (lazar) for the prediction of rodent carcinogenicity and Salmonella mutagenicity. Mol Divers 10:147–158
98. de Cerqueira LP, Golbraikh A, Oloff S, Xiao Y, Tropsha A (2006) Combinatorial QSAR Modeling of P-Glycoprotein Substrates. J Chem Inf Model 46:1245–1254
99. Sachs L (1984) Handbook of statistics. Springer
100. Irwin JJ, Shoichet BK (2005) ZINC – a free database of commercially available compounds for virtual screening. J Chem Inf Model 45:177–182
101. Medina-Franco JL, Golbraikh A, Oloff S, Castillo R, Tropsha A (2005) Quantitative structure-activity relationship analysis of pyridinone HIV-1 reverse transcriptase inhibitors using the k nearest neighbor method and QSAR-based database mining. J Comput Aided Mol Des 19:229–242
102. Zhang S, Wei L, Bastow K, Zheng W, Brossi A, Lee KH, Tropsha A (2007) Antitumor Agents 252. Application of validated QSAR models to database mining: discovery of novel tylophorine derivatives as potential anticancer agents. J Comput Aided Mol Des 21:97–112
103. Hsieh JH, Wang XS, Teotico D, Golbraikh A, Tropsha A (2008) Differentiation of AmpC beta-lactamase binders vs. decoys using classification kNN QSAR modeling and application of the QSAR classifier to virtual screening. J Comput Aided Mol Des
104. Tropsha A, Cho SJ, Zheng W (1999) New Tricks for an Old Dog: Development and Application of Novel QSAR Methods for Rational Design of Combinatorial Chemical Libraries and Database Mining. In: Parrill AL, Reddy MR (eds) Rational Drug Design: Novel Methodology and Practical Applications. American Chemical Society, Washington DC, pp 198–211
105. Gussio R, Pattabiraman N, Kellogg GE, Zaharevitz DW (1998) Use of 3D QSAR methodology for data mining the National Cancer Institute Repository of Small Molecules: application to HIV-1 reverse transcriptase inhibition. Methods 14: 255–263
106. Tropsha A, Zheng W (2001) Identification of the descriptor pharmacophores using variable selection QSAR: applications to database mining. Curr Pharm Des 7:599–612
107. Hengstler JG, Foth H, Kahl R, Kramer PJ, Lilienblum W, Schulz T, Schweinfurth H (2006) The REACH concept and its impact on toxicological sciences. Toxicology 220:232–239
108. Richard AM (2006) Future of toxicology–predictive toxicology: An expanded view of: chemical toxicity. Chem Res Toxicol 19:1257–1262
109. Klopman G, Zhu H, Fuller MA, Saiakhov RD (2004) Searching for an enhanced predictive tool for mutagenicity. SAR QSAR Environ Res 15:251–263
110. Richard AM, Benigni R (2002) AI and SAR approaches for predicting chemical carcinogenicity: Survey and status report. SAR QSAR Environ Res 13:1–19
111. Yang C, Benz RD, Cheeseman MA (2006) Landscape of current toxicity databases and database standards. Curr Opin Drug Discov Dev 9:124–133
112. Zhu H, Tropsha A, Fourches D, Varnek A, Papa E, Gramatica P, Oberg T, Dao P, Cherkasov A, Tetko IV (2008) Combinatorial QSAR Modeling of Chemical Toxicants Tested against Tetrahymena pyriformis. J Chem Inf Model 48:766–784
113. Schultz TW, Netzeva TI (2004) Development and evaluation of QSARs for ecotoxic endpoints: The benzene response-surface model for Tetrahymena toxicity. In: Cronin MTD, Livingstone DJ (eds) Modeling Environmental Fate and Toxicity. CRC Press, Boca Raton, pp 265–284
114. Schultz TW (1999) Structure-toxicity relationships for benzenes evaluated with Tetrahymena pyriformis. Chem Res Toxicol 12:1262–1267
115. Netzeva TI, Schultz TW (2005) QSARs for the aquatic toxicity of aromatic aldehydes from Tetrahymena data. Chemosphere 61:1632–1643
116. Schultz TW, Sinks GD, Miller LA (2001) Population growth impairment of sulfur-containing compounds to Tetrahymena pyriformis. Environ Toxicol 16:543–549
117. Schultz TW, Yarbrough JW, Woldemeskel M (2005) Toxicity to Tetrahymena and abiotic thiol reactivity of aromatic isothiocyanates. Cell Biol Toxicol 21:181–189

118. Aptula AO, Roberts DW, Cronin MTD, Schultz TW (2005) Chemistry-toxicity relationships for the effects of Di-and trihydroxybenzenes to Tetrahymena pyriformis. Chem Res in Toxicol 18:844–854
119. Schultz TW, Hewitt M, Netzeva TI, Cronin MTD (2007) Assessing applicability domains of toxicological QSARs: Definition, confidence in predicted values, and the role of mechanisms of action. QSAR Comb Sci 26:238–254
120. Netzeva TI, Schultz TW (2005) QSARs for the aquatic toxicity of aromatic aldehydes from Tetrahymena data. Chemosphere 61:1632–1643
121. Schultz TW, Yarbrough JW, Woldemeskel M (2005) Toxicity to Tetrahymena and abiotic thiol reactivity of aromatic isothiocyanates. Cell Biol Toxicol 21:181–189
122. Schultz TW, Sinks GD, Miller LA (2001) Population growth impairment of sulfur-containing compounds to Tetrahymena pyriformis. Environ Toxicol 16:543–549
123. Aptula AO, Roberts DW, Cronin MTD, Schultz TW (2005) Chemistry-toxicity relationships for the effects of Di-and trihydroxybenzenes to Tetrahymena pyriformis. Chem Res Toxicol 18:844–854
124. Schultz TW, Hewitt M, Netzeva TI, Cronin MTD (2007) Assessing applicability domains of toxicological QSARs: Definition, confidence in predicted values, and the role of mechanisms of action. QSAR Comb Sci 26:238–254
125. Schultz TW, Hewitt M, Netzeva TI, Cronin MTD (2007) Assessing applicability domains of toxicological QSARs: Definition, confidence in predicted values, and the role of mechanisms of action. QSAR Comb Sci 26:238–254
126. Schultz TW, Hewitt M, Netzeva TI, Cronin MTD (2007) Assessing applicability domains of toxicological QSARs: Definition, confidence in predicted values, and the role of mechanisms of action. QSAR Comb Sci 26:238–254
127. Aptula AO, Roberts DW, Cronin MTD, Schultz TW (2005) Chemistry-toxicity relationships for the effects of Di-and trihydroxybenzenes to Tetrahymena pyriformis. Chem Res Toxicol 18:844–854
128. Netzeva TI, Schultz TW (2005) QSARs for the aquatic toxicity of aromatic aldehydes from Tetrahymena data. Chemosphere 61:1632–1643
129. Schultz TW, Yarbrough JW, Woldemeskel M (2005) Toxicity to Tetrahymena and abiotic thiol reactivity of aromatic isothiocyanates. Cell Biol Toxicol 21:181–189
130. Schultz TW, Sinks GD, Miller LA (2001) Population growth impairment of sulfur-containing compounds to Tetrahymena pyriformis. Environ Toxicol 16:543–549
131. Gramatica P (2007) Principles of QSAR models validation: internal and external. QSAR Comb Sci 26:694–701
132. Schultz TW (1999) Structure-toxicity relationships for benzenes evaluated with Tetrahymena pyriformis. Chem Res Toxicol 12:1262–1267
133. Yang C, Richard AM, Cross KP (2006) The Art of Data Mining the Minefields of Toxicity Databases to Link Chemistry to Biology. Curr Comput-Aided Drug Des 2:135–150
134. Maggiora GM (2006) On outliers and activity cliffs–why QSAR often disappoints. J Chem Inf Model 46:1535
135. Stouch TR, Kenyon JR, Johnson SR, Chen XQ, Doweyko A, Li Y (2003) In silico ADME/Tox: why models fail. J Comput Aided Mol Des 17:83–92
136. Austin CP, Brady LS, Insel TR, Collins FS (2004) NIH Molecular Libraries Initiative. Science 306:1138–1139

Quantum Algorithms

Michele Mosca[1,2,3]

[1] Institute for Quantum Computing and Dept. of Combinatorics and Optimization, University of Waterloo, Waterloo, Canada
[2] St. Jerome's University, Waterloo, Canada
[3] Perimeter Institute for Theoretical Physics, Waterloo, Canada

Article Outline

Glossary

Quantum circuit model One of the standard and most commonly used models of quantum computation which generalizes the classical model of acyclic circuits and closely models most of the proposed physical implementations of quantum computers. When studying algorithms for a problem with an infinite number of possible inputs, one usually restricts attention to *uniform* families of circuits, which are families of circuits in which the circuit for inputs of size n can be generated efficiently as a function of n. For example, one might require that there is a classical Turing machine that can generate the nth circuit in time polynomial in n.

Black box model A model of computation where the input to the problem includes a "black-box" that can be applied (equivalently, an "oracle" that can be "queried"). This is the only way to extract information from the black-box. For example, the black-box could accept inputs $j \in \{0,1\}^n$ and output a value $X_j \in \{0,1\}$. In this particular case, we can think of the black-box as a means for querying the bits of the

string $\mathbf{X} = X_1X_2X_3 \ldots X_{2^n}$. In the black-box model, one usually measures complexity in terms of the number of applications of the black-box.

Computational complexity When referring to an algorithm, the computational complexity (often just called the complexity) is a measure of the resources used by the algorithm (which we can also refer to as the *cost* of the algorithm) usually measured as a function of the size of the input to the algorithm. The complexity for input size n is taken to be the cost of the algorithm on a worst-case input of size n to the problem. This is also referred to as *worst-case complexity*. When referring to a problem, the computational complexity is the minimum amount of resources required by any algorithm to solve the problem. See ▶ Quantum Computational Complexity for an overview.

Query complexity When referring to a black-box algorithm, the query complexity is the number of applications of the black-box or oracle used by the algorithm. When referring to a black-box problem, the query complexity is the minimum number of applications of the black-box required by any algorithm to solve the problem.

Definition of the Subject

The strong Church–Turing thesis states that a probabilistic Turing machine can efficiently simulate any realistic model of computation. By "efficiently", we mean that there is a polynomial p such that the amount of resources used by the Turing machine simulation is not more than $p(M)$ where M is the amount of resources used by the given realistic model of computation.

Since a computer is a physical device, any reasonable model of computation must be cast in a realistic physical framework, hence the condition that the model be "realistic" is very natural. The probabilistic Turing machine is implicitly cast in a classical framework for physics, and it appears to hold as long as the competing model of computation is also cast in a classical framework. However, roughly a century ago, a new framework for physics was developed: quantum mechanics. The impact of this new framework on the theory of computation was not taken very seriously until the early 1970's by Stephen Wiesner [177] for cryptographic purposes (and later by Bennett and Brassard [27]). Benioff [23] proposed using quantum devices in order to implement reversible computation. Feynman [79] noted that a classical computer seems incapable of efficiently simulating the dynamics of rather simple quantum mechanical systems, and proposed that a "quantum" computer, with components evolving according to quantum mechanical rules, should be able to perform such simulations efficiently (see Sect. "Simulation of Quantum Mechanical Systems"). Manin made a similar observation independently [136]. Deutsch [64] worked on proving the original Church–Turing thesis (which was only concerned about effective computability, and not efficient computability) in a quantum mechanical framework, and defined two models of quantum computation; he also gave the first quantum algorithm. One of Deutsch's ideas is that quantum computers could take advantage of the computational power present in many "parallel universes" and thus outperform conventional classical algorithms. While thinking of parallel universes is sometimes a convenient way for researchers to invent quantum algorithms, the algorithms and their successful implementation are independent of any particular interpretation of standard quantum mechanics.

Quantum algorithms are algorithms that run on any realistic model of quantum computation. The most commonly used model of quantum computation is the circuit model (more strictly, the model of uniform families of acyclic quantum circuits), and the *quantum strong* Church–Turing thesis states that the quantum circuit model can efficiently simulate any realistic model of computation. Several other models of quantum computation have been developed, and indeed they can be efficiently simulated by quantum circuits. Quantum circuits closely resemble most of the currently pursued approaches for attempting to construct scalable quantum computers.

The study of quantum algorithms is very important for several reasons. Computationally secure cryptography is widely used in society today, and relies on the believed difficulty of a small number of computational problems. Quantum computation appears to redefine what is a tractable or intractable problem, and one of the first breakthroughs in the development of quantum algorithms was Shor's discovery of efficient algorithms [165] for factoring and finding discrete logarithms. The difficulty of factoring and finding discrete logarithms was (and still is!) at the core of currently-used public-key cryptography, and his results showed that if and when a quantum computer is built, then any messages that had been previously encrypted using our current public-key cryptographic tools could be compromised by anyone who had recorded the ciphertext and public keys. Furthermore the vast public-key infrastructure currently in place would be compromised with no clear alternative to replace it; also, any alternative will take many years to deploy. There was a sudden rush to answer two fundamental questions. Firstly, can we actually build a sufficiently large quantum computer? Perhaps this isn't a reasonable model of computation. Subse-

quent work on quantum fault-tolerant error correction indicates that the answer is "yes", and experimental progress has steadily grown. Secondly, what other interesting and important problems can quantum computers solve more efficiently than the best known classical algorithms? The following sections survey the state of the art on the second question.

As we gain confidence about which problems are still hard in a quantum mechanical framework, we can start to rebuild confidence in a secure public-key cryptographic infrastructure that is robust in the presence of quantum technologies. Although the cryptographic implications are dramatic and of immediate relevance, in the longer term, the most important significance of quantum algorithms will be for a wider range of applications, where important problems cannot be solved because there are no known (or possible) efficient classical algorithms, but there are efficient quantum mechanical solutions. At present, we know of applications such as searching and optimizing (Sect. "Algorithms Based on Amplitude Amplification") and simulating physical systems (see Sect. "Simulation of Quantum Mechanical Systems"), but the full implications are still unknown and very hard to predict. The next few sections give an overview of the current state of quantum algorithmics.

Introduction and Overview

There are several natural models of quantum computation. The most common one is a generalization of the classical circuit model. A detailed description of the circuit model of quantum computation can be found in several textbooks [122,129,149]. One can also define continuous models of computation, where one specifies the Hamiltonian $H(t)$ of the system at time t, where $H(t)$ is a "reasonable" Hamiltonian (e. g. a sum Hamiltonians involving a constant number of nearby subsystems), and one evolves the system for a period of time T. A reasonable measure of the total cost might be $\int_{t=0}^{T} \|H(t)\| \, \mathrm{d}t$.

Most algorithmic work in quantum computing has been developed in discrete models of computation, that is, with a discrete state space and with discrete time steps. In Subsect. "Continuous Time Quantum Walk Algorithms" and Sect. "Adiabatic Algorithms", we discuss algorithms developed in a continuous-time model of computation. Even if, as in the case of classical computers, implementations of scalable fault-tolerant quantum computers have discrete time steps and state spaces, these algorithms are still very useful since there are efficient simulations using any universal discrete model of quantum computation. Note that if no such efficient simulation exists, then either the continuous model of computation in question is physically unrealistic, or the quantum strong Church–Turing thesis is incorrect.

Discrete state spaces can be used to approximate continuous state spaces in order to solve problems normally posed with continuous state spaces. One must choose appropriate discretizations, analyzing errors in the approximation, and quantify the scaling of the algorithm as the overall approximation error gets arbitrarily small. Quantum algorithms for such continuous problems are surveyed in ▶ Quantum Algorithms and Complexity for Continuous Problems.

Many of the key ideas that led to the development of quantum computation emerged from earlier work on reversible computation [24]. Many facts from the theory of reversible computing are fundamental tools in the development of quantum algorithms. For example, suppose we have a classical algorithm for computing a function $f\colon \{0,1\}^n \to \{0,1\}^m$ (we use binary encoding for convenience). The details of the computing model are not so important, as long as it's a realistic model. For concreteness, suppose we have a reasonable encoding of a circuit of size C, using gates from a finite set, that takes $\mathbf{x} \in \{0,1\}^n$ as input and outputs $\mathbf{y} \in \{0,1\}^m$ (discarding any additional information it might have computed). Then this circuit can be efficiently converted into a circuit, composed only of reversible gates, that maps $|\mathbf{x}\rangle\,|\mathbf{y}\rangle\,|\mathbf{0}\rangle \mapsto |\mathbf{x}\rangle\,|\mathbf{y} \oplus f(\mathbf{x})\rangle\,|\mathbf{0}\rangle$, where $\oplus$ denotes the bitwise XOR (addition modulo 2), and the third register of 0s is ancilla workspace that is reset to all 0s by reversible operations. This new circuit uses $O(C)$ reversible gates from a finite set, so the overhead is modest. A basic introduction to this and other important facts about reversible computing can be find in most standard textbooks on quantum computing. For example, in Sect. "Algorithms Based on Amplitude Amplification" on quantum searching, we use the fact that we can convert any classical heuristic algorithm that successfully guesses a solution with probability p into a reversible quantum algorithm that guesses a solution with probability amplitude $\sqrt{p}$.

Most of the known quantum algorithms can be phrased as black-box algorithms solving black-box problems. A black-box, or oracle, is subroutine or subcircuit that implements some operation or function. It does so in a way that provides no other information other than simply taking an input and giving the prescribed output. One cannot, for example, look at the inner workings of the circuit or device implementing the black-box to extract additional information. For example, Shor's factoring algorithm can be viewed as an algorithm that finds the order of an element in a black-box group (that is, a group for which

the group operations are computed by a black-box), or the period of a black-box function, where the black-box is substituted with a subcircuit that implements exponentiation modulo N. The quantum search algorithm is described as a black-box algorithm, but it is straightforward to substitute in a subcircuit that checks if a given input is a valid certificate for some problem in NP.

If we take a black-box algorithm that uses T applications of the black-box and A other computational steps, and replace each black-box with a subcircuit that uses B elementary gates, then we get an algorithm that uses $TB + A$ gates. Thus if T and A are both polynomial in size, then an efficient black-box algorithm yields an efficient algorithm whenever we replace the black-box with a polynomial time computable function.

Many lower bounds have been found in the black-box model. The query complexity of a black-box problem is the number of applications of the black-box (or queries to the oracle) that a black-box algorithm must make in order to solve the problem. If we try to solve a problem that has query complexity T, where the black-box is implemented by some subcircuit, then we can conclude that any algorithm that treats the subcircuit as a black-box must apply the subcircuit a total of T times and thus use $\Omega(T)$ gates (we use the fact that any implementation of the black-box uses at least one gate; if we had a lower bound on the complexity of implementing the black-box, then we could derive a better lower bound in this case). However, this does not imply that an $\Omega(T)$ lower bound applies to any algorithm that uses the subcircuit, since it might exploit the information within the subcircuit in a way other than just applying the subcircuit. A discussion of the black-box model and its practical relevance can be found in [52,122,173].

In the literature, one can often see a progression from studying basic algorithmic primitives (such as the convergence properties of a generic quantum walk), to the application to solve a black-box problem (such as element distinctness), to solving some concrete computational problem (like factoring an integer) that doesn't involve a black-box.

This survey will include both black-box and non-black-box results. It is infeasible to detail all the known quantum algorithms, so a representative sample is given in this article. For a subset of this sample, there is an explicit definition of the problem, a discussion of what the best known quantum algorithm can do, and a comparison to what can be achieved with classical algorithms. For black-box problems, we attempt to give the number of queries and the number of non-query operations used by the algorithm, as well as the best-known lower bounds on the query complexity. In some cases, all of this information is not readily available in the literature, so there will be some gaps.

As a small technical note, when we refer to a real number r as an input or output to a problem, we are referring to a finite description of a real number from which, for any integer n, one can efficiently compute (in time polynomial in n) an approximation of r with error at most $1/2^n$.

In this article, we start with a brief sketch of the very early quantum algorithms, and then in the subsequent sections the algorithms are grouped according to the kind of problems they solve, or the techniques or algorithmic paradigms used.

Section "Early Quantum Algorithms" summarizes the early quantum algorithms. Section "Factoring, Discrete Logarithms and the Abelian Hidden Subgroup Problem" describes the Abelian hidden subgroup algorithms, including Shor's factoring and discrete logarithm algorithms. Section "Algorithms Based on Amplitude Amplification" describes quantum searching and amplitude amplification and some of the main applications. Section "Simulation of Quantum Mechanical Systems" describes quantum algorithms for simulating quantum mechanical systems, which are another important class of algorithms that appear to offer an exponential speed-up over classical algorithms. Section "Generalizations of the Abelian Hidden Subgroup Problem" describes several non-trivial generalizations of the Abelian hidden subgroup problem, and related techniques. Section "Quantum Walk Algorithms" describes the quantum walk paradigm for quantum algorithms and summarizes some of the most interesting results and applications. Section "Adiabatic Algorithms" describes the paradigm of adiabatic algorithms. Section "Topological Algorithms" describes a family of "topological" algorithms. Section "Quantum Algorithms for Quantum Tasks" describes algorithms for quantum tasks which cannot be done by a classical computer. In Sect. "Future Directions" we conclude with a discussion.

Early Quantum Algorithms

The first explicitly defined quantum algorithm was the one described by David Deutsch in his landmark paper [64] where he defined the model of quantum computation, including a circuit and Turing machine model.

The problem was to decide if a given function $f\colon \{0,1\} \mapsto \{0,1\}$ is constant or "balanced" (the function f is balanced if there are an equal number of 0 and 1 outputs). In other words, output $f(0) \oplus f(1)$, where $\oplus$ denotes addition modulo 2. One is given a circuit that implements $|x\rangle|0\rangle \mapsto |x\rangle|f(x)\rangle$. Deutsch showed that us-

ing one application of the circuit, we can compute

$$\frac{1}{\sqrt{2}}\,|\,0\rangle\,|\,f(0)\rangle + \frac{1}{\sqrt{2}}\,|\,1\rangle\,|\,f(1)\rangle\ .$$

If $f(0) = f(1)$, then the first qubit is in the state $\frac{1}{\sqrt{2}}\,|\,0\rangle + \frac{1}{\sqrt{2}}\,|\,1\rangle$ and thus applying a Hadamard transformation to the first qubit will output $|\,0\rangle$ with certainty. However, if $f(0) \neq f(1)$, then applying a Hadamard gate and measuring will output $|\,0\rangle$ with probability $\frac{1}{2}$ and $|\,1\rangle$ with probability $\frac{1}{2}$. Thus if we obtain $|\,1\rangle$, we are certain that the function is balanced, and thus we would output "balanced" in this case. If we obtain $|\,0\rangle$, we cannot be certain; however, by guessing "constant" with probability $\frac{2}{3}$ in this case, we get an algorithm that guesses correctly with probability $\frac{2}{3}$. This is not possible with a classical algorithm that only evaluates f once.

If one knows, for example, that the unitary transformation also maps $|\,x\rangle\,|\,1\rangle \mapsto |\,x\rangle\,|\,1 \oplus f(x)\rangle$, then one can solve this problem with certainty with only one query [54]. A common tool that is used in this and many other quantum algorithms, is to note that if one applies such an implementation of f on the input $|\,x\rangle\,(1/\sqrt{2}\,|\,0\rangle - 1/\sqrt{2}\,|\,1\rangle)$, then the output can be written as $(-1)^{f(x)}\,|\,x\rangle\,(1/\sqrt{2}\,|\,0\rangle - 1/\sqrt{2}\,|\,1\rangle)$. This technique can be used to replace the 2-step process of first mapping $|\,x\rangle\,|\,0\rangle \mapsto |\,x\rangle\,|\,f(x)\rangle$, then applying a Z gate $|\,x\rangle\,|\,f(x)\rangle \mapsto |\,x\rangle\,(-1)^{f(x)}\,|\,f(x)\rangle = (-1)^{f(x)}\,|\,x\rangle\,|\,f(x)\rangle$, and then applying the function evaluation a second time in order to "uncompute" the value of $f(x)$ from the second register: $(-1)^{f(x)}\,|\,x\rangle\,|\,f(x)\rangle \mapsto (-1)^{f(x)}\,|\,x\rangle\,|\,0\rangle$.

This problem was generalized to the Deutsch–Jozsa problem [65] which asks the same constant versus "balanced" question, but for a function $f\colon \{0,1\}^n \to \{0,1\}$, with the promise that the function is either constant or balanced. The notion of a promise problem appears frequently in the study of quantum algorithms and complexity. One can either think of the condition as a promise, or alternatively, one can accept any answer as correct in the case that the promise is not satisfied.

This problem was solved with two queries with a similar algorithm, and the two queries can be reduced to one if the oracle evaluating f has the form $|\,x\rangle\,|\,b\rangle \mapsto |\,x\rangle\,|\,b \oplus f(x)\rangle$. A classical deterministic algorithm requires $2^{n-1} + 1$ evaluations of the function in order to decide if f is constant or balanced with certainty. However, a classical randomized algorithm can decide if f is constant or balanced with error probability ϵ with only $O(\log \frac{1}{\epsilon})$ queries. Other oracle separations were given in [30,31].

Bernstein and Vazirani [28] defined a special family of functions that are either constant or balanced, in particular, for any $\mathbf{a} \in \{0,1\}^n$, $b \in \{0,1\}$, let $f_{\mathbf{a}}(\mathbf{x}) = \mathbf{a} \cdot \mathbf{x} \oplus b$. They showed how to find $\mathbf{a}$ using only 2 evaluations of a black-box implementing $|\,\mathbf{x}\rangle\,|\,0\rangle \mapsto |\,\mathbf{x}\rangle\,|\,f(\mathbf{x})\rangle$. This can be reduced to 1 evaluation given the standard black-box $U_f\colon |\,\mathbf{x}\rangle\,|\,b\rangle \mapsto |\,\mathbf{x}\rangle\,|\,b \oplus f(\mathbf{x})\rangle$. This algorithm can be generalized quite naturally to finding the hidden matrix M in the affine map $\mathbf{x} \mapsto M\mathbf{x} + \mathbf{b}$, where $\mathbf{x} \in \mathbb{Z}_N^n$, $\mathbf{b} \in \mathbb{Z}_N^m$ and M is an $m \times n$ matrix of values in $\mathbb{Z}_N$ (with element-wise addition modulo N), using only m queries [104,143]. In some sense, the quantum algorithm takes a black-box for right-multiplying by M and turns it into a black-box for left-multiplying by M. This allows us to determine M with only m queries instead of $n + 1$ queries, which can be advantageous if $m < n$, although no practical application has been developed to date.

Bernstein and Vazirani also give the first instance of a black-box problem (recursive Fourier sampling) where a quantum algorithm gives a exponential improvement over bounded error randomized classical algorithms. Simon [167] gave a problem where there is an exponential gap. The problem is somewhat simpler to state, and is a special case of a broad family of important problems for which efficient quantum algorithms were subsequently found (using similar methods).

Simon's Problem

Input: A black-box U_f implementing $|\,\mathbf{x}\rangle\,|\,0\rangle \mapsto |\,\mathbf{x}\rangle\,|\,f(\mathbf{x})\rangle$ for a function $f\colon \{0,1\}^n \to \{0,1\}^n$ with the property that $f(\mathbf{x}) = f(\mathbf{y})$ if and only if $\mathbf{x} \oplus \mathbf{y} \in K = \{\mathbf{0}, \mathbf{s}\}$ for some $\mathbf{s} \in \{0,1\}^n$.

Problem: Find $\mathbf{s}$.

Note that f is one-to-one if $\mathbf{s} = \mathbf{0}$, the string of 0s, and f is two-to-one if $\mathbf{s} \neq \mathbf{0}$.

Simon's algorithm is very simple and elegant. We sketch it here, since many of the algorithms presented in the upcoming sections follow the same overall structure. We give a modified version that has a definite running time (versus letting it run indefinitely and analyzing the expected running time).

Simon's Algorithm

1. Set $i = 1$.
2. Prepare the $2n$ qubit state $|\,00\ldots0\rangle\,|\,00\ldots0\rangle$.
3. Apply a Hadamard gate to the first n qubits to obtain the state $\frac{1}{\sqrt{2^n}}\,|\,\mathbf{x}\rangle\,|\,\mathbf{0}\rangle$.
4. Apply U_f to create the state $\frac{1}{\sqrt{2^n}}\,|\,\mathbf{x}\rangle\,|\,f(\mathbf{x})\rangle$.
5. Apply a Hadamard gate to the first n qubits.
6. Measure the first n qubits to obtain a string $\mathbf{y}_i \in \{0,1\}^n$.
7. If $i = n + 3$, go to the next step. Otherwise, increment i and go to step 2.
8. Let M be the $(n+3) \times n$ matrix whose ith row is the vector $\mathbf{y}_i$. Solve the system $M\mathbf{x} = \mathbf{0}$ (where we treat $\mathbf{x}$

and **0** as column vectors). If there is a unique non-zero solution $\mathbf{x} = \mathbf{s}$, then output **s**. If the only solution is $\mathbf{x} = \mathbf{0}$, then output **0**. Otherwise, output "FAIL".

Note that in step 3, since we can partition the space $\mathbb{Z}_2^n$ into a disjoint union of cosets $\mathbb{Z}_2^n/K$ of the subgroup $K = \{\mathbf{0}, \mathbf{s}\}$, where $f(\mathbf{x})$ is constant on each coset, we can rewrite the state of the system as

$$\frac{1}{\sqrt{2^n}} |\,\mathbf{x}\rangle\,|\,f(\mathbf{x})\rangle = \frac{1}{\sqrt{2^{n-1}}} \sum_{\{\mathbf{z},\mathbf{z}\oplus\mathbf{s}\}\in\mathbb{Z}_2^n/K} \frac{1}{\sqrt{2}}(|\,\mathbf{z}\rangle + |\,\mathbf{z}\oplus\mathbf{s}\rangle)\,|\,f(\mathbf{z})\rangle \ .$$

Since the rest of the algorithm ignores the second register, we can assume that in the first register we have a random "coset state" $\frac{1}{\sqrt{2}}(|\,\mathbf{z}\rangle + |\,\mathbf{z}\oplus\mathbf{s}\rangle)$ selected uniformly at random from all the coset states.

After applying the Hadamard gates, the coset state gets mapped to a equally weighted superposition of states that are orthogonal to K,

$$\frac{1}{\sqrt{2^{n-1}}} \sum_{\mathbf{y}\in K^\perp} (-1)^{\mathbf{y}\cdot\mathbf{z}}\,|\,\mathbf{y}\rangle$$

where $K^\perp = \{\mathbf{y} \mid \mathbf{y} \in \mathbb{Z}_2^n, \mathbf{y}\cdot\mathbf{s} = 0\}$. Thus, with $n + O(1)$ random sampled from $K^\perp$, the samples vectors $\mathbf{y}_i$ will generated $K^\perp$ with high probability. Then, using linear algebra, we can efficiently compute generators for K. In the generalizations of Simon's algorithm that we see in the next few sections, we'll see a more general formulation of $K^\perp$ for a hidden subgroup K.

Shortly after Simon came up with his black-box algorithm, Shor [165] used similar ideas to derive his famous algorithms for factoring and finding discrete logarithms.

Quantum Algorithms for Simon's Problem Simon's algorithm solves this problem with bounded error using $n + O(1)$ applications of U_f and $O(n)$ other elementary quantum operations and $O(n^3)$ elementary classical operations.

Brassard and Høyer [35] combined Simon's algorithm with amplitude amplification order to make the algorithm exact.

Classical Algorithms for Simon's Problem Simon [167] showed a lower bound of $\Omega(2^{\frac{n}{4}})$ queries, and this can be improved to $\Omega(2^{\frac{n}{2}})$.

This was a historical moment, since the field of quantum algorithms progressed from work on black-box algorithms and complexity, to having an algorithm without black-boxes and of broad practical importance. It is important to note the fundamental role of the foundational complexity-theoretic work and black-box algorithms to the development of the practical algorithms.

Factoring, Discrete Logarithms and the Abelian Hidden Subgroup Problem

Factoring, Finding Orders and Periods, and Eigenvalue Estimation

The most well known quantum algorithm is Shor's algorithm [165,166] for the integer factorization problem.

Integer Factorization Problem
Input: An integer N.
Problem: Output positive integers $p_1, p_2, \ldots, p_l, r_1, r_2, \ldots, r_l$ where the p_i are distinct primes and $N = p_1^{r_1} p_2^{r_2} \cdots p_l^{r_l}$.

This problem can be efficiently reduced (that is, in time polynomial in $\log N$), by a probabilistic classical algorithm, to $O(l)$ instances (note that $l \in O(\log N)$) of the problem of finding the multiplicative order of an element modulo N, which can be solved efficiently on a quantum computer.

Order Finding Problem
Input: Positive integers a and N, such that $\mathrm{GCD}(a, N) = 1$ (i. e. a is relatively prime to N).
Problem: Find the order of a modulo N.

Essentially the same quantum algorithm can efficiently find the order of an element a in a finite group G given a black-box for performing the group arithmetic. Shor's algorithm in fact solves the more general problem of finding the period of a periodic function f.

Period Finding Problem
Input: A black-box implementing a periodic function $f\colon \mathbb{Z} \mapsto X$ for some finite set X, where $f(x) = f(y)$ if and only if $r \mid x - y$.
Problem: Find the period r.

Shor described his algorithm for the specific function $f(x) = a^x \mod N$, where N was the integer to be factored; however the algorithm will find the period of any such periodic function (note the assumption that the values of $f(1), f(2), \ldots, f(r)$ are distinct; one can also analyze the case where the values are not entirely distinct [32,145]).

Kitaev later showed that the problem of finding the order of $a \in G$ can alternatively be reduced to the problem of estimating the eigenvalues of the operator that multiplies by a, and he described a efficient quantum algorithm for estimating such eigenvalues [126]. His method is to reduce the problem to phase estimation. Although Kitaev's algo-

rithm was qualitatively different [116], it was shown that using an improved phase estimation algorithm yields an order-finding circuit that is essentially equivalent to Shor's factoring algorithm based on period-finding [54].

Both Shor's and Kitaev's approaches find r by finding good estimates of a random integer multiple of $\frac{1}{r}$ and then applying the continued fractions algorithm to find r.

Sampling Estimates to an Almost Uniformly Random Integer Multiple of $\mathbf{1/r}$
Input: Integers a, and N such that $\mathrm{GCD}(a, N) = 1$. Let r denote the (unknown) order of a.
Problem: Output a number $x \in \{0, 1, 2, \ldots, 2^n - 1\}$ such that for each $k \in \{0, 1, \ldots, r-1\}$ we have

$$\Pr\left(\left|\frac{x}{2^n} - \frac{k}{r}\right| \leq \frac{1}{2r^2}\right) \geq c\frac{1}{r}$$

for some constant $c > 0$.

Shor's analysis of the algorithm works by creating a "periodic state", which is done by preparing $\sum_x |x\rangle |a^x \mod N\rangle$ and measuring the second register (in fact, just ignoring or tracing out the second register suffices, since one never uses the value of the measurement outcome). One then applies the quantum Fourier transform, or its inverse, to estimate the period.

Kitaev's algorithm works by estimating a random eigenvalue of the operator U_a that multiplies by a.

Eigenvalue Estimation Problem
Input: A quantum circuit implementing the controlled-U, for some unitary operator U, and an eigenstate $|\psi\rangle$ of U with eigenvalue $e^{2\pi i\omega}$.
Problem: Obtain a good estimate for ω.

He solves the eigenvalue estimation problem by solving a version of the well-studied phase estimation problem, and uses the crucial fact (as did Shor) that one can efficiently implement a circuit that computes $U_a^{2^j} = U_{a^{2^j}}$ using j group multiplications instead of 2^j multiplications. Thus one can efficiently reduce the eigenvalue estimation problem to the following phase estimation problem.

Phase Estimation Problem
Input: The states $\frac{1}{\sqrt{2}}(|0\rangle + e^{2\pi i\omega y}|1\rangle)$, for $y = 1, 2, 4, \ldots, 2^n$, for some $\omega \in [0, 2\pi)$.
Problem: Obtain a good estimate of the phase parameter ω.

Kitaev's phase estimation algorithm used $O(n)$ copies of the states $\frac{1}{\sqrt{2}}(|0\rangle + e^{2\pi i\omega y}|1\rangle)$ for $y = 1, 8, 64, \ldots$, and provides an estimate with error at most $\frac{1}{2^n}$ with high probability. Although there are slightly more efficient phase estimation algorithms, one advantage of his algorithm is that it does not require a quantum Fourier transform, and instead performs some efficient classical post-processing of estimates of $y\omega \mod 1$. This might be advantageous experimentally. The phase estimation problem has been studied for several decades [100,102] recently with some focus on the algorithmic complexity of the optimal or near-optimal estimation procedures.

Quantum Algorithms for Order Finding Finding the order of a random element in $\mathbb{Z}_N^*$

- Quantum complexity is in $O((\log N)^2 \log\log(N) \log\log\log(N))$.

Order finding in a black-box group

- Quantum black-box complexity (for groups with unique encodings of group elements) is $O(\log r)$ black-box multiplications and $O(n + \log^2 r)$ other elementary operations.

Classical Algorithms for Order Finding Finding the order of a random element in $\mathbb{Z}_N^*$

- Best known rigorous probabilistic classical algorithm has complexity in $e^{O(\sqrt{\log N \log\log N})}$.
- Best known heuristic probabilistic classical algorithm has complexity in $e^{O((\log N)^{\frac{1}{3}}(\log\log N)^{\frac{2}{3}})}$.

Order finding in a black-box group

- Classical black-box multiplication complexity is in $\Theta(\sqrt{r})$ [53].

By "heuristic" algorithm, we mean the proof of its running time makes some plausible but unproven assumptions.

Discrete Logarithms

Shor [165,166] also solved the problem of finding discrete logarithms in the multiplicative group of a finite field.

The Discrete Logarithm Problem
Input: Elements a and $b = a^t$ in $\mathbb{Z}_p^*$, where t is an integer from $\{0, 1, 2, \ldots, r-1\}$ and r is the order of a.
Problem: Find t. (The number t is called the discrete logarithm of b with respect to the base a.)

Shor solved this problem by defining $f\colon \mathbb{Z}_r \times \mathbb{Z}_r \mapsto \mathbb{Z}_p^*$ as $f(x, y) = a^x b^y$. Note that $f(x_1, y_1) = f(x_2, y_2)$ if and only if $(x_1 - x_2, y_1 - y_2)$ is in the additive subgroup of $\mathbb{Z}_r \times \mathbb{Z}_r$ generated by $(t, -1)$. Since r is known, there is no need for the continued fractions algorithm. There is also an analogous eigenvalue estimation version of this algorithm.

The algorithm can be defined in general for any group G where we have a black-box for computing the group operation. Given $a, b \in G$, output the smallest positive t such that $b = a^t$. For example, one can apply this algorithm to finding discrete logarithms in the additive group of an elliptic curve over a finite field [121,150], a group widely used in public key cryptography [137]. In this case, the group operation is described as addition, and thus the problem is to find the smallest positive integer t such that $b = ta$, where b and a are points on some elliptic curve.

Quantum Complexities of the Discrete Logarithm Problem

- Finding discrete logarithms in $\mathbb{F}_q^*$
  - Quantum complexity is in $O((\log q)^2 \log\log(q) \log\log\log(q))$.
- Discrete logarithms in a black-box group represented with strings of length n (including elliptic curve groups discussed above)
  - Quantum black-box complexity (for groups with unique encodings of group elements) is $O(\log r)$ black-box multiplications and $O(n + \log^2 r)$ other elementary operations.

Classical Complexities of the Discrete Logarithm Problem

- Finding discrete logarithms in $\mathbb{F}_q^*$
  - Best known rigorous probabilistic classical algorithm has complexity in $e^{O(\sqrt{\log q \log\log q})}$ for certain values of q (including $q = 2^n$ and prime q).
  - Best known heuristic probabilistic classical algorithm has complexity in $e^{O((\log q)^{\frac{1}{3}}(\log\log q)^{\frac{2}{3}})}$.
- Discrete logarithms in a black-box group represented with strings of length n
  - Classical black-box complexity is in $\Theta(\sqrt{r})$. For a large class of elliptic curves, the best known classical algorithms have complexity in $O(\sqrt{r})$ group additions. There are sub-exponential algorithms for special families of curves.

Abelian Hidden Subgroup Problem

Notice how we can rephrase most of the problems we have already discussed, along with some other ones, as a special case of the following problem.

The Abelian Hidden Subgroup Problem Let $f: G \to X$ map an Abelian group G to some finite set X with the property that there exists some subgroup $K \leq G$ such that for any $x, y \in G$, $f(x) = f(y)$ if and only if $x + K = y + K$. In other words f is constant on cosets of K and distinct on different cosets.

Deutsch's problem $G = \mathbb{Z}_2$, $X = \{0, 1\}$, and $K = \{0\}$ if f is balanced, and $K = \{0, 1\}$ if f is constant.

Generalized Simon's problem $G = \mathbb{Z}_2^n$, $X = \{0, 1\}^n$, and K is any subgroup of $\mathbb{Z}_2^n$.

Finding orders $G = \mathbb{Z}$, X is any finite group H, r is the order of $a \in H$. The subgroup $K = r\mathbb{Z}$ is the hidden subgroup of G, and a generator for K reveals r.

Finding the period of a function $G = \mathbb{Z}$, X is any set, r is the period of f. The subgroup $K = r\mathbb{Z}$ is the hidden subgroup of G, and a generator for K reveals the period r.

Discrete logarithms in any group $G = \mathbb{Z}_r \times \mathbb{Z}_r$, X is any group H. Let a be an element of H with $a^r = 1$ and suppose $b = a^k$. Consider the function $f(x_1, x_2) = a^{x_1} b^{x_2}$. We have $f(x_1, x_2) = f(y_1, y_2)$ if and only if $(x_1, x_2) - (y_1, y_2) \in \{(tk, -t), t = 0, 1, \ldots, r-1\}$. The hidden subgroup K is the subgroup generated by $(k, -1)$, where k is the discrete logarithm.

Hidden linear functions [34] $G = \mathbb{Z} \times \mathbb{Z}$. Let g be some permutation of $\mathbb{Z}_N$ for some integer N. Let h be a function from $\mathbb{Z} \times \mathbb{Z}$ to $\mathbb{Z}_N$ defined by $h(x, y) = x + ay \mod N$. Let $f = g \circ h$. The subgroup K is the hidden subgroup generated by $(-a, 1)$, and the generator reveals the hidden linear function h.

Self-shift-equivalent polynomials [83] Given a polynomial P in l variables $X_1, X_2, \ldots, X_l$ over $\mathbb{F}_q$ (the finite field with q elements), the function f which maps $(a_1, a_2, \ldots, a_l) \in \mathbb{F}_q^l$ to $P(X_1 - a_1, X_2 - a_2, \ldots, X_l - a_l)$ is constant on cosets of a subgroup K of $\mathbb{F}_q^l$. This subgroup K is the set of self-shift-equivalences of the polynomial P.

Abelian stabilizer problem [119] Let G be any group acting on a finite set X. That is, each element of G acts as a map from X to X in such a way that for any two elements $a, b \in G$, $a(b(x)) = (ab)(x)$ for all $x \in X$. For a particular element $x \in X$, the set of elements which fix x (that is the elements $a \in G$ such that $a(x) = x$) form a subgroup. This subgroup is called the stabilizer of x in G, denoted $St_G(x)$. Let f_x denote the function from G to X which maps $g \in G$ to $g(x)$. The hidden subgroup of f_x is $St_G(x)$.

If we restrict attention to finite Abelian groups, or more generally, finitely generated Abelian groups, then we can efficiently solve the hidden subgroup problem, by generalizations of the algorithms for factoring, finding discrete logarithms, and Simon's problem.

The Abelian hidden subgroup problem can also be used to decompose a finite Abelian group into a direct sum of cyclic groups if there is a unique representative for each group element [44,143]. For example, the multiplicative group of integers modulo N is an Abelian group, and we

can efficiently perform computations in the group. However, having a decomposition of the group would imply an efficient algorithm for factoring N. The class group of a number field is another Abelian group for which a decomposition is believed to be hard to find on a classical computer. For example, such a decomposition would easily give the size of the class group, which is also known to be as hard as factoring, assuming the Generalized Riemann Hypothesis. Computing the class group of a complex number field is a simple consequence of the algorithm for decomposing an Abelian group into a direct sum of cyclic groups, since there are techniques for computing unique representatives in this case. A unique classical representation is sufficient, but a quantum state could also be used to represent a group element. For example, a uniform superposition over classical representatives of the same group element would also work (this technique was applied by Watrous in the case of solvable groups [175]). Computing the class number of a real number field is not as straightforward, since there is no known way to efficiently compute a unique classical representative for each group element. However Hallgren [95] used the techniques outlined in Subsect. "Lattice and Number Field Problems" to show how to compute quantum states to represent the group elements, assuming the Generalized Riemann Hypothesis, in the case of class group of a real number field of constant degree, and thus is able to compute the class group in these cases as well .

Quantum Algorithms for the Abelian Hidden Subgroup Problem There exists a bounded-error quantum algorithm for finding generators for the hidden subgroup $K \leq G = \mathbb{Z}_{N_1} \times \mathbb{Z}_{N_2} \times \cdots \times \mathbb{Z}_{N_l}$ of f using $O(l)$ evaluations of f and $O(\log^3 N)$ other elementary operations, where $N = N_1 N_2 \ldots N_l = |G|$.

It can be shown that $\Omega(l)$ queries are needed for worst-case K.

Classical Algorithms for the Abelian Hidden Subgroup Problem In the black-box model $\Omega(\sqrt{|G/K|})$ queries are necessary in order to even decide if the hidden subgroup is trivial.

Generalizations are discussed in Sect. "Generalizations of the Abelian Hidden Subgroup Problem".

Algorithms Based on Amplitude Amplification

In 1996 Grover [90] developed a quantum algorithm for solving the search problem that gives a quadratic speed-up over the best possible classical algorithm.

Search Problem
Input: A black-box U_f for computing an unknown function $f\colon \{0,1\}^n \to \{0,1\}$.
Problem: Find a value $\mathbf{x} \in \{0,1\}^n$ satisfying $f(\mathbf{x}) = 1$, if one exists. Otherwise, output "NO SOLUTION".

The *decision version* of this problem is to output 1 if there is a solution, and 0 if there is no solution.

This problem is stated in a very general way, and can be applied to solving a wide range of problems, in particular any problem in *NP*. For example, suppose one wanted to find a solution to an instance Φ of the 3-SAT problem. The Boolean formula Φ is in "3-conjunctive normal form" (3-CNF), which means that it is a conjunction (logical AND) of clauses, each of which is a disjunction (logical OR) of three Boolean variables (or their negations). For example, the following is a 3-CNF formula in the variables $b_1, b_2, \ldots, b_5$:

$$\Phi = (b_1 \vee \bar{b_2} \vee b_5) \wedge (b_1 \vee \bar{b_4} \vee \bar{b_5}) \wedge (b_4 \vee b_2 \vee b_3)\,,$$

where we let $\bar{b}$ denote the logical NOT of b. A "satisfying assignment" of a particular 3-CNF formula Φ is an assignment of 0 or 1 values to each of the n variables such that the formula evaluates to 1. Given a satisfying assignment, it is easy to check if it satisfies the formula. Define f_Φ to be the function that, for any $\mathbf{x} = x_1 x_2 \ldots x_n \in \{0,1\}^n$, maps $\mathbf{x} \mapsto 1$ if the assignment $b_i = x_i$, $i = 1, 2, \ldots, n$ satisfies Φ and $\mathbf{x} \mapsto 0$ otherwise. Solving the search problem for f_Φ is equivalent to finding a satisfying assignment for Φ.

If we only learn about f by applying the black-box U_f, then any algorithm that finds a solution with high probability for any f (even if we just restrict to functions f with at most one solution) requires $\Omega(\sqrt{2^n})$ applications of U_f [25].

The basic intuition as to why a quantum algorithm might provide some speed-up is that a quantum version of an algorithm that guesses the solution with probability $\frac{1}{2^n}$ will in fact produce the correct answer with probability amplitude $1/\sqrt{2^n}$. The hope is that additional guesses increase the amplitude of finding a correction solution by $\Omega(1/\sqrt{2^n})$, and thus $O(\sqrt{2^n})$ guesses and applications of U_f might suffice in order to get the total probability amplitude close to 1. Of course, one needs to find a unitary algorithm for implementing this, and obvious approaches do not permit the amplitudes to add constructively. Lov Grover discovered a quantum algorithm that does permit the amplitudes to add up in such a way. This algorithm was analyzed and generalized to the technique known as "amplitude amplification" [34,35,38]. Further generalizations

of these search algorithms are discussed in Sect. "Quantum Walk Algorithms".

Quantum Algorithms for Searching: There exists a quantum algorithm that finds a solution with probability at least $\frac{2}{3}$ if one exists, otherwise outputs "NO SOLUTION", and uses $O(\sqrt{2^n})$ applications of U_f. The algorithm uses $O(n\sqrt{2^n})$ elementary gates. Note that $\Omega(\sqrt{2^n})$ applications of U_f are necessary in order to find a solution for any f, even if we restrict attention to functions f with at most one solution.

If, as an additional input, we are also given an algorithm A that will find a solution to $f(x) = 1$ with probability p, then amplitude amplification will find a solution with high probability using $O(1/\sqrt{p})$ applications of U_f and of unitary versions of A and A^{-1}.

This more general statement implies that if there are $m \geq 1$ solutions, then there is an algorithm which makes an expected number of queries in $O(\sqrt{2^n/m})$ (since guessing uniformly at random will succeed with probability $\frac{m}{2^n}$). This algorithm works even if the unitary black-box U_f computes f with a small bounded error [108].

Classical Algorithms for the Search Problem: Any classical algorithm must make $\Omega(2^n)$ queries in order to succeed with probability at least $\frac{2}{3}$ on any input f, even if we restrict attention to functions with at most one solution. Exhaustive searching will find a solution using $O(2^n)$ applications of U_f.

If we are also given a black-box A that successfully guesses a solution to $f(x) = 1$ with probability $p \geq \frac{1}{2^n}$, then $\Theta(\frac{1}{p})$ applications of U_f are needed and also sufficient (using random sampling).

If instead of a black-box for U_f, we are given a circuit for f or some other description of f (such as a description of the 3-SAT formula Φ in the case of f_Φ) then there might be more efficient algorithms that uses this additional information to do more than just use it to apply U_f. One can directly use amplitude amplification to get a quadratic speed-up for any such heuristic algorithm that guesses a solution with probability p, and then repeats until a solution is found.

The quantum searching algorithm has often been called a 'database' search algorithm, but this term can be misleading in practice. There is a quadratic speed-up if there is an implicit database defined by an efficiently computable function f. However, if we are actually interested in searching a physical database, then a more careful analysis of the physical set-up and the resources required is necessary. For example, if a database is arranged in a linear array, then to query an N-element database will take time proportional to N. If the data is arranged in a square grid, then a general lookup could be possibly be done in time proportional to $\sqrt{N}$. This limitation is true for both classical and quantum searching; however the hardware assumptions in the two cases are often different, so one needs to be careful when making a comparison. (e. g. [158,180] and Chap. 6 of [149], discuss these issues).

Local searching [2] restricts attention to architectures or models where subsystems can only interact locally, and does not allow access to any memory location in one unit of time regardless of the memory size (which is not truly possible in any case, but can sometimes be appropriate in practice). The algorithms used include quantum walk algorithms (e. g. [182]) and can achieve quadratic or close to quadratic speed-ups depending on details such as the spatial dimension of the database.

It is also important to note that there has been much work studying the quantum query complexity of searching an *ordered* list. It is known that $\Theta(\log N)$ queries are necessary and sufficient, but the constant factor is not yet known [49,98,107].

Other Applications of Amplitude Amplification

Apart from the application to the searching problem, one of the first applications of the quantum searching algorithm was to counting [37], and more generally amplitude amplification can be used to estimate amplitudes [38]. The quantum algorithm for amplitude estimation combines the techniques of the order finding algorithm with amplitude amplification. Bounds on optimal phase estimation translate to bounds on optimal amplitude estimation. It has several applications, including approximate and exact counting [37,38], and approximating the mean (or, in the continuous case ▶ Quantum Algorithms and Complexity for Continuous Problems, the integral) of a function [172]. These applications offer up to a quadratic speed-up over classical algorithms.

We have already mentioned the straight-forward, and very useful, application of amplitude amplification to searching: providing a quadratic speed-up up for any algorithm that consists of repeating a subroutine that guesses a solution, until a solution is found. However, there are many applications of amplitude amplification that are not so straightforward, and require some additional nontrivial algorithmic insights. Other applications of amplitude amplification include the collision finding problem, which has applications to breaking hash functions in cryptography.

Collision Finding Problem:
Input: A black-box U_f for computing a function

$f\colon \{0,1\}^n \to \{0,1\}^m$. The function f is r to 1, for some positive integer r.
Problem: Find two distinct $\mathbf{x}, \mathbf{y} \in \{0,1\}^n$ with $f(\mathbf{x}) = f(\mathbf{y})$, if $r > 1$. Otherwise output "NO COLLISION".

Quantum Algorithms for Collision Finding There exists a quantum algorithm [36] that uses $O((2^n/r)^{1/3})$ applications of U_f, and $O((2^n/r)^{1/3}$ polylog $(2^n/r))$ other elementary operations, and $O((2^n/r)^{1/3})$ space, and outputs a collision with probability at least $\frac{2}{3}$ if $r > 1$, and outputs "NO COLLISION" otherwise.

It can be shown that $\Omega((2^n/r)^{1/3})$ applications of U_f are needed.

Classical Algorithms for Collision Finding There exists a classical probabilistic algorithm that uses $O((2^n/r)^{1/2})$ applications of U_f, and $O((2^n/r)^{1/2}$ polylog $(2^n/r))$ other elementary operations, and $O(n)$ space, and outputs a collision with probability at least 2/3 if $r > 1$, and outputs "NO COLLISION" otherwise. It can be shown that $\Omega((2^n/r)^{1/2})$ applications of U_f are needed.

Other applications include finding claws [37], finding the maximum (or minimum) value of a function [106], string matching [153], estimating the median of a function [91,148] and many others [67,84]. There is also a non-trivial application to the element distinctness problem [42], which we define in Sect. "Quantum Walk Algorithms", since there is a better, optimal, quantum walk algorithm. Amplitude amplification also a basic tool in making algorithms exact [35,146].

Simulation of Quantum Mechanical Systems

Feynman's reason for considering a quantum computer was to simulate other quantum mechanical systems [79]. This general idea was described in more detail by Lloyd [131], Wiesner [178] and Zalka [180], and later by many other authors (e. g. [149]). More detailed situations have been studied by numerous authors, including [43,45,119,170].

There are other notions of what one what one might mean by simulating a physical system, such as computing properties of the ground state of the Hamiltonian, or other properties of the spectrum of the Hamiltonian. In this section, we focus on the problem of simulating the evolution of a quantum mechanical system given the initial state (or a description of it) and a description of the Hamiltonian of the system.

For convenience, we will restrict attention to a discrete Hilbert space. In practice, when modelling systems with a continuous state space, the state space will need to be approximated with a discrete state space [180] with enough states so that the accumulated error is below the desired error threshold. This general problem of dealing with continuous problems is discussed in ▶ Quantum Algorithms and Complexity for Continuous Problems.

We will also restrict attention to time-independent Hamiltonians, since well-behaved time-dependent Hamiltonian evolution can be approximated by a sequence of time-dependent Hamiltonian evolutions.

There are two natural ways that have been studied for representing the Hamiltonian. The first way, is to represent $H = \sum_{k=1}^{M} H_k$, where H_k is a simple enough Hamiltonian that we know how to efficiently simulate its evolution. For convenience, we assume the simulation of the H_k term for a time t takes unit cost and is exact. In practice, if the simulation can be done with error ϵ in time polynomial in $\frac{1}{\epsilon}$, $\tau = ||H_k||t$, and the number of qubits n, then the effect on the overall cost is by a factor that is polynomial in these factors. In many situations, the simulation is polynomial in $\log(\frac{1}{\epsilon})$ and $\log \tau$, and independent of n and ϵ. This leads to the following formulation:

Hamiltonian Simulation Problem 1
Input: An integer n and black-boxes $A_1, A_2, \ldots, A_M$, where A_j takes a non-negative real number r and executes $e^{iH_j r}$, for a set of Hamiltonians H_j acting on n qubits.
The value of $\|H\|$, the trace norm of $H = \sum_{k=1}^{M} H_k$.
A positive real number t.
A positive real number $\epsilon < 1$.
An n-qubit state $|\psi\rangle$.
Output: A state $|\psi_f\rangle$ satisfying $\| |\psi_f\rangle - e^{iHt}|\psi\rangle \| < \epsilon$.

In practice, the input will likely be a classical description of how to prepare the state $|\psi\rangle$. It suffices to have an upper bound on $\|H\|$ (if the bound is within a constant factor, this won't affect the stated complexities). In particular, it suffices that we can efficiently compute or approximate the eigenvalue λ_k of a given an eigenvector $|E_k\rangle$ of H_k. This is the case, for example, when the Hilbert space is a tensor product of n finite sized subsystems and H_k acts non-trivially only on a finite number of subsystems, say c of them. In other words, up to a reordering of the subsystem labels, $H_k = I^{n-c} \otimes \tilde{H}_k$. Since

$$e^{iH_k t} = I^{n-c} \otimes e^{i\tilde{H}_k t},$$

in order to simulate the evolution of H_k for a time interval of size t, one only needs to simulate $\tilde{H}_k$ on the relevant c qubits for a time interval of size t. In this case, since we know $\tilde{H}_k$, one can use "brute-force" methods to approximately implement the map $|E_{\lambda_k}\rangle \mapsto e^{2\pi i \lambda_k t}|E_{\lambda_k}\rangle$, for any time interval t.

An easy example is when the state space is n qubits, and H_k is a tensor product of Pauli operators. This means that

we can easily diagonalize $\tilde{H}_k$ as $\tilde{H}_k = (P_1 \otimes P_2 \otimes \ldots \otimes P_c)Z^c(P_1^\dagger \otimes P_2^\dagger \otimes \ldots \otimes P_c^\dagger)$ for some one-qubit unitary operations $P_1, P_2, \ldots, P_c$. Thus we have

$$e^{i\tilde{H}_k t} = (P_1 \otimes P_2 \otimes \ldots \otimes P_c)e^{iZ^c t}(P_1^\dagger \otimes P_2^\dagger \otimes \ldots \otimes P_c^\dagger).$$

Since

$$e^{iZ^c t} \mid x_1 x_2 \ldots x_c\rangle = e^{itf(x_1 \ldots x_c)} \mid x_1 x_2 \ldots x_c\rangle$$

where $f(x_1, \ldots, x_c) = x_1 \oplus \ldots \oplus x_c$, this simulation can be done easily. In fact, as pointed out in [149], in the case that H_k is such a product Hamiltonian, c does not need to be constant.

Another example, [180], is where the eigenvectors of H_k are of the form

$$\mid E_k\rangle = \sum_{j=0}^{2^n-1} e^{2\pi i jk/2^n} \mid j\rangle$$

(i. e. "momentum" eigenstates), in which case the inverse quantum Fourier transform will map $\mid E_k\rangle \mapsto \mid k\rangle$, and then one can easily approximate $\mid k\rangle \mapsto e^{i\lambda_k t} \mid k\rangle$ for any efficiently computable function λ_k, and then apply the quantum Fourier transform to map $\mid k\rangle \mapsto \mid E_k\rangle$ and thus effectively compute the transformation $\mid E_k\rangle \mapsto e^{i\lambda_k t} \mid E_k\rangle$.

Thus we can study any situation where $H = \sum_k H_k$, and we have some means of efficiently simulating the time evolution of H_k. If the H_k commute, then

$$e^{iHt} = e^{iH_1 t}e^{iH_2 t} \ldots e^{iH_n t},$$

and the problem is straightforward. For the non-trivial case when the H_k do not commute pairwise, we can take advantage of approximations derived from Trotter formulas, like

$$e^{iHt} = \left(e^{iH_1 t/n}e^{iH_2 t/n} \ldots e^{iH_n t/n}\right)^n + O(t^2/n)$$

and other improved versions. A good estimate of the overall cost of this family of simulations is the number of terms of the form $e^{iH_j r}$ that are used, for any choice of r (we can assume $0 \le r \le t$, and that r is some efficiently computable number).

Quantum Complexity of Hamiltonian Simulation Problem 1 There is a quantum algorithm that simulates e^{iHt} on a given input $\mid \psi\rangle$ with trace distance error at most $\epsilon < 1$ that uses a number of exponential terms $N_{\exp}$ satisfying

$$N_{\exp} \in (M\tau + 1)M^{1+o(1)}\tau^{o(1)}\left(\frac{1}{\epsilon}\right)^{O(1/\sqrt{s})},$$

where $\tau = \|H\|\, t$ and $\epsilon = 1/2^s$. Since the running time of simulating each exponential term is assumed to be polynomial in n, the overall running time is polynomial in n.

In general, there are Hamiltonians for which $\Omega(\tau)$ time is necessary.

The product $\tau = \|H\|\, t$ is the relevant parameter since one can effectively speed up time by a factor of s by rescaling the Hamiltonian by a multiplicative factor of s, for any $s > 0$ (e. g. one can perform an operation in half the time by doubling all the energies).

It is hard to give an explicit comparison to the best known classical complexity of this general problem. There are classical algorithms and heuristics for special cases, but in the worst-case, the best solutions known require time exponential in n and polynomial in τ (e. g. in $\tau^{1+o(1)}$).

Another natural formulation of the problem [5,29] considers the case of sparse Hamiltonians, where for any basis state $\mid x\rangle$ there are at most d basis states $\mid y\rangle$ such that $\langle x \mid H \mid y\rangle \neq 0$ and one can efficiently compute this neighborhood for any input x. This leads to the following black-box formulation of the problem.

Hamiltonian Simulation Problem 2
Input: Integers n and d, and a black-box U_H that maps $\mid x, i\rangle \mid 0\rangle \mapsto \mid x, i\rangle \mid y_i, H_{x,y_i}\rangle$, where y_i is the index of the ith non-zero entry in column x of the Hermitian matrix H (if there are $d' < d$ non-zero entries, then U_f can output any index for $i > d'$). The values $H_{x,y} = \langle x \mid H \mid y\rangle$ are the (x, y) entries of H represented in the computational basis.
A positive real number t.
A positive real number $\epsilon < 1$.
An n-qubit state $\mid \psi\rangle$.
Output: A state $\mid \psi_f\rangle$ satisfying $\| \mid \psi_f\rangle - e^{iHt} \mid \psi\rangle \| < \epsilon$.

The best known general solution was shown in [29].

Quantum Complexity of Hamiltonian Simulation Problem 2 There is a quantum algorithm that simulates e^{iHt} on a given input $\mid \psi\rangle$ with trace distance error at most $\epsilon < 1$ that uses a number of black-box calls, N_{bb}, satisfying, for any positive integer k,

$$N_{\text{bb}} \in O\left(\tau^{(1+\frac{1}{2k})}d^{(4+\frac{1}{k})}5^{2k}\left(\frac{1}{\epsilon}\right)^{\frac{1}{2k}}(\log^* n)\right),$$

where $\tau = ||H||t$ and $\log^* n$ is the smallest positive integer r such that $\log_2^{(r)} n < 2$, where $\log_2^{(r)}$ refers to iterating the $\log_2$ function r times.

In general, there are Hamiltonians that require $\Omega(\tau)$ black-box evaluations.

For example, setting $k = \sqrt{s}$ where $\epsilon = \frac{1}{2^s}$, we get

$$N_{\mathrm{bb}} \in \tau^{(1+O(\frac{1}{\sqrt{s}}))} d^{(4+O(\frac{1}{\sqrt{s}}))} \left(\frac{1}{\epsilon}\right)^{O\left(\frac{1}{\sqrt{s}}\right)} (\log^* n) .$$

Some applications involve trying to study properties of the ground state of a Hamiltonian by combining simulation of the dynamics with some approach for generating the ground state with non-trivial initial amplitude. For example, if one had a procedure A for generating a ground state $|E_0\rangle$ of H with probability p, then one can combine the algorithm for simulating H with eigenvalue estimation, and then amplitude amplification on the eigenvalue estimates in order to generate the ground state with high fidelity. The algorithm would use $O(1/\sqrt{p})$ applications of A and A^{-1} and $O(1/\sqrt{p})$ eigenvalue estimations. The eigenvalue estimations would have to be precise enough to distinguish $|E_0\rangle$ from the next highest energy eigenstate produced by A. If the gap between these two energies is Δ, then the eigenvalue estimation would involve simulating the evolution of H for a time in $\Omega(\frac{1}{\Delta})$.

Several papers address techniques for generating the ground states for various problems of interest including [3]. Another application for such simulations is to implement algorithms that are designed in the continuous-time model in a discrete-time quantum computation model.

Generalizations of the Abelian Hidden Subgroup Problem

Non-Abelian Hidden Subgroup Problem

One of the most natural ways to generalize the Abelian hidden subgroup problem is to non-Abelian groups. The problem definition is the same, apart from letting G be non-Abelian. One natural problem in this class of problems is the graph automorphism problem.

Graph automorphism problem: Consider $G = S_n$, the symmetric group on n elements, which corresponds to the permutations of $\{1, 2, \ldots, n\}$. Let $\mathbb{G}$ be a graph on n vertices labeled $\{1, 2, \ldots, n\}$. For any permutation $\sigma \in S_n$, let $f_{\mathbb{G}}$ map S_n to the set of n-vertex graphs by mapping $f_{\mathbb{G}}(\sigma) = \sigma(\mathbb{G})$, where $\sigma(\mathbb{G})$ is the graph obtained by permuting the vertex labels of $\mathbb{G}$ according to σ. For the function $f_{\mathbb{G}}$, the hidden subgroup of G is the automorphism group of $\mathbb{G}$ (i. g. the permutations σ such that $\sigma(\mathbb{G}) = \mathbb{G}$).

The graph isomorphism problem (deciding if two graphs $\mathbb{G}_1$ and $\mathbb{G}_2$ are isomorphic) can be reduced to solving the graph automorphism problem. The best known classical algorithm takes time in $e^{O(\sqrt{n \log n})}$ to decide if two graphs are isomorphic, and there is no substantially better quantum algorithm known.

There has been much work attacking the non-Abelian hidden subgroup problem. Ettinger, Høyer and Knill [71] showed the following.

Query Complexity of the Non-Abelian Hidden Subgroup Problem For any finite group G, the non-Abelian hidden subgroup problem can be solved with high probability using $O(\log |G|)$ queries to U_f.

Thus, the main question remaining is whether it is possible for the entire algorithm, including black-box queries and other elementary operations, to be efficient.

Quantum Fourier Transform Approaches One natural approach is to mimic the algorithm for the Abelian HSP, which starts by computing

$$\frac{1}{\sqrt{|G|}} \sum_{x \in G} |x\rangle |f(x)\rangle$$

and noticing that this equals

$$\sum_{y \in f(G)} \frac{1}{\sqrt{|f(G)|}} \left(\frac{1}{\sqrt{|K|}} \sum_{x \in f^{-1}(y)} |x\rangle \right) |y\rangle .$$

For convenience, we suppose we measure the second register (it suffices just to trace out the 2nd register) and get some random outcome y, and thereby project the first register into the state

$$|a + K\rangle = \frac{1}{\sqrt{|K|}} \sum_{x \in K} |a + x\rangle$$

where a is any element of G such that $f(a) = y$. In other words, as in the Abelian HSP algorithm, the first register is an equal superposition of all the elements in some coset of K.

The question is: how do we extract information about K given a random coset state of K? In the Abelian case, a quantum Fourier transform of the coset state allowed us to sample elements that were orthogonal to K, which we illustrated in the case of Simon's algorithm.

A more general way of looking at the quantum Fourier transform of an Abelian group $G = \mathbb{Z}_{N_1} \times \mathbb{Z}_{N_2} \times \cdots \times \mathbb{Z}_{N_l}$ is in terms of representation theory. The Abelian group G can be represented by homomorphisms ρ that maps G to the complex numbers. There are in fact $|G|$ such homomorphisms, in a natural one-to-one correspondence with the elements of G. For each $g = (a_1, a_2, \ldots, a_l) \in G$, define ρ_g to be the homomor-

phism that maps any $h = (b_1, b_2, \ldots, b_l) \in G$ according to

$$\rho_g(h) = \mathrm{e}^{2\pi i \sum_{i=1}^{l} \frac{a_i b_i}{N_i}} .$$

Using these definitions, we derive the following quantum Fourier transform maps:

$$|\, g\rangle \mapsto \frac{1}{\sqrt{|G|}} \sum_{h \in G} \rho_g(h)\, |\, h\rangle .$$

It is easy to verify that the quantum Fourier transform $\mathrm{QFT}_{N_1} \otimes \mathrm{QFT}_{N_2} \otimes \ldots \otimes \mathrm{QFT}_{N_l}$ maps a coset state of a subgroup K to a superposition of labels h where K is in the kernel of ρ_h, that is, $\rho_h(k) = 1$ for all $k \in K$. Thus, after applying the quantum Fourier transform, one will only measure labels $h = (b_1, b_2, \ldots, b_l)$ such that

$$\rho_g(h) = \mathrm{e}^{2\pi i \sum_{i=1}^{l} \frac{a_i b_i}{N_i}} = 1$$

which generalizes the notion of orthogonality defined in the explanation of Simon's algorithm in a very natural way, since it means $\sum_{i=1}^{l} \frac{a_i b_i}{N_i} = 0 \mod 1$. This gives us a linear equation that must be satisfied by all $h = (b_1, b_2, \ldots, b_l) \in K$, and thus after $l + O(1)$ such random equations are collected, we can efficiently find generators for K by solving a linear system of equations.

One can also define representations for finite non-Abelian groups G, except in order to fully capture the structure of G, we allow homomorphisms ρ to invertible matrices over $\mathbb{C}$ (in the Abelian case, we only need 1×1 matrices). For any finite group G, one can define a finite set of such homomorphisms $\rho_1, \rho_2, \ldots, \rho_k$ that map elements of G to unitary matrices of dimension $d_1 \times d_1$, $d_2 \times d_2, \ldots, d_k \times d_k$, respectively, with the property that $\sum_i d_i^2 = |G|$, and any other such representation is equivalent to a representation that can be factored into a collection of some number of these representations. More precisely, any representation $\rho\colon G \to M_{d\times d}(\mathbb{C})$ has the property that there exists some invertible $P \in M_{d\times d}(\mathbb{C})$, and a list of $\rho_{j_1}, \rho_{j_2}, \ldots$ such that $\sum_i d_i = d$ and for every $g \in G$

$$P\rho(g)P^{-1} = \oplus_i \rho_{j_i}(g)$$

(that is, $P\rho(g)P^{-1}$ is block-diagonal, with the matrices $\rho_{j_i}(g)$ along the diagonals). Furthermore, the representations ρ_j are irreducible in the sense that they cannot be decomposed into the sum of two or more smaller representations in this way. They are also unique up to conjugation.

Since $\sum_i d_i^2 = |G|$, one can define a unitary transformation from G to the vector space spanned by the labels of all the entries $\rho_i(j,k)$ of these irreducible representations ρ_i. In particular, we can map

$$|\, g\rangle \mapsto \sum_i \sqrt{\frac{d_i}{|G|}} \sum_{0 \le j,k \le d_i} \rho_i(j,k)(g)\, |\, i, j, k\rangle$$

where $\rho_i(j,k)(g)$ is the value of the (j,k) entry in the matrix $\rho_i(g)$. Such a mapping is unitary and is called a quantum Fourier transform for the group G. There is freedom in the specific choice of ρ_i within the set of unitary representations that are conjugate to ρ_i, so there is more than one quantum Fourier transform.

There has been much study of such quantum Fourier transforms for non-Abelian groups, which are sometimes possible to implement efficiently [21,96,140,151,156], but efficient constructions are not known in general. It appears they are of limited use in solving the non-Abelian hidden subgroup, except in special cases [68,87,96,156] such as when K is a normal subgroup of G.

In the next sections we discuss several other lines of attack on the non-Abelian hidden subgroup that have yielded some partial progress on the problem.

"Sieving" Kuperberg [130] introduced a method for attacking the hidden subgroup problem for the dihedral group that leads to a sub-exponential algorithm.

The dihedral group D_N is a non-Abelian group of order $2N$, which corresponds to the set of symmetries of a regular N-gon. It can be defined as the set of elements $\{(x,d) \mid x \in \{0,1,2,\ldots,N-1\}, d \in \{0,1\}\}$, where $\{(x,0) \mid x \in \{0,1,2,\ldots,N-1\}\}$ is the Abelian subgroup of D_N, corresponding to the rotations (satisfying $(x,0)+(y,0) = (x+y \mod N, 0)$), and $\{(0,0),(y,1)\}$ are Abelian subgroups of order 2 corresponding to reflections. In general, $(x,0)+(y,1) = (y-x,1)$, $(y,1)+(x,0) = (x-y,1) = -((x,0)+(y,1))$. If the hidden subgroup is a subgroup of the Abelian subgroup of order N, then finding the hidden subgroup easily reduces to the Abelian hidden subgroup problem. However, there is no known efficient algorithm for finding hidden subgroups of the form $\{(0,0),(y,1)\}$. So we can focus attention to the following restricted version of the dihedral hidden subgroup.

Dihedral Hidden Subgroup Problem (Hard Case)
Input: An integer N, and a black-box implementing $U_f\colon |\, x,d\rangle\,|\,0\rangle \mapsto |\, x,d\rangle\,|\, f(x,d)\rangle$, where $f(x_1,d_1) = f(x_2,d_1)$ if and only if $(x_1,d_1)-(x_2,d_2) \in \{(0,0),(y,1)\}$ for some y.
Problem: Find y.

As mentioned earlier, the dihedral hidden subgroup problem can be efficiently reduced to the following phase estimation problem [70].

Ettinger-Høyer Phase Estimation Problem for the Dihedral HSP
Input: An integer N, and a black-box O_d that outputs a classical value $k \in \{0, 1, \ldots, N-1\}$ uniformly at random, along with the qubit $1/\sqrt{2}(|0\rangle + e^{i\phi k}|1\rangle)$ where $\phi = (2\pi d)/N$, for some integer $d \in \{0, 1, 2, \ldots, N-1\}$.
Problem: Find d.

Note that if we could sample the values $k = 2^j$ for $j = 1, 2, \ldots, \lceil \log N \rceil$, then the phase estimation problem can be solved directly using the quantum Fourier transform [54].

Regev designed a clever method for generating states of the form $1/\sqrt{2}(|0\rangle + e^{i2^j\phi}|1\rangle)$ using $O(1)$ calls to the black-box O_d, given an oracle that solves the subset sum problem on average. Kuperberg [130] developed a "sieving" method of generating states of the form $1/\sqrt{2}(|0\rangle + e^{i2^j\phi}|1\rangle)$ and the method was refined and improved to use less memory by Regev [154].

Quantum Algorithms for the Dihedral HSP There exists a quantum algorithm that solves the dihedral HSP with running time in $e^{O(\sqrt{\log N})}$ and uses space in $e^{O(\sqrt{\log N})}$. There is also an algorithm with running time in $e^{O(\sqrt{\log N \log\log N})}$ and uses space in $O(\log N)$.

Classical Algorithms for the Dihedral HSP The classical complexity of the dihedral hidden subgroup problem is $\Theta(\sqrt{N})$ evaluations of the function f.

Similar sieving methods were applied [11] to yield a subexponential time algorithm for the HSP over the product groups G^n for a fixed non-Abelian group G. It has also been show that these kinds of quantum sieve algorithms will not give efficient quantum algorithms for graph isomorphism [141].

"Pretty Good Measurements" A natural approach to solving the non-Abelian hidden subgroup problem is to prepare several instances of a random coset state for the hidden subgroup K, and then try to determine what K is. More precisely, after preparing

$$\sum_{x\in G} |x\rangle |f(x)\rangle = \sum_{y+K\in G/K} |y+K\rangle |f(y)\rangle$$

and discarding the second register, we are left with the mixed state

$$\rho_K = \sum_{y+K\in G/K} |y+K\rangle\langle y+K| .$$

Thus one could try to implement or approximate the optimal quantum measurement for identifying the mixed states ρ_K, over all possible hidden subgroups $K \le G$. Furthermore, one could sample the state ρ_K a polynomial number of times t, and try to guess K given $\rho_K \otimes \rho_K \otimes \cdots \otimes \rho_K = \rho_K^t$.

Holevo [102] determined the optimal measurement for the following general state distinguishability problem. Given $\rho \in \{\rho_j\}$, output a label m such that the probability that $\rho = \rho_m$ is maximum. Let p_j denote the probability that $\rho = \rho_j$. Holevo proved that the maximum probability of guessing the correct input state ρ is achieved by a POVM with elements $\{G_j\}$ satisfying the following conditions: $\sum_i p_i\rho_i G_i = \sum_i p_i G_i \rho_i$ and $\sum_i p_i\rho_i G_i \ge p_j\rho_j$. However, it is not in general easy to efficiently find and implement such a POVM.

Hausladen and Wootters [99] defined a 'pretty good' measurement for distinguishing quantum states that is not necessarily optimal, but has a simpler form. The measurement used POVM elements $G_j = T^{-1/2}\rho_j T^{-1/2}$ and, if these don't sum to the identity, also $I - \sum_j G_j$, where $T = \sum_i \rho_i$. For the case of the dihedral hidden subgroup problem, it was shown [19] that the pretty good measurement is in fact optimal; however, in this case, it is still not known how to efficiently implement the pretty good measurement. However, it was later shown how to implement the pretty good measurement for the Heisengroup group HSP [19].

For example, in the case of the dihedral hidden subgroup problem for D_{2N}, after a quantum Fourier transform on each of n coset states, one can view the resulting state of n registers as

$$(|0\rangle + e^{2\pi i \frac{k_1 d}{N}}|1\rangle)|k_1\rangle \otimes (|0\rangle + e^{2\pi i \frac{k_2 d}{N}}|1\rangle)|k_2\rangle \otimes \ldots \otimes (|0\rangle + e^{2\pi i \frac{k_n d}{N}}|1\rangle)|k_n\rangle$$

for k_i selected independently and uniformly at random from $\{0, 1, \ldots, N-1\}$.

The $(|0\rangle + e^{2\pi i \frac{k_1 d}{N}}|1\rangle) \otimes (|0\rangle + e^{2\pi i \frac{k_2 d}{N}}|1\rangle) \otimes \ldots \otimes (|0\rangle + e^{2\pi i \frac{k_n d}{N}}|1\rangle)$ part of the state can be rewritten as

$$\sum_r \alpha_r |S_r\rangle$$

where r spans all the possible sums of the form $\sum_i b_i k_i$, $b_i \in \{0, 1\}$ (the 'subset sums'), $|S_r\rangle$ is the uniform superposition of the strings $|b_1 b_2 \ldots b_n\rangle$ that satisfy $\sum_{b_i k_i} = r$, and α_r is the appropriate normalization factor (i. e. $\sqrt{n_r/2^n}$ where n_r is the number of solutions to $\sum_{b_i k_i} = r$).

The optimal measurement [19] (in the restricted case of order two subgroups) can be thought of as mapping $|S_r\rangle \mapsto |r\rangle$, performing an inverse quantum Fourier transform and then measuring a value $\tilde{d}$, which will be the guess of the value d (interestingly, a similar measurement is optimal [61] in the case of trying to optimally estimate an arbitrary phase parameter $\phi \in [0, 2\pi)$ given the state $(|0\rangle + e^{ik_1\phi}|1\rangle) \times (|0\rangle + e^{i\phi k_2}|1\rangle) \times \ldots \times (|0\rangle + e^{i\phi k_n}|1\rangle)$). Note that implementing such a basis change in reverse would solve the subset sum problem (which is NP-hard). In fact, it suffices to solve the subset sum problem on average [154]. A nice discussion of this connection can be found in [19].

For groups that are semidirect products of an Abelian group and a cyclic group, the pretty good measurement corresponds to solving what is referred to [19] as a 'matrix sum problem', which naturally generalizes the subset sum problem. They also show that the pretty good measurements are optimal in these cases, and similarly relate their implementation to solving the matrix sum problem to certain average-case algebraic problems. They show that the pretty good measurement can be implemented for several groups, including semidirect product groups of the form $\mathbb{Z}_p^r \rtimes \mathbb{Z}_p$ for constant r (when $r = 2$, this is the Heisenberg group), and of the form $\mathbb{Z}_N \rtimes \mathbb{Z}_p$ with p prime (which are *metacyclic*) and where the ratio N/p is polynomial in $\log N$.

Other Methods and Results There are also some algorithms for solving other cases of the non-Abelian hidden subgroup problem that don't use any of the above techniques, e. g. [82,110,111]. These results use sophisticated classical and quantum group theoretic techniques to reduce an instance of a non-Abelian HSP to instances of HSP in Abelian groups.

One of the most recent results [112] shows that such a reduction is possible for *nil-2* groups, which are nilpotent groups of *class* 2. The group G is nilpotent of class n if the following holds[1]. Let $A_1 = G$, and let $A_{i+1} = [A_i, G]$, for $i > 0$. A group G is nilpotent if $A_{n+1} = \{1\}$, for some integer n, and the class of a nilpotent group is the smallest positive integer n for which $A_{n+1} = \{1\}$.

One of their techniques is to generalize Abelian HSP to a slightly more general problem, where the hidden subgroup K is hidden by a quantum procedure with the following properties. For every $g_1, g_2, \ldots, g_N \in G$ the algorithm maps

$$|g_1\rangle|g_2\rangle \ldots |g_N\rangle|0\rangle|0\rangle \ldots |0\rangle \mapsto |g_1\rangle|g_2\rangle \ldots |g_N\rangle \left|\psi_{g_1}^1\right\rangle \left|\psi_{g_2}^2\right\rangle \ldots \left|\psi_{g_N}^N\right\rangle$$

where the set of states $\{\left|\psi_g^i\right\rangle \mid g \in G\}$ is a *hiding set* for K, for each $i = 1, 2, \ldots, N$. A set of normalized states $\{\left|\psi_g\right\rangle \mid g \in G\}$ is a *hiding set* for the subgroup K of G if

- If g and h are in the same left coset of K then $|\psi_g\rangle = |\psi_h\rangle$.
- If g and h are in different left cosets of K then $\langle\psi_g|\psi_j\rangle = 0$.

Generators for the subgroup K of G can be found in time polynomial in $\log|G|$ using a quantum hiding procedure with $N \in O(\log|G|)$. They find a series of non-trivial reductions of the standard HSP in nil-2 groups to instances of the Abelian HSP with a quantum hiding function.

Lattice and Number Field Problems

The Abelian hidden subgroup problem also works for finitely generated groups G. We can, thus, define the hidden subgroup problem on $G = \mathbb{Z} \times \mathbb{Z} \times \ldots \times \mathbb{Z} = \mathbb{Z}^n$. The hidden subgroup K will be generated by some n-tuples in $\mathbb{Z}^n$. We can equivalently think of G as a lattice and K as a sublattice. The function $f: G \to X$, for some finite set X, that satisfies $f(\mathbf{x}) = f(\mathbf{y})$ if and only if $\mathbf{x} - \mathbf{y} \in K$, can be thought of as hiding the sublattice K.

By generalizing the problem to hiding sublattices of $\mathbb{R}^n$, one can solve some interesting and important number theoretic problems. The solutions in these cases were not a simple extension of the Abelian hidden subgroup algorithm.

Hallgren [93,95] found a quantum algorithm for finding the integer solutions x, y to Pell's equation $x^2 - dy^2 = 1$, for any fixed integer d. He also found an efficient quantum algorithm for the principal ideal problem, and later generalized it to computing the unit group of a number field of constant degree [94]. Solving Pell's equation is known to be at least as hard as factoring integers. We don't have room to introduce all the necessary definitions, but we'll sketch some of the main ideas.

A number field F can be defined as a subfield $Q(\theta)$ of the complex numbers that is generated by the rationals Q together with a root θ of an irreducible polynomial with rational coefficients; the degree of F is the degree of the irreducible polynomial. The "integers" of F are the elements of F that are roots of *monic* polynomials (i. e. polynomials with leading coefficient equal to 1, such as $x^2 + 5x + 1$). The integers of F form a ring, denoted $\mathcal{O}$.

[1] For any two elements g, h of a group, we define their commutator, denoted $[g, h]$ to be $[g, h] = g^{-1}h^{-1}gh$, and for any two subgroups $H, K \leq G$ we define $[H, K]$ to be the (normal) subgroup of G generated by all the commutators $[h, k]$ where $h \in H, k \in K$.

One can define a parameter Δ called the *discriminant* of F (we won't define it here), and an algorithm is considered efficient if its running time is polynomial in $\log \Delta$ and n. The unit group of the ring F, denoted $\mathcal{O}^*$, is the set of elements α in $\mathcal{O}$ that have an inverse $\alpha_{-1} \in \mathcal{O}$. "Computing" the unit group corresponds to finding a description of a system of "fundamental" units, $\epsilon_1, \epsilon_2, \ldots, \epsilon_r$ that generate $\mathcal{O}^*$ in the sense that every unit $\epsilon \in \mathcal{O}^*$ can be written as $\epsilon = \mu \epsilon_1^{k_1} \epsilon_2^{k_2} \ldots \epsilon_r^{k_r}$ for some $k_1, k_2, \ldots, k_r \in \mathbb{Z}$ and some root of unity μ. However, in general, an exact description of a fundamental unit requires an exponential number of bits. There are are some finite precision representations of the elements of the unit group, such as the "Log" embedding into $\mathbb{R}^r$. This representation describes a unit α by an element of $\mathbb{R}^r$ where some finite precision representation of each coordinate suffices. This Log representation of the unit group $\mathcal{O}^*$, corresponds to a sublattice $L = \mathcal{O}^*$ of $\mathbb{R}^r$. Hence, we have a relationship between several important computational number field problems, and lattice problems.

By the above correspondence between $\mathcal{O}^*$ and the lattice $L \subset \mathbb{R}^r$, we can formulate [94,162] the problem of computing the unit group as the problem of finding elements that approximate generators of the sublattice L of $\mathbb{R}^r$. One important non-trivial step is defining a function $f\colon \mathbb{R}^r \to X$ (for some infinite set X) such that $f(x) = f(y)$ if and only if $x - y \in L$ as well as appropriate discrete approximations to this function. The definition of these functions involves substantial knowledge of algebraic number theory, so we will not describe them here.

By designing quantum algorithms for approximating generators of the lattice L, [94,162] one can find polynomial time algorithms for computing the unit group $\mathcal{O}^*$ of an algebraic number field $F = \mathbb{Q}(\theta)$.

A corollary of this result, is a somewhat simpler solution to the principal ideal problem (in a constant degree number field) that had been found earlier by Hallgren [93]. An ideal $\mathcal{I}$ of the ring $\mathcal{O}$ is a subset of elements of $\mathcal{O}$ that is closed under addition, and is also closed under multiplication by elements of $\mathcal{O}$. An ideal $\mathcal{I}$ is *principal* if it can be generated by one element $\alpha \in \mathcal{I}$; in other words $\mathcal{I} = \alpha\mathcal{O} = \{\alpha\beta \mid \beta \in \mathcal{O}\}$. The principal ideal problem is, given generators for $\mathcal{I}$, to decide if $\mathcal{I}$ is principal, and if it is, to find a generator α.

As mentioned in Subsect. "Abelian Hidden Subgroup Problem", the tools developed can also be applied to find unique (quantum) representatives of elements of the class group for constant degree number fields [95] (assuming the generalized Riemann hypothesis), and thus allow for the computation of the class group in these cases.

Hidden Non-linear Structures

Another way to think of the Abelian hidden subgroup problem is as an algorithm for finding a hidden linear structure within a vector space. For simplicity, let's consider the Abelian HSP over the additive group $G = \mathbb{F}_q \times \mathbb{F}_q \times \ldots \times \mathbb{F}_q = \mathbb{F}_q^n$, where $q = p^m$ is a prime power. The elements of G can also be regarded as a vector space over $\mathbb{F}_q$. A hidden subgroup $H \leq G$ corresponds to a subspace of this vector space and its cosets correspond to parallel affine subspaces or *flats*. The function f is constant on these linear structures within the vector space $\mathbb{F}_q^n$.

A natural way to generalize this [51] is to consider functions that are constant on sets that are defined by non-linear equations. One problem they study is the *hidden radius problem*. The circle of radius $r \in \mathbb{F}_q$ centered at $t = (t_1, t_2, \ldots, t_n) \in \mathbb{F}_q^n$ is the set of points $x = (x_1, x_2, \ldots, x_n) \in \mathbb{F}_q^n$ that satisfy $\sum_i (x_i - t_i)^2 = r$. The point x on a circle centered at t will be represented as either $(x, \sigma(s))$, where $s = x - t$ and $\sigma(s)$ is a random permutation of s, or as $(x, \tau(t))$ where $\tau(t)$ is a random permutation of t. We define f_1 and f_{-1} be the functions satisfying $f_1(x, \sigma(s)) = \tau(t)$ and $f_{-1}(x, \tau(t)) = \sigma(s)$.

Hidden Radius Problem
Input: A black box U_{f_1} that implements $|x\rangle\,|\sigma(s)\rangle\,|0\rangle \mapsto |x\rangle\,|\sigma(s)\rangle\,|f_1(x, \sigma(s))\rangle$, and a black box $U_{f_{-1}}$ that implements $|x\rangle\,|\tau(t)\rangle\,|0\rangle \mapsto |x\rangle\,|\tau(t)\rangle\,|f_{-1}(x, \tau(t))\rangle$.
Problem: Find r.

Quantum Algorithms for the Hidden Radius Problem For odd d, there is a bounded error quantum algorithm that makes polylog(q) queries to U_{f_1} and $U_{f_{-1}}$ and finds r. However, there is no known polynomial bound on the non-query operations.

There is a bounded-error quantum algorithm that also makes polylog(q) operations in total, and determines $\chi(r)$, where $\chi(r) = 1$ if r is a quadratic residue (that is, $r = u^2$ for some $u \in \mathbb{F}_q$) and 0 otherwise.

Classical Algorithms for the Hidden Radius Problem It was shown in [51] that the expected number of queries needed to be able to guess any bit of information about r correctly with probability greater than $\frac{1}{2} + \frac{1}{\text{poly}(d \log q)}$ is exponential in $d \log q$.

A number of other black-box problems of this kind were defined in [51] with quantum algorithms that are exponentially more efficient than any classical algorithm, in some cases just in terms of query complexity, and other times in terms of all operations. These problems fit into the frameworks of *shifted subset* problems and *hidden polynomial* problems. They use a variety of non-trivial tech-

niques for these various problems, including the quantum Fourier transform, quantum walks on graphs, and make some non-trivial connections to various *Kloosterman sums*. Further work along these lines has been done in [63].

Hidden Shifts and Translations

There have been a variety of generalizations of the hidden subgroup problem to the problem of finding some sort of hidden shift or translation.

Grigoriev [89] addressed the problem of the shift-equivalence of two polynomials (the self-shift-equivalence problem is a special case of the Abelian hidden subgroup problem). Given two polynomials P_1, P_2 in l variables $X_1, X_2, \ldots, X_l$ over $\mathbb{F}_q$ (the finite field with q elements), does there exist an element $(a_1, a_2, \ldots, a_l) \in \mathbb{F}_q^l$ such that $P_1(X_1 - a_1, X_2 - a_2, \ldots, X_l - a_l) = P_2(X_1, X_2, \ldots, X_l)$. More generally, if there is a group G ($\mathbb{F}_q^l$ in this case) acting on a set X (the set of polynomials in l variables over $\mathbb{F}_q$ in this case), one can ask if two elements $x, y \in X$ are in the same orbit of the action of G on X (that is, if there is a $g \in G$ such that $g(x) = y$). In general, this seems like a hard problem, even for a quantum computer.

The dihedral hidden subgroup problem [70] is a special case of the *hidden translation problem* [82], where there is a finite group G, with unique representation of the elements of G, and two injective functions f_0 and f_1 from G to some finite set X.

Hidden Translation Problem:
Input: Two black boxes U_{f_0} and U_{f_1} that, for any $x \in G$, implement the maps $U_{f_0}: |x\rangle|0\rangle \mapsto |x\rangle|f_0(x)\rangle$ and $U_{f_1}: |x\rangle|0\rangle \mapsto |x\rangle|f_1(x)\rangle$.
A promise that $f_1(x) = f_0(ux)$, for some $u \in G$.
Problem: Find u.

The same problem expressed with additive group notation has been called the *hidden shift problem*, and instances of this problem were solved efficiently by van Dam, Hallgren and Ip [59]. For example, they find an efficient solution in the case that $f_1(x) = \chi(x + s)$ where $f_0 = \chi$ is a multiplicative character function over a finite field, which implies a method for breaking a class of "algebraically homomorphic" cryptosystems. They also describe a more general *hidden coset problem*.

In [82] it is shown how to solve the hidden translation problem in $G = \mathbb{Z}_p^n$ in polynomial time, and then show how to use this to solve the problem for any group that they call "smoothly solvable". Let us briefly define what this means.

The *derived subgroup* of a group G is $G^{(1)} = [G, G]$. In general, we define $G^{(0)} = G, G^{(n+1)} = [G^{(n)}, G^{(n)}]$, for $n \geq 1$. A group G is solvable if $G^{(n)} = \{1\}$, the trivial group, for some positive integer n, and the series of subgroups is called the *derived series* of G. A group G is called smoothly solvable if m is bounded above by a constant, and if the factor groups $G^{(j+1)}/G^{(j)}$ are isomorphic to a direct product of a group of bounded exponent (the exponent of a group is the smallest positive integer r such that $g^r = 1$ for all g in the group) and a group of size polynomial in $\log |G|$.

The algorithm for smoothly solvable groups works by solving a more general *orbit coset* problem, for which they prove a "self-reducibility" property. In particular, orbit coset problem for a finite group G is reducible to the orbit coset problem in G/N and N, for any solvable normal subgroup N of G.

Quantum Algorithms for the Hidden Translation Problem: For groups $G = \mathbb{Z}_p^n$, the hidden translation problem can be solved with bounded probability using $O(p(n + p)^{p-1})$ queries and $(n + p)^{O(p)}$ other elementary operations.

In general, for *smoothly solvable* groups G, the hidden translation problem can also be solved with a polynomial number of queries and other elementary operations. Another consequence of the tools they developed is a polynomial time solution to the hidden subgroup for such smoothly solvable groups.

Classical Algorithms for the Hidden Translation Problem: In general, including the case $G = \mathbb{Z}_p^n$, the hidden translation problem requires $\Omega(\sqrt{|G|})$ queries on a classical computer.

Another natural generalization of the hidden translation or hidden shift problem and the Abelian hidden subgroup problem is the *generalized hidden shift problem* introduced in [46]. There is a function $f: \{0, 1, 2, \ldots, M - 1\} \times \mathbb{Z}_N \to X$ for some finite set X, with the property that for a fixed $b \in \{0, 1, \ldots, M - 1\}$, the mapping $x \mapsto f(b, x)$ is one-to-one, and there is some hidden value $s \in \mathbb{Z}_N$ such that $f(b, x) = f(b + 1, x + s)$ for all $b \in \{0, 1, \ldots, M - 2\}$. Note that for $M = 2$, this is equivalent to the dihedral hidden subgroup problem for the group D_N, and for $M = N$, this problem is equivalent to the Abelian hidden subgroup problem for the hidden subgroup $\langle(1, s)\rangle$ of the group $\mathbb{Z}_N \times \mathbb{Z}_N$.

Generalized Hidden Shift Problem:
Input: Positive integers M and N.
A black boxes U_f that maps $|b, x\rangle|0\rangle \mapsto |b, x\rangle|f(b, x)\rangle$ for all $b \in \{0, 1, \ldots, M - 1\}$ and $x \in \mathbb{Z}_N$, where f satisfies the properties defined above.
Problem: Find s.

Quantum Algorithms for the Generalized Hidden Shift Problem: There is a quantum algorithm that, for any fixed $\epsilon > 0$, and $M \geq N^{\epsilon}$, solves the generalized hidden shift problem in time polynomial in $\log N$.

The algorithm uses a "pretty good measurement" that involves solving instances of the following matrix sum problem. Given $\mathbf{x} \in \mathbb{Z}_N^k$ and $w \in Z_N$ chosen uniformly at random, find $\mathbf{b} \in \{0, 1, \ldots, M-1\}^k$ such that $\sum b_i x_i = w \mod N$. Note how this generalizes the subset sum problem, which was shown to be related to the dihedral hidden subgroup problem [154]. While there is no efficient solution known for small M (even an average case solution suffices for the dihedral HSP), for $M \geq N^{\epsilon}$, Lenstra's integer programming algorithm allows for an efficient solution to the matrix sum problem.

Classical Algorithms for the Generalized Hidden Shift Problem: Any classical algorithm requires $\Omega(\sqrt{N})$ evaluations of the function f.

Other Related Algorithms

There are a variety of other problems that aren't (as far as we know) generalizations of the hidden subgroup problem, and arguably deserve a separate section. We'll mention them here since various parts of the algorithms for these problems use techniques related to those discussed in one of the other subsections of this section.

Van Dam and Seroussi [56] give an efficient quantum algorithm for estimating Gauss sums. Consider a finite field $\mathbb{F}_{p^r}$ (where p is prime, and r is a positive integer). The multiplicative characters are homomorphisms of the multiplicative group, $\mathbb{F}_{p^r}^*$, to the complex numbers $\mathbb{C}$, and also map $0 \mapsto 0$. Each multiplicative character can be specified by an integer $\alpha \in \{0, 1, \ldots, p^r - 2\}$ by defining $\chi_\alpha(g^j) = \xi^{\alpha j}$, where g is a generator for $\mathbb{F}_{p^r}^*$ and $\xi = e^{2\pi i/(p^r-1)}$.

The additive characters are homomorphisms of the additive group of the field to $\mathbb{C}$, and can be specified by a value $\beta \in \mathbb{F}_{p^r}$ according to $e_\beta(x) = \zeta^{Tr(\beta x)}$, where $Tr(y) = \sum_{j=0}^{r-1} y^{p^j}$ and $\zeta = e^{2\pi i/p}$.

The Gauss sum $G(\mathbb{F}_{p^r}, \chi_\alpha, e_\beta)$ is defined as

$$G(\mathbb{F}_{p^r}, \chi, e_\beta) = \sum_{x \in \mathbb{F}_{p^r}} \chi(x) e_\beta(x) .$$

It is known that the norm of the Gauss sum is $|G(\mathbb{F}_{p^r}, \chi, e_\beta)| = \sqrt{p^r}$, and thus the hard part is determining, or approximating, the parameter γ in the equation $G(\mathbb{F}_{p^r}, \chi, e_\beta) = e^{i\gamma}\sqrt{p^r}$.

Gauss Sum Problem for Finite Fields:
Input: A prime number p, positive integer r and a standard specification of $\mathbb{F}_{p^r}$ (including a generator g).
A positive integer $\alpha \in \{0, 1, \ldots, p^r - 2\}$.
An element $\beta \in \mathbb{F}_{p^r}$.
A parameter ϵ, $0 < \epsilon < 1$.
Problem: Output an approximation, with error at most ϵ, to γ in the equation $G(\mathbb{F}_{p^r}, \chi_\alpha, e_\beta) = e^{i\gamma}\sqrt{p^r}$.

One noteworthy feature of this problem is that it is not a black-box problem.

Quantum Algorithms for the Finite Field Gauss Sum Problem: There is a quantum algorithm running in time $O(\frac{1}{\epsilon}\text{polylog}(p^r))$ that outputs a value $\tilde{\gamma}$ such that $|\gamma - \tilde{\gamma}| < \epsilon$ with probability at least $\frac{2}{3}$.

Classical Complexity of the Finite Field Gauss Sum Problem: It was shown that solving this problem is at least as hard as the discrete logarithm problem in the multiplicative group of $\mathbb{F}_{p^r}$ (see Subsect. "Discrete Logarithms").

Various generalizations of this problem were also studied in [56]. Other examples include [57] which studies the problem of finding solutions to equations of the form $af^x + bg^y = c$, where a, b, c, f, g are elements of a finite field, and x, y are integers.

Quantum Walk Algorithms

Quantum walks, sometimes called quantum random walks, are quantum analogues of (classical) random walks, which have proved to be a very powerful algorithmic tool in classical computer science. The quantum walk paradigm is still being developed. For example, the relationship between the continuous time and discrete time models of quantum walks is still not fully understood. In any case, the best known algorithms for several problems are some type of quantum walk.

Here we restrict attention to walks on discrete state spaces. Because of the quantum strong Church–Turing thesis, we expect that any practical application of a walk on a continuous state space will have an efficient (up to polynomial factors) simulation on a discrete system.

In general, any walk algorithm (classical or quantum), consists of a discrete state space, which is usually finite in size, but sometimes infinite state spaces are also considered when it is convenient to do so. The state space is usually modeled as being the vertices of a graph G, and the edges of the graph denote the allowed transitions. In classical discrete time walks, the system starts in some initial state, v_i. Then at every time step the system moves to a random neighbor w of the current vertex v, according

to some probability distribution $p(v, w)$. Let M denote the matrix where the (v, w) entry is $p(v, w)$. Let $\mathbf{v_0}$ be the column vector with the value p_i in the ith position, where p_i is the probability that the initial vertex is v_i. Then the vector $\mathbf{v_t} = M^t\mathbf{v_0}$ describes the probability distribution of the system after t time steps after starting in a state described by the probability distribution $\mathbf{v_0}$.

The walks are usually analyzed as abstract walks on a graph. In practice, the vertices are representing more sophisticated objects. For example, suppose one wishes to solve a 3-SAT formula Φ on n Boolean variables. One could define a random walk on the 2^n possible assignments of the Boolean variables. So the vertices of the graph would represent the 2^n Boolean strings of length n. One could start the walk on a random vertex (which corresponds to a random assignment of the n-Boolean variables). At every step of the walk, if the current vertex v corresponds to a satisfying assignment, then $p(v, v) = 1$ and the walk should not leave the vertex. Otherwise, a random clause should be picked, and one of the variables in that clause should be picked uniformly at random and flipped. This implicitly defines a probability distribution $p(v, w)$.

In a quantum walk, instead of just having classical probability distributions of the vertices $v_i \in V(G)$, one can have superpositions $\sum_{v_i \in V(G)} \alpha_i \mid v_i\rangle$, and more generally any quantum mixed state of the vertices. If we restrict to unitary transitions, then there is a unitary matrix U that contains the transition amplitudes $\alpha(v, w)$ of going from vertex v to vertex w, and if the systems starts in initial state $\mid \psi_0\rangle$, then after t time steps the state of the system is $U^t \mid \psi_0\rangle$. These unitary walks are not really "random" since the evolution is deterministic. More generally, the transition function could be a completely positive map $\mathcal{E}$, and if the system starts in the initial state $\rho = \mid \psi_0\rangle \langle \psi_0 \mid$, then after t time steps the state of the system will be $\mathcal{E}^t(\rho)$.

One cannot in general define a unitary walk on any graph [164]; however if one explicitly adds a "coin" system of dimension as a large as the maximum degree d of the vertices (i. e. the new state space consists of the states $\mid v_i\rangle \mid c\rangle$, $v_i \in V(G)$ and $c \in \{0, 1, \ldots, d-1\}$) then one can define a unitary walk on the new graph one would derive from the combined graph-coin system. In particular, the state of the coin system indicates to which neighbor of a given vertex the system should evolve. More generally, one can define a unitary walk on states of the form (v_i, v_j), where $\{v_i, v_j\}$ is an edge of G.

A continuous version of quantum walks was introduced by Farhi and Gutmann [72]. The idea is to let the adjacency matrix of the graph be the Hamiltonian driving the evolution of the system. Since the adjacency matrix is Hermitian, the resulting evolution will be unitary. The reason such a unitary is possible even for a graph where there is no unitary discrete time evolution is that in this continuous time Hamiltonian model, for any non-zero time evolution, there is some amplitude with which the walk has taken more than one step.

In classical random walks, one is often concerned with the "mixing time", which is the time it takes for the system to reach its equilibrium distribution. In a purely unitary (and thus reversible) walk, the system never reaches equilibrium, but there are alternative ways of arriving at an effective mixing time (e. g. averaging over time). In general, quantum walks offer at most a quadratically faster mixing. Another property of random walks is the "hitting time", which is the time it takes to reach some vertex of interest. There are examples where quantum walks offer exponentially faster hitting times.

The study of what are essentially quantum walks has been around for decades, and the algorithmic applications have been developed for roughly 10 years. Much of the early algorithmic work developed the paradigm and discovered the properties of quantum walks on abstract graphs, such as the line or circle, and also on general graphs (e. g. [6,17]). There have also been applications to more concrete computational problems, and we will outline some of them here.

Element Distinctness Problem
Input: A black-box U_f that maps $\mid i\rangle \mid b\rangle \mapsto \mid i\rangle \mid b \oplus f(i)\rangle$ for some function $f\colon \{0, 1, \ldots, N-1\} \to \{0, 1, \ldots, M\}$.
Problem: Decide whether there exist inputs i and j, $i \neq j$, such that $f(i) = f(j)$.

Prior to the quantum walk algorithm of Ambainis, the best known quantum algorithm used $O(N^{\frac{3}{4}} \log N)$ queries [42].

Quantum Algorithms for Element Distinctness Problem The quantum walk algorithm in [15] uses $O(N^{2/3})$ evaluations of U_f, $O(N^{2/3}\text{polylog}N)$ non-query operations and $O(N^{2/3}\text{polylog}N)$ space.

Classical Algorithms for Element Distinctness A classical computer requires $N - O(1)$ applications of U_f in order to guess correctly with bounded error for worst-case instances of f.

As is often the case with classical random walk algorithms, the graph is only defined implicitly, and is usually exponentially large in the size of the problem instance. For the element distinctness algorithm, the graph is defined as follows. The vertices are subsets of $\{1, 2, \ldots, N\}$ of size $\lceil N^{2/3} \rceil$. Two vertices are joined if the subsets differ in exactly two elements. A detailed description and analysis of this walk is beyond the scope of this survey.

Szegedy [171] extended the approach of Ambainis to develop a powerful general framework for quantizing classical random walks in order to solve search problems. Suppose we wish to search a solution space of size N and there are ϵN solutions to $f(x) = 1$. Furthermore, suppose there is a classical random walk with transition matrix M, with the property that $p(v, w) = p(w, v)$ (known as a 'symmetric' walk). It can be shown that the matrix M has maximum eigenvalue 1, and suppose the next highest eigenvalue is $1 - \delta$, for $\delta > 0$. The classical theory of random walks implies the existence of a bounded-error classical random walk search algorithm with query complexity in $O(\frac{1}{\delta\epsilon})$. Szegedy developed a "$\sqrt{\delta\epsilon}$-rule" that gives a quantum version of the classical walk with query complexity in $O(1/\sqrt{\delta\epsilon})$. This technique was generalized further in [134] and summarized nicely in [160].

Quantum walk searching has been applied to other problems such as triangle-finding [135], commutativity testing [133], matrix product verification [40], associativity testing when the range is restricted [66], and element k-distinctness [15]. A survey of results in quantum walks can be found in [13,14,123,160].

Continuous Time Quantum Walk Algorithms

In this section we describe two very well-known continuous time quantum walk algorithms. The first algorithm [48] illustrates how a quantum walk algorithm can give an exponential speed-up in the black-box model.

A problem instance of size n corresponds to an oracle O_{G_n} that encodes a graph G_n on $O(2^n)$ vertices in the following way. The graph G_n is the graph formed by taking 2 binary trees of depth n, and then "gluing" the two trees together by adding edges that create a cycle that alternates between the leaves of the first tree (selected at random) and the leaves of the second tree (selected at random). The two root vertices are called the "ENTRANCE" vertex, labeled with some known string, say, the all zeroes string $000\ldots0$ of length $2n$, and "EXIT" vertex, which is labeled with a random string of length $2n$. The remaining vertices of the graph are labeled with distinct random bit strings of length $2n$. The oracle O_{G_n} encodes the graph G_n in the following way. For $|x\rangle$ where $x \in \{0,1\}^{2n}$ encodes a vertex label of G_n, O_{G_n} maps $|x\rangle\,|00\ldots0\rangle$ to $|x\rangle\,|n_1(x), n_2(x), n_3(x)\rangle$ where $n_1(x), n_2(x), n_3(x)$ are the labels of the neighbors of x in any order (for the exit and entrance vertex, there will only be two distinct neighbors).

"Glued-Trees" Problem
Input: A black-box implementing O_{G_n} for a graph G_n of the above form.
Problem: Output the label of the EXIT vertex.

Quantum Algorithms for the "Glued-Trees" Problem There is a continuous time quantum walk which starts at the ENTRANCE vertex (in this case $|00\ldots0\rangle$) and evolves according to the Hamiltonian defined by the adjacency matrix of the graph G_n for an amount of time t selected uniformly at random in $[0, n^4/(2\epsilon)]$ where $0 < \epsilon < 1$. Measuring will then yield the EXIT label with probability at least $\frac{(1-\epsilon)}{2n}$.

The authors show how to efficiently simulate this continuous time quantum walk using a universal quantum computer that makes a polynomial number of calls to O_{G_n}.

Classical Algorithms for the "Glued-Trees" Problem Any classical randomized algorithm must evaluate the black-box O_{G_n} an exponential number of times in order to output the correct EXIT vertex label with non-negligible probability. More precisely, any classical algorithm that makes $2^{n/6}$ queries to O_{G_n} can only find the EXIT with probability at most $4 \cdot 2^{-n/6}$.

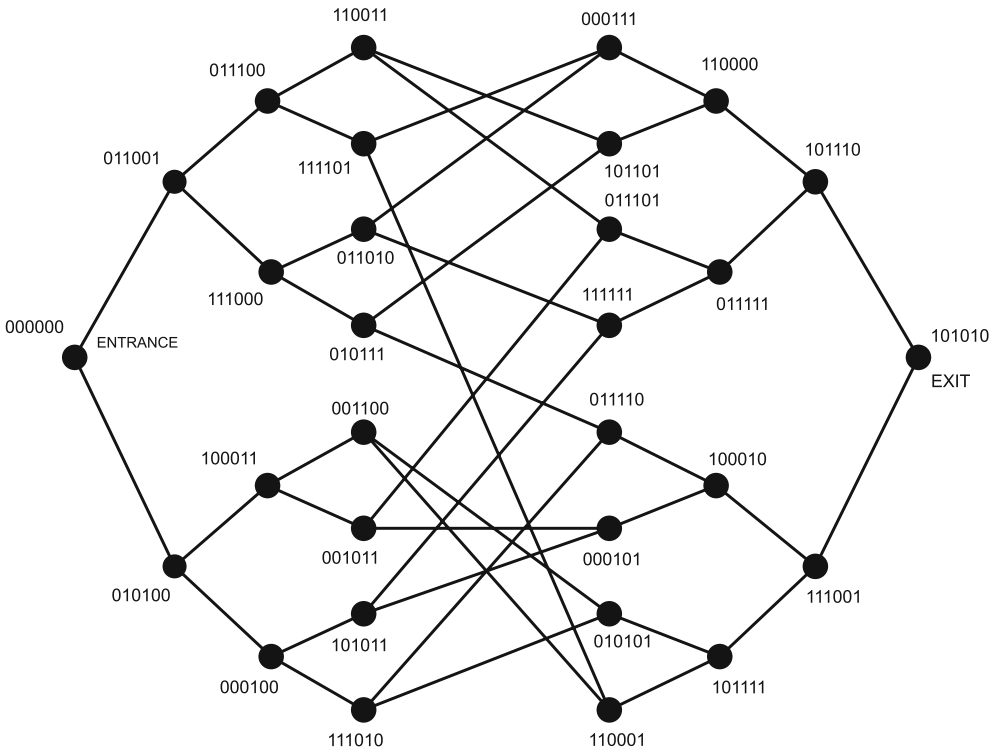
Another very interesting and recent problem for which a quantum walk algorithm has given the optimal algorithm is the problem of evaluating a NAND-tree (or AND-OR tree). The problem is nicely described by a binary tree of depth n whose leaves are labeled by the integers $i \in \{1, 2, \ldots, 2^n\}$. The input is a black-box $O_{\mathbf{X}}$ that encodes a binary string $\mathbf{X} = X_1X_2\ldots X_N$, where $N = 2^n$. The ith leaf vertex is assigned value X_i, and the parent of any pair of vertices takes on the value which is the NAND of the value of its child vertices (the NAND of two input bits is 0 if both inputs are 1 and 1 if either bit is 0). Thus, given the assignment of values to the leaves of a binary tree, one can compute the values of the remaining vertices in the tree, including the root vertex. The value of the NAND tree for a given assignment is the value of the root vertex.

NAND-Tree Evaluation
Input: A black-box $O_{\mathbf{X}}$ that encodes a binary string $\mathbf{X} = X_1X_2\ldots X_N \in \{0,1\}^N$, $N = 2^n$.
Problem: Output the value of the binary NAND tree whose ith leaf has value X_i.

Classical Algorithms for NAND-Tree Evaluation The best known classical randomized algorithm uses $O(N^{0.753\cdots})$ evaluations of the black-box, and it is also known that $\Omega(N^{0.753\cdots})$ evaluations are required for any classical randomized algorithm.

Until recently, no quantum algorithm worked better.

Quantum Algorithms for NAND-Tree Evaluation Farhi, Goldstone and Gutmann [76] showed a continuous time walk that could solve this it time $O(\sqrt{N})$ using a continuous version of the black-box, and it was subsequently



Quantum Algorithms, Figure 1
This is an example of a "glued-trees" graph with random labelings for all vertices (except the "ENTRANCE" vertex). The goal is to find the label of the "EXIT" vertex (in this case, it is 101010), given a black-box that outputs the vertex labels of the neighbors of a given vertex label

showed that $O(N^{1/2+\epsilon})$ queries to the discrete oracle suffice, for any real constant $\epsilon > 0$, and discrete walk versions of the algorithm and other generalizations were developed [18,50]

This was a very interesting breakthrough in solving a fundamental problem that had stumped quantum algorithms experts for a number of years. The general idea is inspired from techniques in particle physics and scattering theory. They consider a graph formed by taking a binary tree and making two additions to it. Firstly, for each leaf vertex where $X_i = 1$, add another vertex and join it to that leaf. Secondly, attach the root vertex to the middle of a long path of length in $\Omega(\sqrt{N})$. Then evolve the system according to the Hamiltonian equal to the adjacency matrix of this graph. Then one should start the system in a superposition of states on the left side of the line graph with phases defined so that if the NAND-tree were not attached, the "wave packet" would move from left to right along the line. If the packet gets reflected with non-negligible amplitude, then the NAND tree has value 1, otherwise, if the packet gets mostly transmitted, then the NAND tree has value 0. Thus one measures the system, and if one obtains a vertex to the left of the NAND-tree, one guesses "1", and if one obtains a vertex to the right of the NAND-tree, one guesses "0". This algorithm outputs the correct answer with high probability.

In a discrete query model, one can carefully simulate the continuous time walk [50] (as discussed in Sect. "Simulation of Quantum Mechanical Systems"), or one can apply the results of Szegedy [171] to define a discrete-time coined walk with the same spectral properties [18] and thus obtain a discrete query complexity of $N^{\frac{1}{2}+\epsilon}$ for any constant $\epsilon > 0$.

This algorithm has also been applied to solve MIN–MAX trees with a similar improvement [55]. The NAND-tree and related problems are related to deciding the winner of two-player games. Another very recent new class of quantum algorithms for evaluating a wider class of formulas, based on "span" programs, was developed in [155].

Adiabatic Algorithms

It is possible to encode the solution to a hard problem into the ground state of an efficiently simulatable Hamiltonian.

For example, in order to try to solve 3-SAT for a formula on n Boolean variables, one could define a Hamiltonian

$$H_1 = \sum_{\mathbf{x}\in\{0,1\}^n} f_{\Phi}(\mathbf{x}) \,|\, \mathbf{x}\rangle \,\langle\, \mathbf{x}|$$

where $f_{\Phi}(\mathbf{x})$ is the number of clauses of Φ that are violated by the assignment $\mathbf{x}$. Then one could try to define algorithms to find such a ground state, such as quantum analogues of classical annealing or other heuristics (e. g. [62,74,101]).

Adiabatic algorithms (also known as adiabatic optimization algorithms) are a new paradigm for quantum algorithms invented by Farhi, Goldstone, Gutmann and Sipser [74]. The paradigm is based on the fact that, under the right conditions, a system that starts in the ground state of a Hamiltonian $H(0)$, will with high probability remain in the ground state of the Hamiltonian $H(t)$ of the system at a later time t, if the Hamiltonian of the system changes "slowly enough" from $H(0)$ to $H(t)$. This fact is called the adiabatic theorem (see e. g. [115]).

This theorem inspires the following algorithmic paradigm:

- Convert your problem to generating the ground state of some easy-to-simulate Hamiltonian H_1.
- Initialize your quantum computer in an easy-to-prepare state $|\,\psi_0\rangle$ of an easy-to-simulate Hamiltonian H_0.
- On the quantum computer, simulate a time-dependent Hamiltonian $H(t) = (1 - t/T)H_0 + t/TH_1$, for t going from 0 to T.

An important detail is how slowly to transition from H_0 to H_1 (in other words, how large T should be). This related to two important parameters. Let $\lambda_0(t)$ be the smallest eigenvalue of $H(t)$, and assume that the corresponding eigenspace is non-degenerate. Let $\lambda_1(t)$ be the second smallest eigenvalue of $H(t)$, and define $g(t) = \lambda_1(t) - \lambda_0(t)$ to be the gap between the two lowest eigenvalues. The norm of the Hamiltonian is also relevant. This is to be expected, since one can effectively speed-up time by a factor of s by just multiplying the Hamiltonian of the system by s. In any realistic implementation of the Hamiltonian one pays for such a speed-up by at least a factor of s in some resource. For example, if we simulate the Hamiltonian using quantum circuits (e. g. as described in Sect. "Simulation of Quantum Mechanical Systems"), the overhead is a factor of $s^{1+o(1)}$ in the circuit depth; thus there is no actual speed-up. Furthermore, the norm of the derivatives of the Hamiltonian is also relevant.

There are a variety of theorems and claims in the literature proving, or arguing, that a value of T polynomial in the operator norm of $\mathrm{d}H(t)/\mathrm{d}t$ (or even some higher derivative) and in the inverse of the minimum gap (i. e. the minimum $g(t)$, for $0 \le t \le T$), and in $\frac{1}{\delta}$, is sufficient in order to generate the final ground state with probability at least $1 - \delta$. The general folklore is that with the right assumptions the dependence on the minimum gap $g_{\min}$ is $\Omega(1/g_{\min}^2)$, and one can find examples when this is the case; however more sophisticated descriptions of the dependence on $g(t)$ are known (see e. g. [115]).

Assuming there is exactly one satisfying assignment $\mathbf{w}$, then $|\,\mathbf{w}\rangle\,\langle\,\mathbf{w}|$ is the unique ground state of H_1. This algorithm has been studied numerically and analytically [58,75,157] and variations have been introduced as well to work around various lower bounds that were proved [77]. Unfortunately, it is not known what the worst-case complexity is for these algorithms on such NP-hard problems, since it has proved very hard to provide rigorous or even convincing heuristic bounds on the minimum gap for such problems of interest. It is widely believed that these algorithms will not solve an NP-hard problem in worst-case polynomial time, partly because it is believed that no quantum algorithm can do this.

A slight generalization of adiabatic algorithms, called adiabatic computation (where the final Hamiltonian does not need to be diagonal in the computational basis) was shown to be polynomially equivalent to general quantum computation [7].

Topological Algorithms

The standard models of quantum computation (e. g. quantum Turing machine, quantum acyclic circuits) are known to be equivalent in power (up to a polynomial factor), and the quantum strong Church–Turing thesis states that any realistic model of computation can be efficiently simulated by such a quantum computer. If this were not the case, then one should seek to define a stronger model of computation that encapsulates the full computational power that the laws of physics offer.

Freedman [80] proposed defining a computing model based on topological quantum field theories. The main objective was that such a computer might naturally solve an NP-hard or #P-hard topological problem, in particular, evaluating the Jones polynomial at certain points. A natural family of such topological quantum field theory computers was shown to be equivalent in power to the standard model of quantum computation, thus such a new model of computation would not provide additional computational power, but it was hoped that this new paradigm might inspire new quantum algorithms. In fact, it has been shown that "topological" algorithms can approximate the value of the Jones polynomial at cer-

tain points more efficiently than any known classical algorithm [8,81]. The known approximations are not good enough to solve an NP-hard problem. Several other generalizations and related problems and algorithms have been found recently [10,85,132,179]. We will briefly sketch some of the definitions, results and techniques.

A *knot* is a closed non-intersecting curve embedded in $\mathbf{R}^3$, usually represented via a knot diagram, which is a projection of the knot into the plane with additional information at each cross-over to indicate which strand goes over and which goes under. Two knots are considered equivalent if one can be manipulated into the other by an isotopy (i. e. by a transformations one could make to an actual knot that can be moved and stretched but not broken or passed through itself). A link is a collection of non-intersecting knots embedded in $\mathbf{R}^3$. They can be represented by similar diagrams, and there is a similar notion of equivalence.

The Jones polynomial of a link L is a polynomial $V_L(t)$ that is a link invariant; in other words it has the property that if two links L_1 and L_2 are equivalent, then $V_{L_1}(t) = V_{L_2}(t)$. Computing the Jones polynomial is in general #P-hard for all but a finite number of values of t. In particular, it is #P-hard to evaluate the Jones polynomial exactly at any primitive rth root of unity for any integer $r \geq 5$. However, certain approximations of these values are not known to be #P-hard.

The Jones polynomial is a special case of the Tutte polynomial of a planar graph. For a planar graph $G = (V, E)$, with weights edge weights $\mathbf{v} = \{v_e \mid e \in E\}$ the multivariate Tutte polynomial is defined as

$$Z_G(q; v_{e_1}, v_{e_2}, \ldots) = \sum_{A \subseteq E} q^{k(A)} \prod_{e \in A} v_e$$

where q is another variable and $k(A)$ is the number of connected components in the subgraph (V, A). The standard Tutte polynomial $T_G(x, y)$ is obtained by setting $v_e = v$ for all $e \in E$, $x = 1 + q/v$ and $y = 1 + v$. Connections with physics are discussed, for example, in [120,169,176].

Here we briefly sketch a specific instance of such a problem, and the approach of [10] taken to solve this problem on a quantum computer.

Firstly, for any planar graph G, one can efficiently find its *medial graph* L_G, which is a 4-regular planar graph which can be drawn from a planar embedding of G as follows. Draw a new vertex with weight u_i in the middle of each edge that had label v_i. For each new vertex, on each side of the original edge on which the vertex is placed, draw a new edge going in the clockwise direction joining the new vertex to the next new vertex encountered along the face of the original graph. Do the same in the counter-clockwise direction, and remove all the original edges and vertices.

From this medial graph, one can define another polynomial called the *Kauffman bracket*, denoted $\langle L_G \rangle(d, u_1, u_2, \ldots)$, that satisfies $\langle L_G \rangle(d, u_1, u_2, \ldots) = d^{-|V|} Z_G(d^2; du_1, du_2 \ldots)$. The next section sketches a quantum algorithm that approximates the Kauffman bracket.

Additive Approximation of the Multivariate Tutte Polynomial for a Planar Graph

Input: A description of a planar graph $G = (V, E)$.

Complex valued weights $v_1, v_2, \ldots, v_m$ corresponding to the edges $e_1, e_2, \ldots, e_m \in E$ of G.

A complex number q.

Problem: Output an approximation of $Z_G(q; v_1, v_2, \ldots, v_m)$.

Quantum Algorithms for Approximating the Tutte Polynomial Aharonov et al. [10] give a quantum algorithm that solves the above problem in time polynomial in n with an additive approximate they denote by $\Delta_{\text{alg}}/\text{poly}(m)$, which is described below.

The value Δ_{alg} depends on the embedding of the graph. The results are hard to compare to what his known about classical algorithms for this problem (see [33] for a discussion), but there are special cases that are BQP-hard.

Classical Algorithms for Approximating the Tutte Polynomial It was shown [10] that for certain ranges of parameter choices, the approximations given by the quantum algorithms are BQP-hard, and thus we don't expect a classical algorithm to be able to provide as good of an approximation unless classical computers can efficiently simulate quantum computers.

Sketch of the Structure of the Algorithm

We only have room to give a broad overview of the algorithm. One of the main points is to emphasize that this algorithm looks nothing any of the other algorithms discussed in the previous sections.

At a very high level, the idea is that these medial graphs T_G can be represented by a linear operation Q_G such that $\langle L_G \rangle(d, u_1, u_2, \ldots) = \langle 1 | Q_G | 1 \rangle$. The quantum algorithm approximates the inner product between $|1\rangle$ and $Q_G |1\rangle$, and therefore gives an approximation to the Kauffman bracket for G and thus the generalized Tutte polynomial for G.

The medial graph L_G will be represented as a product of basic elements $\mathcal{T}_i$ from a generalized Temperley–Lieb algebra. These basic elements $\mathcal{T}_i$ will be represented by simple linear transformations on finite dimensional

Hilbert spaces, which can be implemented on a quantum computer. Below we briefly sketch this decomposition and correspondence.

One can easily draw the medial graph L_G in the plane so that one can slice it with horizontal lines so that in between each consecutive horizontal line there is a diagram with only one of the following: a crossing of two lines, a "cap" or a "cup". One can think of the gluing together of these adjacent diagrams $\mathcal{T}_i$ to form the graph L_G as a product operation. We will sketch how to map each $\mathcal{T}_i$ to a linear operation $\rho(\mathcal{T}_i)$ acting on a finite dimensional Hilbert space.

The state space that the operation $\rho(\mathcal{T}_i)$ will act on is the set of finite walks starting at vertex 1 on the infinite graph G with vertices labeled by the non-negative integers and edges $\{i, i+1\}$ for all $i \geq 0$. For example, the walk $1-2-3-2-3$ is a walk of length 4 from 1 to 3. The linear transformation $\rho(\mathcal{T}_i)$ maps a walk w_1 to a linear combination of walks that are "compatible" with $\mathcal{T}_i$ (we won't explain here the details of defining this linear transformation).

In order to apply this technique to a diagram with a crossing, one eliminates the crossing by replacing it with two non-intersecting lines. This replacement can be done in two different ways, and one can represent the diagram with a crossing as a formal linear combination of these two diagrams one gets by replacing a crossing at a vertex (say with label u) with two non-intersecting lines, where one of the two links gets the coefficient u (by a simple rule that we don't explain here). We can then apply the construction to each of these new diagrams, and combine them linearly to get the linear transformation corresponding to the diagram with the crossing.

These linear transformations Q_G are not necessarily unitary (they were unitary in the earlier work on the Jones polynomial, and other related work also constructs unitary representations); however the authors show how one can use ancilla qubits and unitaries to implement non-unitary transformations and approximate the desired inner product using the "Hadamard test" (see Sect. "Quantum Algorithms for Quantum Tasks").

Quantum Algorithms for Quantum Tasks

At present, when we think of quantum algorithms, we usually think of starting with a classical input, running some quantum algorithm, or series of quantum algorithms, with some classical post-processing in order to get a classical output. There might be some quantum sub-routines that have been analyzed, but the main goal in mind is to solve a classical problem.

For example, quantum error correction (see [124] for a recent survey) can be thought of as an algorithm having a quantum input and a quantum output. There are many algorithms for transferring a qubit of information through a network of qubits under some physical constraints. We might develop a quantum cryptographic infrastructure where objects like money and signatures [88] are quantum states that need to be maintained and manipulated as quantum states for long periods of time.

Several of the algorithms described in the previous sections have versions which have a quantum input or a quantum output or both. For example, the amplitude amplification algorithm can be rephrased in terms of producing a quantum state $|\phi\rangle$ given a black-box U_ϕ that recognizes $|\phi\rangle$ by mapping $|\phi\rangle \mapsto -|\phi\rangle$ and acting as the identity on all states orthogonal to $|\phi\rangle$. The end result of such a quantum search is the quantum state $|\phi\rangle$. Amplitude estimation is estimating a transition probability of a unitary operator.

The topological algorithms require as a basic subroutine a quantum algorithm for approximating the inner product of $|0\rangle$ and $U|00\ldots0\rangle$, which the authors call the "Hadamard test". The algorithm consists of using a controlled-U operation to create the state $\frac{1}{\sqrt{2}}|0\rangle|00\ldots0\rangle + \frac{1}{\sqrt{2}}|1\rangle U|00\ldots0\rangle$. Note that if we apply the Hadamard gate to the first qubit, we get

$$\sqrt{\frac{1+\mathrm{Re}\langle 00\ldots0|U|00\ldots0\rangle}{2}}\,|0\rangle|\psi_0\rangle + \sqrt{\frac{1-\mathrm{Re}\langle 00\ldots0|U|00\ldots0\rangle}{2}}\,|1\rangle|\psi_1\rangle$$

for normalized states $|\psi_0\rangle$ and $|\psi_1\rangle$. We can thus estimate the real part of the $\langle 00\ldots0|U|00\ldots0\rangle$ by repeating several times, or applying the quadratically more efficient amplitude estimation algorithm algorithm described earlier (the goal is a superpolynomial speed-up, so a quadratic improvement is not substantial in this case). We can also estimate the complex part similarly.

Another example is the *coset orbit problem* [82] mentioned in Sect. "Generalizations of the Abelian Hidden Subgroup Problem". The input to this problem consists of two quantum states $|\phi_0\rangle$ and $|\phi_1\rangle$ from a set Γ of mutually orthogonal states and black-boxes for implementing the action of a group G on the set Γ. For a state $|\phi\rangle$, let $|u\cdot\phi\rangle$ denote the state resulting from the action of $u \in G$ on $|\phi\rangle$, and let $G_{|\phi\rangle}$ denote the subgroup of G that stabilizes the state $|\phi\rangle$ (i. e. the set of $u \in G$ such that $|u\cdot\phi\rangle = |\phi\rangle$). The question is whether there exists a $u \in G$ such that $|u\cdot\phi_1\rangle = |\phi_0\rangle$. If the answer is "yes", then the set of u satisfying $|u\cdot\phi_1\rangle = |\phi_0\rangle$ is a left coset of

$G_{|\phi_1\rangle}$. Thus, the algorithm should output a coset representative u along with $O(\log n)$ generators of $G_{|\phi_1\rangle}$. The solution to this problem was an important part of the solution to the hidden translation problem. Several other quantum algorithms have such "quantum" sub-routines.

Other examples of quantum tasks include quantum data compression which was known to be information theoretically possible, and efficient quantum circuits for performing it were also developed [26]. Another example is entanglement concentration and distillation.

Researchers have also developed quantum algorithms for implementing some natural basis changes, which have quantum inputs and quantum outputs. For example, the Clebsch-Gordan transformation [20] or other transformations (e. g. [69,103]).

We've only listed a few examples here. In the future, as quantum technologies develop, in particular quantum computers and reliable quantum memories, quantum states and their manipulation will become an end in themselves.

Future Directions

Roughly 10 years ago, many people said that there were essentially only two quantum algorithms. One serious omission was the simulation of quantum mechanical systems, which was in fact Feynman's initial motivation for quantum computers. Apart from this omission, it was true that researchers were developing a better and deeper understanding of the algorithms of Shor and Grover, analyzing them in different ways, generalizing them, and applying them in non-trivial ways. It would have been feasible to write a reasonably sized survey of all the known quantum algorithms with substantial details included. In addition to this important and non-trivial work, researchers were looking hard for "new" approaches, and fresh ideas like quantum walks, and topological algorithms, were being investigated, as well as continued work on the non-Abelian version of the hidden subgroup problem. The whole endeavor of finding new quantum algorithms was very hard and often frustrating. Fortunately, in the last 10 years, there have been many non-trivial developments, enough to make the writing of a full survey of quantum algorithms in a reasonable number of pages impossible. Some directions in which future progress might be made are listed below.

- The complexity of the non-Abelian hidden subgroup problem will hopefully be better understood. This includes addressing the question: Does there exist a quantum polynomial time algorithm for solving the graph isomorphism problem? Of course, a proof that no such quantum algorithm exists would imply that $P \neq PSPACE$. So, more likely, we might develop a strong confidence that no such algorithm exists, in which case this can form the basis of quantum computationally secure cryptography [142].
- There are many examples where amplitude amplification was used to speed-up a classical algorithm in a non-trivial way. That is, in a way that was more than just treating a classical algorithm as a guessing algorithm A and applying amplitude amplification to it. There are likely countless other algorithms which can be improved in non-trivial ways using amplitude amplification.
- The quantum walk paradigm for quantum algorithms emerged roughly 10 years ago, and has recently been applied to find the optimal black-box algorithm for several problems, and has become a standard approach for developing quantum algorithms. Some of these black-box problems are fairly natural, and the black-boxes can be substituted with circuits for actual functions of interest. For example, collision finding can be applied to find collisions in actual hash functions used in cryptography. We will hopefully see more instances where black-box algorithm can be applied to solve an problem without a black-box, or where there is no black-box in the first place.
- In addition to the development of new quantum walk algorithms, we will hopefully have a more elegant and unified general theory of quantum walks that unites continuous and discrete walks, coined and non-coined walks, and quantum and classical walks.
- The adiabatic algorithm paradigm has not reached the level of success of quantum walks, partly because it is hard to analyze the worst case complexity of the algorithms. To date there is no adiabatic algorithm with a proof that it works more than quadratically faster than the best known classical algorithm. Can we do better with an adiabatic algorithm?
  If and when we have large-scale quantum computers, we will be able to just test these algorithms to see if indeed they do have the conjectured running times on instances of interesting size.
- The topological algorithms have received limited attention to date. This is partly because the initial work in this field was largely inaccessible to researchers without substantial familiarity with topological quantum field theory and related areas of mathematics. The more recent work summarized in this paper and other recent papers is a sign that this approach could mature into a fruitful paradigm for developing new important quantum algorithms.

- The paradigm of measurement based computation (see e. g. [118] for an introduction) has been to date mostly focused on its utility as a paradigm for possibly implementing a scalable fault-tolerant quantum computer. We might see the development of algorithms directly in this paradigm. Similarly for globally controlled architectures.
- There is also a growing group of researchers looking at the computational complexity of various computational problems in physics, in particular of simulating certain Hamiltonian systems, often coming from condensed matter physics. Much of the work has been complexity theoretic, such as proving the *QMA*-hardness of computing ground states of certain Hamiltonians (e. g. [9]). Other work has focused on understanding which quantum systems can be simulated efficiently on a classical computer. This work should lead to the definition of some simulation problems that are not known to be in *BPP*, nor believed to be *NP*-hard or *QMA*-hard, and thus might be good candidates for a quantum algorithm. There has been a language and culture barrier between physicists and theoretical computer scientists when it comes to discussing such problems. However, it is slowly breaking down, as more physicists are becoming familiar with algorithms and complexity, and more quantum computer scientists are becoming familiar with language and notions from physics. This will hopefully lead to more quantum algorithms for computational problems in physics, and new algorithmic primitives that can be applied to a wider range of problems.

In summary, as daunting as it is to write a survey of quantum algorithms at this time, it will be a much harder task in another 10 years. Furthermore, in another 10 years we will hopefully have a better idea of when we might expect to see quantum computers large enough to solve problems faster than the best available classical computers.

Bibliography

1. Aaronson S (2003) Algorithms for Boolean function query properties. SIAM J Comput 32:1140–1157
2. Aaronson S, Ambainis A (2003) Quantum search of spatial regions. In: Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science (FOCS). IEEE Computer Society, Washington, DC, pp 200–209
3. Abrams D, Lloyd S (1999) Quantum algorithm providing exponential speed increase for finding eigenvalues and eigenvectors. Phys Rev Lett 83:5162–5165
4. Aharonov D, Regev O (2005) Lattice problems in NP intersect coNP. J ACM 52:749–765
5. Aharonov D, Ta-Shma A (2007) Adiabatic quantum state generation. SIAM J Comput 37:47–82
6. Aharonov D, Ambainis A, Kempe J, Vazirani U (2001) Quantum walks on graphs. In: Proceedings of ACM Symposium on Theory of Computation (STOC'01). ACM, New York, pp 50–59
7. Aharonov D, van Dam W, Kempe J, Landau Z, Lloyd S, Regev O (2004) Adiabatic auantum computation is equivalent to standard quantum computation. In: Proceedings of the 45th Annual IEEE Symposium on Foundations of Computer Science (FOCS'04). IEEE Computer Society, Washington, DC, pp 42–51
8. Aharonov D, Jones V, Landau Z (2006) A polynomial quantum algorithm for approximating the Jones polynomial. In: Proceedings of the thirty-eighth annual ACM symposium on Theory of computing (STOC). ACM, New York, pp 427–436
9. Aharonov D, Gottesman D, Irani S, Kempe J (2007) The power of quantum systems on a line. In: Proc. 48th IEEE Symposium on the Foundations of Computer Science (FOCS). IEEE Computer Society, Washington, DC, pp 373–383
10. Aharonov D, Arad I, Eban E, Landau Z Polynomial quantum algorithms for additive approximations of the Potts model and other points of the Tutte plane. quant-ph/0702008
11. Alagic G, Moore C, Russell A (2007) Quantum algorithms for Simon's problem over general groups. In: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms (SODA). Society for Industrial and Applied Mathematics, Philadelphia, pp 1217–1224
12. Ambainis A (2002) Quantum lower bounds by quantum arguments. J Comput Syst Sci 64:750–767
13. Ambainis A (2003) Quantum walks and their algorithmic applications. Int J Quantum Inf 1:507–518
14. Ambainis A (2004) Quantum search algorithms. SIGACT News. 35(2):22–35
15. Ambainis A (2004) Quantum walk algorithm for element distinctness. In: Proceedings of the 45th Annual IEEE Symposium on Foundations of Computer Science (FOCS'04). IEEE Computer Society, Washington, DC, pp 22–31
16. Ambainis A, Spalek R (2006) Quantum algorithms for matching and network flows. In: Proceedings of STACS'06. Lecture Notes in Computer Science, vol 3884. Springer, Berlin, pp 172–183
17. Ambainis A, Bach E, Nayak A, Vishwanath A, Watrous J (2001) One-dimensional quantum walks. In: Proceedings of the 33rd ACM Symposium on Theory of Computing. ACM, New York, pp 37–49
18. Ambainis A, Childs A, Reichardt B, Spalek R, Zhang S (2007) Any AND-OR formula of size N can be evaluated in time $N^{1/2+o(1)}$ on a quantum computer. In: 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07). IEEE Computer Society, Washington, DC, pp 363–372
19. Bacon D, Childs A, van Dam W (2005) From optimal measurement to efficient quantum algorithms for the hidden subgroup problem over semidirect product groups. In: Proc 46th IEEE Symposium on Foundations of Computer Science (FOCS 2005). IEEE Computer Society, Washington, DC, pp 469–478
20. Bacon D, Chuang I, Harrow A (2006) Efficient quantum circuits for Schur and Clebsch-Gordan transforms. Phys Rev Lett 97:170502
21. Beals R (1997) Quantum computation of Fourier transforms over symmetric groups. In: Proceedings of the 29th Annual ACM Symposium on Theory of Computing (STOC). ACM, New York, pp 48–53
22. Beals R, Buhrman H, Cleve R, Mosca M, deWolf R (2001) Quantum lower bounds by polynomials. J ACM 48:778–797

23. Benioff P (1982) Quantum mechanical models of Turing machines that dissipate no energy. Phys Rev Lett 48(23): 1581–1585
24. Bennett C (1988) Notes on the history of reversible computation by Charles Bennett. IBM J Res Dev 32(1):16–23
25. Bennett C, Bernstein E, Brassard G, Vazirani U (1997) Strengths and weaknesses of quantum computing. SIAM J Comput 26:1510–1523
26. Bennett C, Harrow A, Lloyd S (2006) Universal quantum data compression via gentle tomography. Phys Rev A 73:032336
27. Bennett CH, Brassard G (1984) Quantum cryptography: Public-key distribution and coin tossing. In: Proceedings of IEEE International Conference on Computers, Systems and Signal Processing, India. IEEE Computer Society, Washington, DC, pp 175–179
28. Bernstein E, Vazirani U (1997) Quantum complexity theory. SIAM J Comput 26:1411–1473
29. Berry DW, Ahokas G, Cleve R, Sanders BC (2007) Efficient quantum algorithms for simulating sparse Hamiltonians. Comm Math Phys 270:359
30. Berthiaume A, Brassard G (1992) The quantum challenge to structural complexity theory. In: Proc 7th Conference Structure Complexity Theory, IEEE Comp Soc Press, pp 132–137
31. Berthiaume A, Brassard G (1994) Oracle quantum computing. J Modern Opt 41(12):2521–2535
32. Boneh D, Lipton R (1995) Quantum cryptanalysis of hidden linear functions (Extended Abstract). In: Proceedings of 15th Annual International Cryptology Conference (CRYPTO'95). Springer, London, pp 424–437
33. Bordewich M, Freedman M, Lovasz L, Welsh DJA (2005) Approximate counting and quantum computation. Comb Probab Comput 14:737–754
34. Boyer M, Brassard G, Høyer P, Tapp A (1998) Tight bounds on quantum searching. Fortschr Phys 56(5-5):493–505
35. Brassard G, Høyer P (1997) An exact quantum polynomial-time algorithm for Simon's problem. In: Proc of Fifth Israeli Symposium on Theory of Computing and Systems (ISTCS'97). IEEE Computer Society, Washington, DC, pp 12–23
36. Brassard G, Høyer P, Tapp A (1997) Cryptology column–quantum algorithm for the collision problem. ACM SIGACT News 28:14–19
37. Brassard G, Høyer P, Tapp A (1998) Quantum Counting. In: Proceedings of the ICALP'98. Lecture notes in computer science. Springer, Berlin, pp 1820–1831
38. Brassard G, Høyer P, Mosca M, Tapp A (2002) Quantum amplitude amplification and estimation. In: Quantum Computation and Quantum Information Science. AMS Contemp Math Ser 35:53–74
39. Brown M (2003) Classical cryptosystems in a quantum setting. Master Thesis, University of Waterloo
40. Buhrman H, Špalek B (2006) Quantum verification of matrix products. In: Proceedings of the 14th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA). IEEE Computer Society, Washington, DC, pp 880–889
41. Buhrman H, Fortnow L, Newman I, Röhrig H (2003) Quantum property testing. In: Proceedings of the 14th Annual ACM-SIAM Symposium on Discrete Algorithms. Society for Industrial and Applied Mathematics, Philadelphia, pp 480–488
42. Buhrman H, Dürr C, Heiligman M, Høyer P, Magniez F, Santha M, de Wolf R (2005) Quantum algorithms for element distinctness. SIAM J Comput 34:1324–1330
43. Byrnes T, Yamamoto Y (2006) Simulating lattice gauge theories on a quantum computer. Phys Rev A 73:022328
44. Cheung K, Mosca M (2001) Decomposing finite Abelian groups. Quantum Inf Comput 1(2):26–32
45. Childs A (2002) Quantum information processing in continuous time. Ph D thesis, MIT
46. Childs A, van Dam W (2007) Quantum algorithm for a generalized hidden shift problem. In: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA). ACM, New York, pp 1225–1232
47. Childs A, Farhi E, Gutmann S (2002) An example of the difference between quantum and classical random walks. Quantum Inf Process 1:35–43
48. Childs A, Cleve R, Deotto E, Farhi E, Gutmann S, Spielman D (2003) Exponential algorithmic speedup by a quantum walk. In: Proceedings of the 35th Annual ACM Symposium on Theory of Computing. IEEE Computer Society, Washington, DC
49. Childs A, Landahl A, Parrilo P (2007) Improved quantum algorithms for the ordered search problem via semidefinite programming. Phys Rev A 75:032335
50. Childs AM, Cleve R, Jordan SP, Yeung D Discrete-query quantum algorithm for NAND trees. quant-ph/0702160
51. Childs A, Schulman L, Vazirani U (2007) Quantum algorithms for hidden nonlinear structures. In: Proceeding of 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2007). IEEE Computer Society, Washington, DC
52. Cleve R (2000) An introduction to quantum complexity theory. In: Macchiavello C, Palma GM, Zeilinger A (eds) Collected Papers on Quantum Computation and Quantum Information Theory. World Scientific, Singapore, pp 103–127
53. Cleve R (2000) The query complexity of order-finding. In: IEEE Conference on Computational Complexity. IEEE Computer Society, Washington, DC, p 54
54. Cleve R, Ekert A, Macchiavello C, Mosca M (1998) Quantum algorithms revisited. Proc Royal Soc Lond A 454:339–354
55. Cleve R, Gavinsky D, Yeung D (2008) Quantum algorithms for evaluating Min-Max trees. In: Proceedings of TQC 2008. Lecture Notes in Computer Science, vol 5106. Springer, Berlin
56. van Dam W, Seroussi G Efficient quantum algorithms for estimating Gauss sums. quant-ph/0207131
57. van Dam W, Shparlinski I (2008) Classical and quantum algorithms for exponential congruences. In: Proceedings of TQC 2008. Lecture Notes in Computer Science, vol 5106. Springer, Berlin, pp 1–10
58. van Dam W, Mosca M, Vazirani U (2001) How powerful is adiabatic quantum computation? In: Proceedings 46th IEEE Symposium on Foundations of Computer Science (FOCS'01). IEEE Computer Society, Washington, DC, pp 279–287
59. van Dam W, Hallgren S, Ip L (2003) Quantum algorithms for some hidden shift problems. In: Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA'03). Society for Industrial and Applied Mathematics, Philadelphia, pp 489–498
60. van Dam W, D'Ariano GM, Ekert A, Macchiavello C, Mosca M (2007) General optimized schemes for phase estimation. Phys Rev Lett 98:090501
61. van Dam W, D'Ariano GM, Ekert A, Macchiavello C, Mosca M (2007) Optimal phase estimation in quantum networks. J Phys A: Math Theor 40:7971–7984
62. Das A, Chakrabarti BK (2008) Quantum annealing and analog quantum computation. Rev Mod Phys 80:1061

63. Decker T, Draisma J, Wocjan P (2009) Efficient quantum algorithm for identifying hidden polynomials. Quantum Inf Comput (to appear)
64. Deutsch D (1985) Quantum theory, the Church–Turing principle and the universal quantum computer. Proc Royal Soc Lond A 400:97–117
65. Deutsch D, Jozsa R (1992) Rapid solutions of problems by quantum computation. Proc Royal Soc Lond A 439:553–558
66. Dörn S, Thierauf T (2007) The quantum query complexity of algebraic properties. In: Proceedings of the 16th International Symposium on Fundamentals of Computation Theory (FCT). Lecture Notes in Computer Science, vol 4639. Springer, Berlin, pp 250–260
67. Durr C, Heiligman M, Høyer P, Mhalla M (2004) Quantum query complexity of some graph problems. In: Proceedings of 31st International Colloquium on Automata, Languages, and Programming (ICALP'04). Lecture Notes in Computer Science, vol 3142. Springer, Berlin pp 481–493
68. Ettinger JM (1998) On noncommutative hidden subgroups. Lecture at AQIP'98. http://www.brics.dk/~salvail/aqip/abstlist.html#upload-paper.10480.1
69. Ettinger JM Quantum time-frequency transforms. quant-ph/0005134
70. Ettinger M, Høyer P (2000) On quantum algorithms for noncommutative hidden subgroups. Adv Appl Math 25(3): 239–251
71. Ettinger M, Høyer P, Knill E (2004) The quantum query complexity of the hidden subgroup problem is polynomial. Inf Process Lett 91:43–48
72. Farhi E, Gutmann S (1998) Quantum computation and decision trees. Phys Rev A 58:915–928
73. Farhi E, Gutmann S An analog analogue of a digital quantum computation. quant-ph/9612026
74. Farhi E, Goldstone J, Gutmann S, Sipser M (2000) Quantum computation by adiabatic evolution. quant-ph/0001106
75. Farhi E, Goldstone J, Gutmann S, Lapan J, Lundgren A, Preda D (2001) A quantum adiabatic evolution algorithm applied to random instances of an NP. Science 20 April:472
76. Farhi E, Goldstone J, Gutmann S A quantum algorithm for the Hamiltonian NAND tree. quant-ph/0702144v2
77. Farhi E, Goldstone J, Gutmann S Quantum adiabatic evolution algorithms with different paths. quant-ph/0208135
78. Fenner S, Zhang Y (2005) Quantum algorithms for a set of group theoretic problems. In: Proceedings of the Ninth ICEATCS Italian Conference on Theoretical Computer Science. Lecture Notes in Computer Science, vol 3701. Springer, Berlin, pp 215–227
79. Feynman R (1982) Simulating physics with computers. Int J Theor Phys 21(6,7):467–488
80. Freedman M (1998) P/NP, and the quantum field computer. Proc Natl Acad Sci 95(1):98–101
81. Freedman M, Kitaev A, Wang Z (2002) Simulation of topological Field theories by quantum computers. Comm Math Phys 227(3):587–603
82. Friedl K, Ivanyos G, Magniez F, Santha M, Sen P (2003) Hidden translation and orbit coset in quantum computing. In: Proceedings of the thirty-fifth annual ACM symposium on Theory of computing (STOC'03). IEEE Computer Society, Washington, DC, pp 1–9
83. Friedl K, Ivanyos G, Santha M (2005) Efficient testing of groups. In: Proceedings of the Thirty-seventh Annual ACM Symposium on Theory of Computing (STOC). IEEE Computer Society, Washington, DC, pp 157–166
84. Furrow B (2008) A panoply of quantum algorithms. Quantum Inf Comput 8(8–9):0834–0859
85. Geraci J, Lidar D (2008) On the exact evaluation of certain instances of the potts Partition function by quantum computers. Comm Math Phys 279:735
86. Gerhardt H, Watrous J (2003) Continuous-time quantum walks on the symmetric group. 6th International Workshop on Approximation Algorithms for Combinatorial Optimization Problems, (APPROX) and 7th International Workshop on Randomization and Approximation Techniques in Computer Science, (RANDOM), pp 290–301
87. Grigni M, Schulman L, Vazirani M, Vazirani U (2001) Quantum mechanical algorithms for the nonabelian hidden subgroup problem. In: Proceedings of the thirty-third annual ACM symposium on Theory of computing (SODA'03), pp 68–74
88. Gottesman D, Chuang I Quantum digital signatures. quant-ph/0105032
89. Grigoriev D (1997) Testing shift-equivalence of polynomials by deterministic, probabilistic and quantum machines. Theor Comput Sci 180:217–228
90. Grover L (1996) A fast quantum mechanical algorithm for database search. In: Proceedings of the 28th Annual ACM Symposium on the Theory of Computing (STOC 1996), pp 212–219
91. Grover L (1998) A framework for fast quantum mechanical algorithms. In: Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing (STOC), pp 53–62
92. Hales L, Hallgren S (2000) An Improved Quantum Fourier Transform Algorithm and Applications. FOCS 2000, pp 515–525
93. Hallgren S (2002) Polynomial-time quantum algorithms for Pell's equation and the principal ideal problem. STOC 2002, pp 653–658
94. Hallgren S (2005) Fast quantum algorithms for computing the unit group and class group of a number field. In: Proceedings of the 37th ACM Symposium on Theory of Computing (STOC 2005), pp 468–474
95. Hallgren S (2007) Polynomial-time quantum algorithms for Pell's equation and the principal ideal problem. J ACM 54(1):653–658
96. Hallgren S, Russell A, Ta-Shma A (2000) Normal subgroup reconstruction and quantum computation using group representations. In: Proceedings of the Thirty-Second Annual ACM Symposium on Theory of Computing, pp 627–635
97. Hallgren S, Moore C, Roetteler M, Russel A, Sen P (2006) Limitations of quantum coset states for graph isomorphism. In: Proceedings of the 38th Annual ACM Symposium on Theory of Computing. IEEE Computer Society, Washington, DC
98. Hassidim A, Ben-Or M (2008) The Bayesian learner is optimal for noisy binary search (and pretty good for quantum as well). In: Proceedings of the 49th Symposium on Foundations of Computer Science. IEEE Computer Society, Washington, DC
99. Hausladen P, Wootters WK (1994) A pretty good measurement for distinguishing quantum. states. J Mod Opt 41:2385
100. Helstrom CW (1976) Quantum detection and estimation theory. Academic Press, New York
101. Hogg T (2000) Quantum search heuristics. Phys Rev A 61:052311

102. Holevo AS (1982) Probabilistic and statistical aspects of quantum theory. North Holland, Amsterdam
103. Høyer P Efficient quantum transforms. quant-ph/9702028
104. Høyer P (1999) Conjugated operators in quantum algorithms. Phys Rev A 59(5):3280–3289
105. Høyer P (2001) Introduction to recent quantum algorithms. In: Proceedings of 26th International Symposium on Mathematical Foundations of Computer Science (MFCS'01). Lecture Notes in Computer Science, vol 2136. Springer, Berlin, pp 62–73
106. Høyer P, Dürr C A quantum algorithm for finding the minimum. quant-ph/9607014
107. Høyer P, Neerbek J, Shi Y (2002) Quantum complexities of ordered searching, sorting, and element distinctness. Algorithmica 34(4):429–448
108. Høyer P, Mosca M, de Wolf R (2003) Quantum search on bounded-error inputs. In: Proceedings of the Thirtieth International Colloquium on Automata, Languages and Programming (ICALP'03), pp 291–299
109. Inui Y, Le Gall F (2008) Quantum property testing of group solvability. In: Proceedings of the 8th Latin American Symposium (LATIN'08: Theoretical Informatics). Lecture Notes in Computer Science, vol 4957. Springer, Berlin, pp 772–783
110. Ivanyos G, Magniez F, Santha M (2003) Efficient quantum algorithms for some instances of the non-Abelian hidden subgroup problem. Int J Found Comput Sci 14(5):723–739
111. Ivanyos G, Sanselme L, Santha M (2007) An efficient quantum algorithm for the Hidden subgroup problem in extraspecial groups. 24th STACS. Lecture Notes in Computer Science, vol 4393. Springer, Berlin, pp 586–597
112. Ivanyos G, Sanselme L, Santha M (2008) An efficient quantum algorithm for the hidden subgroup problem in nil-2 groups. In: Proceedings of the 8th Latin American Symposium (LATIN'08: Theoretical Informatics). Lecture Notes in Computer Science, vol 4957. Springer, Berlin, pp 759–771
113. Ioannou L (2002) Continuous-time quantum algorithms: Searching and adiabatic computation. Master Math Thesis, University of Waterloo
114. Jaeger F, Vertigan DL, Welsh DJA (1990) On the Computational complexity of the Jones and Tutte polynomials. Math Proc Cambridge Phil Soc 108(1):5–53
115. Jansen S, Ruskai M, Seiler R (2007) Bounds for the adiabatic approximation with applications to quantum computation. J Math Phys 48:102111
116. Jozsa R (1998) Quantum algorithms and the Fourier transform. Proc Royal Soc Lond A 454:323–337
117. Jozsa R (2003) Notes on Hallgren's efficient quantum algorithm for solving Pell's equation. quant-ph/0302134
118. Jozsa R (2006) An introduction to measurement based quantum computation. NATO Science Series, III: Computer and Systems Sciences. Quantum Information Processing - From Theory to Experiment, vol 199, pp 137–158
119. Kassal I, Jordan S, Love P, Mohseni M, Aspuru-Guzik A (2008) Quantum algorithms for the simulation of chemical dynamics. arXiv:0801.2986
120. Kauffman L (2001) Knots and Physics. World Scientific, Singapore
121. Kaye P (2005) Optimized quantum implementation of elliptic curve arithmetic over binary fields. Quantum Inf Comput 5(6):474–491
122. Kaye P, Laflamme R, Mosca M (2007) An introduction to quantum computation. Oxford University Press, Oxford
123. Kempe J (2003) Quantum random walks - an introductory overview. Contemp Phys 44(4):307–327
124. Kempe J (2007) Approaches to quantum error correction. Quantum Decoherence, Poincare Seminar 2005. Progress in Mathematical Physics. Birkhäuser, Basel, pp 85–123
125. Kempe J, Shalev A (2005) The hidden subgroup problem and permutation group theory. In: Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms (SODA'05), pp 1118–1125
126. Kitaev A (1995) Quantum measurements and the Abelian stabilizer problem. quant-ph/9511026
127. Kitaev A (1996) Quantum measurements and the Abelian stabilizer problem. Electronic Colloquium on Computational Complexity (ECCC), p 3
128. Kitaev A (1997) Quantum computations: Algorithms and error correction. Russ Math Surv 52(6):1191–1249
129. Kitaev A, Shen A, Vyalvi M (2002) Classical and quantum computation. American Mathematical Society
130. Kuperberg G (2005) A subexponential-time quantum algorithm for the dihedral hidden subgroup problem. SIAM J Comput 35:170–188
131. Lloyd S (1996) Universal quantum simulators. Science 273:1073–1078
132. Lomonaco S, Kauffman L (2006) Topological quantum computing and the Jones polynomial. In: Proc SPIE 6244
133. Magniez F, Nayak A (2007) Quantum complexity of testing group commutativity. Algorithmica 48(3):221–232
134. Magniez F, Nayak A, Roland J, Santha M (2007) Search via quantum walk. 39th ACM Symposium on Theory of Computing (STOC), pp 575–584
135. Magniez F, Santha M, Szegedy M (2007) Quantum algorithms for the triangle problem. SIAM J Comput 37(2):413–424
136. Manin Y (2000) Classical computing, quantum computing, and Shor's factoring algorithm. Séminaire Bourbaki, 41 (1998-1999), 862, pp 375–404. The appendix translates an excerpt about quantum computing from a 1980 paper (in Russian)
137. Menezes A, van Oorschot P, Vanstone S (1996) Handbook of Applied Cryptography. CRC Press, Boca Raton
138. Moore C, Russell A (2007) For distinguishing conjugate hidden dubgroups, the pretty good measurement is as hood as it gets. Quantum Inf Comput 7(8):752–765
139. Moore C, Rockmore D, Russell A, Schulman L (2004) The power of basis selection in fourier sampling: hidden subgroup problems in affine groups. In: Proceedings of the fifteenth annual ACM-SIAM symposium on discrete algorithms (SODA'04), pp 1113–1122
140. Moore C, Rockmore D, Russell A (2006) Generic Quantum Fourier Transforms. ACM Trans Algorithms 2:707–723
141. Moore C, Russell A, Sniady P (2007) On the impossibility of a quantum sieve algorithm for graph isomorphism: unconditional results. In: Proceedings of the thirty-ninth annual ACM symposium on Theory of computing (STOC), pp 536–545
142. Moore C, Russell A, Vazirani U A classical one-way function to confound quantum adversaries. quant-ph/0701115
143. Mosca M (1999) Quantum computer algorithms. Ph D thesis, Oxford
144. Mosca M (2001) Counting by quantum eigenvalue estimation. Theor Comput Sci 264:139–153
145. Mosca M, Ekert A (1998) The hidden subgroup problem and

eigenvalue estimation on a quantum computer. In: Proceedings 1st NASA International Conference on Quantum Computing & Quantum Communications. Lecture Notes in Computer Science, vol 1509. Springer, Berlin, pp 174–188
146. Mosca M, Zalka C (2004) Exact quantum Fourier transforms and discrete logarithm algorithms. Int J Quantum Inf 2(1): 91–100
147. Motwani R, Raghavan P (1995) Randomized Algorithms. Cambridge University Press, Cambridge
148. Nayak A, Wu F (1999) The quantum query complexity of approximating the median and related statistics. In: Proceedings of the Thirty-first Annual ACM Symposium on Theory of Computing (STOC), pp 384–393
149. Nielsen M, Chuang I (2000) Quantum computation and quantum information. Cambridge University Press, Cambridge
150. Proos J, Zalka C (2003) Shor's discrete logarithm quantum algorithm for elliptic curves. Quantum Inf Comput 3:317–344
151. Püschel M, Rötteler M, Beth T (1999) Fast quantum Fourier transforms for a class of non-Abelian groups. In: Proceedings of the 13th International Symposium on Applied Algebra, Algebraic Algorithms and Error-Correcting Codes, pp 148–159
152. Radhakrishnan J, Roetteler M, Sen P (2005) On the power of random bases in Fourier sampling: Hidden subgroup problem in the Heisenberg group. In: Proceedings of the 32nd International Colloquium on Automata, Languages and Programming (ICALP), pp 1399–1411
153. Ramesh H, Vinay V (2003) String matching in $\tilde{O}(\sqrt{n} + \sqrt{m})$ quantum time. J Discret Algorithms 1:103–110
154. Regev O (2004) Quantum computation and lattice problems. SIAM J Comput 33:738–760
155. Reichardt B, Spalek R Span-program-based quantum algorithm for evaluating formulas. In: Proceedings of the fortieth annual ACM symposium on Theory of computing (STOC 2008), pp 103–112
156. Roetteler M, Beth T Polynomial-time solution to the hidden subgroup problem for a class of non-Abelian groups. quant-ph/9812070
157. Roland J, Cerf N (2002) Quantum search by local adiabatic evolution. Phys Rev A 65(4):042308
158. Rudolph T, Grover L (2004) How significant are the known collision and element distinctness quantum algorithms? Quantum Inf Comput 4:201–206
159. Russell A, Shparlinski I (2004) Classical and quantum function reconstruction via character evaluation. J Complex 20: 404–422
160. Santha M (2008) Quantum walk based search algorithms. In: Proceedings of the 5th Annual International Conference on Theory and Applications of Models of Computation (TAMC 2008). Lecture Notes in Computer Science, vol 4978. Springer, Berlin, pp 31–46
161. Sasaki M, Carlini A, Jozsa R (2001) Quantum template matching. Phys Rev A 64:022317
162. Schmidt A, Vollmer U (2005) Polynomial time quantum algorithm for the computation of the unit group of a number field. In: Proceedings of the thirty-seventh annual ACM symposium on Theory of computing, pp 475–480
163. Schumacher B (1995) Quantum coding. Phys Rev A 51:2738–2747
164. Severini S (2003) On the digraph of a unitary matrix. SIAM J Matrix Anal Appl (SIMAX) 25(1):295–300
165. Shor P (1994) Algorithms for quantum computation: discrete logarithms and factoring. In: Proceedings of the 35th Annual Symposium on Foundations of Computer Science, pp 124–134
166. Shor P (1997) Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. SIAM J Comput 26:1484–1509
167. Simon D (1994) On the power of quantum computation. In: Proceedings of the 35th IEEE Symposium on the Foundations of Computer Science (FOCS), pp 116–123
168. Simon D (1997) On the power of quantum computation. SIAM J Comput 26:1474–1483
169. Sokal A (2005) The multivariate Tutte polynomial (alias Potts model) for graphs and matroids. Surveys in Combinatorics. Cambridge University Press, Cambridge, pp 173–226
170. Somma R, Ortiz G, Gubernatis JE, Knill E, Laflamme R (2002) Simulating physical phenomena by quantum networks. Phys Rev A 65:042323
171. Szegedy M (2004) Quantum speed-Up of Markov chain based algorithms. In: Proceedings of the 45th IEEE Symposium on the Foundations of Computer Science (FOCS), pp 32–41
172. Terhal B (1999) Quantum algorithms and quantum entanglement. Ph D thesis, University of Amsterdam
173. Vazirani U (1998) On the power of quantum computation. Phil Trans Royal Soc Lond A 356:1759–1768
174. Vazirani U (1997) Berkeley Lecture Notes. Lecture 8 http://www.cs.berkeley.edu/~vazirani/qc.html
175. Watrous J (2001) Quantum algorithms for solvable groups. In: Proceedings of the thirty-third annual ACM symposium on Theory of computing (STOC), pp 60–67
176. Welsh DJA (1993) Complexity: knots, colourings and counting. Cambridge University Press, Cambridge
177. Wiesner S (1983) Conjugate coding. Sigact News 15(1):78–88
178. Wiesner S (1996) Simulations of many-body quantum systems by a quantum computer. quant-ph/9603028
179. Wocjan P, Yard J (2008) The Jones polynomial: quantum algorithms and applications in quantum complexity theory. Quantum Inf Comput 8(1–2):147–180
180. Zalka C (1998) Efficient simulation of quantum systems by quantum computers. Proc Roy Soc Lond A 454:313–322
181. Zalka C (2000) Using Grover's quantum algorithm for searching actual databases. Phys Rev A 62(5):052305
182. Zhang S (2006) New upper and lower bounds for randomized and quantum local search. In: Proceedings of the thirty-eighth annual ACM symposium on Theory of computing (STOC), pp 634–643

Quantum Algorithms and Complexity for Continuous Problems

ANARGYROS PAPAGEORGIOU, JOSEPH F. TRAUB
Department of Computer Science,
Columbia University, New York, USA

Article Outline

Glossary

Black box model This model assumes we can collect knowledge about an input f through queries without knowing how the answer to the query is computed. A synonym for black box is oracle.

Classical computer A computer which does not use the principles of quantum computing to carry out its computations.

Computational complexity In this article, complexity for brevity. The minimal cost of solving a problem by an algorithm. Some authors use the word complexity when cost would be preferable. An upper bound on the complexity is given by the cost of an algorithm. A lower bound is given by a theorem which states there cannot be an algorithm which does better.

Continuous problem A problem involving real or complex functions of real or complex variables. Examples of continuous problem are integrals, path integrals, and partial differential equations.

Cost of an algorithm The price of executing an algorithm. The cost depends on the model of computation.

Discrete problem A problem whose inputs are from a countable set. Examples of discrete problems are integer factorization, traveling salesman and satisfiability.

ε-Approximation Most real-world continuous problems can only be solved numerically and therefore approximately, that is to within an error threshold ε. The definition of ε-approximation depends on the setting. See worst-case setting, randomized setting, quantum setting.

Information-based complexity The discipline that studies algorithms and complexity of continuous problems.

Model of computation The rules stating what is permitted in a computation and how much it costs. The model of computation is an abstraction of a physical computer. Examples of models are Turing machines, real number model, quantum circuit model.

Optimal algorithm An algorithm whose cost equals the complexity of the problem.

Promise A statement of what is known about a problem a priori before any queries are made. An example in quantum computation is the promise that an unknown 1-bit function is constant or balanced. In information-based complexity a promise is also called global information.

Quantum computing speedup The amount by which a quantum computer can solve a problem faster than a classical computer. To compute the speedup one must know the classical complexity and it is desirable to also know the quantum complexity. Grover proved quadratic speedup for search in an unstructured database. Its only conjectured that Shor's algorithm provides exponential speedup for integer factorization since the classical complexity is unknown.

Query One obtains knowledge about a particular input through queries. For example, if the problem is numerical approximation of $\int_0^1 f(x)\,\mathrm{d}x$ a query might be the evaluation of f at a point. In information-based complexity the same concept is called an information operation.

Quantum setting There are a number of quantum settings. An example is a guarantee of error at most ε with probability greater than 1/2.

Qubit complexity The minimal number of qubits to solve a problem.

Query complexity The minimal number of queries required to solve the problem.

Randomized setting In this setting the expected error with respect to the probability measure generating the random variables is at most ε. The computation is randomized. An important example of a randomized algorithm is the Monte Carlo method.

Worst-case setting In this setting an error of at most ε is guaranteed for all inputs satisfying the promise. The computation is deterministic.

Definition of the Subject

Most continuous mathematical formulations arising in science and engineering can only be solved numerically

and therefore approximately. We shall always assume that we are dealing with a numerical approximation to the solution.

There are two major motivations for studying quantum algorithms and complexity for continuous problems.

1. Are quantum computers more powerful than classical computers for important scientific problems? How much more powerful? This would answer the question posed by Nielsen and Chuang (p. 47 in [48]).
2. Many important scientific and engineering problems have continuous formulations. These problems occur in fields such as physics, chemistry, engineering and finance. The continuous formulations include path integration, partial differential equations (in particular, the Schrödinger equation) and continuous optimization.

To answer the first question we must know the classical computational complexity (for brevity, complexity) of the problem. There have been decades of research on the classical complexity of continuous problems in the field of information-based complexity; see, for example, the monographs [49,57,59,67,68,71,75]. The reason we know the complexity of many continuous problems is that we can use adversary arguments to obtain their query complexity. See, for example, [54] for an exposition. This may be contrasted with the classical complexity of discrete problems where we have only conjectures such as $P \neq NP$. Even the classical complexity of the factorization of large integers is unknown. Knowing the classical complexity of a continuous problem we obtain the quantum computation speedup if we know the quantum complexity. If we know an upper bound on the quantum complexity through the cost of a particular quantum algorithm then we can obtain a lower bound on the quantum speedup. To state and prove complexity theorems, the mathematical formulation of the problem, the promise about the class of inputs, and the model of computation must be precisely specified. See [66] for a discussion of the model of computation for continuous problems.

Regarding the second motivation, in this article we will report on high-dimensional integration, path integration, Feynman path integration, the smallest eigenvalue of a differential equation, approximation, partial differential equations, ordinary differential equations and gradient estimation. We will also briefly report on the simulation of quantum systems on a quantum computer.

Introduction

We provide a summary of the contents of the article.

Section "Overview of Quantum Algorithms" We define basic concepts and notation including quantum algorithm, continuous problem, query complexity and qubit complexity.

Section "Integration" High-dimensional integration, often in hundreds or thousands of variables, is one of the most commonly occurring continuous problems in science. In Subsect. "Classical Computer" we report on complexity results on a classical computer. For illustration we begin with a one-dimensional problem and give a simple example of an adversary argument. We then move on the d-dimensional case and indicate that in the worst case the complexity is exponential in d; the problem suffers the curse of dimensionality. The curse can be broken by the Monte Carlo algorithm. In Subsect. "Quantum Computer" we report on the algorithms and complexity results on a quantum computer. Under certain assumptions on the promise the quantum query complexity enjoys exponential speedup over classical worst case query complexity.

A number of the problems we will discuss enjoy exponential quantum query speedup. This does not contradict Beals et al. [6] who prove that speedup can only be polynomial. The reason is that [6] deals with problems concerning total Boolean functions.

For many classes of integrands there is quadratic speedup over the classical randomized query complexity. This is the same speedup as enjoyed by Grover's search algorithm of an unstructured database. To obtain the quantum query complexity one has to give matching upper and lower bounds. As usual the upper bound is given by an algorithm, the lower bound by a theorem. The upper bound is given by the amplitude amplification algorithm of Brassard et al. [12]. We outline a method for approximating high-dimensional integrals using this algorithm. The quantum query complexity lower bounds for integration are based on the lower bounds of Nayak and Wu [47] for computing the mean of a Boolean function.

Section "Path Integration" We define a path integral and provide an example due to Feynman. In Subsect. "Classical Computer" we report on complexity results on a classical computer. If the promise is that the class of integrands has finite smoothness, then path integration is intractable in the worst case. If the promise is that the class of integrands consists of entire functions the query complexity is tractable even in the worst case. For smooth functions intractability is broken by Monte Carlo. In Subsect. "Quantum Computer" we report on the algorithm and complexity on a quantum computer. The quantum query complexity enjoys Grover-type speedup over clas-

sical randomized query complexity. We outline the quantum algorithm.

Section "Feynman–Kac Path Integration" The Feynman–Kac path integral provides the solution to the diffusion equation. In Subsect. "Classical Computer" we report on algorithms and complexity on a classical computer. In the worst case for a d-dimensional Feynman–Kac path integral the problem again suffers the curse of dimensionality which can be broken by Monte Carlo. In Subsect. "Quantum Computer" we indicate the algorithm and query complexity on a quantum computer.

Section "Eigenvalue Approximation" One of the most important problems in physics and chemistry is approximating the ground state energy governed by a differential equation. Typically, numerical algorithms on a classical computer are well known. Our focus is on the Sturm–Liouville eigenvalue (SLE) problem where the first complexity results were recently obtained. The SLE equation is also called the time-independent Schrödinger equation. In Subsect. "Classical Computer" we present an algorithm on a classical computer. The worst case query complexity suffers the curse of dimensionality. We also state a randomized algorithm. The randomized query complexity is unknown for $d > 2$ and is an important open question. In Subsect. "Quantum Computer" we outline an algorithm for a quantum computer. The quantum query complexity is not known when $d > 1$. It has been shown that it is not exponential in d; the problem is tractable on a quantum computer.

Section "Qubit Complexity" So far we've focused on query complexity. For the foreseeable future the number of qubits will be a crucial computational resource. We give a general lower bound on the qubit complexity of continuous problems. We show that because of this lower bound there's a problem that cannot be solved on a quantum computer but that's easy to solve on a classical computer using Monte Carlo.

A definition of a quantum algorithm is given by (1); the queries are deterministic. Woźniakowski [77] introduced the quantum setting with randomized queries for continuous problems. For path integration there is an exponential reduction in the qubit complexity.

Section "Approximation" Approximating functions of d variables is a fundamental and generally hard problem. The complexity is sensitive to the norm, p, on the class of functions and to several other parameters. For example, if $p = \infty$ approximation suffers the curse of dimensionality in the worst and randomized classical cases. The problem remains intractable in the quantum setting.

Section "Partial Differential Equations" Elliptic partial differential equations have many important applications and have been extensively studied. In particular, consider an equation of order $2m$ in d variables. In the classical worst case setting the problem is intractable. The classical randomized and quantum settings were only recently studied. The conclusion is that the quantum may or may not provide a polynomial speedup; it depends on certain parameters.

Section "Ordinary Differential Equations" Consider a system of initial value ordinary equations in d variables. Assume that the right hand sides satisfy a Hölder condition. The problem is tractable even in the worst case with the exponent of ε^{-1} depending on the Hölder class parameters. The complexity of classical randomized and quantum algorithms have only recently been obtained. The quantum setting yields a polynomial speedup.

Section "Gradient Estimation" Jordan [37] showed that approximating the gradient of a function can be done with a single query on a quantum computer although it takes $d + 1$ function evaluations on a classical computer. A simplified version of Jordan's algorithm is presented.

Section "Simulation of Quantum Systems on Quantum Computers" There is a large and varied literature on simulation of quantum systems on quantum computers. The focus in these papers is typically on the cost of particular classical and quantum algorithms without complexity analysis and therefore without speedup results. To give the reader a taste of this area we list some sample papers.

Section "Future Directions" We briefly indicate a number of open questions.

Overview of Quantum Algorithms

A quantum algorithm consists of a sequence of unitary transformations applied to an initial state. The result of the algorithm is obtained by measuring its final state. The quantum model of computation is discussed in detail in [6,7,8,17,27,48] and we summarize it here as it applies to continuous problems.

The initial state $|\psi_0\rangle$ of the algorithm is a unit vector of the Hilbert space $\mathcal{H}_\nu = \mathbb{C}^2 \otimes \cdots \otimes \mathbb{C}^2$, ν times, for some appropriately chosen integer ν, where $\mathbb{C}^2$ is the two dimensional space of complex numbers. The dimension of

$\mathcal{H}_\nu$ is 2^ν. The number ν denotes the number of qubits used by the quantum algorithm.

The final state $|\psi\rangle$ is also a unit vector of $\mathcal{H}_\nu$ and is obtained from the initial state $|\psi_0\rangle$ through a sequence of unitary $2^\nu \times 2^\nu$ matrices, i. e.,

$$|\psi\rangle_f := U_T Q_f U_{T-1} Q_f \cdots U_1 Q_f U_0 |\psi_0\rangle \,. \tag{1}$$

The unitary matrix Q_f is called a quantum query and is used to provide information to the algorithm about a function f. Q_f depends on n function evaluations $f(t_1), \ldots, f(t_n)$, at deterministically chosen points, $n \le 2^\nu$. The $U_0, U_1, \ldots, U_T$ are unitary matrices that do not depend on f. The integer T denotes the number of quantum queries.

For algorithms solving discrete problems, such as Grover's algorithm for the search of an unordered database [26], the input f is considered to be a Boolean function. The most commonly studied quantum query is the *bit* query. For a Boolean function $f\colon \{0, \ldots, 2^m - 1\} \to \{0, 1\}$, the bit query is defined by

$$Q_f |j\rangle |k\rangle = |j\rangle |k \oplus f(j)\rangle \,. \tag{2}$$

Here $\nu = m + 1$, $|j\rangle \in \mathcal{H}_m$, and $|k\rangle \in \mathcal{H}_1$ with $\oplus$ denoting addition modulo 2. For a real function f the query is constructed by taking the most significant bits of the function f evaluated at some points t_j. More precisely, as in [27], the bit query for f has the form

$$Q_f |j\rangle |k\rangle = |j\rangle |k \oplus \beta(f(\tau(j)))\rangle \,, \tag{3}$$

where the number of qubits is now $\nu = m' + m''$ and $|j\rangle \in \mathcal{H}_{m'}$, $|k\rangle \in \mathcal{H}_{m''}$. The functions β and τ are used to discretize the domain $\mathcal{D}$ and the range $\mathcal{R}$ of f, respectively. Therefore, $\beta\colon \mathcal{R} \to \{0, 1, \ldots, 2^{m''} - 1\}$ and $\tau\colon \{0, 1, \ldots, 2^{m'} - 1\} \to \mathcal{D}$. Hence, we compute f at $t_j = \tau(j)$ and then take the m'' most significant bits of $f(t_j)$ by $\beta(f(t_j))$, for the details and the possible use of ancillary qubits see [27].

At the end of the quantum algorithm, the final state $|\psi_f\rangle$ is measured. The measurement produces one of M outcomes, where $M \le 2^\nu$. Outcome $j \in \{0, 1, \ldots, M-1\}$ occurs with probability $p_f(j)$, which depends on j and the input f. Knowing the outcome j, we classically compute the final result $\phi_f(j)$ of the algorithm.

In principle, quantum algorithms may have measurements applied between sequences of unitary transformations of the form presented above. However, any algorithm with multiple measurements can be simulated by a quantum algorithm with only one measurement [8].

Let S be a linear or nonlinear operator such that

$$S\colon \mathcal{F} \to \mathcal{G} \,. \tag{4}$$

Typically, $\mathcal{F}$ is a linear space of real functions of several variables, and $\mathcal{G}$ is a normed linear space. We wish to approximate $S(f)$ to within ε for $f \in \mathcal{F}$. We approximate $S(f)$ using n function evaluations $f(t_1), \ldots, f(t_n)$ at deterministically and a priori chosen sample points. The quantum query Q_f encodes this information, and the quantum algorithm obtains this information from Q_f.

Without loss of generality, we consider algorithms that approximate $S(f)$ with probability $p \ge 3/4$. We can boost the success probability of an algorithm to become arbitrarily close to one by repeating the algorithm a number of times. The success probability becomes at least $1 - \delta$ with a number of repetitions proportional to $\log \delta^{-1}$.

The local error of the quantum algorithm (1) that computes the approximation $\phi_f(j)$, for $f \in \mathcal{F}$ and the outcome $j \in \{0, 1, \ldots, M-1\}$, is defined by

$$e(\phi_f, S) = \min\left\{\alpha\colon \sum_{j\colon \|S(f) - \phi_f(j)\| \le \alpha} p_f(j) \ge \frac{3}{4}\right\} \,, \tag{5}$$

where $p_f(j)$ denotes the probability of obtaining outcome j for the function f. The worst case error of a quantum algorithm ϕ is defined by

$$e^{\mathrm{quant}}(\phi, S) = \sup_{f \in \mathcal{F}} e(\phi_f, S) \,. \tag{6}$$

The query complexity $\mathrm{comp}^{\mathrm{query}}(\varepsilon, S)$ of the problem S is the minimal number of queries necessary for approximating the solution with accuracy ε, i. e.,

$$\mathrm{comp}^{\mathrm{query}}(\varepsilon) = \min\{T\colon \exists\, \phi \text{ such that } e^{\mathrm{quant}}(\phi, S) \le \varepsilon\} \,. \tag{7}$$

Similarly, the qubit complexity of the problem S is the minimal number of qubits necessary for approximating the solution with accuracy ε, i. e.,

$$\mathrm{comp}^{\mathrm{qubit}}(\varepsilon) = \min\{\nu\colon \exists\, \phi \text{ such that } e^{\mathrm{quant}}(\phi, S) \le \varepsilon\} \,. \tag{8}$$

Integration

Integration is one of the most commonly occurring mathematical problems. One reason is that when one seeks the expectation of a continuous process one has to compute an integral. Often the integrals are over hundreds or thousands of variables. Path integrals are taken over an infinite number of variables. See Sect. "Path Integration".

Classical Computer

We begin with a one dimensional example to illustrate some basic concepts before moving to the d-dimen-

sional case (Number of dimensions and number of variables are used interchangeably.) Our simple example is to approximate

$$I(f) = \int_0^1 f(x)\,\mathrm{d}x\,.$$

For most integrands we can not use the fundamental theorem of calculus to compute the integral analytically; we have to approximate numerically (most real world continuous problems have to be approximated numerically). We have to make a promise about f. Assume

$$F_1 = \{f\colon [0,1] \to \mathbb{R} \mid \text{continuous and } |f(x)| \le 1,\ x \in [0,1]\}\,.$$

Use queries to compute

$$y_f = [f(x_1), \ldots, f(x_n)]\,.$$

We show that with this promise one cannot guarantee an ε-approximation on a classical computer. We use a simple adversary argument. Choose arbitrary numbers $x_1, \ldots, x_n \in [0,1]$. The adversary answers all these queries by answering $f(x_i) = 0$, $i = 1, \ldots, n$.

What is f? It could be $f_1 \equiv 0$ and $\int_0^1 f_1(x)\,\mathrm{d}x = 0$ or it could be the f_2 shown in Fig. 1. The value of $\int_0^1 f_2(x)\,\mathrm{d}x$ can be arbitrarily close to 1. Since $f_1(x)$ and $f_2(x)$ are indistinguishable with these query answers and this promise, it is impossible to guarantee an ε-approximation on a classical computer with $\varepsilon < 1/2$. We will return to this example in Sect. "Qubit Complexity".

We move on to the d-dimensional case. Assume that the integration limits are finite. For simplicity we assume we are integrating over the unit cube $[0,1]^d$. So our problem is

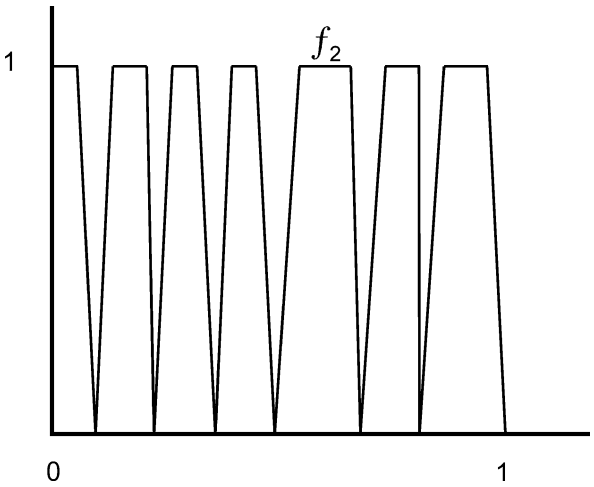
$$I(f) = \int_{[0,1]^d} f(x)\,\mathrm{d}x\,.$$

If our promise is that $f \in F_d$, where

$$F_d = \{f\colon [0,1]^d \to \mathbb{R} \mid \text{continuous and } |f(x)| \le 1,\ x \in [0,1]^d\},$$

then we can not compute an ε-approximation regardless of the value of d. Our promise has to be stronger. Assume that integrand class has *smoothness* r. There are various ways to define smoothness r for functions of d variables. See, for example, the definition on p. 25 in [67]. (For other definitions see [71].) With this definition Bakhvalov [4] showed that the query complexity is

$$\mathrm{comp}^{\mathrm{query}}_{\mathrm{clas-wor}}(\varepsilon) = \Theta\left(\varepsilon^{-d/r}\right). \tag{9}$$



Quantum Algorithms and Complexity for Continuous Problems, Figure 1
All the function evaluations are equal to zero but the integral is close to one

This is the worst case query complexity on a classical computer.

What does this formula imply? If $r = 0$ the promise is that the functions are only continuous but have no smoothness. Then the query complexity is ∞, that is, we cannot guarantee an ε-approximation. If r and ε are fixed the query complexity is an exponential function of d. We say the problem is intractable. Following R. Bellman this is also called the *curse of dimensionality*. In particular, let $r = 1$. Then the promise is that the integrands have one derivative and the query complexity is $\Theta(\varepsilon^{-d})$.

The curse of dimensionality is present for many for continuous problems in the worst case setting. Breaking the curse is one of the central issues of information-based complexity. For high-dimensional integration the curse can be broken for F_d by the Monte Carlo algorithm which is a special case of a randomized algorithm. The Monte Carlo algorithm is defined by

$$\phi^{\mathrm{MC}}(f) = \frac{1}{n}\sum_{i=1}^{n} f(x_i)\,, \tag{10}$$

where the x_i are chosen independently from a uniform distribution. Then the expected error is

$$e^{\mathrm{MC}}(f) = \frac{\sqrt{\mathrm{var}(f)}}{\sqrt{n}}\,,$$

where

$$\mathrm{var}(f) = \int_{[0,1]^d} f^2(x)\,\mathrm{d}x - \left\{\int_{[0,1]^d} f(x)\,\mathrm{d}x\right\}^2$$

is the variance of f. For the promise $f \in F_d$ the query complexity is given by

$$\mathrm{comp}^{\mathrm{query}}_{\mathrm{clas-ran}}(\varepsilon) = \Theta(\varepsilon^{-2})\,. \tag{11}$$

This is the randomized query complexity on a classical computer.

Thus Monte Carlo breaks the curse of dimensionality for the integration problem; the problem is tractable. Why should picking sample points at random be much better than picking them deterministically in the optimal way? This is possible because we have replaced the guarantee of the worst case setting by the stochastic assurance of the randomized setting. There is no free lunch!

The reader will note that (11) is a complexity result even though it is the cost of a particular algorithm, the Monte Carlo algorithm. The reason is that for integrands satisfying this promise Monte Carlo has been proven optimal. It is known that if the integrands are smoother Monte Carlo is not optimal; see p. 32 in [67].

In the worst case setting (deterministic) the query complexity is infinite if $f \in F_d$. In the randomized setting the query complexity is independent of d if $f \in F_d$. This will provide us guidance when we introduce Monte Carlo sampling into the model of quantum computation in Sect. "Qubit Complexity".

Generally pseudo-random numbers are used in the implementation of Monte Carlo on a classical computer. The quality of a pseudo-random number generator is usually tested statistically; see, for example, Knuth [41]. Will (11) still hold if a pseudo-random number generator is used? An affirmative answer is given by [69] provided some care is taken in the choice of the generator and f is Lipschitz.

Quantum Computer

We have seen that Monte Carlo breaks the curse of dimensionality for high-dimensional integration on a classical computer and that the query complexity is of order ε^{-2}. Can we do better on a quantum computer?

The short answer is yes. Under certain assumptions on the promises, which will be made precise below, the quantum query complexity enjoys exponential speedup over the classical worst case query complexity and quadratic speedup over the classical randomized query complexity. The latter is the same speedup as enjoyed by Grover's search algorithm of an unstructured database [26].

To show that the quantum query complexity is of order ε^{-1} we have to give matching, or close to matching upper and lower bounds. Usually, the upper bound is given by an algorithm, the lower bound by a theorem. The upper bound is given by the amplitude amplification algorithm of Brassard et al. [12] which we describe briefly.

The amplitude amplification algorithm of Brassard et al. computes the mean

$$\mathrm{SUM}(f) = \frac{1}{N}\sum_{i=0}^{N-1} f(i)\,,$$

of a Boolean function $f\colon \{0,1,\ldots,N-1\} \to \{0,1\}$, where $N = 2^k$, with error ε and probability at least $8/\pi^2$, using a number of queries proportional to $\min\{\varepsilon^{-1}, N\}$. Moreover, Nayak and Wu [47] show that the order of magnitude of the number of queries of this algorithm is optimal. Without loss of generality we can assume that $\varepsilon^{-1} \ll N$.

Perhaps the easiest way to understand the algorithm is to consider the operator

$$G = (I - |\psi\rangle\langle\psi|)O_f,$$

that is used in Grover's search algorithm. Here

$$O_f|x\rangle = (-1)^{f(x)}|x\rangle, \quad x \in \{0,1\}^k\,,$$

denotes the query, which is slightly different yet equivalent [48] to the one we defined in (2). Let

$$|\psi\rangle = \sqrt{\frac{1}{N}}\sum_x |x\rangle$$

be the equally weighted superposition of all the states.

If M denotes the number of assignments for which f has the value 1 then

$$\mathrm{SUM}(f) = \frac{M}{N}\,.$$

Without loss of generality $1 \le M \le N-1$. Consider the space $\mathcal{H}$ spanned by the states

$$|\psi_0\rangle = \sqrt{\frac{1}{N-M}}\sum_{x\colon f(x)=0} |x\rangle \quad \text{and}$$

$$|\psi_1\rangle = \sqrt{\frac{1}{M}}\sum_{x\colon f(x)=1} |x\rangle\,.$$

Then

$$|\psi\rangle = \cos(\theta/2)|\psi_0\rangle + \sin(\theta/2)|\psi_1\rangle$$

where $\sin(\theta/2) = \sqrt{M/N}$, and $\theta/2$ is the angle between the states $|\psi\rangle$ and $|\psi_0\rangle$. Thus

$$\sin^2\left(\frac{\theta}{2}\right) = \frac{M}{N}\,.$$

Now consider the operator G restricted to $\mathcal{H}$ which has the matrix representation

$$G = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} .$$

Its eigenvalues are $\lambda_\pm = e^{\pm i\theta}$, $i = \sqrt{-1}$, and let $|\xi_\pm\rangle$ denote the corresponding eigenvectors.

We can express $|\psi\rangle$ using the $|\xi_\pm\rangle$ to get

$$|\psi\rangle = a|\xi_-\rangle + b|\xi_+\rangle ,$$

with $a, b \in \mathbb{C}$, $|a|^2 + |b|^2 = 1$. This implies that phase estimation [48] with G^p, $p = 2^j$, $j = 0, \ldots, t-1$, and initial state $|0\rangle^{\otimes t}|\psi\rangle$ can be used to approximate either θ or $2\pi - \theta$ with error proportional to 2^{-t}. Note that the first t qubits of the initial state determine the accuracy of phase estimation. Indeed, let $\tilde{\phi}$ be the result of phase estimation. Since $\sin^2(\theta/2) = \sin^2(\pi - \theta/2)$,

$$\left|\sin^2(\pi\tilde{\phi}) - \frac{N}{M}\right| = O\left(2^{-t}\right) ,$$

with probability at least $8/\pi^2$; see [12,48] for the details. Setting $t = \Theta(\log \varepsilon^{-1})$ and observing that phase estimation requires a number of applications of G (or queries O_f) proportional to ε^{-1} yields the result. (The complexity of quantum algorithms for the average case approximation of the Boolean mean has also been studied [35,52]).

For a real function $f\colon \{0, 1, \ldots, N-1\} \to [0, 1]$, we can approximate

$$\mathrm{SUM}(f) = \frac{1}{N}\sum_{i=0}^{N-1} f(i) ,$$

by reducing the problem to the computation of the Boolean mean. One way to derive this reduction is to truncate $f(i)$ to the desired number of significant bits, typically, polynomial in $\log \varepsilon^{-1}$, and then to use the bit representation of the function values to derive a Boolean function whose mean is equal to the mean of the truncated real function, see, e. g. [77]. The truncation of the function values is formally expressed through the mapping β in (3). Variations of this idea have been used in the literature [3,27,50].

Similarly, one discretizes the domain of a function $f\colon [0, 1]^d \to [0, 1]$ using the function τ in (3) and then uses the amplitude amplification algorithm to compute the average

$$\mathrm{SUM}(f) = \frac{1}{N}\sum_{i=0}^{N-1} f(x_i) , \tag{12}$$

for $x_i \in [0, 1]^d$, $i = 0, \ldots, N-1$.

The quantum query complexity of computing the average (12) is of order ε^{-1}. On the other hand, the classical deterministic worst case query complexity is proportional to N (recall that $\varepsilon^{-1} \ll N$), and the classical randomized query complexity is proportional to ε^{-2}.

We now turn to the approximation of high-dimensional integrals and outline an algorithm for solving this problem. Suppose $f\colon [0, 1]^d \to [0, 1]$ is a function for which we are given some promise, for instance, that f has smoothness r. The algorithm integrating f with accuracy ε has two parts. First, using function evaluations $f(x_i)$, $i = 1, \ldots, n$, at deterministic points, it approximates f classically, by a function $\hat{f}$ with error ε_1, i. e.,

$$\|f - \hat{f}\| \le \varepsilon_1 ,$$

where $\|\cdot\|$ is the L_∞ norm. The complexity of this problem has been extensively studied and there are numerous results [49,67,71] specifying the optimal choice of n and the points x_i that must be used to achieve error ε_1. Thus

$$\int_{[0,1]^d} f(x)\,\mathrm{d}x = \int_{[0,1]^d} \hat{f}(x)\,\mathrm{d}x + \int_{[0,1]^d} g(x)\,\mathrm{d}x ,$$

where $g = f - \hat{f}$. Since $\hat{f}$ is known and depends linearly on the $f(x_i)$ the algorithm proceeds to integrate it exactly. So it suffices to approximate $\int_{[0,1]^d} g(x)\,\mathrm{d}x$ knowing that $\|g\| \le \varepsilon_1$.

The second part of the algorithm approximates the integral of g using the amplitude amplification algorithm to compute

$$\mathrm{SUM}(g) = \frac{1}{N}\sum_{i=0}^{N-1} g(y_i) ,$$

for certain points $y_i \in [0, 1]^d$, with error ε_2. Once more, there are many results, see [67,71] for surveys specifying the optimal N and the points y_i, so that $\mathrm{SUM}(g)$ approximates $\int_{[0,1]^d} g(x)\,\mathrm{d}x$ with error ε_2. Finally, the algorithm approximates the original integral by the sum of the results of its two parts. The overall error of the algorithm is proportional to $\varepsilon = \varepsilon_1 + \varepsilon_2$.

Variations of the integration algorithm we described are known to have optimal query complexity for a number of different promises [27,29,34,50]. The quantum query complexity lower bounds for integration are based on the lower bounds of Nayak and Wu [47] for computing the Boolean mean. The quantum algorithms offer an exponential speedup over classical deterministic algorithms and a polynomial speedup over classical randomized algorithms for the query complexity of high-dimensional integration. Tabel 1 summarizes the query complexity (up to polylog factors) of high-dimensional integration in the

Quantum Algorithms and Complexity for Continuous Problems, Table 1
Query complexity of high-dimensional integration

	Worst case	Randomized	Quantum
$F_d^{r,\alpha}$	$\varepsilon^{-d/(r+\alpha)}$	$\varepsilon^{-2d/(2(r+\alpha)+d)}$	$\varepsilon^{-d/(r+\alpha+d)}$
$W_{p,d}^r$, $2 \le p \le \infty$	$\varepsilon^{-d/r}$	$\varepsilon^{-2d/(2r+d)}$	$\varepsilon^{-d/(r+d)}$
$W_{p,d}^r$, $1 \le p \le 2$	$\varepsilon^{-d/r}$	$\varepsilon^{-pd/(rp+pd-d)}$	$\varepsilon^{-d/(r+d)}$
$W_{1,d}^r$	$\varepsilon^{-d/r}$	$\varepsilon^{-d/r}$	$\varepsilon^{-d/(r+d)}$

worst case, randomized and quantum setting for functions belonging to Hölder classes $F_d^{r,\alpha}$ and Sobolev spaces $W_{p,d}^r$. Heinrich obtained most of the quantum query complexity results in a series of papers, which we cited earlier. Heinrich summarized his results in [28] where a corresponding table showing error bounds can be found.

Abrams and Williams [3] were the first to apply the amplitude amplification algorithm to high-dimensional integration. Novak [50] was the first to spell out his promises and thus obtained the first complexity results for high-dimensional integration.

Path Integration

A path integral is defined as

$$I(f) = \int_X f(x)\,\mu(\mathrm{d}x)\,, \tag{13}$$

where μ is a probability measure on an infinite-dimensional space X. It can be viewed as an infinite-dimensional integral. For illustration we give an example due to R. Feynman. In classical mechanics a particle at a certain position at time t_0 has a unique trajectory to its position at time t_1. Quantum mechanically there are an infinite number of possible trajectories which Feynman called histories, see Fig. 2. Feynman summed over the histories. If one goes to the limit one gets a path integral. Setting $t_0 = 0$, $t_1 = 1$ this integral is

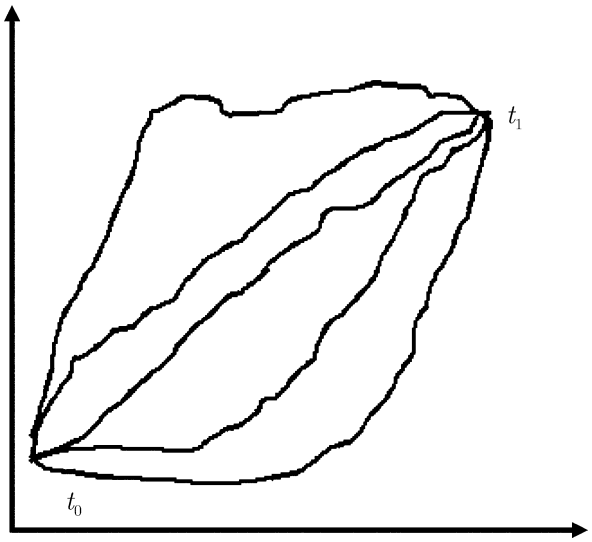
$$I(f) = \int_{C[0,1]} f(x)\,\mu(\mathrm{d}x)\,,$$

which is a special case of (13).

Path integration occurs in numerous applications including quantum mechanics, quantum chemistry, statistical mechanics, and mathematical finance.

Classical Computer

The first complexity analysis of path integration is due to Wasilkowski and Woźniakowski [72]; see also Chap. 5 in [67]. They studied the deterministic worst case and randomized settings and assume that μ is a Gaussian measure; an important special case of a Gaussian measure is a Wiener measure. They make the promise that F, the class of integrands, consists of functions $f\colon X \to \mathbb{R}$ whose rth Fréchet derivatives are continuous and uniformly bounded by unity. If r is finite then path integration is intractable. Curbera [20] showed that the worst case query complexity is of order $\varepsilon^{\varepsilon^{-\beta}}$ where β is a positive number depending on r.



Quantum Algorithms and Complexity for Continuous Problems, Figure 2
Different trajectories of a particle

Wasilkowski and Woźniakowski [72] also considered the promise that F consists of entire functions and that μ is the Wiener measure. Then the query complexity is a polynomial in ε^{-1}. More precisely, they provide an algorithm for calculating a worst case ε-approximation with cost of order ε^{-p} and the problem is tractable with this promise. The exponent p depends on the particular Gaussian measure; for the Wiener measure $p = 2/3$.

We return to the promise that the smoothness r is finite. Since this problem is intractable in the worst case, Wasilkowski and Woźniakowski [72] ask whether settling for a stochastic assurance will break the intractability. The obvious approach is to approximate the infinite-dimensional integral by a d-dimensional integral where d may be large (or even huge) since d is polynomial in ε^{-1}. Then Monte Carlo may be used since its speed of convergence does not depend on d. Modulo an assumption that the nth eigenvalue of the covariance operator of μ does not decrease too fast the randomized query complexity is roughly ε^{-2}. Thus Monte Carlo is optimal.

Quantum Computer

Just as with finite dimensional integration Monte Carlo makes path integration tractable on a classical computer and the query complexity is of order ε^{-2}. Again we ask whether we can do better on a quantum computer and again the short answer is yes. Under certain assumptions on the promises, which will be made precise below, the quantum query complexity is of order ε^{-1}. Thus quantum query complexity enjoys exponential speedup over the classical worst case query complexity and quadratic speedup over the classical randomized query complexity. Again the latter is the same speedup as enjoyed by Grover's search algorithm of an unstructured database.

The idea for solving path integration on a quantum computer is fairly simple but the analysis is not so easy. So we outline the algorithm without considering the details. We start with a classical deterministic algorithm that uses an average as in (12) to approximate the path integral $I(f)$ with error ε in the worst case. The number of terms N of the this average is an exponential function of ε^{-1}. Nevertheless, on a quantum computer we can approximate the average, using the amplitude amplification algorithm as we discussed in Sect. "Integration (Quantum Computer)", with cost that depends on $\log N$ and is, therefore, a polynomial in ε^{-1}, Sect. 6 in [70].

A summary of the promises and the results in [70] follows. The measure μ is Gaussian and the eigenvalues of its covariance operator is of order j^{-h}, $h > 1$. For the Wiener measure occurring in many applications we have $h = 2$. The class of integrands consists of functions f whose rth Fréchet derivatives are continuous and uniformly bounded by unity. In particular assume the integrands are at least Lipschitz. Then

- Path integration on a quantum computer is tractable.
- Query complexity on a quantum computer enjoys exponential speedup over the worst case and quadratic speedup over the classical randomized query complexity. More precisely, the number of quantum queries is at most $4.22\varepsilon^{-1}$.

Results on the qubit complexity of path integration will be given in Sect. "Qubit Complexity". Details of an algorithm for computing an ε-approximation to a path integral on a quantum computer may be found in Sect. 6 in [70].

Feynman–Kac Path Integration

An important special case of a path integral is a Feynman–Kac path integral. Assume that X is the space C of continuous functions and that the measure μ is the Wiener measure w. Feynman–Kac path integrals occur in many applications; see [23]. For example consider the diffusion equation

$$\frac{\partial z}{\partial t}(u,t) = \frac{1}{2}\frac{\partial^2 z}{\partial u^2}(u,t) + V(u)\,z(u,t)$$
$$z(u,0) = v(u)\,,$$

where $u \in \mathbb{R}$, $t > 0$, V is a potential function, and v is an initial condition function. Under mild conditions on v and V the solution is given by the Feynman–Kac path integral

$$z(u,t) = \int_C v(x(t)+u)\mathrm{e}^{\int_0^t V(x(s)+u)\,\mathrm{d}s}\,w(\mathrm{d}x)\,. \qquad (14)$$

The problem generalizes to the multivariate case by considering the diffusion equation

$$\frac{\partial z}{\partial t}(u,t) = \frac{1}{2}\Delta z(u,t) + V(u)\,z(u,t)$$
$$z(u,0) = v(u)\,,$$

with $u \in \mathbb{R}^d$, $t > 0$, and $V, v\colon \mathbb{R}^d \to \mathbb{R}$, the potential and the initial value function, respectively. As usual, Δ denotes the Laplacian. The solution is given by the Feynman–Kac path integral

$$z(u,t) = \int_C v(x(t)+u)\mathrm{e}^{\int_0^t V(x(s)+u)\,\mathrm{d}s}\,w(\mathrm{d}x)\,,$$

where C is the set of continuous functions $x\colon \mathbb{R}_+ \to \mathbb{R}^d$ such that $x(0) = 0$.

Note that there are two kinds of dimension here. A Feynman–Kac path integral is infinite dimensional since we are integrating over continuous functions. Furthermore u is a function of d variables.

Classical Computer

We begin with the case when u is a scalar and then move to the case where u is a multivariate function. There have been a number of papers on the numerical solution of (14); see, for example [14].

The usual attack is to solve the problem with a stochastic assurance using randomization. For simplicity we make the promise that $u = 1$ and V is four times continuously differentiable. Then by Chorin's algorithm [16], the total cost is of order $\varepsilon^{-2.5}$.

The first complexity analysis may be found in Plaskota et al. [58] where a new algorithm is defined which enjoys certain optimality properties. They construct an algorithm which computes an ε-approximation at cost of order $\varepsilon^{-.25}$ and show that the worst case complexity is of the same order. Hence the exponent of ε^{-1} is an order of magnitude smaller and with a worst case rather than a stochastic guarantee. However, the new algorithm requires a nu-

merically difficult precomputation which may limit its applicability.

We next report on multivariate Feynman–Kac path integration. First consider the worst case setting with the promise that v and V are r times continuously differentiable with r finite. Kwas and Li [43] proved that the query complexity is of order $\varepsilon^{-d/r}$. Therefore in the worst case setting the problem suffers the curse of dimensionality.

The randomized setting was studied by Kwas [42]. He showed that the curse of dimensionality is broken by using Monte Carlo using a number of queries of order ε^{-2}. The number of queries can be further improved to $\varepsilon^{-2/(1+2r/d)}$, which is the optimal number of queries, by a bit more complicated algorithm. The randomized algorithms require extensive precomputing [42,43].

Quantum Computer

Multivariate multivariate Feynman–Kac path integration on a quantum computer was studied in [42]. With the promise as in the worst and randomized case, Kwas presents an algorithm and complexity analysis. He exhibits a quantum algorithm that uses a number of queries of order ε^{-1} that is based on the Monte Carlo algorithm. He shows that the query complexity is of order $\varepsilon^{-1/(1+r/d)}$ and is achieved by a bit more complicated quantum algorithm. Just as in the randomized case the quantum algorithms require extensive precomputing [42,43].

Eigenvalue Approximation

Eigenvalue problems for differential operators arising in physics and engineering have been extensively studied in the literature; see, e. g. [5,18,19,22,25,40,63,65]. Typically, the mathematical properties of the eigenvalues and the corresponding eigenfunctions are known and so are numerical algorithms approximating them on a classical computer. Nevertheless, the complexity of approximating eigenvalues in the worst, randomized and quantum settings has only recently been addressed for the Sturm–Liouville eigenvalue problem [53,55] (see also [10] for quantum lower bounds with a different kind of query than the one we discuss in this article). In some cases we have sharp complexity estimates but there are important questions that remain open.

Most of our discussion here concerns the complexity of approximating the smallest eigenvalue of a Sturm–Liouville eigenvalue problem. We will conclude this section by briefly addressing quantum algorithms for other eigenvalue problems.

In the physics literature this problem is called the time-independent Schrödinger equation. The smallest eigenvalue is the energy of the ground state. In the mathematics literature it is called the Sturm–Liouville eigenvalue problem.

Let $I_d = [0,1]^d$ and consider the class of functions

$$\mathbf{Q} = \left\{ q\colon I_d \to [0,1] \,\middle|\, q\,,\ D_j q\colon = \frac{\partial q}{\partial x_j} \in C(I_d)\,,\ \|D_j q\|_\infty \le 1\,,\ \|q\|_\infty \le 1 \right\},$$

where $\|\cdot\|_\infty$ denotes the supremum norm. For $q \in \mathbf{Q}$, define $\mathbb{L}_q\colon = -\Delta + q$, where $\Delta = \sum_{j=1}^d \partial^2/\partial x_j^2$ is the Laplacian, and consider the eigenvalue problem

$$\mathbb{L}_q u = \lambda u,\ x \in (0,1)^d\,, \tag{15}$$

$$u(x) \equiv 0,\ x \in \partial I_d\,. \tag{16}$$

In the variational form, the smallest eigenvalue $\lambda = \lambda(q)$ of (15), (16) is given by

$$\lambda(q) = \min_{0 \ne u \in H_0^1} \frac{\int_{I_d} \sum_{j=1}^d [D_j u(x)]^2 + q(x) u^2(x)\,\mathrm{d}x}{\int_{I_d} u^2(x)\,\mathrm{d}x}\,, \tag{17}$$

where H_0^1 is the space of all functions vanishing on the boundary of I_d having square integrable first order partial derivatives. We consider the complexity of classical and quantum algorithms approximating $\lambda(q)$ with error ε.

Classical Computer

In the worst case we discretize the differential operator on a grid of size $h = \Theta(\varepsilon^{-1})$ and obtain a matrix $M_\varepsilon = -\Delta_\varepsilon + B_\varepsilon$, of size proportional to $\varepsilon^{-d} \times \varepsilon^{-d}$. The matrix Δ_ε is the $(2d+1)$-point finite difference discretization of the Laplacian. The matrix B_ε is a diagonal matrix containing evaluations of q at the grid points. The smallest eigenvalue of M_ε approximates $\lambda(q)$ with error $O(\varepsilon)$ [73,74]. We compute the smallest eigenvalue of M_ε using the bisection method (p. 228 in [22]). The resulting algorithm uses a number of queries proportional to ε^{-d}. It turns out that this number of queries is optimal in the worst case, and the problem suffers from the curse of dimensionality.

The query complexity lower bounds are obtained by reducing the eigenvalue problem to high-dimensional integration. For this we use the perturbation formula [53,55]

$$\lambda(q) = \lambda(\tilde{q}) + \int_{I_d} \big(q(x) - \tilde{q}(x)\big)\, u_{\tilde{q}}^2(x)\,\mathrm{d}x + O\left(\|q - \tilde{q}\|_\infty^2\right), \tag{18}$$

where $q, \tilde{q} \in \mathbf{Q}$ and $u_{\tilde{q}}$ is the normalized eigenfunction corresponding to $\lambda(\tilde{q})$.

The same formula is used for lower bounds in the randomized setting. Namely, the query complexity is

$$\Omega\left(\varepsilon^{-2d/(d+2)}\right).$$

Moreover, we can use (18) to derive a randomized algorithm. First we approximate q by a function $\tilde{q}$ and then approximate $\lambda(q)$ by approximating the first two terms on the right hand side of (18); see Papageorgiou and Woźniakowski [55] for $d = 1$, and Papageorgiou [53] for general d. However, this algorithm uses

$$O\left(\varepsilon^{-\max(2/3,d/2)}\right),$$

queries. So, it is optimal only when $d \leq 2$. Determining the randomized complexity for $d > 2$ and determining if the randomized complexity is an exponential function of d are important open questions.

Quantum Computer

The perturbation formula (18) can be used to show that the quantum query complexity is

$$\Omega\left(\varepsilon^{-d/(d+1)}\right).$$

As in the randomized case, we can use (18) to derive an algorithm that uses $O(\varepsilon^{-d/2})$ quantum queries. The difference between the quantum algorithm and the randomized algorithm is that the former uses the amplitude amplification algorithm to approximate the integral on the right hand side of (18) instead of Monte Carlo. The algorithm is optimal only when $d = 1$ [53,55].

For general d the query complexity is not known exactly, we only have the upper bound $O(\varepsilon^{-p})$, $p \leq 6$. The best quantum algorithm known is based on phase estimation. In particular, we discretize the problem as in the worst case and apply phase estimation to the unitary matrix

$$\mathrm{e}^{i\gamma M_\varepsilon},$$

where γ is chosen appropriately so that the resulting phase belongs to $[0, 2\pi)$. We use a splitting formula to approximate the necessary powers of the matrix exponential. The largest eigenvalue of Δ_ε is $O(\varepsilon^{-2})$ and $\|q\|_\infty \leq 1$. This implies that the resulting number of queries does not grow exponentially with d.

Finally, there are a number of papers providing quantum algorithms for eigenvalue approximation without carrying out a complete complexity analysis. Abrams and Lloyd [2] have written an influential paper on eigenvalue approximation of a quantum mechanical system evolving with a given Hamiltonian. They point out that phase estimation, which requires the corresponding eigenvector as part of its initial state, can also be used with an approximate eigenvector. Jaksch and Papageorgiou [36] give a quantum algorithm which computes a *good* approximation of the eigenvector at low cost. Their method can be generally applied to the solution of continuous Hermitian eigenproblems on a discrete grid. It starts with a classically obtained eigenvector for a problem discretized on a coarse grid and constructs an approximate eigenvector on a fine grid.

We describe this algorithm briefly for the case $d = 1$. Suppose $N = 2^k$ and $N_0 = 2^{k_0}$ are the number of points in the fine and coarse grid, respectively. Given the eigenvector for the coarse grid $|U^{(N_0)}\rangle$, we approximate the eigenvector for the fine grid $|U^{(N)}\rangle$ by

$$\left|\tilde{U}^{(N)}\right\rangle = \left|U^{(N_0)}\right\rangle \left(\frac{|0\rangle + |1\rangle}{\sqrt{2}}\right)^{\otimes(k-k_0)}.$$

The effect of this transformation is to replicate the coordinates of $|U^{(N_0)}\rangle 2^{k-k_0}$ times. The resulting approximation is good enough in the sense that the success probability of phase estimation with initial state $|\tilde{U}_h\rangle$ is greater than 1/2.

Szkopek et al. [64] use the algorithm of Jaksch and Papageorgiou in the approximation of low order eigenvalues of a differential operator of order $2s$ in d dimensions. Their paper provides an algorithm with cost polynomial in ε^{-1} and generalizes the results of Abrams and Lloyd [2]. However, [64] does not carry out a complexity analysis but only considers known classical algorithms in the worst case for comparison.

Qubit Complexity

For the foreseeable future the number of qubits will be a crucial computational resource. We give a general lower bound on the qubit complexity of continuous problems.

Recall that in (1) we defined a quantum algorithm as

$$|\psi_f\rangle = U_T Q_f U_{T-1} Q_f \ldots U_1 Q_f U_0 |\psi_0\rangle.$$

where $|\psi_0\rangle$ and $|\psi_f\rangle$ are the initial and final state vectors, respectively. They are column vectors of length 2^ν. The query Q_f, a $2^\nu \times 2^\nu$ unitary matrix, depends on the values of f at $n \leq 2^\nu$ deterministic points. That is

$$Q_f = Q_f(f(x_1), \ldots, f(x_n)).$$

It is important to note that in the standard model the evolution is completely deterministic. The only probabilistic element is in the measurement of the final state.

For the qubit complexity (8) we have the following lower bound

$$\mathrm{comp}^{\mathrm{qubit}}_{\mathrm{std}}(\varepsilon, S) = \Omega\left(\log \mathrm{comp}^{\mathrm{query}}_{\mathrm{clas}}(2\varepsilon, S)\right) . \tag{19}$$

Here S specifies the problem, see (4), and $\mathrm{comp}^{\mathrm{qubit}}_{\mathrm{std}}(\varepsilon, S)$ is the qubit complexity in the standard quantum setting. On the right hand side $\mathrm{comp}^{\mathrm{query}}_{\mathrm{clas}}(\varepsilon, S)$ is the query complexity on a classical computer in the worst case setting. See [77] for details.

We provide an intuition about (19). Assume 2^ν function evaluations are needed to solve the problem specified by S to within ε. Note that ν qubits are needed to generate the Hilbert space $\mathcal{H}_\nu$ of size 2^ν to store the evaluations.

Equation (19) can be interpreted as a certain limitation of the standard quantum setting, which considers queries using function evaluations at deterministic points. We will show why this is a limitation and show we can do better.

Consider multivariate integration which was discussed in Sect. "Integration". We seek to approximate the solution of

$$S(f) = \int_{0,1]^d} f(x)\,\mathrm{d}x .$$

Assume our promise is $f \in F_0$ where

$$F_d = \Big\{ f\colon [0,1]^d \to \mathbb{R} \mid \text{continuous and } |f(x)| \le 1, x \in [0,1]^d \Big\} .$$

For the moment let $d = 1$. We showed that with this promise we cannot guarantee an ε-approximation on a classical computer with $\varepsilon < 1/2$. That is,

$$\mathrm{comp}^{\mathrm{query}}_{\mathrm{clas}}(\varepsilon, S) = \infty .$$

By (19)

$$\mathrm{comp}^{\mathrm{qubit}}_{\mathrm{std}}(\varepsilon, S) = \infty .$$

If the qubit complexity is infinite even for $d = 1$ its certainly infinite for general d. But we saw that if classical randomization (Monte Carlo) is used

$$\mathrm{comp}^{\mathrm{query}}_{\mathrm{clas-ran}}(\varepsilon, S) = \Theta(\varepsilon^{-2}) .$$

Thus we have identified a problem which is easy to solve on a classical computer and is impossible to solve on a quantum computer using the standard formulation of a quantum algorithm (1). Our example motivates extending the notion of quantum algorithm by permitting *randomized queries*. The quantum setting with randomized queries was introduced by Woźniakowski [77]. The idea of using randomized queries is not entirely new. Shor's algorithm [60] uses a special kind of randomized query, namely,

$$Q_x|x\rangle = |jx \bmod N\rangle ,$$

with $j = 0, \ldots, N-1$ and a random x from the set $\{2, 3, \ldots, N-1\}$.

In our case, the randomization affects only the selection of sample points and the number of queries. It occurs prior to the implementation of the queries and the execution of the quantum algorithm. In this extended setting, we define a quantum algorithm as

$$|\psi_{f,\omega}\rangle = U_{T_\omega} Q_{f,\omega} U_{T_\omega - 1} Q_{f,\omega} \ldots U_1 Q_{f,\omega} U_0 |\psi_0\rangle , \tag{20}$$

where ω is a random variable and

$$Q_{f,\omega} = Q_{f,\omega}(f(x_{1,\omega}), \ldots, f(x_{n,\omega})) ,$$

and the $x_{j,\omega}$ are random points. The number of queries T_ω and the points $x_{j,\omega}$ are chosen at random initially and then remain fixed for the remainder of the computation. Note that (20) is identical to (1) except for the introduction of randomization. Randomized queries require that we modify the criterion (6) by which we measure the error of an algorithm. One possibility is to consider the expected error and another possibility is to consider the probabilistic error with respect to the random variable ω. Both cases have been considered in the literature [77] but we will avoid the details here because they are rather technical.

A test of the new setting is whether it buys us anything. We compare the qubit complexity of the standard and randomized settings for integration and path integration.

Integration

We make the same promise as above, namely, $f \in F_d$.

- Quantum setting with deterministic queries. We remind the reader that

$$\mathrm{comp}^{\mathrm{query}}_{\mathrm{std}}(\varepsilon) = \infty$$
$$\mathrm{comp}^{\mathrm{qubit}}_{\mathrm{std}}(\varepsilon) = \infty .$$

- Quantum setting with randomized queries. Then [77]

$$\mathrm{comp}^{\mathrm{query}}_{\mathrm{ran}}(\varepsilon) = \Theta\left(\varepsilon^{-1}\right)$$
$$\mathrm{comp}^{\mathrm{qubit}}_{\mathrm{ran}}(\varepsilon) = \Theta\left(\log \varepsilon^{-1}\right) .$$

Therefore, there is infinite improvement in the randomized quantum setting over the standard quantum setting.

Path Integration

- Quantum setting with deterministic queries. With an appropriate promise it was shown [77] that

$$\mathrm{comp}^{\mathrm{query}}_{\mathrm{std}}(\varepsilon) = \Theta\left(\varepsilon^{-1}\right)$$
$$\mathrm{comp}^{\mathrm{qubit}}_{\mathrm{std}}(\varepsilon) = \Theta\left(\varepsilon^{-2}\log\varepsilon^{-1}\right) .$$

  Thus, modulo the log factor, the qubit complexity of path integration is a second degree polynomial in ε^{-1}. That seems pretty good but we probably will not have enough qubits for a long time to do new science, especially with error correction.
- Quantum setting with randomized queries. Then [77]

$$\mathrm{comp}^{\mathrm{query}}_{\mathrm{ran}}(\varepsilon) = \Theta\left(\varepsilon^{-1}\right)$$
$$\mathrm{comp}^{\mathrm{qubit}}_{\mathrm{ran}}(\varepsilon) = \Theta\left(\log\varepsilon^{-1}\right) .$$

  Thus there is an exponential improvement in the randomized quantum setting over the standard quantum setting.

As the analogue of (19) we have the following lower bound on the randomized qubit complexity [77]

$$\mathrm{comp}^{\mathrm{qubit}}_{\mathrm{ran}}(\varepsilon, S) = \Omega\left(\log \mathrm{comp}^{\mathrm{query}}_{\mathrm{clas-ran}}(\varepsilon, S)\right) , \qquad (21)$$

where $\mathrm{comp}^{\mathrm{query}}_{\mathrm{clas-ran}}(\varepsilon, S)$ is the query complexity on a classical computer in the randomized setting.

Approximation

Approximating functions of d variables is a fundamental and generally hard problem. Typically, the literature considers the approximation of functions that belong to the Sobolev space $W_p^r([0,1]^d)$ in the norm of $L_q([0,1]^d)$. The condition $r/d > 1/p$ ensures that functions in $W_p^r([0,1]^d)$ are continuous, which is necessary for function values to be well defined. Thus, when $p = \infty$ the dimension d can be arbitrarily large while the smoothness r can be fixed, which cannot happen when $p < \infty$.

For $p = \infty$ the problem suffers the curse of dimensionality in the worst and the randomized classical cases [49,71]. Recently, Heinrich [30] showed that quantum computers do not offer any advantage relative to classical computers since the problem remains intractable in the quantum setting.

For different values of the parameters p, q, r, d the classical and quantum complexities are also known [30, 49,71]. In some cases quantum computers can provide a roughly quadratic speedup over classical computers, but there are also cases where the classical and quantum complexities coincide. Table 2 summarizes the order of the query complexity (up to polylog factors) of approximation in the worst case, randomized and quantum setting, and is based on a similar table in [30] describing error bounds.

Elliptic Partial Differential Equations

Elliptic partial differential equations have many important applications and have been extensively studied in the literature, see [75] and the references therein. A simple example is the Poisson equation, for which we want to find a function $u\colon \bar{\Omega} \to \mathbb{R}$, that satisfies

$$-\Delta u(x) = f(x) , \quad x \in \Omega$$
$$u(x) = 0 , \quad x \in \partial\Omega ,$$

where $\Omega \subset \mathbb{R}^d$,

More generally we consider elliptic partial differential equations of order $2m$ on a smooth bounded domain $\Omega \subset \mathbb{R}^d$ with smooth coefficients and homogeneous boundary conditions with the right hand side belonging to $C^r(\Omega)$ and the error measured in the L_∞ norm; see [32] for details.

In the worst case the complexity is proportional to $\varepsilon^{-d/r}$ [75] and the problem is intractable. The randomized complexity of this problem was only recently studied along with the quantum complexity by Heinrich [31,32]. In particular, the randomized query complexity (up to polylog factors) is proportional to

$$\varepsilon^{-\max\{d/(r+2m),\, 2d/(2r+d)\}} ,$$

and the quantum query complexity is proportional to

$$\varepsilon^{-\max\{d/(r+2m),\, d/(r+d)\}} .$$

Quantum Algorithms and Complexity for Continuous Problems, Table 2
Query complexity of approximation

	Worst case	Randomized	Quantum
$1 \le p < q \le \infty, \quad r/d \ge 2/p - 2/q$	$\varepsilon^{-dpq/(rpq-d(q-p))}$	$\varepsilon^{-dpq/(rpq-d(q-p))}$	$\varepsilon^{-d/r}$
$1 \le p < q \le \infty, \quad r/d < 2/p - 2/q$	$\varepsilon^{-dpq/(rpq-d(q-p))}$	$\varepsilon^{-dpq/(rpq-d(q-p))}$	$\varepsilon^{-dpq/(2rpq-2d(q-p))}$
$1 \le q \le p \le \infty$	$\varepsilon^{-d/r}$	$\varepsilon^{-d/r}$	$\varepsilon^{-d/r}$

Thus the quantum setting may provide a polynomial speedup over the classical randomized setting but not always. Moreover, for fixed m and r and for $d > 4m$ the problem is intractable in all three settings.

Ordinary Differential Equations

In this section we consider the solution of a system of ordinary differential equations with initial conditions

$$z'(t) = f(z(t)), \quad t \in [a,b], \quad z(a) = \eta\,,$$

where $f\colon \mathbb{R}^d \to \mathbb{R}^d$, $z\colon [a,b] \to \mathbb{R}^d$ and $\eta \in \mathbb{R}^d$ with $f(\eta) \neq 0$. For the right hand side function $f = [f_1, \ldots, f_d]$, where $f_j\colon \mathbb{R}^d \to \mathbb{R}$, we assume that the f_j belong to the Hölder class $F_d^{r,\alpha}$, $r + \alpha \geq 1$. We seek to compute a bounded function on the interval $[a,b]$ that approximates the solution z.

Kacewicz [38] studied the classical worst case complexity of this problem and found it to be proportional to $\varepsilon^{-1/(r+\alpha)}$. Recently he also studied the classical randomized and quantum complexity of the problem and derived algorithms that yield upper bounds that from the lower bounds by only an arbitrarily small positive parameter in the exponent [39]. The resulting randomized and quantum complexity bounds (up to polylog factors) are

$$O\big(\varepsilon^{-1/(r+\alpha+1/2-\gamma)}\big)$$

and

$$O\big(\varepsilon^{-1/(r+\alpha+1-\gamma)}\big)\,,$$

where $\gamma \in (0,1)$ is arbitrarily small, respectively. Observe that the randomized and quantum complexities (up to polylog factors) satisfy

$$\Omega\big(\varepsilon^{-1/(r+\alpha+1/2)}\big)$$

and

$$\Omega\big(\varepsilon^{-1/(r+\alpha+1)}\big)\,,$$

respectively. Even more recently, Heinrich and Milla [33] showed that the upper bound for the randomized complexity holds with $\gamma = 0$, thereby establishing tight upper and lower randomized complexity bounds, up to polylog factors.

Once more, the quantum setting provides a polynomial speedup over the classical setting.

Gradient Estimation

Approximating the gradient of a function $f\colon \mathbb{R}^d \to \mathbb{R}$ with accuracy ε requires a minimum of $d+1$ function evaluations on a classical computer. Jordan [37] shows how this can be done using a single query on a quantum computer.

We present Jordan's algorithm for the special case where the function is a plane passing through the origin, i. e., $f(x_1, \ldots, x_d) = \sum_{j=1}^d a_j x_j$, and is uniformly bounded by 1. Then $\nabla f = (a_1, \ldots, a_d)^T$. Using a single query and *phase kickback* we obtain the state

$$\frac{1}{\sqrt{N^d}} \sum_{j_1=0}^{N-1} \cdots \sum_{j_d=0}^{N-1} \mathrm{e}^{2\pi i f(j_1,\ldots,j_d)} |j_1\rangle \cdots |j_d\rangle\,,$$

where N is a power of 2. Equivalently, we have

$$\frac{1}{\sqrt{N^d}} \sum_{j_1=0}^{N-1} \cdots \sum_{j_d=0}^{N-1} \mathrm{e}^{2\pi i (a_1 j_1 + \cdots + a_d j_d)} |j_1\rangle \cdots |j_d\rangle\,.$$

This is equal to the state

$$\frac{1}{\sqrt{N}} \sum_{j_1=0}^{N-1} \mathrm{e}^{2\pi i a_1 j_1} |j_1\rangle \ldots \frac{1}{\sqrt{N}} \sum_{j_d=0}^{N-1} \mathrm{e}^{2\pi i a_d j_d} |j_d\rangle\,.$$

We apply the Fourier transform to each of the d registers and then measure each register in the computational basis to obtain $m_1, \ldots, m_d$. If a_j can be represented with finitely many bits and N is sufficiently large then $m_j/N = a_j$, $j = 1, \ldots, d$.

For functions with second order partial derivatives not identically equal to zero the analysis is more complicated and we refer the reader to [37] for the details.

Simulation of Quantum Systems on Quantum Computers

So far this article has been devoted to work on algorithms and complexity of problems where the query and qubit complexities are known or have been studied. In a number of cases, the classical complexity of these problems is also known and we know the quantum computing speedup.

The notion that quantum systems could be simulated more efficiently by quantum computers than by classical computers was first mentioned by Manin [44], see also [45], and discussed thoroughly by Feynman [24].

There is a large and varied literature on simulation of quantum systems on quantum computers. The focus in these papers is typically on the cost of particular quantum and classical algorithms without complexity analysis and therefore without speedup results. To give the reader a taste of this area we list some sample papers:

- Berry et al. [9] present an efficient quantum algorithm for simulating the evolution of a sparse Hamiltonian.

- Dawson, Eisert and Osborne [21] introduce a unified formulation of variational methods for simulating ground state properties of quantum many-body systems.
- Morita and Nishimori [46] derive convergence conditions for the quantum annealing algorithm.
- Brown, Clark, Chuang [13] establish limits of quantum simulation when applied to specific problems.
- Chen, Yepez and Cory [15] report on the simulation of Burgers equation as a type-II quantum computation.
- Paredes, Verstraete and Cirac [56] present an algorithm that exploits quantum parallelism to simulate randomness.
- Somma et al. [61] discuss what type of physical problems can be efficiently simulated on a quantum computer which cannot be simulated on a Turing machine.
- Yepez [78] presents an efficient algorithm for the many-body three-dimensional Dirac equation.
- Nielsen and Chuang [48] discuss simulation of a variety of quantum systems.
- Sornborger and Stewart [62] develop higher order methods for simulations.
- Boghosian and Taylor [11] present algorithms for efficiently simulating quantum mechanical systems.
- Zalka [79] shows that the time evolution of the wave function of a quantum mechanical many particle system can be simulated efficiently.
- Abrams and Lloyd [1] provide fast algorithms for simulating many-body Fermi systems.
- Wisner [76] provides two quantum many-body problems whose solution is intractable on a classical computer.

Future Directions

The reason there is so much interest in quantum computers is to solve important problems fast. The belief is that we will be able to solve scientific problems, and in particular quantum mechanical systems, which cannot be solved on a classical computer. That is, that quantum computation will lead to new science.

Research consists of two major parts. The first is identification of important scientific problems with substantial speedup. The second is the construction of machines with sufficient number of qubits and long enough decoherence times to solve the problems identified in the first part. Abrams and Lloyd [2] have argued that with 50 to 100 qubits we can solve interesting classically intractable problems from atomic physics. Of course this does not include qubits needed for fault tolerant computation.

There are numerous important open questions. We will limit ourselves here to some open questions regarding the problems discussed in this article.

1. In Sect. "Path Integration" we reported big wins for the qubit complexity for integration and path integration. Are there big wins for other problems?
2. Are there problems for which we get big wins for query complexity using randomized queries?
3. Are there tradeoffs between the query complexity and the qubit complexity?
4. What are the classical and quantum complexities of approximating the solution of the Schrödinger equation for a many-particle system? How do they it depend on the number of particles? What is the quantum speedup?

Acknowledgments

We are grateful to Erich Novak, University of Jena, and Henryk Woźniakowski, Columbia University and University of Warsaw, for their very helpful comments. We thank Jason Petras, Columbia University, for checking the complexity estimates appearing in the tables.

Bibliography

1. Abrams DS, Lloyd S (1997) Simulation of Many-Body Fermi Systems on a Universal Quantum Computer. Phys Rev Lett 79(13):2586–2589; http://arXiv.org/quant-ph/9703054
2. Abrams DS, Lloyd S (1999) Quantum Algorithm Providing Exponential Speed Increase for Finding Eigenvalues and Eigenvectors. Phys Rev Lett 83:5162–5165
3. Abrams DS, Williams CP (1999) Fast quantum algorithms for numerical integrals and stochastic processes. http://arXiv.org/quant-ph/9908083
4. Bakhvalov NS (1977) Numerical Methods. Mir Publishers, Moscow
5. Babuska I, Osborn J (1991) Eigenvalue Problems. In: Ciarlet PG, Lions JL (eds) Handbook of Numerical Analysis, vol II. North-Holland, Amsterdam, pp 641–787
6. Beals R, Buhrman H, Cleve R, Mosca R, de Wolf R (1998) Quantum lower bounds by polynomials. Proceedings FOCS'98, pp 352–361. http://arXiv.org/quant-ph/9802049
7. Bennett CH, Bernstein E, Brassard G, Vazirani U (1997) Strengths and weaknesses of quantum computing. SIAM J Comput 26(5):1510–1523
8. Bernstein E, Vazirani U (1997) Quantum complexity theory. SIAM J Comput 26(5):1411–1473
9. Berry DW, Ahokas G, Cleve R, Sanders BC (2007) Efficient quantum algorithms for simulating sparse Hamiltonians. Commun Math Phys 270(2):359–371; http://arXiv.org/quant-ph/0508139
10. Bessen AJ (2007) On the complexity of classical and quantum algorithms for numerical problems in quantum mechanics. Ph D thesis. Department of Computer Science, Columbia University

11. Boghosian BM, Taylor W (1998) Simulating quantum mechanics on a quantum computer. Physica D 120:30–42 http://arXiv.org/quant-ph/9701019
12. Brassard G, Hoyer P, Mosca M, Tapp A (2002) Quantum Amplitude Amplification and Estimation. Contemporary Mathematics, vol 305. Am Math Soc, Providence, pp 53–74. http://arXiv.org/quant-ph/0005055
13. Brown KR, Clark RJ, Chuang IL (2006) Limitations of Quantum Simulation Examined by Simulating a Pairing Hamiltonian using Magnetic Resonance. Phys Rev Lett 97(5):050504; http://arXiv.org/quant-ph/0601021
14. Cameron RH (1951) A Simpson's rule for the numerical evaluation of Wiener's integrals in function space. Duke Math J 8:111–130
15. Chen Z, Yepez J, Cory DG (2006) Simulation of the Burgers equation by NMR quantum information processing. Phys Rev A 74:042321; http://arXiv.org/quant-ph/0410198
16. Chorin AJ (1973) Accurate evaluation of Wiener integrals. Math Comp 27:1–15
17. Cleve R, Ekert A, Macchiavello C, Mosca M (1996) Quantum Algorithms Revisited. Proc R Soc Lond A 454(1969):339–354
18. Collatz L (1960) The Numerical Treatment of Differential Equations. Springer, Berlin
19. Courant C, Hilbert D (1989) Methods of Mathematical Physics, vol I. Wiley Classics Library. Wiley-Interscience, New York
20. Curbera F (2000) Delayed curse of dimension for Gaussian integration. J Complex 16(2):474–506
21. Dawson CM, Eisert J, Osborne TJ (2007) Unifying variational methods for simulating quantum many-body systems. http://arxiv.org/abs/0705.3456v1
22. Demmel JW (1997) Applied Numerical Linear Algebra. SIAM, Philadelphia
23. Egorov AD, Sobolevsky PI, Yanovich LA (1993) Functional Integrals: Approximate Evaluation and Applications. Kluwer, Dordrecht
24. Feynman RP (1982) Simulating physics with computers. Int J Theor Phys 21:476
25. Forsythe GE, Wasow WR (2004) Finite-Difference Methods for Partial Differential Equations. Dover, New York
26. Grover L (1997) Quantum mechanics helps in searching for a needle in a haystack. Phys Rev Lett 79(2):325–328; http://arXiv.org/quant-ph/9706033
27. Heinrich S (2002) Quantum Summation with an Application to Integration. J Complex 18(1):1–50; http://arXiv.org/quant-ph/0105116
28. Heinrich S (2003) From Monte Carlo to Quantum Computation. In: Entacher K, Schmid WC, Uhl A (eds) Proceedings of the 3rd IMACS Seminar on Monte Carlo Methods MCM2001, Salzburg. Special Issue of Math Comput Simul 62:219–230
29. Heinrich S (2003) Quantum integration in Sobolev spaces. J Complex 19:19–42
30. Heinrich S (2004) Quantum Approximation II. Sobolev Embeddings. J Complex 20:27–45; http://arXiv.org/quant-ph/0305031
31. Heinrich S (2006) The randomized complexity of elliptic PDE. J Complex 22(2):220–249
32. Heinrich S (2006) The quantum query complexity of elliptic PDE. J Complex 22(5):691–725
33. Heinrich S, Milla B (2007) The randomized complexity of initial value problems. Talk presented at First Joint International Meeting between the American Mathematical Society and the Polish Mathematical Society, Warsaw, Poland
34. Heinrich S, Novak E (2002) Optimal summation by deterministic, randomized and quantum algorithms. In: Fang KT, Hickernell FJ, Niederreiter H (eds) Monte Carlo and Quasi-Monte Carlo Methods 2000. Springer, Berlin
35. Heinrich S, Kwas M, Woźniakowski H (2004) Quantum Boolean Summation with Repetitions in the Worst-Average Setting. In: Niederreiter H (ed) Monte Carlo and Quasi-Monte Carlo Methods, 2002. Springer, New York, pp 27–49
36. Jaksch P, Papageorgiou A (2003) Eigenvector approximation leading to exponential speedup of quantum eigenvalue calculation. Phys Rev Lett 91:257902; http://arXiv.org/quant-ph/0308016
37. Jordan SP (2005) Fast Quantum Algorithm for Numerical Gradient Estimation. Phys Rev Lett 95:050501; http://arXiv.org/quant-ph/0405146
38. Kacewicz BZ (1984) How to increase the order to get minimal-error algorithms for systems of ODEs. Numer Math 45:93–104
39. Kacewicz BZ (2006) Almost optimal solution of initial-value problems by randomized and quantum algorithms. J Complex 22(5):676–690
40. Keller HB (1968) Numerical methods for two-point boundary-value problems. Blaisdell Pub Co, Waltham
41. Knuth DE (1997) The Art of Computer Programming, vol 2: Seminumerical Algorithms, 3rd edn. Addison-Wesley Professional, Cambridge
42. Kwas M (2005) Quantum algorithms and complexity for certain continuous and related discrete problems. Ph D thesis. Department of Computer Science, Columbia University
43. Kwas M, Li Y (2003) Worst case complexity of multivariate Feynman–Kac path integration. J Complex 19:730–743
44. Manin Y (1980) Computable and Uncomputable. Sovetskoye Radio, Moscow (in Russian)
45. Manin YI (1999) Classical computing, quantum computing, and Shor's factoring algorithm. http://arXiv.org/quant-ph/9903008
46. Morita S, Nishimori H (2007) Convergence of Quantum Annealing with Real-Time Schrödinger Dynamics. J Phys Soc Jpn 76(6):064002; http://arXiv.org/quant-ph/0702252
47. Nayak A, Wu F (1999) The quantum query complexity of approximating the median and related statistics. In: Proc STOC 1999, Association for Computing Machinery, New York, pp 384–393. http://arXiv.org/quant-ph/9804066
48. Nielsen MA, Chuang IL (2000) Quantum Computation and Quantum Information. Cambridge University Press, Cambridge
49. Novak E (1988) Deterministic and Stochastic Error Bounds in Numerical Analysis. Lecture Notes in Mathematics, 1349. Springer, Berlin
50. Novak E (2001) Quantum complexity of integration. J Complex 17:2–16; http://arXiv.org/quant-ph/0008124
51. Ortiz G, Gubernatis JE, Knill E, Laflamme R (2001) Quantum algorithms for fermionic simulations. Phys Rev A 64(2):022319; http://arXiv.org/cond-mat/0012334
52. Papageorgiou A (2004) Average case quantum lower bounds for computing the boolean mean. J Complex 20(5):713–731
53. Papageorgiou A (2007) On the complexity of the multivariate Sturm–Liouville eigenvalue problem. J Complex 23(4–6):802–827

54. Papageorgiou A, Traub JF (2005) Qubit complexity of continuous problems. http://arXiv.org/quant-ph/0512082
55. Papageorgiou A, Woźniakowski H (2005) Classical and Quantum Complexity of the Sturm–Liouville Eigenvalue Problem. Quantum Inf Process 4(2):87–127; http://arXiv.org/quant-ph/0502054
56. Paredes B, Verstraete F, Cirac JI (2005) Exploiting Quantum Parallelism to Simulate Quantum Random Many-Body Systems. Phys Rev Lett 95:140501; http://arXiv.org/cond-mat/0505288
57. Plaskota L (1996) Noisy Information and Computational Complexity. Cambridge University Press, Cambridge
58. Plaskota L, Wasilkowski GW, Woźniakowski H (2000) A new algorithm and worst case complexity for Feynman–Kac path integration. J Comp Phys 164(2):335–353
59. Ritter K (2000) Average-Case Analysis of Numerical Problems. Lecture Notes in Mathematics, 1733. Springer, Berlin
60. Shor PW (1997) Polynomial-time algorithms for prime factorization and discrete logarithm on a quantum computer. SIAM J Comput 26(5):1484–1509
61. Somma R, Ortiz G, Knill E, Gubernatis (2003) Quantum Simulations of Physics Problems. In: Pirich AR, Brant HE (eds) Quantum Information and Computation. Proc SPIE 2003, vol 5105. The International Society for Optical Engineering, Bellingham, pp 96–103. http://arXiv.org/quant-ph/0304063
62. Sornborger AT, Stewart ED (1999) Higher Order Methods for Simulations on Quantum Computers. Phys Rev A 60(3):1956–1965; http://arXiv.org/quant-ph/9903055
63. Strang G, Fix GJ (1973) An Analysis of the Finite Element Method. Prentice-Hall, Englewood Cliffs
64. Szkopek T, Roychowdhury V, Yablonovitch E, Abrams DS (2005) Egenvalue estimation of differential operators with a quantum algorithm. Phys Rev A 72:062318
65. Titschmarsh EC (1958) Eigenfunction Expansions Associated with Second-Order Differential Equations, Part B. Oxford University Press, Oxford
66. Traub JF (1999) A continuous model of computation. Phys Today May:39–43
67. Traub JF, Werschulz AG (1998) Complexity and Information. Cambridge University Press, Cambridge
68. Traub JF, Woźniakowski H (1980) A general theory of optimal algorithms. ACM Monograph Series. Academic Press, New York
69. Traub JF, Woźniakowski H (1992) The Monte Carlo algorithm with a pseudorandom generator. Math Comp 58(197):323–339
70. Traub JF, Woźniakowski H (2002) Path integration on a quantum computer. Quantum Inf Process 1(5):365–388; http://arXiv.org/quant-ph/0109113
71. Traub JF, Wasilkowski GW, Woźniakowski H (1988) Information-Based Complexity. Academic Press, New York
72. Wasilkowski GW, Woźniakowski H (1996) On tractability of path integration. J Math Phys 37(4):2071–2088
73. Weinberger HF (1956) Upper and Lower Bounds for Eigenvalues by Finite Difference Methods. Commun Pure Appl Math IX:613–623
74. Weinberger HF (1958) Lower Bounds for Higher Eigenvalues by Finite Difference Methods. Pacific J Math 8(2):339–368
75. Werschulz AG (1991) The Computational Complexity of Differential and Integral Equations. Oxford University Press, Oxford
76. Wisner S (1996) Simulations of Many-Body Quantum Systems by a Quantum Computer. http://arXiv.org/quant-ph/96
77. Woźniakowski H (2006) The Quantum Setting with Randomized Queries for Continuous Problems. Quantum Inf Process 5(2):83–130
78. Yepez J (2002) An efficient and accurate quantum algorithm for the Dirac equation. http://arXiv.org/quant-ph/0210093
79. Zalka C (1998) Simulating quantum systems on a quantum computer. Proc Royal Soc Lond A 454(1969):313–322; http://arXiv.org/quant-ph/9603026

Quantum Bifurcations

Boris Zhilinskií
Université du Littoral, Dunkerque, France

Article Outline

Glossary

Classical limit The classical limit is the classical mechanical problem which can be constructed from a given quantum problem by some limiting procedure. During such a construction the classical limiting manifold should be defined which plays the role of classical phase space. As soon as quantum mechanics is more general than classical mechanics, going to the classical limit from a quantum problem is much more reasonable than discussing possible quantizations of classical theories [73].

Energy-momentum map In classical mechanics for any problem which allows the existence of several integrals of motion (typically energy and other integrals often named as momenta) the Energy-Momentum (EM) map gives the correspondence between the phase space

of the initial problem and the space of values of all independent integrals of motion. The energy-momentum map introduces the natural foliation of the classical phase space into common levels of values of energy and momenta [13,35]. The image of the EM map is the region of the space of possible values of integrals of motion which includes regular and critical values. The quantum analog of the image of the energy-momentum map is the joint spectrum of mutually commuting quantum observables.

Joint spectrum For each quantum problem a maximal set of mutually commuting observables can be introduced [16]. A set of quantum wave functions which are mutual eigenfunctions of all these operators exists. Each such eigenfunction is characterized by eigenvalues of all mutually commuting operators. The representation of mutual eigenvalues of n commuting operators in the n-dimensional space gives the geometrical visualization of the joint spectrum.

Monodromy In general, the monodromy characterizes the evolution of some object after it makes a close path around something. In classical Hamiltonian dynamics the Hamiltonian monodromy describes for completely integrable systems the evolution of the first homology group of the regular fiber of the energy-momentum map after a close path in the regular part of the base space [13].

For a corresponding quantum problem the quantum monodromy describes the modification of the local structure of the joint spectrum after its propagation along a close path going through a regular region of the lattice.

Quantum bifurcation Qualitative modification of the joint spectrum of the mutually commuting observables under the variation of some external (or internal) parameters and associated in the classical limit with the classical bifurcation is named quantum bifurcation [59]. In other words the quantum bifurcation is the manifestation of the classical bifurcation presented in the classical dynamic system in the quantum version of the same system.

Quantum-classical correspondence Starting from any quantum problem the natural question consists of defining the corresponding classical limit, i. e. the classical dynamic variables forming the classical phase space and the associated symplectic structure. Whereas in simplest quantum problems defined in terms of standard position and momentum operators with commutation relation $[q_i, p_j] = i\hbar\delta_{ij}$, $[q_i, q_j] = [p_i, p_j] = 0$ $(i, j = 1 \ldots n)$ the classical limit phase space is the $2n$-dimensional Euclidean space with standard symplectic structure on it, the topology of the classical limit manifold in many other important for physical applications cases can be rather non-trivial [73,87].

Quantum phase transition Qualitative modifications of the ground state of a quantum system occurring under the variation of some external parameters at zero temperature are named quantum phase transitions [65]. For finite particle systems the quantum phase transition can be considered as a quantum bifurcation [60].

Spontaneous symmetry breaking Qualitative modification of the system of quantum states caused by perturbation which has the same symmetry as the initial problem. Local symmetry of solutions decreases but the number of solutions increases. In the energy spectra of finite particle systems the spontaneous symmetry breaking produces an increase of the "quasidegeneracy", i. e. formation of clusters of quasi-degenerate levels whose multiplicity can be much higher than the dimension of the irreducible representations of the global symmetry group [51].

Symmetry breaking Qualitative changes in the properties (dynamical behavior, and in particular in the joint spectrum) of quantum systems which are due to modifications of the global symmetry of the problem caused by external (less symmetrical than original problem) perturbation can be described as symmetry breaking effects. Typical effects consist of splitting of degenerate energy levels classified initially according to irreducible representation of the initial symmetry group into less degenerate groups classified according to irreducible representation of the subgroup (the symmetry group of the perturbation) [47].

Definition of the Subject

Quantum bifurcations (QB) are qualitative phenomena occurring in quantum systems under the variation of some internal or external parameters. In order to make this definition a little more precise we add the additional requirement: The qualitative modification of the "behavior" of a quantum system can be described as QB if it consists of the manifestation for the quantum system of the classical bifurcation presented in classical dynamic systems which is the classical analog of the initial quantum system. Quantum bifurcations are typical elementary steps leading from the simplest in some way effective Hamiltonian to more complicated ones under the variation of external or internal parameters. As internal parameters one may consider the values of exact or approximate integrals of motion. The construction of an effective Hamiltonian

is typically based on the averaging and/or reduction procedure which results in the appearance of "good" quantum numbers (or approximate integrals of motion). The role of external parameters can be played by forces of external champs, phenomenological constants in the effective Hamiltonians, particle masses, etc. In order to limit the very broad field of qualitative changes and of possible quantum bifurcations in particular, we restrict ourselves mainly to quantum systems whose classical limit is associated with compact phase space and is nearly integrable. This means that for quantum problems the set of mutually commuting observables can be constructed within a reasonable physical approximation almost everywhere at least locally.

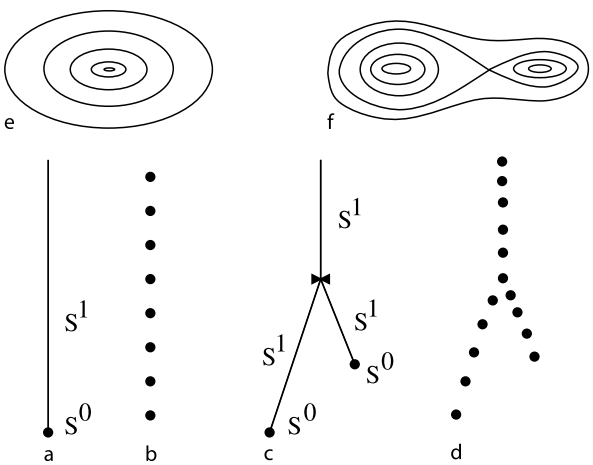
Quantum bifurcations are supposed to be universal phenomena which appear in generic families of quantum systems and explain how relatively simple behavior becomes complicated under the variation of some physical parameters. To know these elementary bricks responsible for increasing complexity of quantum systems under control parameter modifications is extremely important in order to make the extrapolation to regimes unaccessible to experimental study.

Introduction

In order to better understand the manifestations of quantum bifurcations and their significance for concrete physical systems we start with the description of several simple model physical problems which exhibit in some sense the simplest (but nevertheless) generic behavior. Let us start with the harmonic oscillator. A one-dimensional harmonic oscillator has an equidistant system of eigenvalues. All eigenvalues can be labeled by consecutive integer quantum numbers which have the natural interpretation in terms of the number of zeros of eigenfunctions. The classical limit manifold (classical phase space) is a standard Euclidean 2-dimensional space with natural variables $\{p, q\}$. The classical Hamiltonian for the harmonic oscillator is an example of a Morse-type function which has only one stationary point $p = q = 0$ and all non-zero energy levels of the Hamiltonian are topological circles. If now we deform slightly the Hamiltonian in such a way that its classical phase portrait remains topologically the same, the spectrum of the quantum problem changes but it can be globally described as a regular sequence of states numbered consecutively by one integer and such description remains valid for any mass parameter value. Note, that for this problem increasing mass means increasing quantum state density and approaching classical behavior (classical limit).

More serious modification of the harmonic oscillator can lead, for example, to creation of new stationary points of the Hamiltonian. In classical theory this phenomenon is known as fold bifurcation or fold catastrophe [3,31]. The phase portrait of the classical problem changes qualitatively. As a function of energy the constant level set of the Hamiltonian has different topological structure (one circle, two circles, figure eight, circle and a point, or simply point). The quantum version of the same problem shows the existence of three different sequences of states which become clearly visible in the limit of the high density of states which can be reached by increasing the mass value parameter [32]. Such qualitative modification of the energy spectrum of the 1D-quantum Hamiltonian gives the simplest example of the phenomenon which can be described as a quantum bifurcation. Figure 1 shows a schematic representation of quantum bifurcations for a model system with one degree-of-freedom in parallel in quantum and classical mechanics.

After looking for one simple example we can formulate a more general question which concerns the appearance in more general quantum systems of qualitative phenomena which can be characterized as quantum bifurcations.



Quantum Bifurcations, Figure 1

Classical and quantum bifurcations for a one degree-of-freedom system. Situations before (a,b,e) and after (c,d,f) the bifurcation are shown. a Energy map for harmonic oscillator-type system. Inverse images of each point are indicated. b Quantum state lattice for harmonic oscillator-type system. c Energy map after the bifurcation. Inverse images of each point are indicated. d Quantum state lattice after bifurcation represented as composed of three regular parts glued together. e Phase portrait for harmonic oscillator-type system. Inverse images are S^1 (generic inverse image) and S^0 (inverse image for minimal energy value). f Phase portrait after bifurcation

Simplest Effective Hamiltonians

We turn now to several models which describe some specific classes of relatively simple real physical quantum systems formed by a finite number of particles (atoms, molecules, ...). Spectra of such quantum objects are studied nowadays with very high accuracy and this allows us to compare the behavior predicted by quantum bifurcations with the precise information about energy level structure found, for example, from high-resolution molecular spectroscopy.

Typically, the intra-molecular dynamics can be split into electronic, vibrational, and rotational ones due to important differences in characteristic energy excitations or in time scales. The most classical is the rotational motion and probably due to that the quantum bifurcations as a counterpart to classical bifurcations were first studied for purely rotational problems [59,61].

Effective rotational Hamiltonians describe the internal structure of rotational multiplets formed by isolated finite particle systems (atoms, molecules, nuclei) [36]. For many molecular systems in the ground electronic state any electronic and vibrational excitations are much more energy consuming as compared with rotational excitations. Thus, to study the molecular rotation the simplest physical assumption is to suppose that all electronic and all vibrational degrees-of-freedom are frozen. This means that a set of quantum numbers is given which have the sense of approximate integrals of motion specifying the character of vibrational and electronic motions in terms of these "good" quantum numbers. At the same time for a free molecule in the absence of any external fields due to isotropy of the space the total angular momentum J and its projection J_z on the laboratory fixed frame are strict integrals of motion. Consequently, to describe the rotational motion for fixed values of J and J_z it is sufficient to analyze the effective problem with only one degree-of-freedom. The dimension of classical phase space in this case equals two and the two classical conjugate variables are: the projection of the total angular momentum on the body fixed frame and conjugate angle variable. The classical phase space is topologically a two-dimensional sphere, S^2. There is a one-to-one correspondence between the points on a sphere and the orientation of the angular momentum in the body-fixed frame. Such a representation gives a clear visualization of a classical rotational Hamiltonian as a function defined over a sphere [36,49].

In quantum mechanics the rotation of molecules is traditionally described in terms of an effective rotational Hamiltonian which is constructed as a series in rotational operators J_x, J_y, J_z, the components of the total angular momentum $\mathbf{J}$. In a suitably chosen molecular fixed frame the effective Hamiltonian has the form

$$H_{\text{eff}} = AJ_x^2 + BJ_y^2 + CJ_z^2 + \sum c_{\alpha\beta\gamma} J_x^\alpha J_y^\beta J_z^\gamma + \cdots , \quad (1)$$

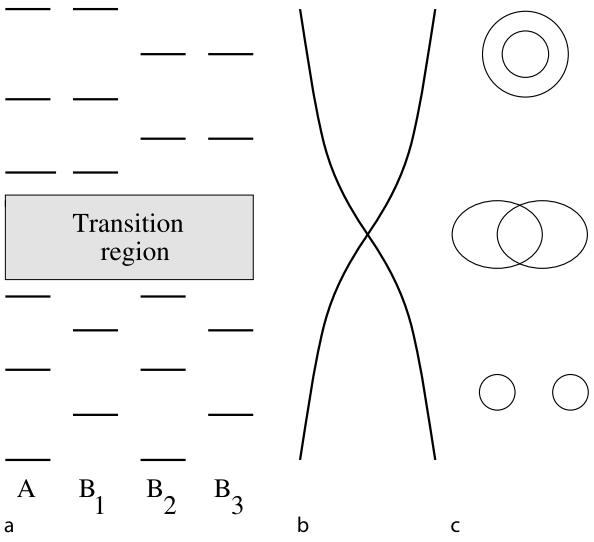
where A, B, C and $c_{\alpha\beta\gamma}$ are constants. To relate quantum and classical pictures we note that $\mathbf{J}^2$ and energy are integrals of Euler's equations of motion for dynamic variables J_x, J_y, J_z. The phase space of the classical rotational problem with constant $|\mathbf{J}|$ is S^2, the two-dimensional sphere, and it can be parametrized with spherical angles (θ, ϕ) in such a way that the points on S^2 define the orientation of $\mathbf{J}$, i. e. the position of the axis and the direction of rotation. To get the classical interpretation of the quantum Hamiltonian we introduce the classical analogs of the operators J_x, J_y, J_z

$$\mathbf{J} \longrightarrow \begin{pmatrix} J_x \\ J_y \\ J_z \end{pmatrix} = \begin{pmatrix} \sin\theta\cos\phi \\ \sin\theta\sin\phi \\ \cos\theta \end{pmatrix} \sqrt{J(J+1)} \quad (2)$$

and consider the rotational energy as a function of the dynamical variables (θ, ϕ) and the parameter J.

Thus, for an effective rotational Hamiltonian the corresponding classical symbol is a function $E_J(\theta, \phi)$ defined over S^2 and named usually the rotational energy surface [36].

Taking into account the symmetry imposed by the initial problem and the topology of the phase space the simplest rotational Hamiltonian can be constructed. In classical mechanics the simplest Hamiltonian can be defined (using Morse theory [55,89]) as a Hamiltonian function with the minimal possible number of non-degenerate stationary points compatible with the symmetry group action of the classical phase space. Morse theory in the presence of symmetry (or equivariant Morse theory) implies important restrictions on the number of minima, maxima, and saddle points. In the absence of symmetry the simplest Morse type function on the S^2 phase space has one minimum and one maximum, as a consequence of Morse inequalities. In the presence of non-trivial symmetry group action the minimal number of stationary points on the sphere increases. For example, many asymmetric top molecules (possessing three different moment of inertia of the equilibrium configuration) have D_{2h} symmetry group [47]. This group includes rotations over π around $\{x, y, z\}$ axes, reflections in $\{xy, yz, zx\}$ planes and inversion as symmetry operations. Any D_{2h} invariant function on the sphere has at least six stationary points (two equivalent minima, two equivalent maxima, and two equivalent saddle points). This means that in quantum



Quantum Bifurcations, Figure 2
a Schematic representation of the energy level structure for asymmetric top molecule. *Vertical axis* corresponds to energy variation. Quantum levels are classified by the symmetry group of the asymmetric top. Two fold clusters at two ends of the rotational multiplet are formed by states with different symmetry. b Foliation of the classical phase space (S^2 sphere) by constant levels of the Hamiltonian given in the form of its Reeb graph. Each *point* corresponds to a connected component of the constant level set of the Hamiltonian (energy). c Geometric representation of the constant energy sections

mechanics the asymmetric top has eigenvalues which form two regular sequences of quasi-degenerate doublets with the transition region between them. The correspondence between the quantum spectrum and the structure of the energy map for the classical problem is shown in Fig. 2. Highly symmetrical molecules which have cubic symmetry, for example, can be described by a simplest Morse-type Hamiltonian with 26 stationary points (6 and 8 minima/maxima and 12 saddle points). As a consequence, the corresponding quantum Hamiltonian shows the presence of six-fold and eight-fold quasi-degenerate clusters of rotational levels.

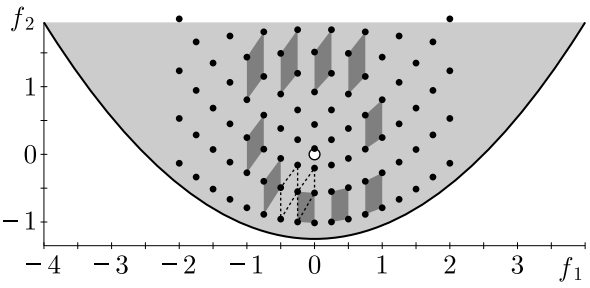
As soon as the simplest classical Hamiltonian is characterized by the appropriate system of stationary points the whole region of possible classical energy values (in the case of dynamical systems with only one degree-of-freedom the energy-momentum map becomes simply the energy map) appears to be split into different regions corresponding to different dynamical regimes, i. e. to different regions of the phase portrait foliated by topologically non-equivalent systems of classical trajectories. Accordingly, the energy spectrum of the corresponding quantum Hamiltonian can be qualitatively described as formed by regular sequences of states within each region of the classical energy map.

Quantum bifurcations are universal phenomena which lead to a new organization of the energy spectrum into qualitatively different regions in accordance with corresponding qualitative modifications of the classical energy-momentum map under the variation of some control parameter.

Simplest Hamiltonians for Two Degree-of-Freedom Systems

When the quantum system has two or larger number of degrees-of-freedom the simplest dynamical regimes often correspond in classical mechanics to a quasi-regular dynamics which can be reasonably well approximated by an integrable model. The integrable model in classical mechanics can be constructed by normalizing the Hamiltonian and by passing to so-called normal forms [2,49]. The quantum counterpart of normalization is the construction of a mutually commuting set of operators which should not be mistaken with quantization of systems in normal form. Corresponding eigenvalues can be used as "good" quantum numbers to label quantum states. A joint spectrum of mutually commuting operators corresponds to the image of the energy-momentum map for the classical completely integrable dynamical problem. In this context the question about quantum bifurcations first of all leads to the question about qualitative classification of the joint spectra of mutually commuting operators. To answer this question we need to start with the qualitative description of foliations of the total phase space of the classical problem by common levels of integrals of motion which are mutually in involution [2,7]. One needs to distinguish the regular and the singular values of the energy-momentum map. For Hamiltonian systems the inverse images of the regular values are regular tori (one or several) [2]. A lot of different singularities are possible. In classical mechanics different levels of the classifications are studied in detail [7]. The diagram which represents the image of the classical EM map together with its stratification into regular and critical values is named the bifurcation diagram. The origin of such a name is due to the fact that the values of integrals of motion can be considered as control parameters for the phase portraits (inverse images of the EM map) of the reduced systems.

For quantum problems the analog of the classical stratification of the EM map for integrable systems is the splitting of the joint spectrum of several commuting observables into regions formed by regular lattices of joint eigenvalues. Any local simply connected neighborhood of a reg-



Quantum Bifurcations, Figure 3
Joint spectrum of two commuting operators together with the image of the classical EM map for the resonant 1 : (−1) oscillator given by (3),(4). Quantum monodromy is seen as a result of transportation of the elementary cell of the quantum lattice along a close path through a non simply connected region of the regular part of the image of the EM map. Taken from [58]

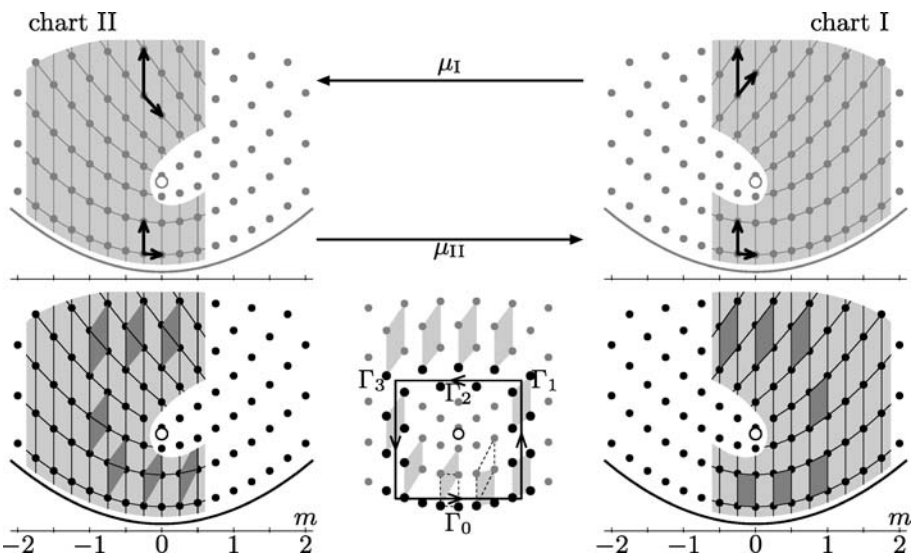
ular point of the lattice can be deformed into part of the regular Z^n lattice of integers. This means that local quantum numbers can be consistently introduced to label states of the joint spectrum. If the regular region is not simply connected it still can be characterized locally by a set of "good" quantum numbers. At the same time this is impossible globally. Likewise in classical mechanics the Hamiltonian monodromy is the simplest obstruction to the existence of the global action-angle variables [17,57], in quantum mechanics the analog notion of quantum monodromy [14,34,68,80] characterizes the global non-triviality of the regular part of the lattice of joint eigenvalues. Figure 3 demonstrates the effect of the presence of a classical singularity (isolated focus-focus point) on the global properties of the quantum lattice formed by joint eigenvalues of two commuting operators for a simple problem with two degrees-of-freedom, which is essentially the 1 : (−1) resonant oscillator [58]. Two integrals of motions in this example are chosen as

$$f_1 = \frac{1}{2}\left(p_1^2 + q_1^2\right) - \frac{1}{2}\left(p_2^2 + q_2^2\right), \tag{3}$$

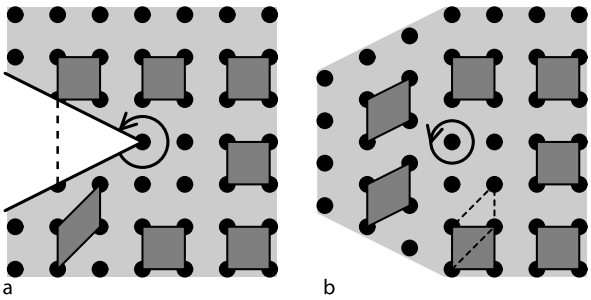
$$f_2 = p_1 q_2 + p_2 q_1 + \frac{1}{4}\left(p_1^2 + q_1^2 + p_2^2 + q_2^2\right)^2. \tag{4}$$

Locally in any simply connected region which does not include the classical singularity of the EM map situated at $f_1 = f_2 = 0$, the joint spectrum can be smoothly deformed to the regular Z^2 lattice [58,90]. Such lattices are shown, for example, in Fig. 4. If somebody wants to use only one chart to label states, it is necessary to take care in respect of the multivaluedness of such a representation. There are two possibilities:

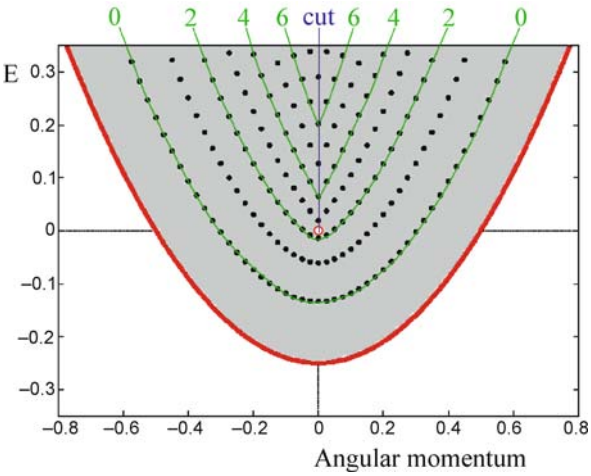
(i) One makes a cut and maps the quantum lattice to a regular Z^2 lattice with an appropriate solid angle removed from it (see Fig. 5 [58,68,90]). Points on the boundary of such a cut should be identified and a spe-



Quantum Bifurcations, Figure 4
Two chart atlas which covers the quantum lattice of the 1 : (−1) resonant oscillator system represented in Fig. 3. *Top plots* show the choice of basis cells and the gluing map between the charts. *Bottom plots* show the transport of the elementary cell (*dark gray quadrangles*) in each chart. *Central bottom panel* shows closed path Γ and its quantum realization (*black dots*) leading to non-trivial monodromy (compare with Fig. 3). Taken from [58]



Quantum Bifurcations, Figure 5
Construction of the 1 : (−1) lattice defect starting from the regular Z^2 lattice. The solid angle is removed from the regular Z^2 lattice and points on the so-obtained boundary are identified by vertical shifting. *Dark gray quadrangles* show the evolution of an elementary lattice cell along a closed path around the defect point. Taken from [58]



Quantum Bifurcations, Figure 6
Representation of the quantum joint spectrum for the "Mexican hat" potential $V(r) = ar^4 - br^2$ with the "cut" along the eigenray. For such a cut the left and the right limits at the cut give the same values of actions (good quantum numbers) but the lines of constant values of actions exhibit a "kink" at the cut (the discontinuity of the first derivative)

cial matching rule explaining how to cross the path should be introduced. Similar constructions are quite popular in solid state physics in order to represent defects of lattices, like dislocations, disclinations, etc. We just note that the "monodromy defect" introduced in such a way is different from standard construction for dislocation and disclination defects [45,50]. The inverse procedure of the construction of the "monodromy defect" [90] from a regular lattice is represented in Fig. 5. Let us note that the width of the solid angle removed depends on the direction of the cut and the direction of the cut itself can be chosen in an ambiguous way.

(ii) An alternative possibility is to make a cut in such a way that the width of the removed angle becomes equal to zero. For focus-focus singularities one such direction always exists and is named an eigenray by Symington [75]. The same construction is used in some physical papers [10,11,85]. The inconvenience of such a procedure is the appearance of discontinuity of the slope of the constant action (quantum number) line at the cut, whereas the values of actions themselves are continued (see Fig. 6). This gives the wrong impression that this eigenray is associated with some special non-regular behavior of the initial problem, whereas there is no singularity except at one focus-focus point.

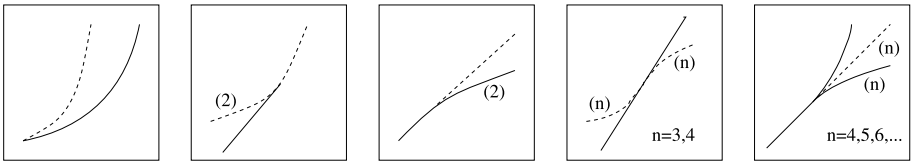
Bifurcations and Symmetry

The general mathematical answer about the possible qualitative modifications of a system of stationary points of functions depending on some control parameters can be found in bifurcation (or catastrophe) theory [3,31,33]. It is important that the answer depends on the number of control parameters and on the symmetry. Very simple classification of possible typical bifurcations of stationary points of a one-parameter family of functions under presence of symmetry can be formulated for dynamical systems with one degree-of-freedom. The situation is particularly simple because the phase space is two-dimensional and the complete list of local symmetry groups (which are the stabilizers of stationary points) includes only 2D-point groups [84]. It should be noted that the global symmetry of the problem can be larger than the local symmetry of the bifurcating stationary points. In such a case the bifurcations occur simultaneously for all points forming one orbit of the global symmetry group [51,52]. We describe briefly here (see Table 1) the classification of the bifurcations of stationary points in the presence of symmetry for families of functions depending on one parameter and associated quantum bifurcations [59,61]. Their notation includes the local symmetry group and several additional indexes which specify creation/annihilation of stationary points and the local or non-local character of the bifurcation. The list of possible bifurcations includes:

$C_1^{\pm}$ A non-symmetrical non-local bifurcation resulting in the appearance (+) or disappearance (−) of a stable-unstable pair of stationary points with the trivial local symmetry C_1. In the quantum problem this bifurcation is associated with the appearance or disappearance of a new regular sequence of states glued at its end with

Quantum Bifurcations, Table 1

Bifurcations in the presence of symmetry. *Solid lines* denote stable stationary points. *Dashed lines* denote unstable stationary points. *Numbers in parenthesis* indicate the multiplicity of stationary points



$C_1^{\pm}$ $C_2^{N\pm}$ $C_2^{L\pm}$ $C_n^N,\ n=3,4$ $C_n^{L\pm},\ n\geq 4$

the intermediate part of another regular sequence of quantum states [32,77].

$\boldsymbol{C_2^{L\pm}}$ A local bifurcation with the broken C_2 local symmetry. This bifurcation results either in appearance of a triple of points (two equivalent stable points with C_1 local symmetry and one unstable point with C_2 local symmetry) instead of one stable point with C_2 symmetry, or in inverse transformation. The number of stationary points in this bifurcation increases or decreases by two. For the quantum problem the result is the transformation of a local part of a regular sequence of states into one sequence of quasi-degenerate doublets.

$\boldsymbol{C_2^{N\pm}}$ A non-local bifurcation with the broken C_2 local symmetry. This bifurcation results in appearance (+) or disappearance (−) of two new unstable points with broken C_2 symmetry and simultaneous transformation of the initially stable (for +) or unstable (for −) stationary point into an unstable/stable one. The number of stationary points in this bifurcation increases or decreases by two. For the quantum problem this means the appearance of a new regular sequence of states near the separatrix between two different regular regions.

$\boldsymbol{C_n^N}$ $(n=3,4)$ A non-local bifurcation corresponding to passage of n unstable stationary points through a stable stationary point with C_n local symmetry which is accompanied with the minimum ↔ maximum change for a stable point with the C_n local symmetry. The number of stationary points remains the same. For the quantum problem this bifurcation corresponds to transformation of the increased sequence of energy levels into a decreased sequence.

$\boldsymbol{C_n^{L\pm}}$ $(n\geq 4)$ A local bifurcation which results in appearance (+) or disappearance (−) of n stable and n unstable stationary points with the broken C_n symmetry and a simultaneous minimum ↔ maximum change of a stable point with the C_n local symmetry. The number of stationary points increases or decreases by $2n$. In the quantum problem after bifurcation a new sequence of n-times quasi-degenerate levels appears/disappears.

Universal quantum Hamiltonians which describe the qualitative modification of the quantum energy level system around the bifurcation point are given in [59,61].

The presence of symmetry makes it much easier to observe the manifestation of quantum bifurcations. Modification of the local symmetry of stable stationary points results in the modification of the cluster structure of energy levels, i. e. the number and the symmetry types of energy level forming quasi-degenerate groups of levels. This phenomenon is essentially the spontaneous breaking of symmetry [51]. Several concrete molecular systems which show the presence of quantum bifurcations in rotational structure under rotational excitation are cited in Table 2. Many other examples can be found in [9,23,59,67,71,88,89,92,93] and references therein. In purely vibrational problems breaking dynamical $SU(N)$ symmetry of the isotrope harmonic oscillator till finite symmetry group results in formation of so-called non-linear normal modes [23,54] or quasimodes [1], or local modes [9,25,39,40,43,46,48]. In the case of two degrees-of-freedom the analysis of the vibrational problem can be reduced to the analysis of the problem similar to the rotational one [36,66] and all the results about possible types of bifurcations found for rotational problems remain valid in the case of intra-molecular vibrational dynamics.

Quantum Bifurcations, Table 2

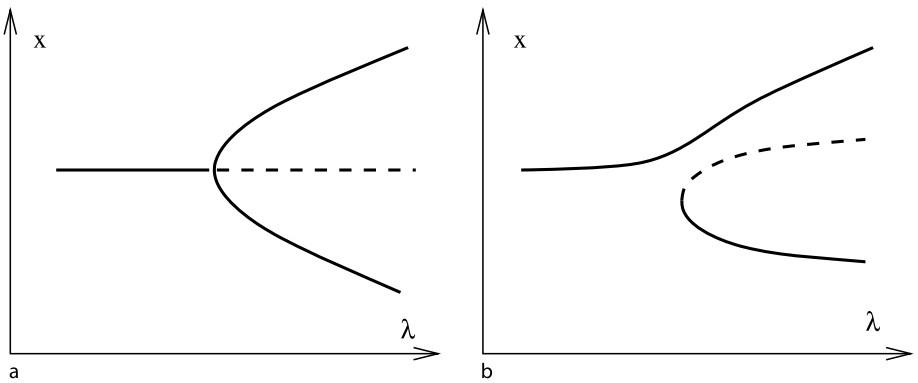
Molecular examples of quantum bifurcations in the rotational structure of individual vibrational components under the variation of the absolute value of angular momentum, *J*. J_c is the critical value corresponding to bifurcation

Molecule	Component	J_c	Bifurcation type
SiH_4	$\nu_2(+)$	12	C_2^{N+}
SnH_4	$\nu_2(-)$	10	$C_2^{N+}, C_3^N, C_4^N, C_2^{N-}$
CF_4	$\nu_2(+)$	50	C_4^{L+}
H_2Se	$\|0\rangle$	20	C_2^{L+}

Imperfect Bifurcations

According to general results the possible types of bifurcations which are generically present (and persist under small deformations) in a family of dynamical systems strictly depend on the number of control parameters. In the absence of symmetry only one bifurcation of stationary points is present for a one-parameter family of Morse-type functions, namely the formation (annihilation) of two new stationary points. This corresponds to saddle-node bifurcation for one degree-of-freedom Hamiltonian systems. The presence of symmetry increases significantly the number of possible bifurcations even for families with only one parameter [31,33]. From the physical point-of-view it is quite natural to study the effect of symmetry breaking on the symmetry allowed bifurcation. Decreasing symmetry naturally results in the modification of the allowed types of bifurcations but at the same time it is quite clear that at sufficient slight symmetry breaking perturbation the resulting behavior of the system should be rather close to the behavior of the original system with higher symmetry.

In the case of a small violation of symmetry the so-called "imperfect bifurcations" can be observed. Imperfect bifurcations, which are well known in the classical theory of bifurcations [33] consist of the appearance of stationary points in the neighborhood of another stationary point which does not change its stability. In some way one can say that imperfect bifurcation mimics generic bifurcation in the presence of higher symmetry by the special organization of several bifurcations which are generic in the presence of lower symmetry. Naturally quantum bifurcations follow the same behavior under the symmetry breaking as classical ones. Very simple and quite natural examples of imperfect quantum bifurcations were demonstrated on the example of the rotational structure modifications under increasing angular momentum [91]. The idea of appearance of imperfect bifurcations is as follows. Let us suppose that some symmetrical molecule demonstrates under the variation of angular momentum a quantum rotational bifurcation allowed by symmetry. The origin of this bifurcation is due, say, to centrifugal distortion effects which depend strongly on J but are not very sensitive to small variation of masses even in the case of symmetry breaking isotopic substitution. In such a case a slight modification of the masses of one or several equivalent atoms breaks the symmetry and this symmetry violation can be made very weak due to the small ratio $\Delta M/M$ under isotope substitution. In classical theory the effect of symmetry breaking can be easily seen through the variation of the position of stationary points with control parameter. For example, instead of a pitchfork bifurcation which is typical for C_2 local symmetry, we get for the unsymmetrical problem (after slight breaking of C_2 symmetry) a smooth evolution of the position of one stationary point and the appearance of two new stationary points in fold catastrophe (see Fig. 7). In associated quantum bifurcations the most important effect is the splitting of clusters. But one should be careful with this interpretation because in quantum mechanics of finite particle systems the clusters are always split due to quantum mechanical tunneling between different equivalent regions of localization of quantum wave functions. Intercluster splitting increases rapidly approaching the region of classical separatrix. The behavior of quantum tunneling was studied extensively in relation to the quantum breathers problem [6,29]. Systematic application of quasi-classical methods to reproduce quantum energy



Quantum Bifurcations, Figure 7
Imperfect bifurcations. a Position *x* of stationary points as a function of control parameter λ during a pitchfork bifurcation in the presence of C_2 local symmetry. b Modifications induced by small symmetry perturbation of lower symmetry. *Solid line*: Stable stationary points. *Dashed lines*: Unstable stationary points

level structure near the singularities of the energy-momentum maps where exponentially small corrections are important is possible but requires special efforts (see for example [12]) and we will not touch upon this problem here.

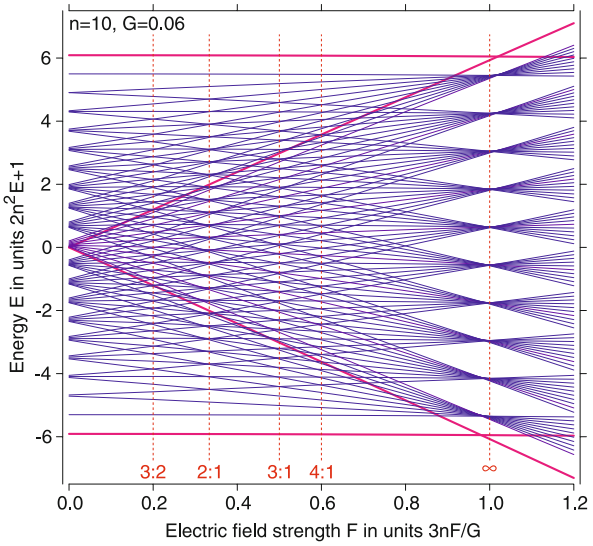
Organization of Bifurcations

The analysis of the quantum bifurcations in concrete examples of rotating molecules have shown that in some cases the molecule undergoes several consecutive qualitative changes which can be interpreted as a sequence of bifurcations which sometimes cannot even be separated into elementary bifurcations for the real scale of the control parameter [89]. One can imagine in principle that successive bifurcations lead to quantum chaos in analogy with classical dynamical systems where the typical scenario for the transition to chaos is through a sequence of bifurcations. Otherwise, the molecular examples were described with effective Hamiltonians depending only on one degree-of-freedom and the result of the sequence of bifurcations was just the crossover of the rotational multiplets [64]. In some sense such a sequence of bifurcations can be interpreted as an imperfect bifurcation assuming initially higher dynamical symmetry, like the continuous $SO(3)$ group.

Later, a similar crossover phenomenon was found in a quite different quantum problem, like the hydrogen atom in external fields [24,53,72]. The general idea of such organization of bifurcations is based on the existence of two different limiting cases of dynamical regimes for the same physical quantum system (often under presence of the same symmetry group) which are qualitatively different. For example, the number of stationary points, or their stability differs. If H_1 and H_2 are two corresponding effective Hamiltonians, the natural question is: Is it possible to transform H_1 into H_2 by a generic perturbation depending on only one parameter? And if so, what is the minimal number of bifurcations to go through?

The simplest quantum system for which such a question becomes extremely natural is the hydrogen atom in the presence of external static electric (F) and magnetic (G) fields. Two natural limits – the Stark effect in the electric field and Zeeman effect in the magnetic field – show quite different qualitative structure even in the extremely low field limit [15,20,63,72,78]. Keeping a small field one can go from one (Stark) limit to another (Zeeman) and this transformation naturally goes through qualitatively different regimes [24,53]. In spite of the fact that the hydrogen atom (even without spin and relativistic corrections) is only a three degree-of-freedom system, the complete description of qualitatively different regimes in a small field limit is still not done and remains an open problem [24].

An example of clearly seen qualitative modifications of the quantum energy level system of the hydrogen atom under the variation of F/G ratio of the strengths of two parallel electric and magnetic fields is shown in Fig. 8. The calculations are done for a two degree-of-freedom system after the normalization with respect to the global action. In quantum mechanics language this means that only energy levels which belong to the same n-shell of the hydrogen atom are treated and the interaction with other n' shells is taken into account only effectively. The limiting classical phase space for this effective problem is the four-dimensional space $S^2 \times S^2$, which is the direct product of two two-dimensional spheres. In the presence of axial symmetry this problem is completely integrable and the Hamiltonian and the angular momentum provide a complete set of mutually commuting operators. Energies of stationary points of classical Hamiltonian limit are shown on the same Fig. 8 along with quantum levels. When one of the characteristic frequencies goes through zero, the so-called collapse phenomena occurs. Some other non-trivial resonance relations between two frequencies are also indicated. These resonances correspond to special orga-



Quantum Bifurcations, Figure 8

Reorganization of the internal structure of the *n*-multiplet of the hydrogen atom in small parallel electric and magnetic fields. Energies of stationary points of the classical Hamiltonian (*red solid lines*) are shown together with quantum energy levels (*blue solid lines*). The figure is done for $n = 10$ (there are $n^2 = 100$ energy levels forming this multiplet). As the ratio F/G of electric F and magnetic G fields varies this two degree-of-freedom system goes through different zones associated with special resonance relations between two characteristic frequencies (shown by *vertical dashed lines*). Taken from [24]

nization of quantum energy levels. At the same time it is not necessary here to go to joint spectrum representation in order to see the reorganization of stationary points of the Hamiltonian function on $S^2 \times S^2$ phase space under the variation of the external control parameter F/G. A more detailed treatment of qualitative features of the energy level systems for the hydrogen atom in low fields is given in [15,20,24].

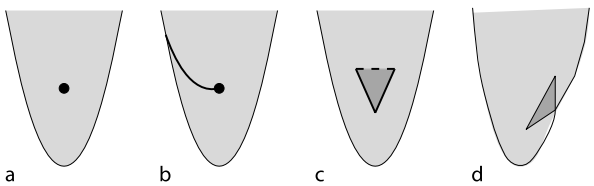
Bifurcation Diagrams for Two Degree-of-Freedom Integrable Systems

Let us consider now the two degree-of-freedom integrable system with compact phase space as a bit more complex but still reasonably simple problem. Many examples of such systems possess EM maps with the stratification of the image formed by the regular part surrounded by the singular boundary. The most naturally arising examples of classical phase spaces, like $S^2 \times S^2$, CP^2, are of that type. All internal points on the image of the EM map are regular in these cases. In practice, real physical problems, even after necessary simplifications and approximations lead to more complicated models. Some examples of fragments of images of the EM map with internal singular points are shown in Fig. 9. In classical mechanics the inverse images of critical values are singular tori of different kinds. Some of them are represented in Fig. 10. Inverse images of critical points situated on the boundary of the EM image have lower dimension. They can be one-dimensional tori (S^1-circles), or zero-dimensional (points).

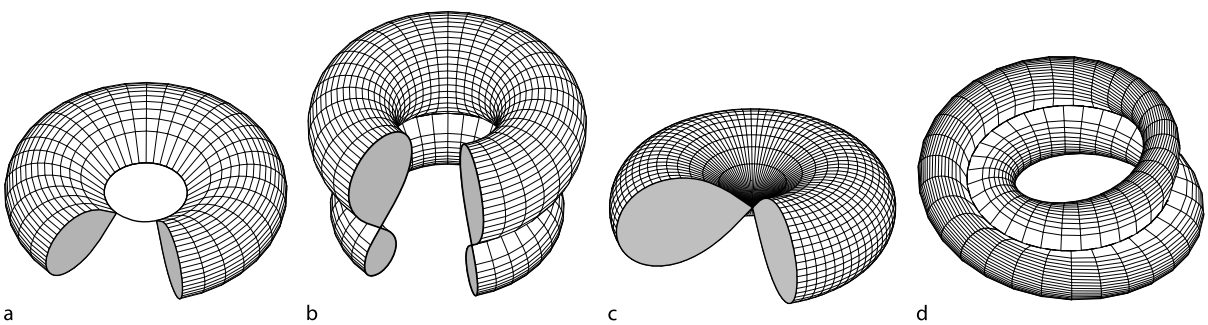
The natural question now is to describe typical generic modifications of the Hamiltonian which lead to qualitative modifications of the EM map image in classical mechanics and to associated modifications of the joint spectrum in quantum mechanics.

The simplest classical bifurcation leading to modification of the image of the EM map is the Hamiltonian Hopf bifurcation [79]. It is associated with the following modification of the image of the EM map. The critical value of the EM map situated on the boundary leaves the boundary and enters the internal domain of regular values (see Fig. 11). As a consequence, the toric fibration over the closed path surrounding an isolated singularity is non-trivial. Its non-triviality can be characterized by the Hamiltonian monodromy which describes the mapping from the fundamental group of the base space into the first homology group of the regular fiber [18]. A typical pattern of the joint spectrum around such a classical singularity is shown in Fig. 3. The joint spectrum manifests the presence of quantum monodromy. Its interpretation in terms of regular lattices is given in Figs. 4 and 5.

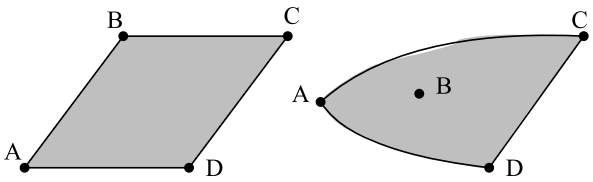
Taking into account additional terms of higher order it is possible to distinguish different types of Hamiltonian



Quantum Bifurcations, Figure 9
Typical images of the energy momentum map for completely integrable Hamiltonian systems with two degrees-of-freedom in the case of: a integer monodromy, b fractional monodromy, c non-local monodromy, and d bidromy. Values in the *light shaded area* lift to single 2-tori; values in the *dark shaded area* lift to two 2-tori. Taken from [69]



Quantum Bifurcations, Figure 10
Two-dimensional singular fibers in the case of integrable Hamiltonian systems with two degrees-of-freedom (*left* to *right*): singular torus, bitorus, pinched and curled tori. Singular torus corresponds to critical values in Fig. 9c, d (ends of bitoris line). Bitorus corresponds to critical values in Fig. 9c, d, which belong to singular line (fusion of two components). Pinched torus corresponds to isolated focus-focus singularity in Fig. 9a. Curled torus is associated with critical values at singular line in Fig. 9b (fractional monodromy). Taken from [69]



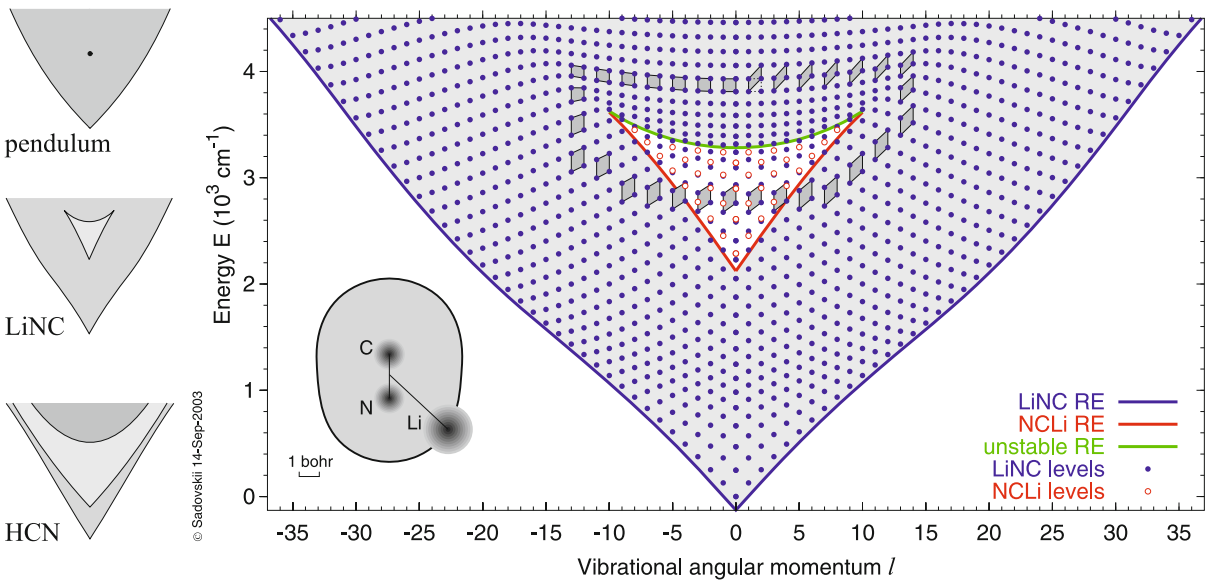
Quantum Bifurcations, Figure 11

Qualitative modification of the image of the EM map due to Hamiltonian Hopf bifurcation. *Left*: Simplest integrable toric fibration over $S^2 \times S^2$ classical phase space. *A*, *B*, *C*, *D*: Critical values corresponding to singular S^0 fibers. Regular points on the boundary correspond to S^1 fibers. Regular internal points: Regular T^2 fibers. *Right*: Appearance of an isolated critical value inside the field of regular values. Critical value *B* corresponds to pinched torus shown in Fig. 10

Hopf bifurcations usually named as subcritical and supercritical [19,79]. New qualitative modification, for example, corresponds to transformation of an isolated singular value of the EM map into an "island", i. e. the region of the EM image filled by points whose inverse images consist of two connected components. Integrable approximation for vibrational motion in the LiCN molecule shows the presence of such an island associated with the non-local quantum monodromy (see Fig. 12) [41]. The monodromy naturally coincides with the quantum monodromy of isolated focus-focus singularity which deforms continuously into the island monodromy. It is interesting to note that in molecule HCN which is rather similar to LiCN, the region with two components in the inverse image of the EM map exists also but the monodromy cannot be defined due to impossibility to go around the island [22].

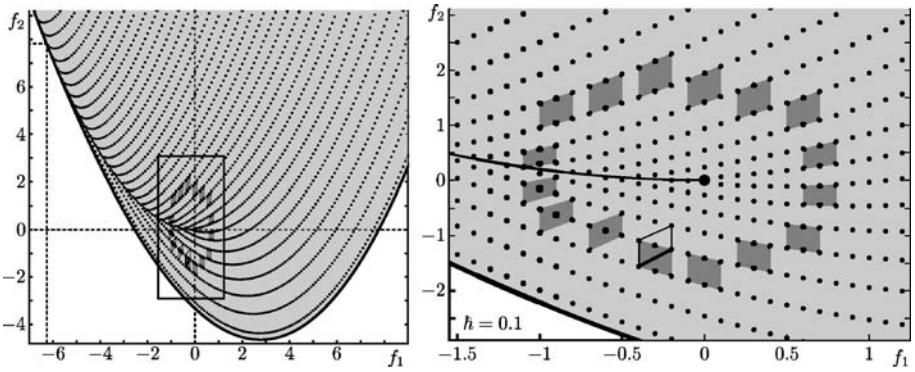
In the quantum problem the presence of "standard" quantum monodromy in the joint spectrum of two mutually commuting observables can be seen through the mapping of a locally regular part of the joint spectrum lattice to an idealized Z^2 lattice. Existence of local actions for the classical problem which are defined almost everywhere and the multivaluedness of global actions from one side and the quantum-classical correspondence from another side allow the interpretation of the joint spectrum with quantum monodromy as a regular lattice with an isolated defect.

Recently, the generalization of the notion of quantum (and classical) monodromy was suggested [21,58]. For quantum problems the idea is based on the possibility to study instead of the complete lattice formed by the joint spectrum only a sub-lattice of finite index. Such a transformation allows one to eliminate certain "weak line sin-

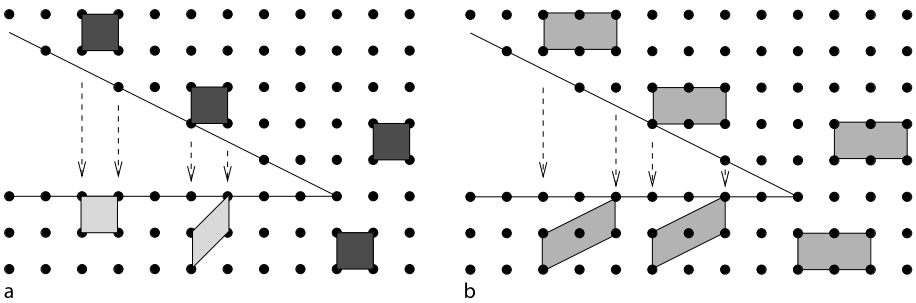


Quantum Bifurcations, Figure 12

Quantum joint spectrum for the quantum model problem with two degrees-of-freedom describing two vibrations in the LiCN molecule. The non-local quantum monodromy is shown by the evolution of the elementary cell of the quantum lattice around the singular line associated with gluing of two regular lattices corresponding in molecular language to two different isomers, LiCN and LiNC. Classical limit (*left*) shows the possible deformation of isolated focus-focus singularity for pendulum to non-local island singularity for LiNC model. In contrast to LiCN, the HCN model has an infinite island which cannot be surrounded by a close path. Taken from [41]



Quantum Bifurcations, Figure 13
Joint quantum spectrum for two-dimensional non-linear 1 : (−2) resonant oscillator (5),(6). The singular line is formed by critical values whose inverse images are curled tori shown in Fig. 10. In order to get the unambiguous result of the propagation of the cell of the quantum lattice along a closed path crossing the singular line, the elementary cell is doubled. Taken from [58]



Quantum Bifurcations, Figure 14
Representation of a lattice with 1 : 2 rational defect by cutting and gluing. *Left*: The elementary cell goes through cut in an ambiguous way. The result depends on the place where the cell crosses the cut. *Right*: Double cell crosses the cut in an unambiguous way. Taken from [58]

gularities" presented in the image of the EM map. The resulting monodromy is named "fractional monodromy" because for the elementary cell in the regular region the formal transformation after a propagation along a close path crossing "weak line singularities" turns out to be represented in a form of a matrix with fractional coefficients.

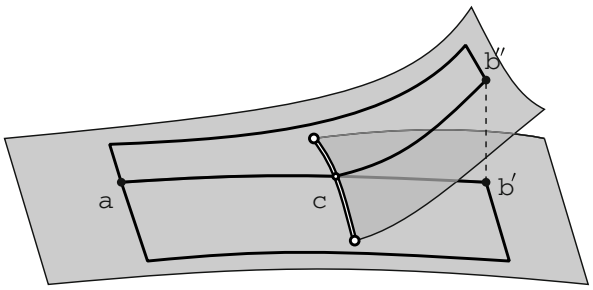
An example of quantum fractional monodromy can be given with a 1 : (−2) resonant oscillator system possessing two integrals of motion f_1, f_2 in involution:

$$f_1 = \frac{\omega}{2}\left(p_1^2 + q_1^2\right) - \frac{2\omega}{2}\left(p_2^2 + q_2^2\right) + R_1(q,p)\,, \qquad (5)$$

$$f_2 = \mathrm{Im}\left[(q_1 + ip_1)^2(q_2 + ip_2)\right] + R_2(q,p)\,. \qquad (6)$$

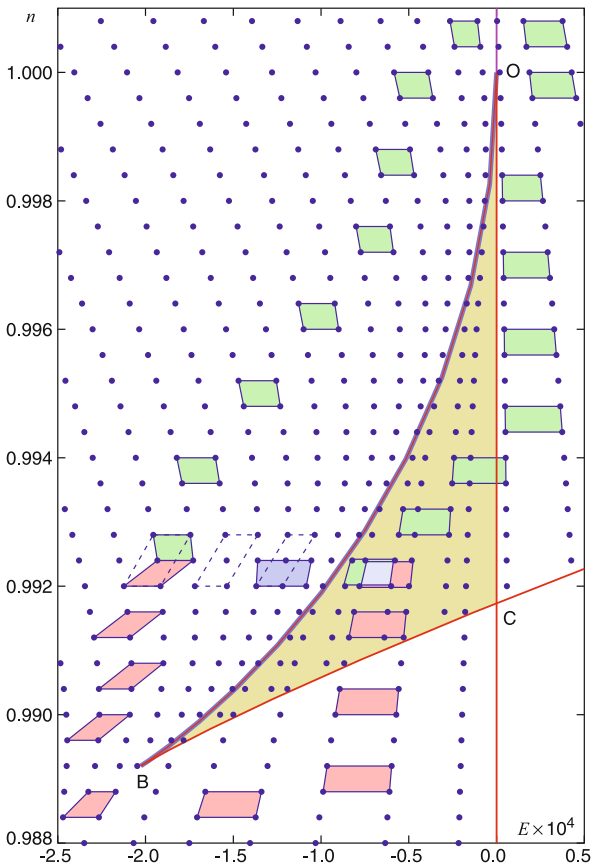
The corresponding joint spectrum for the quantum problem is shown in Fig. 13. It can be represented as a regular Z^2 lattice with a solid angle removed (see Fig. 14). The main difference with the standard integer monodromy representation is due to the fact that even after gluing two sides of the cut we get the one-dimensional singular stratum which can be neglected only after going to a sublattice (to a sub-lattice of index 2 for 1 : 2 fractional singularity).

Another kind of generalization of the monodromy notion is related to the appearance of multi-component inverse images for the EM maps. We have already mentioned such a possibility with the appearance of non-local monodromy and Hamiltonian Hopf bifurcations (see Fig. 12). But in this case two components of the inverse image belong to different regular domains and cannot be joined by a path going only through regular values. Another possibility is suggested in [69,70] and is explained schematically in Fig. 15. This figure shows that the arrangement of fibers can be done in such a way that one connected component can be deformed into another connected component along a path which goes only through regular tori. The existence of a quantum joint spectrum corresponding to such a classical picture was demon-



Quantum Bifurcations, Figure 15
Schematic representation of the inverse images for a problem with bidromy in the form of the unfolded surface. Each connected component of the inverse image is represented as a *single point*. The path $b' - a - b''$ starts and ends at the same point of the space of possible values of integrals of motion but it starts at one connected component and ends at another one. At the same time the path goes only through regular tori. Taken from [70]

strated on the example of a very well-known model problem with three degrees-of-freedom: Three resonant oscillators with 1 : 1 : 2 resonance, axial symmetry and with small detuning between double degenerate and non-degenerate modes [30,70]. The specific behavior of the joint spectrum for this model can be characterized as self-overlapping of a regular lattice. The possibility to propagate the initially chosen cell through a regular lattice from the region of self-overlapping of lattice back to the same region but to another component was named "bidromy". More complicated construction for the same problem allows us to introduce the "bipath" notion. The bipath starts at a regular point of the EM image, and crosses the singular line by splitting itself into two components. Each component belongs to its proper lattice in the self-overlapping region. Two components of the path can go back through the regular region only and fuse together. The behavior of quantum cells along a bipath is shown in Fig. 16. Providing a rigorous mathematical description of such a construction is still an open problem. Although the original problem has three degrees-of-freedom, it is possible to construct a model system with two degrees-of-freedom and with similar properties.



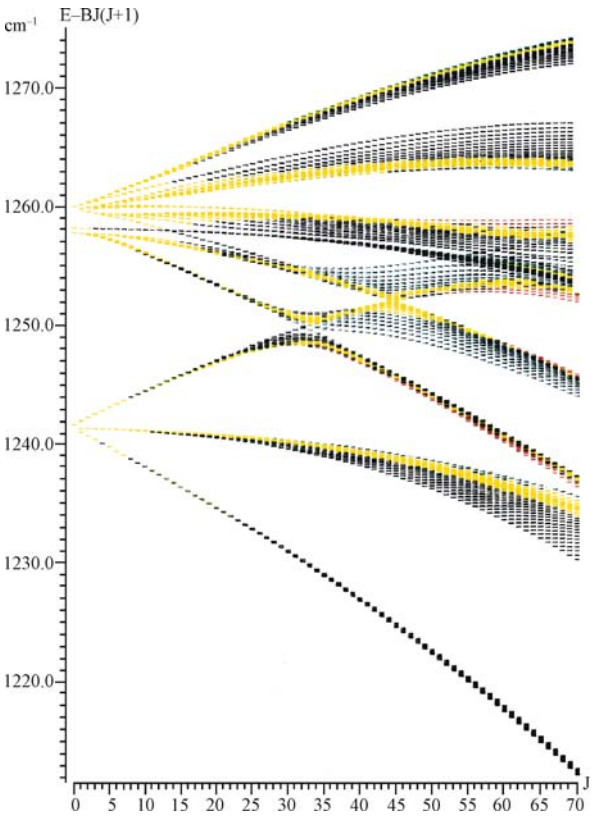
Quantum Bifurcations, Figure 16
Joint quantum spectrum for problem with bidromy. Quantum states are given by two numbers (energy, *E*, and polyad number, *n*) which are the eigenvalues of two mutually commuting operators. Inside the *OAB* curvilinear triangle two regular lattices are clearly seen. One can be continued smoothly through the *OC* boundary whereas another continues through the *BC* boundary. This means that the regular part of the whole lattice can be considered as a one self-overlapping regular lattice. The figure suggests also the possibility to define the propagation of a double cell along a "bipath" through the singular line *BO* which leads to splitting of the cell into two elementary cells fusing at the end into one cell defining in such a way the "bidromy" transformation associated with a bipath. Taken from [70]

Bifurcations of "Quantum Bifurcation Diagrams"

We want now to stress some differences in the role of internal and external control parameters. From one point-of-view a quantum problem, which corresponds in the classical limit to a multidimensional integrable classical model, possesses a joint spectrum qualitatively described by a "quantum bifurcation diagram". This diagram shows that the joint spectrum is formed from several parts of regular lattices through a cutting and gluing procedure. Going from one regular region to another is possible by crossing singular lines. The parameter defined along such a path can be treated as an internal control parameter. It is essentially a function of values of integrals of motion. To cross the singular line is equivalent to passing the quantum bifurcation for a family of reduced systems with a smaller number of degrees of freedom.

On the other side we can ask the following more general question. What kinds of generic modifications of "bifurcation diagrams" are possible for a family of integrable systems depending on some external parameters? Hamiltonian Hopf bifurcation leading to the appearance of a new isolated singular value and as a consequence appearance of monodromy is just one of the possible effects of this kind. Another possibility is the transformation of an isolated focus-focus singular value into the island associated with the presence of a second connected component of the inverse image of the EM map. It is also possible that such an island is born within the regular region of the EM map. In such a case naturally the monodromy transformation associated with a closed path surrounding the so-obtained island should be trivial (identity).

The boundary of the image of the EM map can also undergo transformation which results in the appearance of the region with two components in the inverse image but, in contrast to the previous example of the appearance of an island, these two components can be smoothly deformed one onto another along a continuous path going only through regular values of the EM map. Examples of all such modifications were studied on simple models inspired by concrete quantum molecular systems like the H atom, CO_2, LiCN molecules and so on [24,30,41].



Quantum Bifurcations, Figure 17
System of rovibrational energy levels of $^{13}CF_4$ molecule represented schematically in E, J coordinates. The number of energy levels in each clearly seen band is $2J + 1 + \delta$, where δ is a small integer which remains constant for isolated bands and changes at band intersections. In the semi-quantum model δ is interpreted as the first Chern class, characterizing the non-triviality of the vector bundle formed by eigenfunctions of the "fast" subsystem over the classical phase space of the "slow" subsystem [27]

Semi-Quantum Limit and Reorganization of Quantum Bands

Up to now we have discussed the qualitative modifications of internal structures of certain groups of quantum levels which are typically named bands. Their appearance is physically quite clear in the adiabatic approximation. The existence of fast and slow classical motions manifests itself in quantum mechanics through the formation of so-called energy bands. The big energy difference between energies of different bands correspond to fast classical variables whereas small energy differences between energy levels belonging to the same band correspond to classical slow variables. Typical bands in simple quantum systems correspond to vibrational structure of different electronic states, rotational structure of different vibrational states, etc.

If now we have a quantum problem which shows the presence of bands in its energy spectrum, the natural generalization consists of putting this quantum system in a family, depending on one (or several) control parameters. What are the generic qualitative modifications which can be observed within such a family of systems when control parameters vary? Apart from qualitative modifications of the internal structure of individual bands which can be treated as the earlier discussed quantum bifurcations, another qualitative phenomenon is possible, namely the redistribution of energy levels between bands or more generally, the reorganization of bands under the variation of some control parameters [8,26,28,62,68]. In fact this phenomenon is very often observed in both the numerical simulations and the real experiments with molecular systems exhibiting bands. A typical example of molecular rovibrational energy levels classified according to their energy and angular momentum is shown in Fig. 17. It is important to note that the number of energy levels in bands before and after their "intersection" changes.

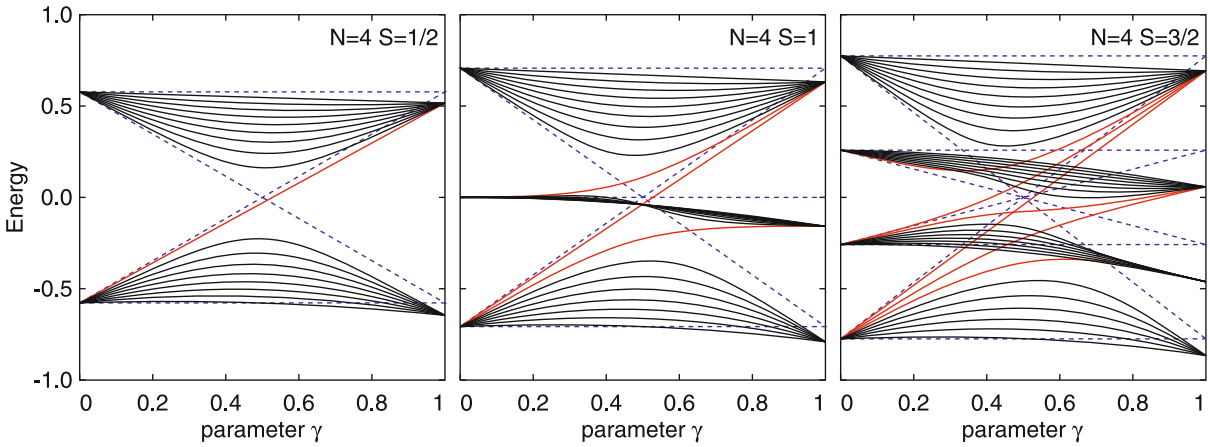
The same phenomenon of the redistribution of energy levels between energy bands can be understood by the example of a much simpler quantum system of two coupled angular momenta, say orbital angular momentum

and spin in the presence of a magnetic field interacting only with spin [62,68].

$$H = \frac{1-\gamma}{\mathbf{S}} S_z + \frac{\gamma}{\mathbf{NS}} (\mathbf{N}\cdot\mathbf{S}) , \quad 0 \le \gamma \le 1 . \tag{7}$$

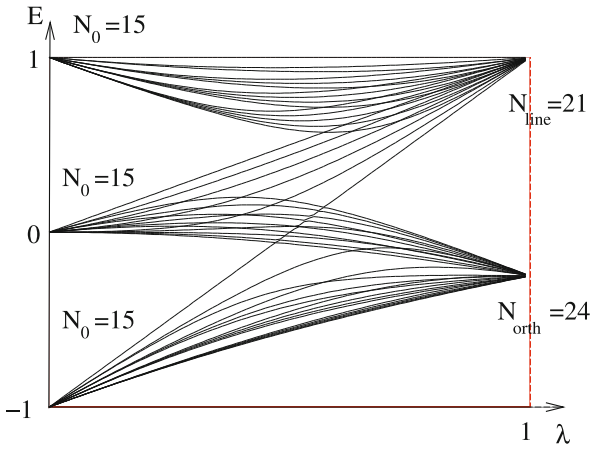
The Hamiltonian for such a system can be represented in the form of a one-parameter family (7) having two natural limits corresponding to uncoupled and coupled angular momenta. The interpolation of eigenvalues between these two limits is shown in Fig. 18 for different values of spin quantum number, $S = 1/2, 1, 3/2$. The quantum number of orbital momentum is taken to be $N = 4$. Although this value is not much larger than the S values, the existence of bands and their reorganization under the variation of the external parameter γ is clearly seen in the figure.

Although the detailed description of this reorganization of bands will take us rather far away from the principal subject it is important to note that in the simplest situations there exists a very close relation between the redistribution phenomenon and the Hamiltonian Hopf bifurcations leading to the appearance of Hamiltonian monodromy [81]. In the semi-quantum limit when part of the dynamical variables are treated as purely classical and all the rest as quantum, the description of the complete system naturally leads to a fiber bundle construction [27]. The role of the base space is taken by the classical phase space for classical variables. A set of quantum wave-functions associated with one point of the base space forms a complex fiber. As a whole the so-obtained vector bundle with complex fibers can be topologically characterized by its rank and Chern classes [56]. Chern classes are related to the number of quantum states in bands formed due to quantum character of the total problem with respect to "classical" variables. Modification of the number of states in bands can occur only at band contact and is associated with the modification of Chern classes of the corresponding fiber bundle [26]. The simplest situation takes place when the number of degrees of freedom associated with classical variables is one. In this case only one topological invariant – the first Chern class is sufficient to characterize the non-triviality of the fiber bundle and the difference in Chern classes is equal to the number of energy levels redistributed between corresponding bands. Moreover, in the generic situation (in the absence of symmetry) the typical behavior consists of the redistribution of only one energy level between two bands. The generic phenomena become more complicated with increasing the number of degrees of freedom for the classical part of variables. The model problem with two slow degrees of freedom (described in classical limit by the CP^2 phase space) and three quantum states was studied in [28]. A new qualitative phenomenon was found, namely, the modification of the number of bands due to formation of topologically coupled bands. Figure 19 shows the evolution of the system of energy levels along with the variation of control parameter λ. Three quantum bands (at $\lambda = 0$) transform into two bands (in the $\lambda = 1$) limit. One of these bands has rank one, i. e. it is associated with one quantum state. Another has rank two. It is associated with two quantum states. Both bands have non-trivial topology (non-trivial Chern classes). Moreover, it is quite important that the newly formed topologically coupled band of rank two can



Quantum Bifurcations, Figure 18

Rearrangement of energy levels between bands for model Hamiltonian (7) with two, three, or four states for "fast" variable. Quantum energy levels are shown by *solid lines*. Classical energies of stationary points for energy surfaces are shown by *dashed lines*. Taken from [68]



Quantum Bifurcations, Figure 19
Rearrangement of three bands into two topologically non-trivially coupled bands. Example of a model with three electronic states and vibrational structure of polyads formed by three quasi-degenerate modes. At $\lambda = 0$ three bands have each the same number of states, namely 15. In the classical limit each initial band has rank one and trivial topology. At $\lambda = 1$ there are only two bands. One of them has rank 2 and non-trivial first and second Chern classes. Taken from [28]

be split into two bands of rank one only if a coupling with the third band is introduced.

The corresponding qualitative modifications of quantum spectra can be considered as natural generalizations of quantum bifurcations and probably should be treated as topological bifurcations. Thus, the description of possible "elementary" rearrangements of energy bands is a direct consequence of topological restrictions imposed by a fiber bundle structure of the studied problem.

It is interesting to mention here the general mathematical problem of finding proper equivalence or better to say correspondence between some construction made over real numbers and their generalizations to complex numbers and quaternions. This paradigm of complexification and quaternization was discussed by Arnold [4,5] on many different examples. The closest to the present subject is the example of complexification of the Wigner–Neumann non-crossing rule resulting in a quantum Hall effect (in physical terms). In fact, the mathematical basis of the quantum Hall effect is exactly the same fiber bundle construction which explains the redistribution of energy levels between bands in the above-mentioned simple quantum mechanical model.

Multiple Resonances and Quantum State Density

Rearrangement of quantum energy states between bands is presented in the previous section as an example of a generic qualitative phenomenon occurring under variation of a control parameter. One possible realization of bands is the sequence of vibrational polyads formed by a system of resonant vibrational modes indexed by the polyad quantum number. In the classical picture this construction corresponds to the system of oscillators reduced with respect to the global action. The reduced classical phase space is in such a case the weighted projective space. In the case of particular $1 : 1 : \ldots : 1$ resonance the corresponding reduced phase space is a normal complex projective space CP^n. The specific resonance conditions impose for a quantum problem specific conditions on the numbers of quantum states in polyads. In the simplest case of harmonic oscillators with $n_1 : n_2 : \ldots : n_k$ resonance the numbers of states in polyads are given by the generating function

$$g = \frac{1}{(1-t^{n_1})(1-t^{n_2})\cdots(1-t^{n_k})} = \sum_N C_N\, t^N\,, \quad (8)$$

where N is the polyad quantum number. Numbers C_N are integers for integer N values, but they can be extended to arbitrary N values and represented in the form of a quasi-polynomial, i.e, a polynomial in N with coefficients being a periodic function whose period equals the least common multiplier of n_i, $i = 1, \ldots, k$. Moreover, the coefficients of the polynomial can be expressed in terms of so-called Todd polynomials which indicates the possibility of topological interpretation of such information [52,89].

Physical Applications and Generalizations

The most clearly seen physical applications of quantum bifurcations is the qualitative modification of the rotational multiplet structure under rotational excitation, i. e. under the variation of the absolute value of the angular momentum. This is related first of all with the experimental possibility to study high J multiplets (which are quite close to the classical limit but nevertheless manifest their quantum structure) and to the possibility to use symmetry arguments, which allow one to distinguish clusters of states before and after bifurcation just by counting the number of states in the cluster, which depends on the order of group of stabilizer. Nuclear rotation is another natural example of quantum rotational bifurcations [60]. Again the interest in corresponding qualitative modifications is due to the fact that rotational bands are extremely well studied up to very high J values. But in contrast to molecular physics examples, in nuclear physics it mostly happens that only ground states (for each value of J) are known. Thus, one speaks more often about qualitative changes of

the ground state (in the absence of temperature) named quantum phase transitions [65].

Internal structure of vibrational polyads is less evident for experimental verifications of quantum bifurcations, but it gives many topologically non-trivial examples of classical phase spaces on which the families of Hamiltonians depending on parameters are defined [25,30,39,42,44,46,66,76,77,86]. The main difficulty here is the small number of quantum states in polyads accessible to experimental observations. But this problem is extremely interesting from the point-of-view of extrapolation of theoretical results to the region of higher energy (or higher polyad quantum numbers) which is responsible as a rule for many chemical intra-molecular processes. Certain molecules, like CO_2, or acetylene (C_2H_2) are extremely well studied and a lot of highly accurate data exist. At the same time the qualitative understanding of the organization of excited states even in these molecules is not yet completed and new qualitative phenomena are just starting to be discovered.

Among other physically interesting systems it is necessary to mention model problems suggested to study the behavior of Bose condensates or quantum qubits [37,38,74,83,84]. These models have a mathematical form which is quite similar to rotational and vibrational models. At the same time their physical origin and the interpretation of results is quite different. This is not an exception. For example, the model Hamiltonian corresponding in the classical limit to a Hamiltonian function defined over S^2 classical phase space is relevant to rotational dynamics, description of internal structure of vibrational polyads formed by two (quasi)degenerate modes, in particular to so-called local-normal mode transition in molecules, interaction of electromagnetic field with a two-level system, the Lipkin–Meshkov–Glick model in nuclear physics, entanglement of qubits, etc.

Future Directions

To date many new qualitative phenomena have been suggested and observed in experimental and numerical studies due to intensive collaboration between mathematicians working in dynamical system theory, classical mechanics, complex geometry, topology, etc., and molecular physicists using qualitative mathematical tools to classify behavior of quantum systems and to extrapolate this behavior from relatively simple (low energy regions) to more complicated ones (high energy regions). Up to now the main accent was placed on the study of the qualitative features of isolated time-independent molecular systems. Specific patterns formed by energy eigenvalues and by common eigenvalues of several mutually commuting observables were the principal subject of study. Existence of qualitatively different dynamical regimes for time-independent problems at different values of exact or approximate integrals of motion were clearly demonstrated. Many of these new qualitative features and phenomena are supposed to be generic and universal although their rigorous mathematical formulation and description is still absent.

On the other side, the analysis of the time-dependent processes should be developed. This step is essential in order to realize at the level of quantum micro-systems the transformations associated with the qualitative modifications of dynamical regimes and to control such time-dependent processes as elementary reactions, information data storage, and so on. From this global perspective the main problem of the future development is to support the adequate mathematical formulation of qualitative methods and to improve our understanding of qualitative modifications occurring in quantum micro-systems in order to use them as real micro-devices.

Bibliography

1. Arnold VI (1972) Modes and quasimodes. Funct Anal Appl 6:94–101
2. Arnold VI (1989) Mathematical methods of classical mechanics. Springer, New York
3. Arnold VI (1992) Catastrophe theory. Springer, Berlin
4. Arnold VI (1995) Remarks on eigenvalues and eigenvectors of Hermitian matrices, Berry phase, adiabatic connections and quantum Hall effect. Selecta Math New Ser 1:1–19
5. Arnold VI (2005) Arnold's problems. Springer, Berlin
6. Aubry S, Flach S, Kladko K, Olbrich E (1996) Manifestation of classical bifurcation in the spectrum of the integrable quantum dimer. Phys Rev Lett 76:1607–1610
7. Bolsinov AV, Fomenko AT (2004) Integrable Hamiltonian systems. Geometry topology classifications. Chapman and Hall/CRC, London
8. Brodersen S, Zhilinskiĭ BI (1995) Transfer of clusters between the vibrational components of CF_4. J Mol Spectrosc 169:1–17
9. Child MS (2000) In: Jensen P, Bunker PR (eds) Computational molecular spectroscopy, chapter 18. Wiley Interscience, Chichester
10. Child MS (2001) Quantum level structure and nonlinear classical dynamics. J Mol Spectrosc 210:157–165
11. Child MS, Weston T, Tennyson J (1999) Quantum monodromy in the spectrum of H_2O and other systems. Mol Phys 96: 371–379
12. Colin de Verdier Y, Vũ Ngoc S (2003) Singular Bohr–Sommerfeld rules for 2D integrable systems. Ann Ec Norm Sup 36:1–55
13. Cushman RS, Bates L (1997) Global aspects of classical integrable systems. Birkhäuser, Basel
14. Cushman RH, Duistermaat JJ (1988) The quantum mechanical spherical pendulum. Bull Am Math Soc 19:475–479
15. Cushman RH, Sadovskiĭ DA (2000) Monodromy in the hydrogen atom in crossed fields. Physica D 142:166–196

16. Dirac PAM (1982) The principles of quantum mechanics. Oxford University Press, Oxford
17. Duistermaat JJ (1980) On global action angle coordinates. Comm Pure Appl Math 33:687–706
18. Duistermaat JJ (1998) The monodromy in the Hamiltonian Hopf bifurcation. Angew Z Math Phys 49:156–161
19. Efstathiou K (2004) Metamorphoses of Hamiltonian systems with symmetry. Lecture Notes in Mathematics, vol 1864. Springer, Heidelberg
20. Efstathiou K, Cushman RH, Sadovskií DA (2004) Hamiltonian Hopf bifurcation of the hydrogen atom in crossed fields. Physica D 194:250–274
21. Efstathiou K, Cushman RH, Sadovskií DA (2007) Fractional monodromy in the $1:-2$ resonance. Adv Math 209:241–273
22. Efstathiou K, Joyeux M, Sadovskií DA (2004) Global bending quantum number and the absence of monodromy in the HCN↔CNH molecule. Phys Rev A 69(3):032504-1–15
23. Efstathiou K, Sadovskií DA, Zhilinskií BI (2004) Analysis of rotation-vibration relative equilibria on the example of a tetrahedral four atom molecule. SIAM J Dyn Syst 3:261–351
24. Efstathiou K, Sadovskií DA, Zhilinskií BI (2007) Classification of perturbations of the hydrogen atom by small static electric and magnetic fields. Proc Roy Soc Lond A 463:1771–1790
25. Ezra GS (1996) Periodic orbit analysis of molecular vibrational spectra: Spectral patterns and dynamical bifurcations in Fermi resonant systems. J Chem Phys 104:26–35
26. Faure F, Zhilinskií BI (2000) Topological Chern indices in molecular spectra. Phys Rev Lett 85:960–963
27. Faure F, Zhilinskií BI (2001) Topological properties of the Born–Oppenheimer approximation and implications for the exact spectrum. Lett Math Phys 55:219–238
28. Faure F, Zhilinskií BI (2002) Topologically coupled energy bands in molecules. Phys Lett A 302:242–252
29. Flach S, Willis CR (1998) Discrete breathers. Phys Rep 295: 181–264
30. Giacobbe A, Cushman RH, Sadovskií DA, Zhilinskií BI (2004) Monodromy of the quantum 1 : 1 : 2 resonant swing spring. J Math Phys 45:5076–5100
31. Gilmore R (1981) Catastrophe theory for scientists and engineers. Wiley, New York
32. Gilmore R, Kais S, Levine RD (1986) Quantum cusp. Phys Rev A 34:2442–2452
33. Golubitsky M, Schaeffer DG (1984) Singularities and groups in bifurcation theory, vol 1. Springer, Berlin
34. Grondin L, Sadovskií DA, Zhilinskií BI (2002) Monodromy in systems with coupled angular momenta and rearrangement of bands in quantum spectra. Phys Rev A 142:012105-1–15
35. Guillemin V (1994) Moment maps and combinatorial invariants of Hamiltonian T^n-spaces. Birkhäuser, Basel
36. Harter W (1988) Computer graphical and semiclassical approaches to molecular rotations and vibrations. Comput Phys Rep 8:319–394
37. Hines AP, McKenzie RH, Milburn GJ (2005) Quantum entanglement and fixed-point bifurcations. Phys Rev A 71:042303-1–9
38. Hou X-W, Chen J-H, Hu B (2005) Entanglement and bifurcation in the integrable dimer. Phys Rev A 71:034302-1–4
39. Joyeux M, Farantos SC, Schinke R (2002) Highly excited motion in molecules: Saddle-node bifurcations and their fingerprints in vibrational spectra. J Phys Chem A 106:5407–5421
40. Joyeux M, Grebenschikov S, Bredenbeck J, Schinke R, Farantos SC (2005) Intramolecular dynamics along isomerization and dissociation pathways. In: Toda M, Komatsuzaki T, Konishi T, Berry RS, Rice SA (eds) Geometric Structures of Phase Space in Multidimensional Chaos: A Special Volume of Advances in Chemical Physics, part A, vol 130. Wiley, pp 267–303
41. Joyeux M, Sadovskií DA, Tennyson J (2003) Monodromy of the LiNC/NCLi molecule. Chem Phys Lett 382:439–442
42. Joyeux M, Sugny D, Tyng V, Kellman ME, Ishikawa H, Field RW, Beck C, Schinke R (2000) Semiclassical study of the isomerization states of HCP. J Chem Phys 112:4162–4172
43. Kellman ME (1995) Algebraic models in spectroscopy. Annu Rev Phys Chem 46:395–422
44. Kellman ME, Lynch ED (1986) Fermi resonance phase space structure from experimental spectra. J Chem Phys 85: 7216–7223
45. Kleman M (1983) Points, lines and walls. Wiley, Chichester
46. Kozin IN, Sadovskií DA, Zhilinskií BI (2005) Assigning vibrational polyads using relative equilibria: Application to ozone. Spectrochim Acta A 61:2867–2885
47. Landau L, Lifschits EM (1981) Quantum mechanics, nonrelativistic theory. Elsevier, Amsterdam
48. Lu Z-M, Kellman ME (1997) Phase space structure of triatomic molecules. J Chem Phys 107:1–15
49. Marsden JE, Ratiu TS (1994) Introduction to mechanics and symmetry. Springer, New York
50. Mermin ND (1979) The topological theory of defects in ordered media. Rev Mod Phys 51:591–648
51. Michel L (1980) Symmetry defects and broken symmetry, configurations, hidden symmetry. Rev Mod Phys 52:617–651
52. Michel L, Zhilinskií BI (2001) Symmetry, invariants, topology, vol I. Basic tools. Phys Rep 341:11–84
53. Michel L, Zhilinskií BI (2001) Symmetry, invariants, topology, vol III. Rydberg states of atoms and molecules. Basic group theoretical and topological analysis. Phys Rep 341:173–264
54. Montaldi J, Roberts R, Stewart I (1988) Periodic solutions near equilibria of symmetric Hamiltonian systems. Philos Trans Roy Soc Lond A 325:237–293
55. Morse M (1925) Relation between the critical points of a real function of n independent variables. Trans Am Math Soc 27:345–396
56. Nakahara M (1990) Geometry, topology and physics. IOP Publishing, Bristol
57. Nekhoroshev NN (1972) Action-angle variables and their generalizations. Trans Moscow Math Soc 26:180–198
58. Nekhoroshev NN, Sadovskií DA, Zhilinskií BI (2006) Fractional Hamiltonian monodromy. Ann Henri Poincaré 7:1099–1211
59. Pavlichenkov I (1993) Bifurcations in quantum rotational spectra. Phys Rep 226:173–279
60. Pavlichenkov I (2006) Quantum bifurcations and quantum phase transitions in rotational spectra. Phys At Nucl 69: 1008–1013
61. Pavlichenkov I, Zhilinskií BI (1988) Critical phenomena in rotational spectra. Ann Phys NY 184:1–32
62. Pavlov-Verevkin VB, Sadovskií DA, Zhilinskií BI (1988) On the dynamical meaning of the diabolic points. Europhys Lett 6:573–578
63. Peters AD, Jaffe C, Gao J, Delos JB (1997) Quantum manifestations of bifurcations of closed orbits in the photodetachment cross section of H^- in parallel fields. Phys Rev A 56:345–355
64. Pierre G, Sadovskií DA, Zhilinskií BI (1989) Organization of quantum bifurcations: Crossover of rovibrational bands in spherical top molecules. Europhys Lett 10:409–414

65. Sachdev S (1999) Quantum phase transitions. Cambridge University Press, Cambridge
66. Sadovskií DA, Fulton NG, Henderson JR, Tennyson J, Zhilinskií BI (1993) Nonlinear normal modes and local bending vibrations of H3+ and D3+. J Chem Phys 99(2):906–918
67. Sadovskií DA, Zhilinskií BI (1993) Group theoretical and topological analysis of localized vibration-rotation states. Phys Rev A 47(4):2653–2671
68. Sadovskií DA, Zhilinskií BI (1999) Monodromy, diabolic points, and angular momentum coupling. Phys Lett A 256:235–244
69. Sadovskií DA, Zhilinskií BI (2006) Quantum monodromy, its generalizations and molecular manifestations. Mol Phys 104:2595–2615
70. Sadovskií DA, Zhilinskií BI (2007) Hamiltonian systems with detuned 1 : 1 : 2 resonance, manifestations of bidromy. Ann Phys NY 322:164–200
71. Sadovskií DA, Zhilinskií BI, Champion JP, Pierre G (1990) Manifestation of bifurcations and diabolic points in molecular energy spectra. J Chem Phys 92:1523–1537
72. Sadovskií DA, Zhilinskií BI, Michel L (1996) Collapse of the Zeeman structure of the hydrogen atom in an external electric field. Phys Rev A 53:4064–4047
73. Simon B (1980) The classical limit of quantum partition functions. Commun Math Phys 71:247–276
74. Somma R, Ortiz G, Barnum H, Knill E, Viola L (2004) Nature and measure of entanglement in quantum phase transitions. Phys Rev A 70:042311-1–21
75. Symington M (2003) Four dimensions from two in symplectic topology. In: Athens GA (ed) Topology and geometry of manifolds. Proc Symp Pure Math, vol 71. AMS, Providence, pp 153–208
76. Tyng V, Kellman ME (2006) Bending dynamics of acetylene: New modes born in bifurcations of normal modes. J Phys Chem B 119:18859–18871
77. Uwano Y (1999) A quantum saddle-node bifurcation in a resonant perturbed oscillator with four parameters. Rep Math Phys 44:267–274
78. Uzer T (1990) Zeeman effect as an asymmetric top. Phys Rev A 42:5787–5790
79. Van der Meer JC (1985) The Hamiltonian Hopf bifurcation. Lect Notes Math, vol 1160. Springer, New York
80. Vũ Ngoc S (1999) Quantum monodromy in integrable systems. Comm Math Phys 203:465–479
81. Vũ Ngoc S (2007) Moment polytopes for symplectic manifolds with monodromy. Adv Math 208:909–934
82. Waalkens H, Dullin HR (2001) Quantum monodromy in prolate ellipsoidal billiards. Ann Phys NY 295:81–112
83. Wang J, Kais S (2004) Scaling of entanglement at a quantum phase transition for a two-dimensional array of quantum dots. Phys Rev A 70:022301-1-4
84. Weyl H (1952) Symmetry. Princeton University Press, Princeton
85. Winnewisser M, Winnewisser B, Medvedev I, De Lucia FC, Ross SC, Bates LM (2006) The hidden kernel of molecular quasi-linearity: Quantum monodromy. J Mol Struct V 798:1–26
86. Xiao L, Kellman ME (1990) Catastrophe map classification of the generalized normal-local transition in Fermi resonance spectra. J Chem Phys 93:5805–5820
87. Zhang W-M, Feng DH, Gilmore R (1990) Coherent states: Theory and some applications. Rev Mod Phys 62:867–927
88. Zhilinskií BI (1996) Topological and symmetry features of intramolecular dynamics through high resolution molecular spectroscopy. Spectrochim Acta A 52:881–900
89. Zhilinskií BI (2001) Symmetry, invariants, and topology, vol II. Symmetry, invariants, and topology in molecular models. Phys Rep 341:85–171
90. Zhilinskií BI (2006) Hamiltonian monodromy as lattice defect. In: Monastyrsky M (ed) Topology in condensed matter. Springer series in solid state sciences, vol 150. Springer, Berlin, pp 165–186
91. Zhilinskií BI, Kozin I, Petrov S (1999) Correlation between asymmetric and spherical top: Imperfect quantum bifurcations. Spectrochim Acta A 55:1471–1484
92. Zhilinskií BI, Pavlichenkov IM (1988) Critical phenomenon in the rotational spectra of water molecule. Opt Spectrosc 64:688–690
93. Zhilinskií BI, Petrov SV (1996) Nonlocal bifurcation in the rotational dynamics of an isotope-substituted $A_2A_2^*$ molecule. Opt Spectrosc 81:672–676

Quantum Cellular Automata

Karoline Wiesner[1,2]
[1] School of Mathematics, University of Bristol, Bristol, UK
[2] Centre for Complexity Sciences, University of Bristol, Bristol, UK

Article Outline

Glossary

Configuration The state of all cells at a given point in time.

Neighborhood All cells with respect to a given cell that can affect this cell's state at the next time step. A neighborhood always contains a finite number of cells.

Space-homogeneous The transition function / update table is the same for each cell.

Time-homogeneous The transition function / update table is time-independent.

Update table Takes the current state of a cell and its neighborhood as an argument and returns the cell's state at the next time step.

Schrödinger picture Time evolution is represented by a quantum state evolving in time according to a time-independent unitary operator acting on it.

Heisenberg picture Time evolution is represented by observables (elements of an operator algebra) evolving in time according to a unitary operator acting on them.

BQP complexity class *Bounded error, quantum probabilistic*, the class of decision problems solvable by a quantum computer in polynomial time with an error probability of at most 1/3.

QMA complexity class *Quantum Merlin–Arthur*, the class of decision problems such that a "yes" answer can be verified by a 1-message quantum interactive proof (verifiable in BQP).

Quantum Turing machine A quantum version of a Turing machine – an abstract computational model able to compute any computable sequence.

Swap operation The one-qubit unitary gate $U = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$

Hadamard gate The one-qubit unitary gate

$$U = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

Phase gate The one-qubit unitary gate $U = \begin{pmatrix} 1 & 0 \\ 0 & e^{i\phi} \end{pmatrix}$

Pauli operator The three Pauli operators are

$$\sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

Qubit 2-state quantum system, representable as vector $a|0\rangle + b|1\rangle$ in complex space with $a^2 + b^2 = 1$.

Definition of the Subject

Quantum cellular automata (QCA) are a generalization of (classical) cellular automata (CA) and in particular of reversible CA. The latter are reviewed shortly. An overview is given over early attempts by various authors to define one-dimensional QCA. These turned out to have serious shortcomings which are discussed as well. Various proposals subsequently put forward by a number of authors for a general definition of one- and higher-dimensional QCA are reviewed and their properties such as universality and reversibility are discussed.

Quantum cellular automata (QCA) are a quantization of classical cellular automata (CA), d-dimensional arrays of cells with a finite-dimensional state space and a local, spatially-homogeneous, discrete-time update rule. For QCA each cell is a finite-dimensional quantum system and the update rule is unitary. CA as well as some versions of QCA have been shown to be computationally universal. Apart from a theoretical interest in a quantized version of CA, QCA are a natural framework for what is most likely going to be the first application of quantum computers – the simulation of quantum physical systems. In particular, QCA are capable of simulating quantum dynamical systems whose dynamics are uncomputable by classical means. QCA are now considered one of the standard models of quantum computation next to quantum circuits and various types of measurement-based quantum computational models[1]. Unlike their classical counterpart, an axiomatic, all-encompassing definition of (higher-dimensional) QCA is still missing.

Introduction

Automata theory is the study of abstract computing devices and the class of functions they can perform on their inputs. The original concept of cellular automata is most strongly associated with John von Neumann (*1903, †1957), a Hungarian mathematician who made major contributions to a vast range of fields including quantum mechanics, computer science, functional analysis and many others. According to Burks, an assistant of von Neumann, [45] von Neumann had posed the fundamental questions: "What kind of logical organization is sufficient for an automaton to reproduce itself?". It was Stanislav Ulam who suggested to use the framework of cellular automata to answer this question. In 1966 von Neumann presented a detailed analysis of the above question in his book *Theory of Self-Reproducing Automata* [45].

Thus, von Neumann initiated the field of cellular automata. He also made central contributions to the mathematical foundations of quantum mechanics and, in a sense von Neumann's quantum logic ideas were an early attempt at defining a computational model of physics. But he did not pursue this, and did not go in the directions that have led to modern ideas of quantum computing in general or quantum cellular automata in particular.

The idea of quantum computation is generally attributed to Feynman who, in his now famous lecture in 1981, proposed a computational scheme based on quantum mechanical laws [19]. A contemporary paper by Benioff contains the first proposal of a quantum Turing machine [6]. The general idea was to devise a computa-

[1] For details on these and other aspects of quantum computation see the article by Kendon in this Encyclopedia.

tional device based on and exploiting quantum phenomena that would outperform any classical computational device. These first proposals were sequentially operating quantum mechanical machines imitating the logical operations of classical digital computation. The idea of parallelizing the operations was found in classical cellular automata. However, how to translate cellular automata into a quantum mechanical framework turned out not to be trivial. And to a certain extent how to do this in general remains an open question until today.

The study of quantum cellular automata (QCA) started with the work of Grössing and Zeilinger who coined the term QCA and provided a first definition [21]. Watrous developed a different model of QCA [46]. His work lead to further studies by several groups [16,17,18]. Independently of this, Margolus developed a parallelizable quantum computational architecture building on Feynman's original ideas [26]. For various reasons to be discussed below, none of these early proposals turned out to be physical. The study of QCA gained new momentum with the work by Richter, Schumacher, and Werner [36,37] and others [3,4,34] who avoided unphysical behavior allowed by the early proposals [4,37]. It is important to notice that in spite of the over two-decade long history of QCA there is no single agreed-upon definition of QCA, in particular of higher-dimensional QCA. Nevertheless, many useful properties have been shown for the various models. Most importantly, quite a few models were shown to be computationally universal, i. e. they can simulate any quantum Turing machine and any quantum circuit efficiently [16,34,35,38,46]. Very recently, their ability to generate and transport entanglement has been illustrated [14].

A comment is in order on a class of models which is often labeled as QCA but in fact are classical cellular automata implemented in quantum mechanical structures. They do not exploit quantum effects for the actual computation. To make this distinction clear they are now called *quantum-dot QCA*. These types of QCA will not be discussed here.

Cellular Automata

Definition (Cellular Automata) A *cellular automaton* (CA) is a 4-tuple $(\mathrm{L}, \Sigma, \mathcal{N}, f)$ consisting of (1) a d-dimensional lattice of cells L indexed $i \in \mathbb{Z}^d$, (2) a finite set of states Σ, (3) a finite neighborhood scheme $\mathcal{N} \subset \mathbb{Z}^d$, and (4) a local transition function $f\colon \Sigma^{\mathcal{N}} \to \Sigma$.

A CA is discrete in time and space. It is *space and time homogeneous* if at each time step the same transition function, or *update rule*, is applied simultaneously to all cells. The update rule is *local* if for a given lattice L and lattice site x, $f(x)$ is localized in $x+\mathcal{N} = \{x+n | x \in \mathrm{L}, n \in \mathcal{N}\}$, where $\mathcal{N}$ is the *neighborhood scheme* of the CA. In addition to the locality constraint the local transition function f must generate a unique global transition function mapping a lattice *configuration* $C_t \in \Sigma^{\mathrm{L}}$ at time t to a new configuration C_{t+1} at time $t+1$: $F\colon \Sigma^{\mathrm{L}} \to \Sigma^{\mathrm{L}}$. Most CA are defined on infinite lattices or, alternatively, on finite lattices with periodic boundary conditions. For finite CA only a finite number of cells is not in a *quiescent* state, i. e. a state that is not effected by the update.

Quantum Cellular Automata, Table 1
Update table for CA rule '110' (the second row is the decimal number '110' in binary notation)

	111	110	101	100	011	010	001	000
$M^{110} =$	0	1	1	0	1	1	1	0

The most studied CA are the so-called *elementary CA* – 1-dimensional lattices with a set of two states and a neighborhood scheme of *radius* 1 (*nearest-neighbor interaction*). i. e. the state of a cell at point x at time $t+1$ only depends on the state of cells $x-1, x$, and $x+1$ at time t. There are 256 such elementary CA, easily enumerated using a scheme invented by Wolfram [48]. As an example and for later reference, the update table of *rule 110* is given in Table 1. CA with update rule '110' have been shown to be computationally universal, i. e. they can simulate any Turing machine in polynomial time [15].

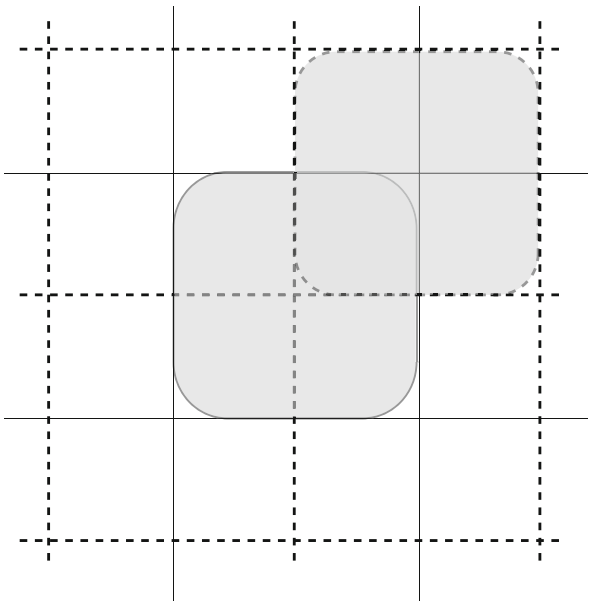
A possible approach to constructing a QCA would be to simply "quantize" a CA by rendering the update rule unitary. There are two problems with this approach. One is, that applying the same unitary to each cell does not yield a well-defined global transition function nor necessarily a unitary one. The second problem is the synchronous update of all cells. "In practice", the synchronous update of, say, an elementary CA can be achieved by storing the current configuration in a temporary register, then update all cells with odd index in the original CA, update all cells with even index in the register and finally splice the updated cells together to obtain the original CA at the next time step. Quantum states, however, cannot be copied in general due to the so-called *no-cloning theorem* [49]. Thus, parallel update of a QCA in this way is not possible. Sequential update on the other hand leads to either an infinite number of time steps for each update or inconsistencies at the boundaries. One solution is a partitioning scheme as it is used in the construction of reversible CA.

Reversible Cellular Automata

Definition (Reversible CA) A CA is said to be *reversible* if for every current configuration there is exactly one previous configuration.

The global transition function F of a reversible CA is *bijective*. In general, CA are not reversible. Only 16 out of the 256 elementary CA rules are reversible. However, one can construct a reversible CA using a partitioning scheme developed by Toffoli and Margolus for 2-dimensional CA [40].

Consider a 2-dimensional CA with nearest neighborhood scheme $\mathcal{N} = \{x \in \mathbb{Z}^2 | \forall |x_i| \leq 1\}$. In the *partitioning scheme* introduced by Toffoli and Margolus each block of 2×2 cells forms a unit cube $\square$ such that the even translates $\square + 2x$ with $x \in \mathbb{Z}^2$ and the odd translates $\square + \mathbf{1} + 2x$, respectively, form a *partition* of the lattice, see Fig. 1. The update rule of a partitioned CA takes as input an entire block of cells and outputs the updated state of the entire block. The rule is then applied alternatingly to the even and to the odd translates. The Margolus partitioning scheme is easily extended to d-dimensional lattices. A *generalized Margolus scheme* was introduced by Schumacher and Werner [37]. It allows for different cell sizes in the intermediate step.



Quantum Cellular Automata, Figure 1
Even (*solid lines*) and odd (*dashed lines*) of a Margolus partitioning scheme in $d = 2$ dimensions using blocks of size 2×2. For each partition one block is shown shaded. Update rules are applied alternatingly to the *solid* and *dashed* partition

A *partitioned CA* is then a CA with a partitioning scheme such that the set of cells are partitioned in some periodic way: Every cell belongs to exactly one block, and any two blocks are connected by a lattice translation. Such a CA is neither time homogeneous nor space homogeneous anymore, but periodic in time and space. As long as the rule for evolving each block is reversible, the entire automaton will be reversible.

Early Proposals

Grössing and Zeilinger were the first to coin the term and formalize a QCA [21]. In the Schrödinger picture of quantum mechanics the state of a system at some time t is described by a state vector $|\psi_t\rangle$ in Hilbert space $\mathcal{H}$. The state vector evolves unitarily,

$$|\psi_{t+1}\rangle = U|\psi_t\rangle \,. \tag{1}$$

U is a unitary operator, i. e. $UU^\dagger = \mathbf{1}$, with the complex conjugate $U^\dagger$ and the identity matrix $\mathbf{1}$. If $\{|\phi_i\rangle\}$ is a computational basis of the Hilbert space $\mathcal{H}$ any state $|\psi\rangle \in \mathcal{H}$ can be written as a superposition $\sum_{|\phi_i\rangle} c_i |\phi_i\rangle$, with coefficients $c_i \in \mathbb{C}$ and $\sum_i c_i c_i^* = 1$. The QCA constructed by Grössing and Zeilinger is an infinite 1-dimensional lattice where at time t lattice site i is assigned the complex amplitude c_i of state $|\psi_t\rangle$. The update rule is given by unitary operator U.

Definition (Grössing–Zeilinger QCA) A *Grössing–Zeilinger QCA* is a 3-tuple $(\mathrm{L}, \mathcal{H}, U)$ which consists of (1) an infinite 1-dimensional lattice $\mathrm{L} \subseteq \mathbb{Z}$ representing basis states of (2) a Hilbert space $\mathcal{H}$ with basis set $\{|\phi_i\rangle\}$, and (3) a band-diagonal unitary operator U.

Band-diagonality of U corresponds to a locality condition. It turns out that there is no Grössing–Zeilinger QCA with nearest-neighbor interaction and nontrivial dynamics. In fact, later on, Meyer showed more generally that "*in one dimension there exists no nontrivial homogeneous, local, scalar QCA. More explicitly, every band r-diagonal unitary matrix U which commutes with the one-step translation matrix T is also a translation matrix T^k for some $k \in \mathbb{Z}$, times a phase*" [27].

Grössing and Zeilinger also introduced QCA where the unitarity constraint is relaxed to only approximate unitarity. After each update the configuration can be normalized which effectively causes non-local interactions.

The properties of Grössing–Zeilinger QCA were studied by Grössing and co-workers in some more detail in following years, see [20] and references therein. This pioneering definition of QCA, however, was not studied much further, mostly because the "non-local" behavior

renders the Grössing–Zeilinger definition non-physical. In addition, it has little in common with the concepts developed in quantum computation later on. The Grössing-Zeilinger definition really concerns what one would call today a quantum random walk (for further details see the review by Kempe [23]).

The first model of QCA researched in depth was that introduced by Watrous [46], whose ideas where further explored by van Dam [16], Dürr, LêThanh, and Santha [17,18], and Arrighi [2]. A Watrous-QCA is defined over an infinite 1-dimensional lattice, a finite set of states including a quiescent state. The transition function maps a neighborhood of cells to a single quantum state instantaneously and simultaneously.

Definition (Watrous-QCA) A *Watrous-QCA* is a 4-tuple $(\mathrm{L}, \Sigma, \mathcal{N}, f)$ which consists of (1) a 1-dimensional lattice $\mathrm{L} \subseteq \mathbb{Z}$, (2) a finite set of cell states Σ including a quiescent state ε, (3) a finite neighborhood scheme $\mathcal{N}$, and (4) a local transition function $f\colon \Sigma^{\mathcal{N}} \to \mathcal{H}_{\Sigma}$.

Here, $\mathcal{H}_{\Sigma}$ denotes the Hilbert space spanned by the cell states Σ. This model can be viewed as a direct quantization of a CA where the set of possible configurations of the CA is extended to include all linear superpositions of the classical cell configurations, and the local transition function now maps the cell configurations of a given neighborhood to a quantum state. One cell is labeled "accept" cell. The quiescent state is used to allow only a finite number of states to be active and renders the lattice effectively finite. This is crucial to avoid an infinite product of unitaries and, thus, to obtain a well-defined QCA.

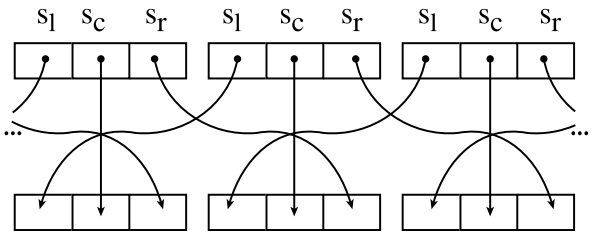
The Watrous QCA, however, allows for non-physical dynamics. It is possible to define transition functions that do not represent unitary evolution of the configuration, either by producing superpositions of configurations which do not preserve the norm, or by inducing a global transition function which is not unitary. This leads to non-physical properties such as super-luminal signaling [37]. The set of Watrous QCA is not closed under composition and inverse [37].

Watrous defined a restricted class of QCA by introducing a partitioning scheme.

Definition (Partitioned Watrous QCA) A *partitioned Watrous QCA* is a Watrous QCA with $\Sigma = \Sigma_l \times \Sigma_c \times \Sigma_r$ for finite sets Σ_l, Σ_c, and Σ_r, and matrix Λ of size $\Sigma \times \Sigma$. For any state $s = (s_l, s_c, s_r) \in \Sigma$ define transition function f as

$$f(s_1, s_2, s_3, s) = \Lambda_{(s_{l_3}, s_{m_2}, s_{r_1}, s)}\,, \tag{2}$$

with matrix element Λ_{s_i, s_j}.



Quantum Cellular Automata, Figure 2
Each cell is divided into three sub-cells labeled l, c, and r for *left*, *center*, and a *right*, respectively. The update rule consists of swapping *left* and *right* sub-cells of neighboring cells and then updating each cell internally using a unitary operation acting on the *left*, *center*, and *right* part of each cell

In a partitioned Watrous QCA each cell is divided into three sub-cells – left, center, and right. The neighborhood scheme is then a nearest-neighbor interaction confined to each cell. The transition function consists of a unitary acting on each partitioned cell and swap operations among sub-cells of different cells. Figure 2 illustrates the swap operation between neighboring cells.

For the class of partitioned Watrous QCA Watrous provides the first proof of computational universality of a QCA by showing that any quantum Turing machine can be efficiently simulated by a partitioned Watrous-QCA with constant slowdown and that any partitioned Watrous-QCA can be simulated by a quantum Turing machine with linear slowdown.

Theorem ([46]) *Given any quantum Turing machine M_{TM}, there exists a partitioned Watrous QCA M_{CA} which simulates M_{TM} with constant slowdown.*

Theorem ([46]) *Given any partitioned Watrous QCA M_{CA}, there exists a quantum Turing machine M_{TM} which simulates M_{CA} with linear slowdown.*

Watrous' model was further developed by van Dam [16], who defined a QCA as an assignment of a product vector to every basis state in the computational basis. Here the quiescent state is eliminated and thus the QCA is made explicitly finite. Van Dam showed that the finite version is also computationally universal. Efficient algorithms to decide whether a given 1-dimensional QCA is unitary was presented by Dürr, LeTanh, and Santha [17,18]. Due to substantial shortcomings such as non-physical behavior, these early proposals were replaced by a second wave of proposals to be discussed below.

Today, there is not a generally accepted QCA model that has all the attributes of the CA model: unique definition, simple to describe, and computationally powerful. In particular, there is no axiomatic definition, contrary to its

classical counterpart, that yields an immediate way of constructing/enumerating all of the instances of this model. Rather, each set of authors defines QCA in their own particular fashion.

The states $s \in \Sigma$ are basis states spanning a finite-dimensional Hilbert space. At each point in time a cell represents a finite-dimensional quantum system in a superposition of basis states. The unitary operators represent the discrete-time evolution of strictly finite propagation speed.

Models of QCA

Reversible QCA

Schumacher and Werner used the Heisenberg picture rather than the Schrödinger picture in their model [37]. Thus, instead of associating a d-level quantum system with each cell they associated an observable algebra with each cell. Taking a *quasi-local* algebra as the tensor product of observable algebras over a finite subset of cells, a QCA is then a homomorphism of the quasi-local algebra, which commutes with lattice translations and satisfies locality on the neighborhood.

The observable-based approach was first used in Ref. [36] with focus on the irreversible case. However, this definition left questions open such as whether the composition of two QCA will again form a QCA. The following definition does avoid this uncertainty.

Consider an infinite d-dimensional lattice $\mathrm{L} \subset \mathbb{Z}^d$ of cells $x \in \mathbb{Z}^d$, where each cell is associated with the observable algebra $\mathcal{A}_x$ and each of these algebras is an isomorphic copy of the algebra of complex $d \times d$-matrices. When $\Lambda \subset \mathbb{Z}^d$ is a finite subset of cells, denote by $\mathcal{A}(\Lambda)$ the algebra of observables belonging to all cells in Λ, i. e. the tensor product $\otimes_{x\in\Lambda}\mathcal{A}_x$. The completion of this algebra is called a *quasi-local* algebra and will be denoted by $\mathcal{A}(\mathbb{Z}^d)$.

Definition (Reversible QCA) A *quantum cellular automaton* with neighborhood scheme $\mathcal{N} \subset \mathbb{Z}^d$ is a homomorphism $T\colon \mathcal{A}(\mathbb{Z}^d) \to \mathcal{A}(\mathbb{Z}^d)$ of the quasi-local algebra, which commutes with lattice translations, and satisfies the locality condition $T(\mathcal{A}(\Lambda)) \subset T(\mathcal{A}(\Lambda + \mathcal{N}))$ for every finite set $\Lambda \subset \mathbb{Z}^d$. The local transition rule of a cellular automaton is the homomorphism $T_0\colon \mathcal{A}_0 \to \mathcal{A}(\mathcal{N})$.

Schumacher and Werner presented and proved the following theorem on one-dimensional QCA.

Theorem (Structure Theorem [37]) *Let T be the global transition homomorphism of a one-dimensional a nearest-neighbor QCA on the lattice $\mathbb{Z}^d$ with single-cell algebra $\mathcal{A}_0 = \mathcal{M}_d$. Then T can be represented in the generalized Margolus partitioning scheme, i. e. T restricts to an isomorphism*

$$T\colon \mathcal{A}(\square) \to \bigotimes_{s\in\Sigma} \mathcal{B}_s \,, \tag{3}$$

where for each quadrant vector $q \in Q$, the subalgebra $\mathcal{B}_q \subset \mathcal{A}(\square + q)$ is a full matrix algebra, $\mathcal{B}_q \mathcal{M}_{n(q)}$. These algebras and the matrix dimensions n(q) are uniquely determined by T.

Theorem (Structure Theorem [37]) does not hold in higher dimensions [47]. A central result obtained in this framework is that almost any [47] 1-dimensional QCA can be represented using a set of local unitary operators and a generalized Margolus partitioning [37], as illustrated in Fig. 3. Furthermore, if the local implementation allows local ancillas, then any QCA, in any lattice dimension can be built from local unitaries [37,47]. In addition, they proved the following Corollary.

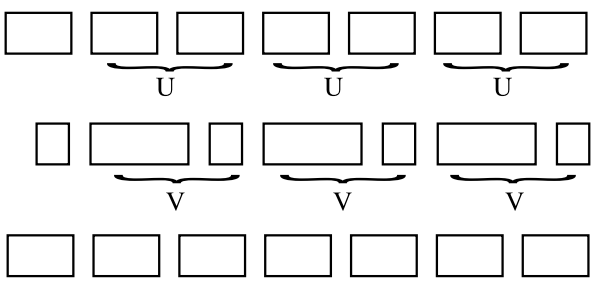
Corollary ([37]) *The inverse of a nearest-neighbor QCA exists, and is a nearest-neighbor QCA.*

The latter result is not true for CA. A similar result for finite configurations was obtained in [4]. Here evidence is presented that the result does not hold for two dimensional QCA. The work by Schumacher and Werner can be considered the first general definition for 1-dimensional QCA. A similar result for many-dimensional QCA does not exist.

Local Unitary QCA

Péres-Delgado and Cheung proposed a *local unitary QCA* [34].

Definition (Local-unitary QCA) A *local-unitary QCA* is a 5-tuple $\{(\mathrm{L}, \Sigma, \mathcal{N}, U_0, V_0)\}$ consisting of (1) a d-dimensional lattice of cells indexed by integer tuples $\mathrm{L} \subset \mathbb{Z}^d$, (2) a finite set of orthogonal basis states Σ, (3) a finite



Quantum Cellular Automata, Figure 3
Generalized Margolus partitioning scheme in 1 dimension using two unitary operations *U* and *V*

neighborhood scheme $\mathcal{N} \subseteq \mathbb{Z}^d$, (4) a local read function $U_0 \colon (\mathcal{H}_\Sigma)^{\otimes \mathcal{N}} \to (\mathcal{H}_\Sigma)^{\otimes \mathcal{N}}$, and (5) a local update function $V_0 \colon \mathcal{H}_\Sigma \to \mathcal{H}_\Sigma$. The read operation carries the further restriction that any two lattice translations U_x and U_y must commute for all $x, y \in \mathrm{L}$.

The product VU is a valid local, unitary quantum operation. The resulting global update rule is well defined and space homogeneous. The set of states includes a *quiescent* state as well as an "ancillary" set of states/subspace which can store the result of the "read" operation. The initial state of a local-unitary QCA consists of identical k^d blocks of cells initialized in the same state. Local-unitary QCA are universal in the sense that for any arbitrary quantum circuit there is a local-unitary QCA which can simulate it. In addition any local-unitary QCA can be simulated efficiently using a family of quantum circuits [34]. Adding an additional memory register to each cell allows this class of QCA to model any reversible QCA of the Schumacher/Werner type discussed above.

Block-Partitioned and Nonunitary QCA

Brennen and Williams introduced a model of QCA which allows for unitary and nonunitary rules [14].

Definition (Block-partitioned QCA) A *Block-partitioned QCA* is a 4-tuple $\{\mathrm{L}, \Sigma, \mathcal{N}, M\}$ consisting of (1) a 1-dimensional lattice of n cells indexed $\mathrm{L} = 0, \ldots, n-1$, (2) a 2-dimensional state space Σ, (3) a neighborhood scheme $\mathcal{N}$, and (4) an update rule M applied over $\mathcal{N}$.

Given a system with nearest-neighbor interactions, the simplest unitary QCA rule has radius $r = 1$ describing a unitary operator applied over a three-cell neighborhood $j-1, j, j+1$:

$$M(u_{00}, u_{01}, u_{10}, u_{11}) = |00\rangle\langle 00| \otimes u_{00} + |01\rangle\langle 01| \otimes u_{01} + |10\rangle\langle 10| \otimes u_{10} + |11\rangle\langle 11| \otimes u_{11} \,, \quad (4)$$

where $|ab\rangle\langle ab| \otimes u_{ab}$ means update the qubit at site j with the unitary u_{ab} if the qubit at the site $j-1$ is in state $|a\rangle$ and the qubit at site $j+1$ is in state $|b\rangle$. M commutes with its own 2-site translation. Thus, a partitioning is introduced by updating simultaneously all even qubits with rule M before updating all odd qubits with rule M. Periodic boundaries are assumed. However, by addressability of the end qubits simulation of a block-partitioned QCA by a QCA with boundaries can be achieved.

Nonunitary update rules correspond to completely positive maps on the quantum states where the neighboring states act as the environment. Take a nearest-neighbor 1-dimensional Block-partitioned QCA. In the density operator formalism each quantum system ρ is given by the probability distribution $\rho = \sum_i p_i |\psi\rangle\langle\psi|$ over outer products of quantum states $|\psi\rangle$. A completely positive map $S(\rho)$ applied to state ρ is represented by a set of Krauss operators F_μ, which are positive operators that sum up to the identity $\sum_\mu F_\mu^\dagger F_\mu = \mathbf{1}$. The map $S_j^{ab}(\rho)$ acting on cell j conditioned on state a of the left neighbor and state b of the right neighbor can then be written as

$$S_j^{ab}(\rho) = |ab\rangle\langle ab| \otimes \sum_\mu F_\mu^{ab} \rho F_\mu^{ab\dagger} \otimes |ab\rangle\langle ab| \,. \quad (5)$$

As an example, the CA rule '110' can now be translated into an update rule for cell j in a block-partitioned nonunitary QCA:

$$F_1^j = |00\rangle\langle 00| \otimes \mathbf{1}^j + |10\rangle\langle 10| \otimes \mathbf{1}^j + |11\rangle\langle 11| \otimes \sigma_x^j + |01\rangle\langle 01| \otimes |1\rangle_{jj}\langle 1| \quad (6)$$

$$F_2^j = |01\rangle\langle 01| \otimes |1\rangle_{jj}\langle 0| \,, \quad (7)$$

where σ_x is the Pauli operator.

The implementation of such a block-partitioned nonunitary QCA is proposed in form of a lattice of even order constructed with an alternating array of two distinguishable species $ABABABAB\ldots$ that are globally addressable and interact via the Ising interaction. Update rules that generate and distribute entanglement were studied in this framework [14].

Continuous-Time QCA

Vollbrecht and Cirac initiated the study of continuous-time QCA [44]. They show that the computability of the ground state energy of a translationally invariant n-neighbor Hamiltonian was QMA-hard. Their QCA model is taken up by Nagaj and Wocjam [31] who used the term *Hamiltonian QCA*.

Definition (Hamiltonian QCA) A *Hamiltonian QCA* is a tuple $\{L, \Sigma = \Sigma_p \times \Sigma_d\}$ consisting of (1) a 1-dimensional lattice of length L, (2) a finite set of orthogonal basis states $\Sigma = \Sigma_p \times \Sigma_d$ containing (2a) a data register Σ_d, and (2b) a program register Σ_p.

The initial state encodes both the program and the data, stored in separate subspaces of the state space:

$$|\phi\rangle = \bigotimes_{j=1}^{L} \left(|p_j\rangle \otimes |d_j\rangle\right)_j \quad (8)$$

The computation is carried out autonomously. Nagaj and Wocjam showed that, if the system is left alone for

a period of time $t = O(L \log L)$, polynomially in the length of the chain, the result of the computation is obtained with probability $p \geq 5/6 - O(1/\log L)$. Hamiltonian QCA are computationally universal, more precisely they are in the complexity class BQP. Two constructions for Hamiltonian QCA are given in [31], one using a 10-dimensional state space, and the resulting system can be thought of as the diffusion of a system of free fermions. The second construction given uses a 20-dimensional state space and can be thought of as a quantum walk on a line.

Examples of QCA

Raussendorf proved an explicit construction of QCA and proves its computational universality [35]. The QCA lives on a torus with a 2×2 Margolus partitioning. The update rule is given by a single 4-qubit unitary acting on 2×2 blocks of qubits. The four-qubit unitary operation consists of swap operations, the Hadamard transformation, and a phase gate. The initial state of the QCA is prepared such that columns encode alternatingly data and program. When the QCA is running the data travel in one direction while the program (encoding classical information in orthogonal states) travels in the opposite direction. Where the two cross the computation is carried out through nearest-neighbor interaction. After a fixed number of steps the computation is done and the result can be read out of a dedicated "data" column. This QCA is computationally universal, more precisely, it is within a constant as efficient as a quantum logic network with local and nearest-neighbor gates.

Shepherd, Franz, and Werner compared *classically controlled* QCA to autonomous QCA [38]. The former is controlled by a classical compiler that selects a sequence of operations acting on the QCA at each time step. The latter operates autonomously, performing the same unitary operation at each time step. The only step that is classically controlled is the measurement (and initialization). They show the computational equivalence of the two models. Their result implies that a particular quantum simulator may be as powerful as a general one.

Computationally Universal QCA

Quite a few models have been shown to be computationally universal, i. e. they can simulate any quantum Turing machine and any quantum circuit efficiently. A Watrous-QCA simulates any quantum Turing machine with constant slowdown [46]. The QCA defined by Van Dam is a finite version of a Watrous QCA and is computationally universal as well [16]. Local-unitary QCA can simulate any quantum circuit and thus are computationally universal [34]. Block-partitioned QCA can simulate a quantum computer with linear overhead in time and space [14]. Continuous-time QCA are in complexity class BQP and thus computationally universal [44]. The explicit constructions of 2-dimensional QCA by Raussendorf is computationally universal, more precisely, it is within a constant as efficient as a quantum logic network with local and nearest-neighbor gates [35]. Shepherd, Franz, and Werner provided an explicit construction of a 12-state 1-dimensional QCA which is in complexity class BQP. It is universally programmable in the sense that it simulates any quantum-gate circuit with polynomial overhead [38]. Arrighi and Fargetton proposed a 1-dimensional QCA capable of simulating any other 1-dimensional QCA with linear overhead [3].

Implementations of computationally universal QCA have been suggested by Lloyd [24] and Benjamin [8].

Modeling Physical Systems

One of the goals in developing QCA is to create a useful modeling tool for physical systems. Physical systems that can be simulated with QCA include Ising and Heisenberg interaction spin chains, solid state NMR, and quantum lattice gases. Spin chains are perhaps the most obvious systems to model with QCA. The simple cases of such 1-dimensional lattices of spins are Hamiltonians which commute with their own lattice translations. Vollbrecht and Cirac showed that the computability of the ground state energy of a translationally invariant n-neighbor Hamiltonian is in complexity class QMA [44]. For simulating noncommuting Hamiltonians a block-wise update such as the Margolus partitioning has to be used (see Sect. "Reversible Cellular Automata"). Here the fact is used that any Hamiltonian can be expressed as the sum of two Hamiltonians, $H = H_a + H_b$. H_a and H_b can then, to a good approximation, be applied sequentially to yield the original Hamiltonian H, even if these do not commute. It has been shown that such 1-dimensional spin chains can be simulated efficiently on a classical computer [43]. It is not known, however, whether higher dimensional spin systems can be simulated efficiently classically.

Quantum Lattice Gas Automata

Any numerical evolution of a discretized partial differential equation can be interpreted as the evolution of some CA, using the framework of *lattice gas automata*. In the continuous time and space limit such a CA mimics the behavior of the partial differential equation. In quantum mechanical lattice gas automata (QLGA) the continuous limit on a set of so called quantum lattice Boltzman equation

recovers the Schrödinger equation [39]. The first formulation of a linear unitary CA was given in Ref. [10]. Meyer coined the term *quantum lattice gas automata* (QLGA) and demonstrated the equivalence of a QLGA and the evolution of a set of quantum lattice Boltzman equations [27,28]. Meyer [29], Boghosian and Taylor [12], and Love and Boghosian [25] explored the idea of using QLGA as a model for simulating physical systems. Algorithms for implementing QLGA on a quantum computer have been presented in [13,30,32].

Implementations

A large effort is being made in many laboratories around the world to implement a model of a quantum computer. So far all of them are confined to a very finite number of elements and are no way near to a quantum Turing machine (which in itself is a purely theoretical construct but can be approximated by a very large number of computational elements). One existing experimental set-up that is very promising for quantum information processing and that does not suffer from this "finiteness" are optical lattices (for a review, see [11]). They possess a translation symmetry which makes QCA a very suitable framework in which to study their computational power. Optical lattices are artificial crystals of light and consist of hundreds of thousands of microtraps. One or more neutral atoms can be trapped in each of the potential minima. If the potential minima are deep enough any tunneling between the traps is suppressed and each site contains the same amount of atoms. A quantum register – here in form of a so-called Mott insulator – has been created. The biggest challenge at the moment is to find a way to address the registers individually to implement quantum gates. For a QCA all that is needed is implementing the unitary operation(s) acting on the entire lattice simultaneously. The internal structure of the QCA guarantees the locality of the operations. This is a huge simplification compared to individual manipulation of the registers. Optical lattices are created routinely by superimposing two or three orthogonal standing waves generated from laser beams of a certain frequency. They are used to study Fermionic and Bosonic quantum gases, nonlinear quantum dynamics, strongly correlated quantum phases, to name a few.

A type of locally addressed architecture by global control was put forward by Lloyd [24]. In this scheme a 1-dimensional array is built out of three atomic species, periodically arranged as $\mathcal{ABCABCABC}$. Each species encodes a qubit and can be manipulated without affecting the other species. The operations on any species can be controlled by the states of the neighboring cells. The end-cells are used for readout, since they are the only individually addressable components. Lloyd showed that such a quantum architecture is universal. Benjamin investigated the minimum physical requirements for such a many-species implementation and found a similar architecture using only two types of species, again arranged periodically $\mathcal{ABABAB}$ [7,8,9]. By giving explicit sequences of operations implementing one-qubit and two-qubit (CNOT) operations Benjamin showed computational universality. But the reduction in spin resources comes with an increase in logical encoding into four spin sites with a buffer space of at least four empty spin sites between each logical qubit.

A continuation of this multi-species QCA architecture is found in the work by Twamley [42]. Twamley constructed a proposal for a QCA architecture based on Fullerene (C_{60}) molecules doped with atomic species ^{15}N and ^{31}P, respectively, arranged alternatingly in a one-dimensional array. Instead of electron spins which would be too sensitive to stray electric charges the quantum information is encoded in the nuclear spins. Twamley constructed sequences of pulses implementing Benjamin's scheme for one- and two-qubit operations. The weakest point of the proposal is the readout operation which is not well-defined.

A different scheme for implementing QCA was suggested by Tóth and Lent [41]. Their scheme is based on the technique of quantum-dot CA. The term quantum-dot CA is usually used for CA implementations in quantum dots (for classical computation). The authors, therefore, called their model a *coherent* quantum-dot CA. They illustrated the usage of an array of N quantum dots as an N-qubit quantum register. However, the set-up and the allowed operations allow for individual control of each cell. This coherent quantum-dot CA is more a hybrid of a quantum circuit with individual qubit control and a QCA with constant nearest-neighbor interaction. The main property of a QCA, operating under global control only, is not taken advantage of.

Future Directions

The field of QCA is developing rapidly. New definitions have appeared very recently. Since QCA are now considered to be one of the standard measurement-based models of quantum computation, further work on a consistent and sufficient definition of higher-dimensional QCA is to be expected. One proposal for such a "final" definition has been put forward in [4,5].

In the search for robust and easily implementable quantum computational architectures QCA are of considerable interest. The main strength of QCA is global con-

trol without the need to address cells individually (with the possible exception of the read-out operation). It has become clear that the global update of a QCA would be a way around practical issues related to the implementation of quantum registers and the difficulty of their individual manipulation.

More concretely, QCA provide a natural framework for describing quantum dynamical evolution of optical lattices, a field in which the experimental physics community has made huge progress in the last decade.

The main focus so far has been on reversible QCA. Irreversible QCA are closely related to measurement-based computation and remain to be explored further.

Bibliography

Primary Literature

1. Aoun B, Tarifi M (2004) Introduction to quantum cellular automata. http://arxiv.org/abs/quant-ph/0401123
2. Arrighi P (2006) Algebraic characterizations of unitary linear quantum cellular automata. In: Mathematical Foundations of Computer Science 2006. Lecture Notes in Computer Science, vol 4162. Springer, Berlin, pp 122–133
3. Arrighi P, Fargetton R (2007) Intrinsically universal one-dimensional quantum cellular automata. 0704.3961. http://arxiv.org/abs/0704.3961
4. Arrighi P, Nesme V, Werner R (2007) One-dimensional quantum cellular automata over finite, unbounded configurations. 0711.3517v1. http://arxiv.org/abs/0711.3517
5. Arrighi P, Nesme V, Werner R (2007) N-dimensional quantum cellular automata. 0711.3975v1. http://arxiv.org/abs/arXiv:0711.3975
6. Benioff P (1980) The computer as a physical system: A microscopic quantum mechanical hamiltonian model of computers as represented by turing machines. J Stat Phys 22:563–591
7. Benjamin SC (2000) Schemes for parallel quantum computation without local control of qubits. Phys Rev A 61:020301–4
8. Benjamin SC (2001) Quantum computing without local control of qubit-qubit interactions. Phys Rev Lett 88(1):017904
9. Benjamin SC, Bose S (2004) Quantum computing in arrays coupled by "always-on" interactions. Phys Rev A 70:032314
10. Bialynicki-Birula I (1994) Weyl, Dirac, and Maxwell equations on a lattice as unitary cellular automata. Phys Rev D 49:6920
11. Bloch I (2005) Ultracold quantum gases in optical lattices. Nature Phys 1:23–30
12. Boghosian BM, Taylor W (1998) Quantum lattice-gas model for the many-particle Schrödinger equation in d dimensions. Phys Rev E 57:54
13. Boghosian BM, Taylor W (1998) Simulating quantum mechanics on a quantum computer. Physica D: Nonlinear Phenomena 120:30–42
14. Brennen GK, Williams JE (2003) Entanglement dynamics in one-dimensional quantum cellular automata. Phys Rev A 68:042311
15. Cook M (2004) Universality in elementary cellular automata. Complex Syst 15:1
16. van Dam W (1996) Quantum cellular automata. Master's thesis, University of Nijmegen
17. Dürr C, Santha M (2002) A decision procedure for unitary linear quantum cellular automata. SIAM J Comput 31:1076–1089
18. Dürr C, LêThanh H, Santha M (1997) A decision procedure for well-formed linear quantum cellular automata. Random Struct Algorithms 11:381–394
19. Feynman R (1982) Simulating physics with computers. Int J Theor Phys 21:467–488
20. Fussy S, Grössing G, Schwabl H, Scrinzi A (1993) Nonlocal computation in quantum cellular automata. Phys Rev A 48:3470
21. Grössing G, Zeilinger A (1988) Quantum cellular automata. Complex Syst 2:197–208
22. Gruska J (1999) Quantum Computing. Osborne/McGraw-Hill. QCA are treated in Section 4.3
23. Kempe J (2003) Quantum random walks: an introductory overview. Contemp Phys 44:307
24. Lloyd S (1993) A potentially realizable quantum computer. Science 261:1569–1571
25. Love P, Boghosian B (2005) From Dirac to diffusion: Decoherence in quantum lattice gases. Quantum Inf Process 4:335–354
26. Margolus N (1991) Parallel quantum computation. In: Zurek WH (ed) Complexity, Entropy, and the Physics of Information, Santa Fe Institute Series. Addison Wesley, Redwood City, pp 273–288
27. Meyer DA (1996) From quantum cellular automata to quantum lattice gases. J Stat Phys 85:551–574
28. Meyer DA (1996) On the absence of homogeneous scalar unitary cellular automata. Phys Lett A 223:337–340
29. Meyer DA (1997) Quantum mechanics of lattice gas automata: One-particle plane waves and potentials. Phys Rev E 55:5261
30. Meyer DA (2002) Quantum computing classical physics. Philos Trans Royal Soc A 360:395–405
31. Nagaj D, Wocjan P (2008) Hamiltonian quantum cellular automata in 1d. 0802.0886. http://arxiv.org/abs/0802.0886
32. Ortiz G, Gubernatis JE, Knill E, Laflamme R (2001) Quantum algorithms for fermionic simulations. Phys Rev A 64:022319
33. Perez-Delgado CA, Cheung D (2005) Models of quantum cellular automata. http://arxiv.org/abs/quant-ph/0508164
34. Perez-Delgado CA, Cheung D (2007) Local unitary quantum cellular automata. Phys Rev A (Atomic, Molecular, and Optical Physics) 76:032320–15
35. Raussendorf R (2005) Quantum cellular automaton for universal quantum computation. Phys Rev A (Atomic, Molecular, and Optical Physics) 72:022301–4
36. Richter W (1996) Ergodicity of quantum cellular automata. J Stat Phys 82:963–998
37. Schumacher B, Werner RF (2004) Reversible quantum cellular automata. quant-ph/0405174. http://arxiv.org/abs/quant-ph/0405174
38. Shepherd DJ, Franz T, Werner RF (2006) Universally programmable quantum cellular automaton. Phys Rev Lett 97:020502–4
39. Succi S, Benzi R (1993) Lattice Boltzmann equation for quantum mechanics. Physica D: Nonlinear Phenomena 69:327–332
40. Toffoli T, Margolus NH (1990) Invertible cellular automata: A review. Physica D: Nonlinear Phenomena 45:229–253
41. Tóth G, Lent CS (2001) Quantum computing with quantum-dot cellular automata. Phys Rev A 63:052315
42. Twamley J (2003) Quantum-cellular-automata quantum computing with endohedral fullerenes. Phys Rev A 67:052318

43. Vidal G (2004) Efficient simulation of one-dimensional quantum many-body systems. Phys Rev Lett 93(4):040502
44. Vollbrecht KGH, Cirac JI (2008) Quantum simulators, continuous-time automata, and translationally invariant systems. Phys Rev Lett 100:010501
45. von Neumann J (1966) Theory of Self-Reproducing Automata. University of Illinois Press, Champaign
46. Watrous J (1995) On one-dimensional quantum cellular automata. In: Proceedings of the 36th Annual Symposium on Foundations of Computer Science, Milwaukee, pp 528–537
47. Werner R Private communication
48. Wolfram S (1983) Statistical mechanics of cellular automata. Rev Mod Phys 55:601
49. Wootters WK, Zurek WH (1982) A single quantum cannot be cloned. Nature 299:802–803

Books and Reviews

Summaries of the topic of QCA can be found in chapter 4.3 of Gruska [21], and in Refs. [1,32].

Quantum Chaos

GIULIO CASATI[1], TOMAŽ PROSEN[2]
[1] Universita dell'Insubria, Como, Italy
[2] Univerza v Ljubljani, Ljubljana, Slovenia

Article Outline

Glossary

Ergodicity The property of a dynamical system, according to which a single trajectory, starting from almost any initial condition, explores (densely covers) the entire available phase space of physical states.

Integrability A classical Hamiltonian dynamical system of N degrees of freedom is said to be integrable (according to Liouville) if there exist N independent conserved quantities. Integrability implies explicit quasi-periodic solution of the equations of motion.

Random matrix theory The statistical theory which allows to describe the fluctuation properties of quantum systems in terms of the sets (ensembles) of random Hermitian matrices with appropriate invariant measures.

Wigner surmise Nearest neighbor energy level spacing distribution based on the simplest 2×2 Gaussian Hermitian random matrix models, accurately approximating spacing distributions in complex quantum systems.

Periodic orbit theory or trace formula A relationship between certain properties of energy spectrum of a quantized chaotic system, and the set of unstable periodic orbits of the corresponding classical chaotic system.

Quantum Loschmidt echo or fidelity A measure of stability of quantum time evolution. It is computed as a Hilbert space inner product of two slightly different quantum time evolutions starting from the same initial state.

Definition of the Subject

As it is now widely recognized, classical dynamical chaos has been one of the major scientific breakthroughs of the past century. Quantum chaos, sometimes called Quantum chaology, studies the manifestations of chaotic motion and related dynamical phenomena in quantum mechanics [1,2].

More abstractly, one may define as quantum chaos those phenomena of *simple* quantum systems which can be described statistically and exhibit some universal (i. e. system independent) features. By the term *simple* we mean here that the system can be specified by a finite set of parameters or, generally, can be described by a finite amount of information. So we can fundamentally distinguish the phenomena of quantum chaos from similar dynamical phenomena in *disordered* systems – specified in terms of random parameters and which therefore contain infinite amount of information in an appropriate (say thermodynamic) limit.

The universal statistical properties of quantum chaotic systems which have been widely studied include statistics of energy level spectra, statistical and semiclassical structures of the wave-functions and statistical distributions of transition matrix elements (matrix elements of certain physical observables in the energy eigenbasis). These properties are of key importance for understanding quantum state transitions, dissipation, ionization and related phenomena. Traditionally quantum chaos has been intimately connected to problems in atomic physics, nuclear physics, mesoscopic solid state physics, and more recently also to the emerging field of quantum information.

The subject of quantum chaos is at the core of the fundamental and general understanding of the correspondence principle according to which classical mechanics

emerges as a limit of quantum mechanics when an 'effective' value of the Planck's constant goes to zero. However, changing perspective, and treating quantum mechanics as a fundamental theory and classical mechanics as its convenient approximation, one can ask what happens to the phenomena of quantum chaos in systems which lack the classical limit, such as systems of spin 1/2 or quantum bits (qubits). Still, many of the successful statistical tools developed or widely used in the field of quantum chaos can be applied to understand the dynamical and statistical properties of interacting qubits.

Introduction

Classical Hamiltonian dynamical systems display a rich variety of behaviors [3]. At one end, we have *strongly chaotic* dynamical systems, which are distinguished by such properties as *algorithmic complexity*, *exponential sensitivity to initial conditions*, *continuous spectrum* and consequent decay of temporal correlations ('loss of memory'), relaxation to equilibrium, and *ergodicity*. Without entering a discussion of the above properties, we notice that rigorous examples of such strongly chaotic systems are billiard balls bouncing inside a table with inward curved boundaries, or point mass particles moving freely on any surfaces of constant *negative curvature*.

At the opposite end, there are *completely integrable* or *regular* dynamical systems, which, according to Liouville, are characterized by the existence of as many independent smooth constants of motion as there are degrees of freedom, and which are distinguished by analytic predictability of time evolutions and absence of algorithmic complexity. Consequently, integrable systems have a discrete spectrum of time evolution, do not display relaxation to equilibrium and, due to existence of non-trivial constants of motion, they are not ergodic.

Nowadays we know few examples of completely integrable systems, such as an arbitrary system of coupled linear (harmonic) oscillators, the general problem of two moving bodies interacting with a centrally symmetric force, and even many-body models such as the famous Toda lattice which is a system of equal point masses in one dimension interacting with exponentially distance dependent force. On the other hand we also know that a generic, or typical Hamiltonian system is not integrable, neither it is strongly chaotic in a rigorous sense. Instead we find a variety of intermediate behaviors between completely integrable and strongly chaotic.

A famous Kolmogorov–Arnold–Moser (KAM) theorem states that a small generic (smooth) perturbation of a completely integrable Hamiltonian systems preserves most of the features of regular motion, such as quasi-periodicity or discrete spectrum, for most of initial conditions. However, in the vicinity of the so-called resonant tori (i. e. for initial conditions for which the motion of the unperturbed system has commensurate frequencies) the motion becomes locally (weakly) chaotic even for arbitrary weak perturbations. Still, the relative overall phase space volume occupied by chaotic trajectories decreases to zero faster than any power of the perturbation strength.

However, KAM theory does not describe the only scenario of integrability breaking in Hamiltonian systems. Other types of perturbations, which do not obey the conditions of KAM theorem are possible which yield physically interesting behavior. One class of such behaviors is the motion in generic polygonal billiards [4] (namely billiard tables with the shape of a generic polygon, say a triangle), or the motion of any number of elastically colliding point masses in one dimension. Such systems are neither integrable nor chaotic in the sense of exponential sensitivity or algorithmic complexity.

The fundamental problem of quantum chaos concerns the manifestations in the quantum world of the various degrees of complexity of classical dynamics, as described by the above hierarchy. Primarily we are interested in *simple*, closed (isolated) and bounded systems for which, as it is known, the spectrum of the Hamilton operator (the generator of the Schrödinger equation) is always discrete. Notice that in classical ergodic theory, discrete spectrum is associated to integrability which is just the opposite of chaos which requires continuous spectrum. In such a situation quantum mechanics, at most, can follow chaotic classical dynamics only up to the so called *breaking* time scale t^* after which the quantum motion becomes fundamentally different from the classical one. Precise understanding of this transition phenomena, scaling of the time scale t^* with various physical properties, is at the heart of quantum chaos and shall be briefly discussed in the subsequent sections.

Quantum Chaos – Stationary Aspects

The problems of quantum chaos can be viewed from two different but closely connected view-points, namely the stationary aspects in the energy domain and the dynamical aspects in the time domain. Let us start by discussing the stationary aspects.

It has been recognized as early as in 1950's by Wigner, and later by Dyson, that many statistical features of energy levels of complex nuclei, or long lived resonance states, can be adequately described by a simple statistical model of random Hermitian matrices. Wigner's idea was that for

a sufficiently complicated quantum system with many degrees of freedom like a heavy nucleus, one assumes that the matrix elements of the Hamiltonian in a *typical* basis can be treated as independent Gaussian random numbers. Such random matrix model has essentially no free parameters and is invariant under almost arbitrary basis changes, and therefore one can obtain many explicit analytical results.

The simplest and perhaps the main result of random matrix theory [5] predicts that the statistical distribution of spacings between adjacent energy levels, denoted by $S_n = E_{n+1} - E_n$, properly scaled or normalized such that the average spacing S equals one, obeys universal distributions which only depend on certain symmetry properties. Roughly speaking, they depend on the existence of time-reversal invariance of the Hamiltonian, and are, to a high level of accuracy, given by a the so-called Wigner surmise

$$P_\beta(S) = AS^\beta \exp(-BS^2)$$

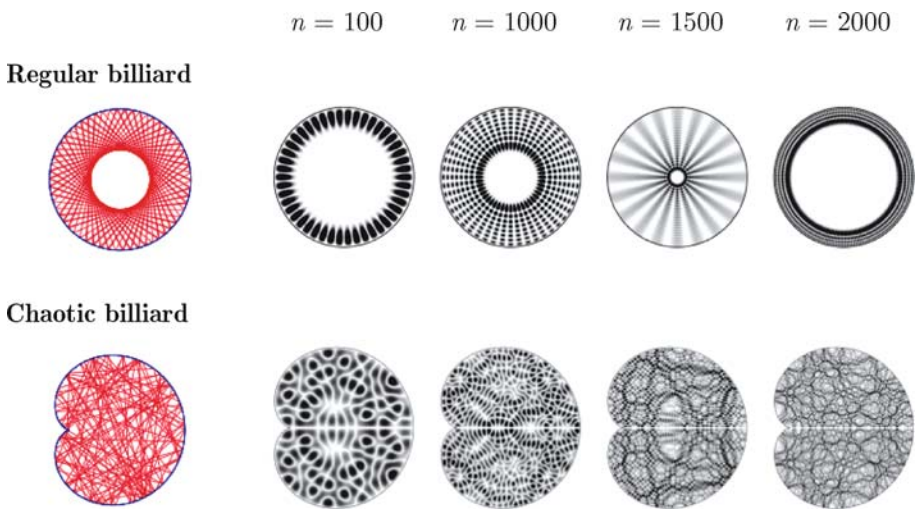
where the constants A and B are determined from normalization conditions, $\int_0^\infty P(S)\,\mathrm{d}S = 1$, $\int_0^\infty SP(S)\,\mathrm{d}S = 1$. The integer β is known as universality index and equals $\beta = 1$ for real Gaussian random matrices applying in the time-reversal invariant case, and $\beta = 2$ for complex Gaussian random matrices applying in the case where time-reversal invariance is broken. Note that β has also an interesting interpretation as a *level repulsion parameter* since $\beta > 0$ implies repulsive correlations between adjacent energy levels.

Of course, random matrix theory would only remain an interesting mathematical model if it would not have been so immensely successful in describing spectral correlations of complex quantum systems. As already mentioned it started with nuclear spectra for which one is ready to believe that due to immense complexity of interactions random matrix description is intuitively adequate. However in beginning of 1980's numerical evidence started to accumulate [6,7] that even spectra of very simple, but yet non-integrable and classically chaotic systems exhibit universal level fluctuations described by random matrix theory.

This observation, namely *that short range spectral correlations of quantum systems which are strongly chaotic in the classical limit obey universal fluctuation laws which are given by ensembles of random matrices without free parameters*, has been known as *quantum chaos conjecture* and, despite of still not being rigorously proven, remains one of the defining and most important results in the field.

One of the standard paradigmatic models where quantum chaos conjecture has been most often and convincingly demonstrated are quantum billiards. The stationary Schrödinger problem for a billiard of a point particle of mass m moving freely inside a planar domain (billiard table) $\mathcal{D}$ and colliding elastically off its boundary is simply the well known Helmholtz (amplitude) equation for the particle wavefunction

$$\nabla^2\Psi(x, y) + k^2\Psi(x, y) = 0\,,$$



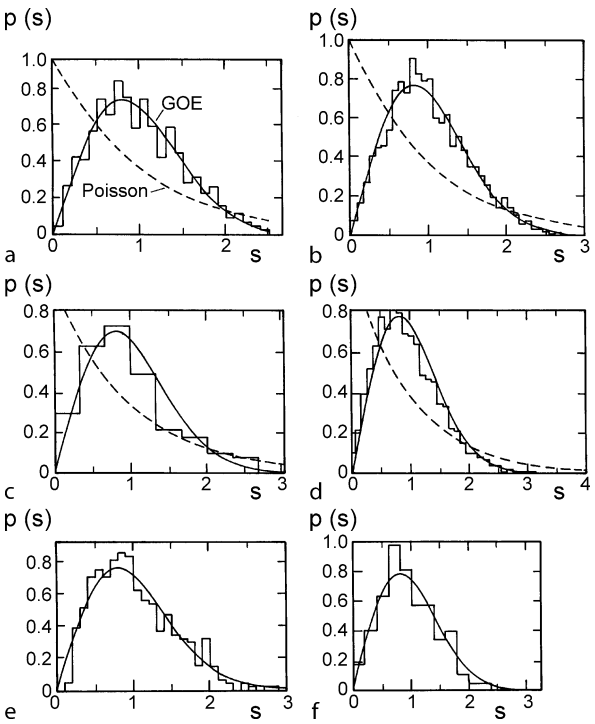
Quantum Chaos, Figure 1

Typical examples of a regular billiard (circular billiard, *above*) and a fully chaotic billiard (cardioid billiard, *below*). In the *left* side of the figure we show two typical trajectories in the billiard table, whereas to the *right* we depict few examples of wave-functions of eigenstates at different sequential level numbers *n*. Courtesy of Bäcker [9]

with Dirichlet boundary condition $\Psi|_{\partial\mathcal{D}} = 0$, having a discrete set of solutions $\{k_n, \Psi_n; n = 1, 2, \dots\}$ with a dispersion relation to the eigenenergies $E_n = \hbar^2 k_n^2/(2m)$. In Fig. 1 we show two examples of a typical integrable billiard (circular billiard), and a typical fully chaotic and ergodic billiard (cardioid billiard proposed by Robnik [8]) – plotting typical trajectories for each and a sequence of few typical chaotic and regular eigenstates.

Solving the Dirichlet Helmholtz problem for a finite domain is one of the most standard problems in linear wave physics and emerges in equivalent forms in gas acoustics, flat electromagnetic (microwave) resonators, transverse waves in optical fibers, etc, the only difference from the stationary quantum problem being the dispersion relation which connects the frequency of stationary waves to the eigenvalues of the wavenumber k_n.

In Fig. 2 one can observe that spectra of chaotic quantum billiard, hydrogen atom in strong magnetic field, excitation spectrum of NO_2 molecule, microwave electromagnetic spectrum of a three dimensional chaotic cavity and even spectra of vibrating elastic solids all exhibit spectral correlations which are given by a Gaussian ensemble of real random matrices.



Quantum Chaos, Figure 2

Level spacing distributions for a the fully chaotic Sinai billiard [7], b a Hydrogen atom in a strong magnetic field, c an NO_2 molecule, d a vibrating quartz block shaped like a three dimensional sinai billiard, e the microwave spectrum of a three-dimensional chaotic cavity, f a vibrating elastic disc shaped like a quarter stadium. Courtesy of Stöckmann [2]

The universality of quantum statistics of classically chaotic systems applies also to other properties, such as the distribution of a chaotic wave-function amplitudes, which is Gaussian and very well modeled by the so-called random plane wave superposition [10], the distribution of matrix elements of typical observables in the eigenbasis of chaotic Hamiltonians [11], which is again Gaussian, or even distribution of chaotic Wigner functions (quantum analogues of phase space densities) which is found again to be Gaussian [12].

An interesting related mathematical question is the problem of *quantum ergodicity*. For a generic physical observable A the question is whether its expectation value in a given eigenstate approaches, with increasing quantum level number, the microcanonical average of the corresponding classical observable evaluated at the corresponding energy, namely whether the sequence

$$\{\langle\Psi_n|A|\Psi_n\rangle - A_{\text{classical}}(E_n); n = 1, 2, 3, \dots\}$$

converges to zero? Mathematicians often relax the above condition to hold for a *subsequence of density 1* meaning that the above property holds for a typical eigenstate, whereas the sequence of exceptions (with semiclassically vanishing statistical weight) correspond to the famous *scar states* discovered by Heller [13]. For quantum billiards, the statement of quantum ergodicity is equivalent to the statement of uniform equidistribution of probability density for eigenstates (see e. g. lower panels of Fig. 1). It has been proven (the proof has been announced by Shnirelman [14], and later worked out by Zelditch [15] and Colin de Verrdiere [16]) that strongly chaotic billiards, and some other rigorous examples of strongly chaotic systems, are quantum ergodic. However, the minimal classical ergodic properties (like ergodicity, weak mixing, etc) sufficient for quantum ergodicity are still under debate.

One can now ask a similar questions for the other extreme of ergodic hierarchy, namely for classically regular systems: Are there some universal features of spectral fluctuations of quantum systems whose classical limit is completely regular? The answer is only partly affirmative, in the sense of an argument which has been originally given by Berry and Tabor [17]. The starting point of this argument is that the eigenenergies of a completely integrable system with d degrees of freedom can be labeled by a d-duple of quantum numbers – integers n_j – namely $E_{n_1 n_2 \dots n_d}$. Berry and Tabor argued that level sub-sequences, where $d - 1$ quantum numbers are fixed and only one of them

is being varied, are *mutually independent*. As the overall spectrum of an integrable system is a superposition of many such almost uncorrelated level sequences, one concludes that any short range level correlations should be absent. For example, the level spacing distribution for an integrable system should be the same as for a Poissonian distribution of uncorrelated events, namely

$$P_{\text{int}}(S) = \exp(-S) \, .$$

One has to note that for integrable systems one should not really expect a universality in the same sense as for chaotic systems. Berry and Tabor's argument cannot in general be turned into a rigorous proof, and for any particular integrable system one can in principle always find statistical measures which deviate from Poissonian predictions [18]. For example, a one dimensional harmonic oscillator has an equidistant (the so-called "picket fence") spectrum so its level spacing distribution can be written in terms of a Dirac's delta distribution $P_{\text{harm.osc.}} = \delta(S-1)$. In other words, as quantum statistical properties are concerned, there is no such thing as *a typical integrable system*. This can be viewed as another manifestation of quantum non-ergodicity of integrable systems.

Even if spectral fluctuations of integrable systems are relatively universal – in the sense as explained above – this is not at all true for other statistical properties of quantized integrable systems, such as wave-function *amplitudes*, sizes and numbers of *nodal domains* of wave functions, *matrix elements* of typical physical observables, etc. There we find system specific features which can hardly be described by some general statistical rules.

The general case of typical quantum systems, say those which can be understood as quantizations of systems with classically mixed phase space with coexisting regular and chaotic orbits, is the most difficult to describe. It is fair to say that quantum chaos of generic systems is still in its infancy. Among the few results which can be mentioned is the semiclassical theory of Berry and Robnik [19] describing level fluctuations of mixed systems as statistical superposition of independent level subsequences corresponding to each invariant classical phase space component: level subsequences corresponding to areas of chaotic motion are modeled by an appropriate ensemble of random matrices with the relative level density which is given by the classical volume of the chaotic component, and a (single) subsequence corresponding to all regular trajectories is modeled by a Poissonian level sequence of the corresponding overall level density. However, in realistic quantum systems *tunneling* between chaotic and regular states has to be taken into account and, in spite of few attempts, we are still lacking a general statistical theory of tunneling amplitudes in mixed phase space systems.

It should be noted that statistical properties of energy level fluctuations can be related to classical recurrent phase space structures such as periodic orbits. In particular, for chaotic dynamical systems, one finds a beautiful relationship between semiclassical approximation to the energy spectrum and isolated unstable classical periodic orbits, which is known as Gutzwiller's *trace formula* [20], and is based on stationary phase approximation to Feynman path integral representation of Green's function of the Schrödinger equation. A similar trace formula has been proposed by Berry and Tabor also for classically regular systems [21]. Related periodic or closed orbit theories have been later developed for describing semiclassical properties of other quantities, for example scars in chaotic wavefunctions [22,23]. The trace formula has been believed to be the theoretical tool to attack the proof of quantum chaos conjecture. Considerable recent progress has been achieved by Müller et al. [24], building upon an idea of Sieber [25], who have shown that the correct resummation of a trace formula indeed yields short time expansion of the universal *spectral form factor* (Fourier transformation of the two-point spectral correlation function) which is identical to the one obtained from random matrix theory.

Quantum Chaos – Dynamical Aspects

Classical chaos is a property of time evolution and, as we have seen, requires continuous spectrum of the motion. However, the spectrum of bounded closed quantum systems is always discrete, therefore genuine quantum chaos which would be a property of asymptotic time evolution is not possible, namely there is an ultimate breaking time scale t^* which is determined by the density of states $\rho = \mathrm{d}E_n/\mathrm{d}n$. Indeed, writing a completely general solution of the time dependent Schrödinger equation as

$$\Psi(t) = \sum_n c_n \exp(-\mathrm{i}E_n t/\hbar)\Psi_n$$

one sees that the quasi-periodic nature of time evolution shows up, on average, when the difference of two adjacent phases grows to 2π, namely

$$t^* = 2\pi\hbar\rho \, .$$

This breaking time scale is also sometimes referred to as Heisenberg time, and after t^* the time evolution in quantum mechanics is dominated by quantum fluctuations.

Thus genuine quantum chaos is possible only within a time scale $t < t^*$, where quantum phenomena with semiclassical description are possible, such as relaxation,

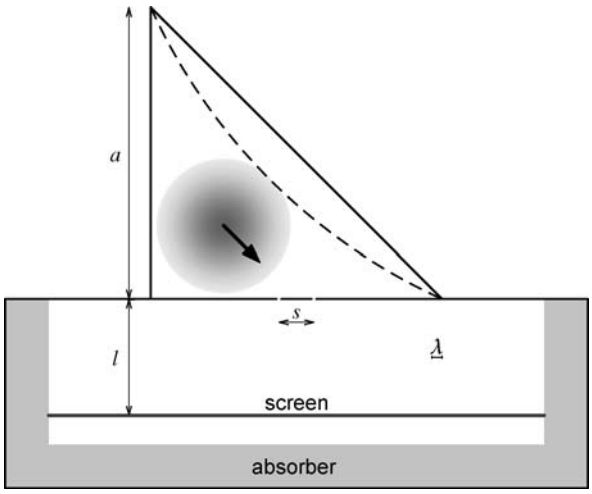
exponential sensitivity etc. However, not all chaotic phenomena can be observed in quantum dynamics up to t^*, some are much more short lived. For example, exponential sensitivity to initial conditions can be mimicked by quantum motion only up to the so-called Ehrenfest time

$$t_E = \frac{\ln(A_0/\hbar)}{\lambda}$$

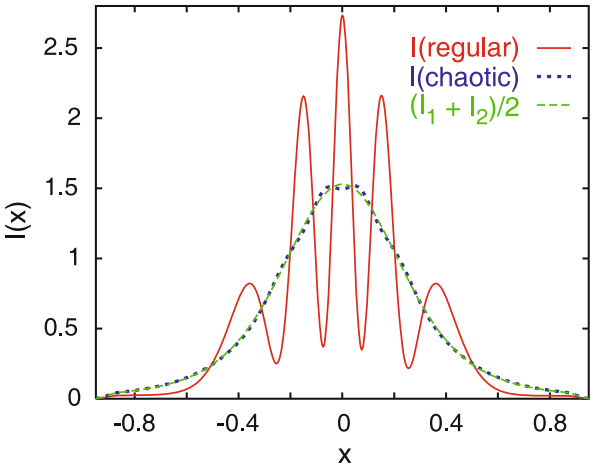
where A_0 is the phase space volume explored by a classical chaotic trajectory, and λ is the classical Lyapunov exponent measuring the exponential rate of divergence $\delta x(t) \sim \exp(\lambda t)\delta x(0)$ of nearby trajectories in classical dynamics. Up to time t_E Gaussian wave-packets centered on classical trajectories can be used for semi-classical description of quantum motion [26]. Ehrenfest time t_E and Heisenberg time t^* are two essential time-scales of quantum chaos.

A simple but fundamental manifestation of quantum chaos is the destruction of quantum interferences by quantum dephasing as a results of chaotic classical dynamics. To illustrate this, let us consider a time-dependent double slit experiment, where the source is closed inside a two dimensional wave resonator in the shape of a chaotic billiard [27]. The setting is depicted in Fig. 3. We take an initial Gaussian wave-packet with average energy corresponding to about the 1600th excited state of the closed billiard and direct it towards two narrow openings – slits – which are a few (about 3) De Broglie wavelengths apart. The wave-packet is taken to be as sharply as possible localized in momentum space, so that, due to Heisenberg uncertainty principle, its spreading in position space is of the order of the diameter of the billiard. Then we solve the time dependent Schrödinger equation and observe the radiation which is transmitted through the two slits to an infinite plane below. We then record the time integrated probability current density $I(x)$ as a function of the horizontal position x on the screen.

The results of such numerical experiment are shown in Fig. 4. We observe almost no sign of quantum interference if the shape of the resonator is chaotic. This is a manifestation of chaotic dynamics of classical rays and consequently phase randomization of multiply reflected waves impacting onto the slits. For comparison we show an analogous result for a regular geometry of the resonator which is represented by an integrable billiard. Here we find the well known interference fringes whose location and visibility is a well controlled function of the initial conditions of the wave-packet. In Fig. 5 we show instant snapshots of the probability density at around half the Heisenberg time for chaotic and regular geometry. The crucial difference is that the jets of probability coming out from the two slits have time-dependent direction in the chaotic case whereas these directions are frozen in time for the regular geometry.

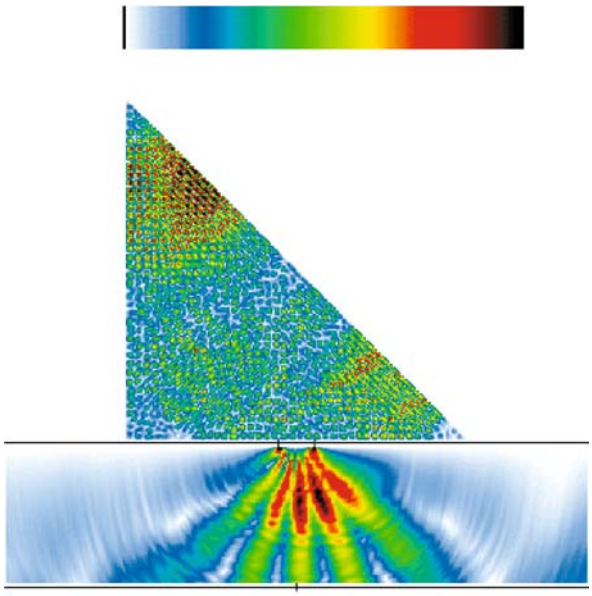


Quantum Chaos, Figure 3
The geometry of the numerical double-slit experiment. All scales are in proper proportions. The two slits are placed at a distance *s* on the lower side of the billiard



Quantum Chaos, Figure 4
The total intensity after the double-slit experiment as a function of the position on the screen. *I*(*x*) is obtained as the perpendicular component of the probability current, integrated in time. The *red full curve* indicates the case of regular billiard, while the *blue dotted curve* indicates the case of chaotic one. The *green dashed curve* indicates the averaged intensity over two 1-slit experiments (where one of the slits is closed), with either the regular or chaotic billiard (with results being practically the same)

We note that such a simple numerical experiment has a well defined analogue (and explanation) in the stationary quantum chaos. Namely the intensity $I(x)$, in the far field



Quantum Chaos, Figure 5
Typical snapshots of the wave-function (plotted is the probability density) for the two cases: a for the regular billiard at $t = 0.325$, and b for the chaotic billiard at $t = 0.275$ (both cases correspond to about half the Heisenberg time). The probability density is normalized separately in both parts of each plot, namely the probability density, in absolute units, in the radiating region is typically less than 1% of the probability density in the billiard domain. The screen, its center, and the positions of the slits are indicated with *thin black lines*. Please note that the color code on the top of the figure is proportional to the square root of probability density

and narrow slit approximations, can be expressed in terms of spatial autocorrelation of the eigenfunctions of the billiard at the two positions of the slit. It is known [10] that chaotic eigenfunctions are characterized by decaying spatial correlations – hence destruction of interference pattern – whereas eigenfunctions of regular systems have long range spatial correlations.

Another important aspect of time-dependent quantum chaos is the study of the so-called Loschmidt echoes or fidelity decay [28]. This quantity has been proposed by Peres [29] as a natural analogy in quantum mechanics of Lyapunov exponents and sensitive dependence to initial conditions. Let $U_0(t)$ represent some unitary quantum mechanical evolution operator, e. g. $U_0(t) = \exp(-\mathrm{i}Ht/\hbar)$ where H is the Hamiltonian, and let $U_\varepsilon(t)$ represent a perturbed evolution, where ε is some small perturbation strength parameter. Quantum Loschmidt echo is defined in terms of fidelity (square modulus of Hilbert space inner product) of a state of perturbed time evolution $|\Psi_\varepsilon(t)\rangle = U_\varepsilon(t)|\Psi(0)\rangle$ with respect to time evolved state of the unperturbed evolution $|\Psi_0(t)\rangle = U_0(t)|\Psi(0)\rangle$, namely

$$F(t) = |\langle\Psi_0(t)|\Psi_\varepsilon(t)\rangle|^2 = |\langle\Psi(0)|U_0(-t)U_\varepsilon(t)|\Psi(0)\rangle|^2 .$$

The first expression (fidelity) can be interpreted as the probability that two nearby quantum evolutions end up in the same state, whereas the second expression (Loschmidt echo) is the probability that the state after forward perturbed time evolution composed with time reversal operation and (backward) unperturbed evolution for the same amount of time, end up in the same (initial) state. The equivalence of the two expressions is a simple consequence of the unitarity of quantum dynamics. The fidelity is a measure of stability of quantum motion. Note that the formalism of Loschmidt echoes can be closely connected to the theory of decoherence in open quantum systems [28]. For example, in many interesting specific situations the Loschmidt echo can be used to bound or estimate certain standard measures of decoherence.

For a semi-classical understanding of Loschmidt echoes it is very useful to write the expression of quantum fidelity in terms of the Wigner phase space function

$$W_\varepsilon(q,p,t) = (2\pi\hbar)^{-d} \cdot \int \mathrm{d}^d r\, \langle q - r/2|\Psi_\varepsilon(t)\rangle\langle\Psi_\varepsilon(t)|q + r/2\rangle \exp(\mathrm{i}p\cdot r/\hbar)$$

corresponding to state $|\Psi_\varepsilon(t)\rangle$, namely

$$F(t) = (2\pi\hbar)^{-d} \int \mathrm{d}^d q\, \mathrm{d}^d p\, W_0(q,p,t) W_\varepsilon(q,p,t) .$$

It is known that the Wigner function of an initial Gaus-

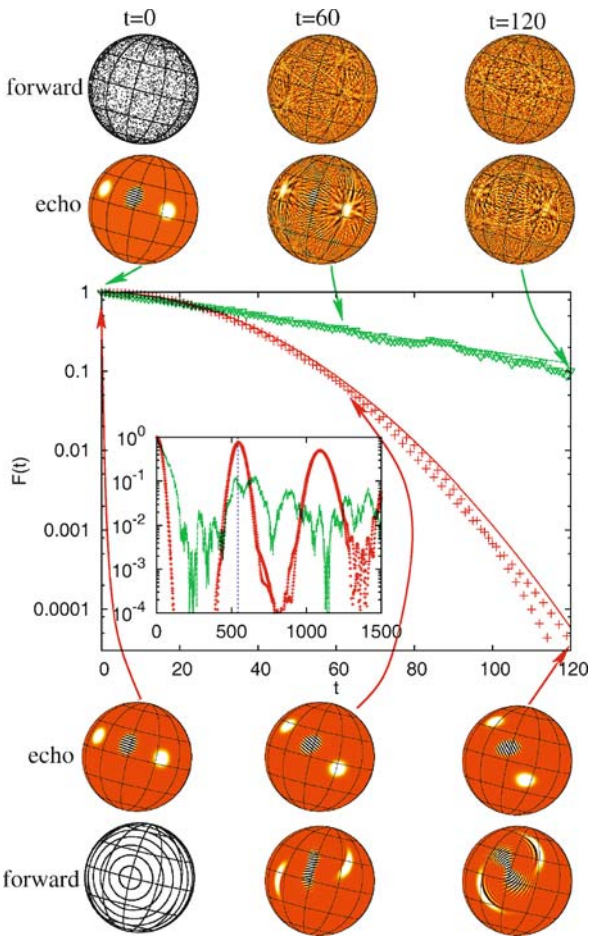
sian wave-packet follows the classical Liouville equation for times below the Ehrenfest time. It can also be easily shown that, for short times, the corresponding *classical fidelity* computed in terms of classical phase space densities decays with the the same rate with which nearby-orbits diverge, namely with the Lyapunov exponent [28]. Therefore, for $t < t_E$, quantum fidelity decays with the perturbation independent rate given by the classical Lyapunov exponent λ, $F_{lyap}(t) \sim \exp(-\lambda t)$.

For times between Ehrenfest and Heisenberg time, $t_E < t < t^*$, and correspondingly small perturbation strength so that the fidelity is appreciable, the decay of Loschmidt echo can be computed using time-dependent perturbation theory, or Fermi golden rule, and one obtains $F_{fgr}(t) \sim \exp(-\varepsilon^2 \sigma t)$ where the coefficient σ is equal to average square of near diagonal matrix elements of the generator of the perturbation. The quantity σ can be semiclassically computed in terms of integrated time-correlation function of the classical perturbation [28].

For even larger times, $t > t^*$, the quantum dynamics is dominated by fluctuations, and fidelity decay, provided ε is small enough, can be computed by static quantum perturbation theory which leads to the Gaussian decay $F_p(t) \sim \exp(-2\varepsilon^2 \sigma t^2/t^*)$.

On the other hand, fidelity decay for classically integrable (regular) systems is somewhat simpler but less universal – namely it depends on the structure of invariant tori of the integrable system and on the initial state. For initial Gaussian wave-packets, the decay of fidelity is typically faster than for chaotic systems, which is a manifestation of *ballistic* nature of regular dynamics as opposed to *diffusive* nature of chaotic dynamics. In Fig. 6 we demonstrate the decay of fidelity for Haake's kicked top model [1] and compare the regimes of regular and chaotic dynamics for the same initial coherent (Gaussian) state and the same value of perturbation parameter. Both cases can be theoretically described and understood [28]. For illustration we plot the Wigner functions of the time evolution as well as the Wigner function of the echo dynamics – perturbed forward evolution composed with unperturbed backward evolution. Note that the Wigner function of a quantum top (spin) is defined over a surface of the sphere.

Quantum chaos has been studied also in unbounded systems with infinite classical phase space, such as the kicked rotator [30], where classical chaos – through decaying temporal correlations – gives rise to deterministic diffusion. It has been shown [30,31,32] that the role of classical deterministic chaos in kicked one-dimensional systems with infinite momentum space is analogous and sometimes even formally equivalent to the role of disorder in one-dimensional tight binding model of a solid. In fact



Quantum Chaos, Figure 6
Fidelity decay for chaotic (*top curve* and *pictures*) and regular (*bottom curve* and *pictures*) kicked top. Initial conditions and the perturbation are the same in both case and theoretical formulae, with the only input given in terms of classical dynamics, are shown with *full curves* (see [28] for details). Wigner functions after forward and echo evolution are shown for shown for illustration of the ballistic versus diffusive mechanisms

the phenomenon of *dynamical localization* has been discovered [31] in full analogy with Anderson localization in disordered one-dimensional solids.

Applications of Quantum Chaos

Theoretical phenomena of quantum chaos have been applied or experimentally observed in a variety of fields. Far from being complete we just mention a few.

Atomic Physics

One of the first successful applications of the ideas of quantum chaos has been a theoretical explanation of

multi-photon ionization experiments with hydrogen in microwave field which were performed by Bayfield and Koch [33]. Single hydrogen atoms prepared in very elongated states with high principal quantum number were injected into microwave cavity and the ionization rate was measured. Even though the microwave frequency was well bellow the ionization energy, in fact even lower than the transition energy to the next excited energy level, it was found very surprisingly that very efficient ionization occurred when the electric field intensity exceeded a certain threshold value. The theoretical analysis (described e. g. in [30]) explained the threshold intensities as critical values for the onset of chaotic diffusion in phase space. Quite interestingly, it has been shown that classical mechanics alone accurately describes the results of experiment, namely in the situation above the delocalization border (which has been predicted theoretically [34]) and in the semiclassical regime of high principal quantum number, where an effective value of the Planck constant is sufficiently small. However, in [34] it has been shown that dynamical localization can take place in hydrogen atom and this has been experimentally observed in [35,36].

A second notable experimental achievement has been the observation of dynamical localization, in fact a realization of quantum mechanical kicked rotor in terms of cold Cesium atoms in a standing wave of kicked electric field by the Raizen's group [37].

Mesoscopic Solid State Physics

The ideas of quantum chaos have been extensively investigated in mesoscopic solid state systems, in particular in the studies of quantum transport in the ballistic regime, where the mean free path of the electrons is much larger than the device.

Perhaps it is worth mentioning the discovery of universal conductance fluctuations [38] which correspond to Ericson fluctuations in nuclear physics. It has been shown – by measuring the so-called magneto-resistance in quantum dots (2d conducting structure of a size of the order of a micrometer) – that the conductance fluctuations have some universal statistical features if the shape of the quantum dot represents a chaotic billiard. Universal conductance fluctuations have been later extensively discussed theoretically and explained in terms of random matrix theory (see e. g. review [39]).

Quantum Information

Chaotic quantum dynamics is effective in producing entanglement, which is a key resource for quantum information processing [40,41]. This has been demonstrated theoretically with many chaotic toy models, see e. g. [42]. Perhaps it deserves to be mentioned that two of these models, namely the quantized Baker map and the quantized sawtooth map have been realized in a real world quantum computer [43]. Dynamical chaos can also be explored to engineer robust and accurate decoupling schemes for processing quantum information in the presence of static noise [28].

Wave Chaos

There are many applications of the ideas and mechanisms of quantum chaos outside of quantum mechanics. For example, essentially all the stationary aspects of quantum chaos, and sometimes even some dynamical aspects provided the change in dispersion relations is taken into account, are being beautifully demonstrated in the planar electromagnetic microwave resonators since early 1990's by the groups of Stöckmann (see e. g. [44]), Richter (see e. g. [45]) and Sridhar (see e. g. [46]). Very impressive "quantum chaos" experimental studies were conducted in acoustics, namely with blocks of vibrating solids, see e. g. [47,48]. It turns out that experimentally feasible quality (Q) factors in elastodynamics resonators can be much larger than in microwave billiards, however the complication here being that the underlying amplitude wave equation is of 4th order and cannot be interpreted as a Schrödinger equation. Still, using the universality of quantum chaos one can argue (and find) that statistical properties of resonance frequencies are described by the same statistical random matrix theory [49]. We may also mention a recent application of quantum chaos to optical fibers [50]. Another impressive extension of quantum chaos is to non-linear wave-optics, namely to engineering of directed stimulated emission of micron-size and chaotic-billiard-shaped semiconductor lasers [51].

Future Directions

By now quantum chaos of a single – or few particles – systems is relatively well understood. Among open future problems we should perhaps mention the case of systems with mixed phase space being neither completely integrable nor fully chaotic. In such systems, one of the most important problems is to derive a quantitative theory of tunneling rates between classically disjoint phase space components.

Most of the results so far obtained in the field of quantum chaos are based on numerical evidence and we are still lacking of rigorous proofs. For example, we are missing the proof of quantum chaos conjecture, the precise conditions (ergodic properties of the underlying classical system) un-

der which it holds, precise statements about quantum ergodicity in general Hamiltonian systems, etc.

However, one of the key challenges for future is to understand and apply dynamical chaos mechanisms to quantum systems of many interacting particles. In particular it seems that complete integrability of many body systems implies anomalous non-equilibrium statistical behaviors, such as e. g. ballistic transport, and it seems necessary to operate in the regime of quantum chaos in order to validate diffusive statistical laws in many body transport. Furthermore, it seems that quantum chaos in many body systems is intimately connected to the impossibility of efficient simulation of such systems on classical computers. See e. g. [52] for a recent review of these topics.

At the general level we recall that, while in classical mechanics there is a very well developed ergodic theory which is important for understanding equilibrium and non-equilibrium properties of classical physical systems, in quantum mechanics there is not, so far, a well established theory. A theory would be required which could explain the asymptotic relaxation process (in the correct order of limits, namely letting the time to infinity at last) in the presence of a purely discrete spectrum and in the absence of exponential instability.

We close on a more speculative note. In quantum mechanics we are always facing with the problem of the measurement device which should be treated as a macroscopic classical system. As for such, classical chaos and exponential instability are present. This is indeed even necessary since, by its purpose, a measurement device must be unstable because a microscopic intervention must produce a macroscopic effect. The importance of chaos in the quantum measurement is that it destroys the coherence of the initial pure state to be measured converting it to an incoherent mixture. In the existing theories of quantum measurement this is described as the effect of external noise. Chaos theory allows to get rid of this unsatisfactory assumption and to develop a purely dynamical theory of loss of quantum coherence. Still, a fundamental problem remains open, namely the redistribution of probability according to the result of the measurement: the so-called collapse of the wave-function which remains to be understood.

Bibliography

1. Haake F (2001) Quantum Signatures of Chaos, 2nd edn. Springer, Heidelberg
2. Stöckmann HJ (1999) Quantum Chaos: An Introduction. Cambridge University Press, Cambridge
3. Cornfeld IP, Fomin SV, Sinai YG (1982) Ergodic theory. Springer, New York
4. Casati G, Prosen T (1999) Phys Rev Lett 83:4279; Casati G, Prosen T (2000) Phys Rev Lett 85:4261
5. Mehta ML (1991) Random matrices. Academic Press, New York
6. Casati G, Guarneri I, Valz-Gris (1980) Lett Nuovo Cimento 28:279
7. Bohigas O, Giannoni MJ, Schmit C (1984) Phys Rev Lett 52:1
8. Robnik M (1983) J Phys A 16:3971
9. Bäcker A (2007) Comput Sci Eng 9:60
10. Berry MV (1977) J Phys A 10:2083
11. Feingold M, Peres A (1986) Phys Rev A 34:591
12. Horvat M, Prosen T (2003) J Phys A 36:4015
13. Heller EJ (1984) Phys Rev Lett 53:1515
14. Shnirelman AI (1974) Usp Math Nauk 29:181
15. Zelditch S (1987) Duke Math J 55:919
16. Colin de Verdiere Y (1985) Comm Math Phys 102:111
17. Berry MV, Tabor M (1977) Proc R Soc A 356:375
18. Casati G, Chirikov BV (1985) Phys Rev Lett 54:1350
19. Berry MV, Robnik M (1984) J Phys A 17:2413
20. Gutzwiller MC (1991) Chaos in Classical and Quantum Mechanics. Springer, New York
21. Berry MV, Tabor M (1977) J Phys A 10:371
22. Bogomolny EB (1988) Physica D 31:169
23. Berry MV (1989) Proc Royal Soc A 423:219
24. Müller S, Heusler S, Braun P, Haake F, Altland A (2004) Phys Rev Lett 93:014 103; Müller S, Heusler S, Braun P, Haake F, Altland A (2005) Phys Rev E 72:046 207
25. Sieber M (2002) J Phys A 35:L613
26. Heller EJ (1991) Chaos and Quantum Physics. In: Giannoni MJ, Voros A, Zinn-Justin J (eds) Les Houches, Session LII, 1989. North Holland, Amsterdam, p 547
27. Casati G, Prosen T (2005) Phys Rev A 72:032111
28. Gorin T, Prosen T, Seligman TH, Žnidarič M (2006) Phys Rep 435:33
29. Peres A (1984) Phys Rev A 30:1610
30. Casati G, Chirikov BV (1995) In: Quantum chaos: between order and disorder. Cambridge University Press, Cambridge, p 3
31. Casati G, Chirikov BV, Izrailev FM, Ford J (1979) In: Casati FG, Ford J (eds) Stochastic Behavior in Classical and Quantum Hamiltonian Systems. Lecture Notes in Physics, vol 93. Springer, Berlin
32. Fishman S, Grempel DR, Prange RE (1982) Phys Rev Lett 49:509
33. Bayfield JE, Koch PM (1974) Phys Rev Lett 33:258
34. Casati G, Chirikov BV, Shepelyanski D (1984) Phys Rev Lett 53:2525
35. Bayfield JE, Casati G, Guarneri I, Sokol DW (1989) Phys Rev Lett 63:364
36. Galvez EJ, Sauer BE, Moorman L, Koch PM, Richards D (1988) Phys Rev Lett 61:2011
37. Klappauf BG, Oskay WH, Steck DA, Raizen MG (1999) Physica D 131:78
38. Marcus CM, Rimberg AJ, Westervelt RM, Hopkins PF, Gossard AC (1992) Phys Rev Lett 69:506
39. Beenakker CWJ (1997) Rev Mod Phys 69:731
40. Nielsen MA, Chuang IL (2000) Quantum Computation and Quantum Information. Cambridge University Press, Cambridge
41. Benenti G, Casati G, Strini G (2004) Principles of Quantum Computation and Information, vol I: Basic concepts. World Scientific, Singapore; Benenti G, Casati G, Strini G (2007) Principles of Quantum Computation and Information, vol II: Basic tools and special topics. World Scientific, Singapore

42. Žnidarič M, Prosen T (2003) J Phys A: Math Gen 36:2463
43. Weinstein YA, Lloyd S, Emerson J, Cory DG (2002) Phys Rev Lett 89:157 902; Henry MK, Emerson J, Martinez R, Cory DG (2006) Phys Rev A 74:032 617
44. Stein J, Stöckmann HJ (1992) Phys Rev Lett 68:2867
45. Richter A (2001) Phys Scr T90:212
46. Sridhar S (1991) Phys Rev Lett 67:785
47. Ellegaard C, Schaadt K, Bertelsen P (2001) Phys Scr T90:223
48. Kuhl U, Stöckmann HJ, Weaver R (2005) J Phys A 38:10 433
49. Fyodorov YV, Savin DV, Sommers HJ (2005) J Phys A 38:10 731
50. Doya V, Legrand O, Mortessagne F, Miniatura C (2001) Phys Rev Lett 88:014 102
51. Harayama T, Fukushima T, Sunada S, Ikeda KS (2003) Phys Rev Lett 91:073 903
52. Prosen T (2007) J Phys A: Math Theor 40:7881

Quantum Computational Complexity

JOHN WATROUS
Institute for Quantum Computing and School of Computer Science, University of Waterloo, Waterloo, Canada

Article Outline

Glossary

Quantum circuit A quantum circuit is an acyclic network of quantum gates connected by wires: the gates represent quantum operations and the wires represent the qubits on which these operations are performed. The quantum circuit model is the most commonly studied model of quantum computation.

Quantum complexity class A quantum complexity class is a collection of computational problems that are solvable by a chosen quantum computational model that obeys certain resource constraints. For example, BQP is the quantum complexity class of all decision problems that can be solved in polynomial time by a quantum computer.

Quantum proof A quantum proof is a quantum state that plays the role of a witness or certificate to a quantum computer that runs a verification procedure. The quantum complexity class QMA is defined by this notion: it includes all decision problems whose yes-instances are efficiently verifiable by means of quantum proofs.

Quantum interactive proof system A quantum interactive proof system is an interaction between a verifier and one or more provers, involving the processing and exchange of quantum information, whereby the provers attempt to convince the verifier of the answer to some computational problem.

Definition of the Subject

The inherent difficulty, or *hardness*, of computational problems is a fundamental concept in computational complexity theory. Hardness is typically formalized in terms of the resources required by different models of computation to solve problems, such as the number of steps of a deterministic Turing machine. A variety of models and resources are often considered, including deterministic, nondeterministic and probabilistic models; time and space constraints; and interactions among models of differing abilities. Many interesting relationships among these different models and resource constraints are known.

One common feature of the most commonly studied computational models and resource constraints is that they are *physically motivated*. This is quite natural, given that computers are physical devices, and to a significant extent it is their study that motivates and directs research on computational complexity. The predominant example is the class of polynomial-time computable functions, which ultimately derives its relevance from physical considerations; for it is a mathematical abstraction of the class of functions that can be efficiently computed without error by physical computing devices.

In light of its close connection to the physical world, it seems only natural that modern physical theories should be considered in the context of computational complexity. In particular, *quantum mechanics* is a clear candidate for a physical theory to have the potential for implications, if not to computational complexity then at least to computation more generally. Given the steady decrease in the size of computing components, it is inevitable that quantum mechanics will become increasingly relevant to the construction of computers—for quantum mechanics provides a remarkably accurate description of extremely small physical systems (on the scale of atoms) where classical physical theories have failed completely. Indeed, an

extrapolation of Moore's Law predicts sub-atomic computing components within the next two decades [74,78]; a possibility inconsistent with quantum mechanics as it is currently understood.

That quantum mechanics should have implications to computational complexity theory, however, is much less clear. It is only through the remarkable discoveries and ideas of several researchers, including Richard Feynman [49], David Deutsch [40,41], Ethan Bernstein and Umesh Vazirani [29,30], and Peter Shor [88,89], that this potential has become evident. In particular, Shor's polynomial-time quantum factoring and discrete-logarithm algorithms [89] give strong support to the conjecture that quantum and classical computers yield differing notions of computational hardness. Other quantum complexity-theoretic concepts, such as the efficient verification of quantum proofs, suggest a wider extent to which quantum mechanics influences computational complexity.

It may be said that the principal aim of quantum computational complexity theory is to understand the implications of quantum physics to computational complexity theory. To this end, it considers the hardness of computational problems with respect to models of quantum computation, classifications of problems based on these models, and their relationships to classical models and complexity classes.

Introduction

This article surveys quantum computational complexity, with a focus on three fundamental notions: polynomial-time quantum computations, the efficient verification of quantum proofs, and quantum interactive proof systems. Based on these notions one defines quantum complexity classes, such as BQP, QMA, and QIP, that contain computational problems of varying hardness. Properties of these complexity classes, and the relationships among these classes and classical complexity classes, are presented. As these notions and complexity classes are typically defined within the quantum circuit model, this article includes a section that focuses on basic properties of quantum circuits that are important in the setting of quantum complexity. A selection of other topics in quantum complexity, including quantum advice, space-bounded quantum computation, and bounded-depth quantum circuits, is also presented.

Two different but closely related areas of study are not discussed in this article: *quantum query complexity* and *quantum communication complexity*. Readers interested in learning more about these interesting and active areas of research may find the surveys of Brassard [33], Cleve [34], and de Wolf [39] to be helpful starting points.

It is appropriate that brief discussions of computational complexity theory and quantum information precede the main technical portion of the article. These discussions are intended only to highlight the aspects of these topics that are non-standard, require clarification, or are of particular importance in quantum computational complexity. In the subsequent sections of this article, the reader is assumed to have basic familiarity with both topics, which are covered in depth by several text books [13,43,60,67,79,82].

Computational Complexity

Throughout this article the binary alphabet $\{0,1\}$ is denoted Σ, and all computational problems are assumed to be encoded over this alphabet. As usual, a function $f\colon \Sigma^* \to \Sigma^*$ is said to be *polynomial-time computable* if there exists a polynomial-time deterministic Turing machine that outputs $f(x)$ for every input $x \in \Sigma^*$. Two related points on the terminology used throughout this article are as follows.

1. A function of the form $p\colon \mathbb{N} \to \mathbb{N}$ (where $\mathbb{N} = \{0,1,2,\ldots\}$) is said to be a *polynomial-bounded function* if and only if there exists a polynomial-time deterministic Turing machine that outputs $1^{f(n)}$ on input 1^n for every $n \in \mathbb{N}$. Such functions are upper-bounded by some polynomial, and are efficiently computable.
2. A function of the particular form $a\colon \mathbb{N} \to [0,1]$ is said to be *polynomial-time computable* if and only if there exists a polynomial-time deterministic Turing machine that outputs a binary representation of $a(n)$ on input 1^n for each $n \in \mathbb{N}$. References to functions of this form in this article typically concern bounds on probabilities that are functions of the length of an input string to some problem.

The notion of *promise problems* [44,52] is central to quantum computational complexity. These are decision problems for which the input is assumed to be drawn from some subset of all possible input strings. More formally, a promise problem is a pair $A = (A_{\text{yes}}, A_{\text{no}})$, where $A_{\text{yes}}, A_{\text{no}} \subseteq \Sigma^*$ are sets of strings satisfying $A_{\text{yes}} \cap A_{\text{no}} = \varnothing$. *Languages* may be viewed as promise problems that obey the additional constraint $A_{\text{yes}} \cup A_{\text{no}} = \Sigma^*$. Although complexity theory has traditionally focused on languages rather than promise problems, little is lost and much is gained in shifting one's focus to promise problems. Karp reductions and the notion of completeness are defined for promise problems in the same way as for languages.

Several classical complexity classes are referred to in this article, and compared with quantum complexity classes when relations are known. The following classical complexity classes, which should hereafter be understood to be classes of promise problems and not just languages, are among those discussed.

P A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in P if and only if there exists a polynomial-time deterministic Turing machine M that accepts every string $x \in A_{\text{yes}}$ and rejects every string $x \in A_{\text{no}}$.

NP A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in NP if and only if there exists a polynomial-bounded function p and a polynomial-time deterministic Turing machine M with the following properties. For every string $x \in A_{\text{yes}}$, it holds that M accepts (x, y) for some string $y \in \Sigma^{p(|x|)}$, and for every string $x \in A_{\text{no}}$, it holds that M rejects (x, y) for all strings $y \in \Sigma^{p(|x|)}$.

BPP A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in BPP if and only if there exists a polynomial-time probabilistic Turing machine M that accepts every string $x \in A_{\text{yes}}$ with probability at least 2/3, and accepts every string $x \in A_{\text{no}}$ with probability at most 1/3.

PP A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in PP if and only if there exists a polynomial-time probabilistic Turing machine M that accepts every string $x \in A_{\text{yes}}$ with probability strictly greater than 1/2, and accepts every string $x \in A_{\text{no}}$ with probability at most 1/2.

MA A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in MA if and only if there exists a polynomial-bounded function p and a probabilistic polynomial-time Turing machine M with the following properties. For every string $x \in A_{\text{yes}}$, it holds that $\Pr[M \text{ accepts } (x, y)] \geq \frac{2}{3}$ for some string $y \in \Sigma^{p(|x|)}$; and for every string $x \in A_{\text{no}}$, it holds that $\Pr[M \text{ accepts } (x, y)] \leq \frac{1}{3}$ for all strings $y \in \Sigma^{p(|x|)}$.

AM A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in AM if and only if there exist polynomial-bounded functions p and q and a polynomial-time deterministic Turing machine M with the following properties. For every string $x \in A_{\text{yes}}$, and at least 2/3 of all strings $y \in \Sigma^{p(|x|)}$, there exists a string $z \in \Sigma^{q(|x|)}$ such that M accepts (x, y, z); and for every string $x \in A_{\text{no}}$, and at least 2/3 of all strings $y \in \Sigma^{p(|x|)}$, there are no strings $z \in \Sigma^{q(|x|)}$ such that M accepts (x, y, z).

SZK A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in SZK if and only if it has a statistical zero-knowledge interactive proof system.

PSPACE A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in PSPACE if and only if there exists a deterministic Turing machine M running in polynomial space that accepts every string $x \in A_{\text{yes}}$ and rejects every string $x \in A_{\text{no}}$.

EXP A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in EXP if and only if there exists a deterministic Turing machine M running in exponential time (meaning time bounded by 2^p, for some polynomial-bounded function p), that accepts every string $x \in A_{\text{yes}}$ and rejects every string $x \in A_{\text{no}}$.

NEXP A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in NEXP if and only if there exists an exponential-time non-deterministic Turing machine N for A.

PL A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in PL if and only if there exists a probabilistic Turing machine M running in polynomial time and logarithmic space that accepts every string $x \in A_{\text{yes}}$ with probability strictly greater than 1/2 and accepts every string $x \in A_{\text{no}}$ with probability at most 1/2.

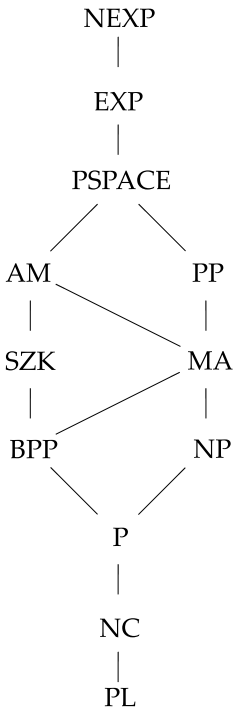
NC A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in NC if and only if there exists a logarithmic-space generated family $C = \{C_n : n \in \mathbb{N}\}$ of poly-logarithmic depth Boolean circuits such that $C(x) = 1$ for all $x \in A_{\text{yes}}$ and $C(x) = 0$ for all $x \in A_{\text{no}}$.

For most of the complexity classes listed above, there is a standard way to attach an *oracle* to the machine model that defines the class, which provides a subroutine for solving instances of a chosen problem $B = (B_{\text{yes}}, B_{\text{no}})$. One then denotes the existence of such an oracle with a superscriptfor example, P^B is the class of promise problems that can be solved in polynomial time by a deterministic Turing machine equipped with an oracle that solves instances of B (at unit cost). When classes of problems appear as superscripts, one takes the union, as in the following example:

$$\mathrm{P}^{\mathrm{NP}} = \bigcup_{B \in \mathrm{NP}} \mathrm{P}^B .$$

Quantum Information

The standard general description of quantum information is used in this article: mixed states of systems are represented by density matrices and operations are represented by completely positive trace-preserving linear maps. The choice to use this description of quantum information is deserving of a brief discussion, for it will likely be less familiar to many non-experts than the simplified picture of quantum information where states are represented by unit vectors and operations are represented by unitary matrices. This simplified picture is indeed commonly used in



Quantum Computational Complexity, Figure 1
A diagram illustrating known inclusions among most of the classical complexity classes discussed in this paper. Lines indicate containments going upward; for example, AM is contained in PSPACE

the study of both quantum algorithms and quantum complexity theory; and it is often adequate. However, the general picture has major advantages: it unifies quantum information with classical probability theory, brings with it powerful mathematical tools, and allows for simple and intuitive descriptions in many situations where this is not possible with the simplified picture.

Classical simulations of quantum computations, which are discussed below in Subsect. "Classical Upper Bounds on BQP", may be better understood through a fairly straightforward representation of quantum operations by matrices. This representation begins with a representation of density matrices as vectors based on the function defined as $\mathrm{vec}(|x\rangle\langle y|) = |x\rangle|y\rangle$ for each choice of $n \in \mathbb{N}$ and $x, y \in \Sigma^n$, and extended by linearity to all matrices indexed by Σ^n. The effect of this mapping is to form a column vector by reading the entries of a matrix in rows from left to right, starting at the top. For example,

$$\mathrm{vec}\begin{pmatrix}\alpha & \beta\\ \gamma & \delta\end{pmatrix} = \begin{pmatrix}\alpha\\ \beta\\ \gamma\\ \delta\end{pmatrix}.$$

Now, the effect of a general quantum operation Φ, represented in the typical Kraus form as

$$\Phi(\rho) = \sum_{j=1}^{k} A_j \rho A_j^*,$$

is expressed as a matrix by means of the equality

$$\mathrm{vec}(\Phi(\rho)) = \left(\sum_{j=1}^{k} A_j \otimes \overline{A_j}\right)\mathrm{vec}(\rho).$$

The matrix

$$K_\Phi = \sum_{j=1}^{k} A_j \otimes \overline{A_j}$$

is sometimes called the *natural representation* (or *linear representation*) of the operation Φ. Although this matrix could have negative or complex entries, one can reasonably view it as being analogous to a stochastic matrix that describes a probabilistic computation.

For example, the complete phase-damping channel for a single qubit can be written

$$D(\rho) = |0\rangle\langle 0|\rho|0\rangle\langle 0| + |1\rangle\langle 1|\rho|1\rangle\langle 1|.$$

The effect of this mapping is to zero-out the off-diagonal entries of a density matrix

$$D\begin{pmatrix}\alpha & \beta\\ \gamma & \delta\end{pmatrix} = \begin{pmatrix}\alpha & 0\\ 0 & \delta\end{pmatrix}.$$

The natural representation of this operation is easily computed

$$K_D = \begin{pmatrix}1 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 1\end{pmatrix}.$$

The natural matrix representation of quantum operations is well-suited to performing computations. For the purposes of this article, the most important observation about this representation is that a composition of operations corresponds simply to matrix multiplication.

The Quantum Circuit Model

General Quantum Circuits

The term *quantum circuit* refers to an acyclic network of *quantum gates* connected by *wires*. The quantum gates represent general quantum operations, involving some

constant number of qubits, while the wires represent the qubits on which the gates act. An example of a quantum circuit having four input qubits and three output qubits is pictured in Fig. 2. In general a quantum circuit may have n input qubits and m output qubits for any choice of integers $n, m \geq 0$. Such a circuit induces some quantum operation from n qubits to m qubits, determined by composing the actions of the individual gates in the appropriate way. The *size* of a quantum circuit is the total number of gates plus the total number of wires in the circuit.

A *unitary quantum circuit* is a quantum circuit in which all of the gates correspond to unitary quantum operations. Naturally this requires that every gate, and hence the circuit itself, has an equal number of input and output qubits. It is common in the study of quantum computing that one works entirely with unitary quantum circuits. The unitary model and general model are closely related, as will soon be explained.

A Finite Universal Gate Set

Restrictions must be placed on the gates from which quantum circuits may be composed if the quantum circuit model is to be used for complexity theory—for without such restrictions it cannot be argued that each quantum gate corresponds to an operation with unit-cost. The usual way in which this is done is simply to fix a suitable finite set of allowable gates. For the remainder of this article, quantum circuits will be assumed to be composed of gates from the following list:

1. *Toffoli gates*. Toffoli gates are three-qubit unitary gates defined by the following action on standard basis states:

$$T\colon |a\rangle|b\rangle|c\rangle \mapsto |a\rangle|b\rangle|c \oplus ab\rangle \,.$$

2. *Hadamard gates*. Hadamard gates are single-qubit unitary gates defined by the following action on standard basis states:

$$H\colon |a\rangle \mapsto \frac{1}{\sqrt{2}}|0\rangle + \frac{(-1)^a}{\sqrt{2}}|1\rangle \,.$$

3. *Phase-shift gates*. Phase-shift gates are single-qubit unitary gates defined by the following action on standard basis states:

$$P\colon |a\rangle \mapsto i^a|a\rangle \,.$$

4. *Ancillary gates*. Ancillary gates are non-unitary gates that take no input and produce a single qubit in the state $|0\rangle$ as output.
5. *Erasure gates*. Erasure gates are non-unitary gates that take a single qubit as input and produce no output. Their effect is represented by the partial trace on the space corresponding to the qubit they take as input.

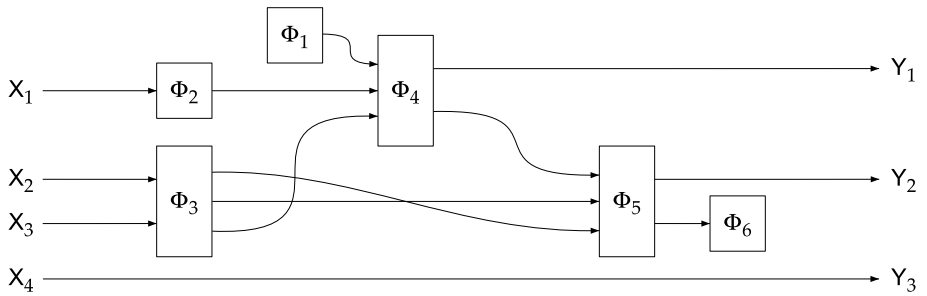
The symbols used to denote these gates in quantum circuit diagrams are shown in Fig. 3. For the sake of simplifying some of the diagrams that appear later in this article, some additional gates are illustrated in Fig. 4 along with their realizations as circuits with gates from the chosen basis set.

The above gate set is *universal* in a strong sense: every quantum operation can be approximated to within any desired degree of accuracy by some quantum circuit. Moreover, the size of the approximating circuit scales well with respect to the desired accuracy. Theorem 1, stated below, expresses this fact in more precise terms, but requires a brief discussion of a specific sense in which one quantum operation approximates another.

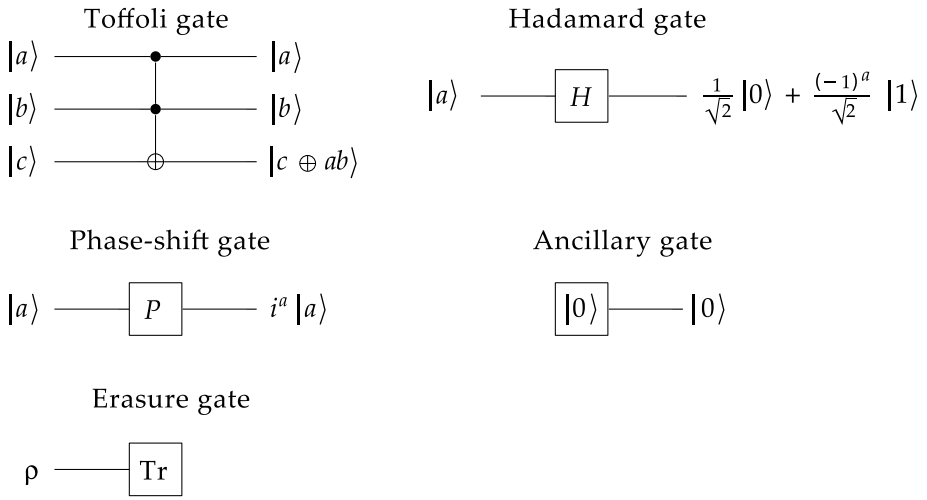
A natural and operationally meaningful way to measure the distance between two given quantum operations Φ and Ψ is given by

$$\delta(\Phi, \Psi) = \frac{1}{2} \| \Phi - \Psi \|_\diamond \,,$$

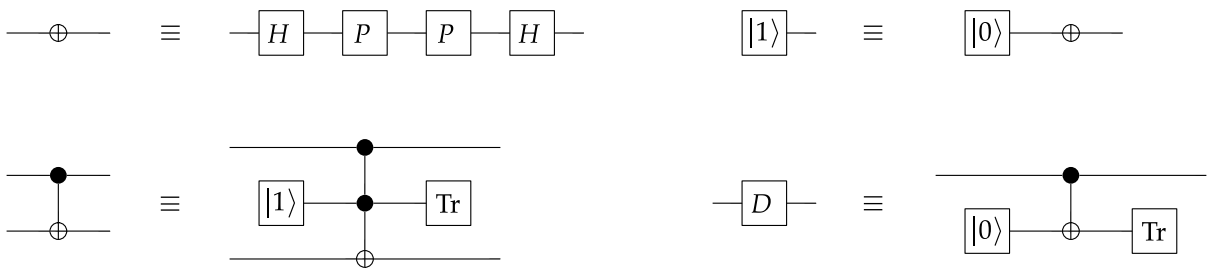
where $\| \cdot \|_\diamond$ denotes a norm usually known as the *diamond norm* [64,67]. A technical discussion of this norm is



Quantum Computational Complexity, Figure 2
An example of a quantum circuit. The input qubits are labeled $X_1, \ldots, X_4$, the output qubits are labeled $Y_1, \ldots, Y_3$, and the gates are labeled by (hypothetical) quantum operations $\Phi_1, \ldots, \Phi_6$



Quantum Computational Complexity, Figure 3
A universal collection of quantum gates: Toffoli, Hadamard, phase-shift, ancillary, and erasure gates



Quantum Computational Complexity, Figure 4
Four additional quantum gates, together with their implementations as quantum circuits. *Top left*: a NOT gate. *Top right*: a constant $|1\rangle$ ancillary gate. *Bottom left*: a controlled-NOT gate. *Bottom right*: a phase-damping (or decoherence) gate

not necessary for the purposes of this article and is beyond its scope. Instead, an intuitive description of the distance measure $\delta(\Phi, \Psi)$ will suffice.

When considering the distance between quantum operations Φ and Ψ, it must naturally be assumed that these operations agree on their numbers of input qubits and output qubits; so assume that Φ and Ψ both map n qubits to m qubits for nonnegative integers n and m. Now, suppose that an arbitrary quantum state on n or more qubits is prepared, one of the two operations Φ or Ψ is applied to the first n of these qubits, and then a general measurement of all of the resulting qubits takes place (including the m output qubits and the qubits that were not among the inputs to the chosen quantum operation). Two possible probability distributions on measurement outcomes arise: one corresponding to Φ and the other corresponding to Ψ. The quantity $\delta(\Phi, \Psi)$ is simply the maximum possible total variation distance between these distributions, ranging over all possible initial states and general measurements. This is a number between 0 and 1 that intuitively represents the *observable difference* between quantum operations.

In the special case that Φ and Ψ take no inputs, the quantity $\delta(\Phi, \Psi)$ is simply one-half the trace norm of the difference between the two output states; a common and useful ways to measure the distance between quantum states.

Now the Universality Theorem, which represents an amalgamation of several results that suits the needs of this article, may be stated. In particular, it incorporates the Solovay–Kitaev Theorem, which provides a bound on the size of an approximating circuit as a function of the accuracy.

Theorem 1 (Universality Theorem) *Let Φ be an arbitrary quantum operation from n qubits to m qubits. Then for every $\varepsilon > 0$ there exists a quantum circuit Q with n input qubits and m output qubits such that $\delta(\Phi, Q) < \varepsilon$.*

Moreover, for fixed n and m, the circuit Q may be taken to satisfy $size(Q) = \text{poly}(\log(1/\varepsilon))$.

Note that it is inevitable that the size of Q is exponential in n and m in the worst case [68]. Further details on the facts comprising this theorem can be found in both Nielsen and Chuang [79] and Kitaev, Shen, and Vyalyi [67].

Unitary Purifications of Quantum Circuits

The connection between general and unitary quantum circuits can be understood through the notion of a *unitary purification* of a general quantum circuit. This may be thought of as a very specific manifestation of the *Stinespring Dilation Theorem* [90], which implies that general quantum operations can be represented by unitary operations on larger systems. It was first applied to the quantum circuit model by Aharonov, Kitaev, and Nisan [9], who gave several arguments in favor of the general quantum circuit model over the unitary model. The term *purification* is borrowed from the notion of a purification of a mixed quantum state, as the process of unitary purification for circuits is similar in spirit. The universal gate set, described in the previous section has the effect of making the notion of a unitary purification of a general quantum circuit nearly trivial at a technical level.

Suppose that Q is a quantum circuit taking input qubits $(X_1, \ldots, X_n)$ and producing output qubits $(Y_1, \ldots, Y_m)$, and assume there are k ancillary gates and l erasure gates among the gates of Q to be labeled in an arbitrary order as $G_1, \ldots, G_k$ and $K_1, \ldots, K_l$, respectively. A new quantum circuit R may then be formed by removing the gates labeled $G_1, \ldots, G_k$ and $K_1, \ldots, K_l$; and to account for the removal of these gates the circuit R takes k additional input qubits $(Z_1, \ldots, Z_k)$ and produces l additional output qubits $(W_1, \ldots, W_l)$. Figure 5 illustrates this process. The circuit R is said to be a unitary purification of Q. It is obvious that R is equivalent to Q, provided the qubits $(Z_1, \ldots, Z_k)$ are initially set to the $|0\rangle$ state and the qubits $(W_1, \ldots, W_l)$ are traced-out, or simply ignored, after the circuit is run—for this is precisely the meaning of the removed gates.

Despite the simplicity of this process, it is often useful to consider the properties of unitary purifications of general quantum circuits.

Oracles in the Quantum Circuit Model

Oracles play an important, and yet uncertain, role in computational complexity theory; and the situation is no different in the quantum setting. Several interesting oracle-related results, offering some insight into the power of quantum computation, will be discussed in this article.

Oracle queries are represented in the quantum circuit model by an infinite family

$$\{R_n : n \in \mathbb{N}\}$$

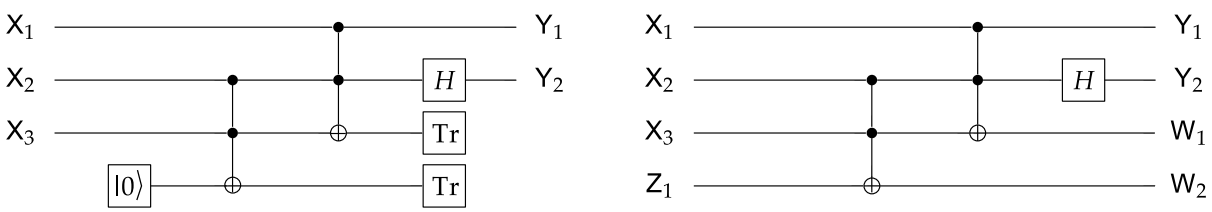
of quantum gates, one for each possible query length. Each gate R_n is a unitary gate acting on $n + 1$ qubits, with the effect on computational basis states given by

$$R_n|x, a\rangle = |x, a \oplus A(x)\rangle \tag{1}$$

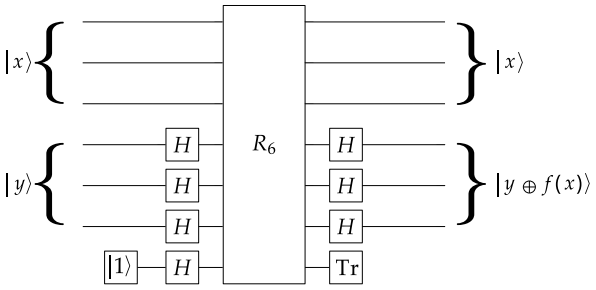
for all $x \in \Sigma^n$ and $a \in \Sigma$, where A is some predicate that represents the particular oracle under consideration. When quantum computations relative to such an oracle are to be studied, quantum circuits composed of ordinary quantum gates as well as the oracle gates $\{R_n\}$ are considered; the interpretation being that each instance of R_n in such a circuit represents one oracle query.

It is critical to many results concerning quantum oracles, as well as most results in the area of quantum query complexity, that the above definition (1) takes each R_n to be unitary, thereby allowing these gates to make queries "in superposition". In support of this seemingly strong definition is the fact, discussed in the next section, that any efficient algorithm (quantum or classical) can be converted to a quantum circuit that closely approximates the action of these gates.

Finally, one may consider a more general situation in which the predicate A is replaced by a function that outputs multiple bits. The definition of each gate R_n is adapted



Quantum Computational Complexity, Figure 5
A general quantum circuit *(left)* and its unitary purification *(right)*



Quantum Computational Complexity, Figure 6
The Bernstein–Vazirani algorithm allows a multiple-bit query to be simulated by a single-bit query. In the example pictured, $f: \Sigma^3 \to \Sigma^3$ is a given function. To simulate a query to this function, the gate R_6 is taken to be a standard oracle gate implementing the predicate $A(x, z) = \langle f(x), z \rangle$, for $\langle \cdot, \cdot \rangle$ denoting the modulo 2 inner product

appropriately. Alternately, one may restrict their attention to single-bit queries as discussed above, and use the Bernstein–Vazirani algorithm [30] to simulate one multiple-bit query with one single-bit query as illustrated in Fig. 6.

Polynomial-Time Quantum Computations

This section focuses on *polynomial-time quantum computations*. These are the computations that are viewed, in an abstract and idealized sense, to be efficiently implementable by the means of a quantum computer. In particular, the complexity class BQP (short for *bounded-error quantum polynomial time*) is defined. This is the most fundamentally important of all quantum complexity classes, as it represents the collection of decision problems that can be efficiently solved by quantum computers.

Polynomial-Time Generated Circuit Families and BQP

To define the class BQP using the quantum circuit model, it is necessary to briefly discuss encodings of circuits and the notion of a polynomial-time generated circuit family.

It is clear that any quantum circuit formed from the gates described in the previous section could be *encoded* as a binary string using any number of different encoding schemes. Such an encoding scheme must be chosen, but its specifics are not important so long as following simple restrictions are satisfied:

1. The encoding is sensible: every quantum circuit is encoded by at least one binary string, and every binary string encodes at most one quantum circuit.
2. The encoding is efficient: there is a fixed polynomial-bounded function p such that every circuit of size N has an encoding with length at most $p(N)$. Specific information about the structure of a circuit must be computable in polynomial time from an encoding of the circuit.
3. The encoding disallows compression: it is not possible to work with encoding schemes that allow for extremely short (e. g., polylogarithmic-length) encodings of circuits; so for simplicity it is assumed that the length of every encoding of a quantum circuit is at least the size of the circuit.

Now, as any quantum circuit represents a finite computation with some fixed number of input and output qubits, quantum algorithms are modeled by *families* of quantum circuits. The typical assumption is that a quantum circuit family that describes an algorithm contains one circuit for each possible input length. Precisely the same situation arises here as in the classical setting, which is that it should be possible to efficiently generate the circuits in a given family in order for that family to represent an efficient, finitely specified algorithm. The following definition formalizes this notion.

Definition 2 Let $S \subseteq \Sigma^*$ be any set of strings. Then a collection $\{Q_x : x \in S\}$ of quantum circuits is said to be *polynomial-time generated* if there exists a polynomial-time deterministic Turing machine that, on every input $x \in S$, outputs an encoding of Q_x.

This definition is slightly more general than what is needed to define BQP, but is convenient for other purposes. For instance, it allows one to easily consider the situation in which the input, or some part of the input, for some problem is hard-coded into a collection of circuits; or where a computation for some input may be divided among several circuits. In the most typical case that a polynomial-time generated family of the form $\{Q_n : n \in \mathbb{N}\}$ is referred to, it should be interpreted that this is a shorthand for $\{Q_{1^n} : n \in \mathbb{N}\}$. Notice that every polynomial-time generated family $\{Q_x : x \in S\}$ has the property that each circuit Q_x has size polynomial in $|x|$. Intuitively speaking, the number of quantum and classical computation steps required to implement such a computation is polynomial; and so operations induced by the circuits in such a family are viewed as representing *polynomial-time quantum computations*.

The complexity class BQP, which contains those promise problems abstractly viewed to be efficiently solvable using a quantum computer, may now be defined. More precisely, BQP is the class of promise problems that can be solved by polynomial-time quantum computations that may have some small probability to make an error. For decision problems, the notion of a polynomial-time

quantum computation is equated with the computation of a polynomial-time generated quantum circuit family $Q = \{Q_n : n \in \mathbb{N}\}$, where each circuit Q_n takes n input qubits, and produces one output qubit. The computation on a given input string $x \in \Sigma$ is obtained by first applying the circuit $Q_{|x|}$ to the state $|x\rangle\langle x|$, and then measuring the output qubit with respect to the standard basis. The measurement results 0 and 1 are interpreted as *yes* and *no* (or *accept* and *reject*), respectively. The events that Q accepts x and Q rejects x are understood to have associated probabilities determined in this way.

BQP Let $A = (A_{\text{yes}}, A_{\text{no}})$ be a promise problem and let $a, b : \mathbb{N} \to [0, 1]$ be functions. Then $A \in \text{BQP}(a, b)$ if and only if there exists a polynomial-time generated family of quantum circuits $Q = \{Q_n : n \in \mathbb{N}\}$, where each circuit Q_n takes n input qubits and produces one output qubit, that satisfies the following properties:

1. if $x \in A_{\text{yes}}$ then $\Pr[Q \text{ accepts } x] \geq a(|x|)$, and
2. if $x \in A_{\text{no}}$ then $\Pr[Q \text{ accepts } x] \leq b(|x|)$.

The class BQP is defined as $\text{BQP} = \text{BQP}(2/3, 1/3)$.

Similar to BPP, there is nothing special about the particular choice of error probability 1/3, other than that it is a constant strictly smaller than 1/2. This is made clear in the next section.

There are several problems known to be in BQP but not known (and generally not believed) to be in BPP. Decision-problem variants of the integer factoring and discrete logarithm problems, shown to be in BQP by Shor [89], are at present the most important and well-known examples.

Error Reduction for BQP

When one speaks of the flexibility, or *robustness*, of BQP with respect to error bounds, it is meant that the class $\text{BQP}(a, b)$ is invariant under a wide range of "reasonable" choices of the functions a and b. The following proposition states this more precisely.

Proposition 3 (Error reduction for BQP) *Suppose that $a, b : \mathbb{N} \to [0, 1]$ are polynomial-time computable functions and $p : \mathbb{N} \to \mathbb{N}$ is a polynomial-bounded function such that $a(n) - b(n) \geq 1/p(n)$ for all but finitely many $n \in \mathbb{N}$. Then for every choice of a polynomial-bounded function $q : \mathbb{N} \to \mathbb{N}$ satisfying $q(n) \geq 2$ for all but finitely many $n \in \mathbb{N}$, it holds that*

$$\text{BQP}(a, b) = \text{BQP} = \text{BQP}\left(1 - 2^{-q}, 2^{-q}\right).$$

The above proposition may be proved in the same standard way that similar statements are proved for classical probabilistic computations: by repeating a given computation some large (but still polynomial) number of times, overwhelming statistical evidence is obtained so as to give the correct answer with an extremely small probability of error. It is straightforward to represent this sort of repeated computation within the quantum circuit model in such a way that the requirements of the definition of BQP are satisfied.

Simulating Classical Computations with Quantum Circuits

It should not be surprising that quantum computers can efficiently simulate classical computers—for quantum information generalizes classical information, and it would be absurd if there were a loss of computational power in moving to a more general model. This intuition may be confirmed by observing the containment $\text{BPP} \subseteq \text{BQP}$. Here an advantage of working with the general quantum circuit model arises: for if one truly believes the Universality Theorem, there is almost nothing to prove.

Observe first that the complexity class P may be defined in terms of Boolean circuit families in a similar manner to BQP. In particular, a given promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in P if and only if there exists a polynomial-time generated family $C = \{C_n : n \in \mathbb{N}\}$ of Boolean circuits, where each circuit C_n takes n input bits and outputs 1 bit, such that

1. $C(x) = 1$ for all $x \in A_{\text{yes}}$, and
2. $C(x) = 0$ for all $x \in A_{\text{no}}$.

A Boolean circuit-based definition of BPP may be given along similar lines: a given promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in BPP if and only if there exists a polynomial-bounded function r and a polynomial-time generated family $C = \{C_n : n \in \mathbb{N}\}$ of Boolean circuits, where each circuit C_n takes $n + r(n)$ input bits and outputs 1 bit, such that

1. $\Pr[C(x, y) = 1] \geq 2/3$ for all $x \in A_{\text{yes}}$, and
2. $\Pr[C(x, y) = 1] \leq 1/3$ for all $x \in A_{\text{no}}$,

where $y \in \Sigma^{r(|x|)}$ is chosen uniformly at random in both cases.

In both definitions, the circuit family C includes circuits composed of constant-size Boolean logic gates—which for the sake of brevity may be assumed to be composed of NAND gates and FANOUT gates. (FANOUT operations must be modeled as gates for the sake of the simulation.) For the randomized case, it may be viewed that the random bits $y \in \Sigma^{r(|x|)}$ are produced by gates that take no input and output a single uniform

random bit. As NAND gates, FANOUT gates, and random bits are easily implemented with quantum gates, as illustrated in Fig. 7, the circuit family C can be simulated gate-by-gate to obtain a quantum circuit family $Q = \{Q_n : n \in \mathbb{N}\}$ for A that satisfies the definition of BQP. It follows that BPP $\subseteq$ BQP.

The BQP Subroutine Theorem

There is an important issue regarding the above definition of BQP, which is that it is not an inherently "clean" definition with respect to the modularization of algorithms. The *BQP subroutine theorem* of Bennett, Brassard, Bernstein and Vazirani [27] addresses this issue.

Suppose that it is established that a particular promise problem A is in BQP, which by definition means that there must exist an efficient quantum algorithm (represented by a family of quantum circuits) for A. It is then natural to consider the use of that algorithm as a subroutine in other quantum algorithms for more complicated problems, and one would like to be able to do this without worrying about the specifics of the original algorithm. Ideally, the algorithm for A should function as an oracle for A, as defined in Subsect. "Oracles in the Quantum Circuit Model".

A problem arises, however, when queries to an algorithm for A are made in superposition. Whereas it is quite common and useful to consider quantum algorithms that query oracles in superposition, a given BQP algorithm for A is only guaranteed to work correctly on classical inputs. It could be, for instance, that some algorithm for A begins by applying phase-damping gates to all of its input qubits, or perhaps this happens inadvertently as a result of the computation. Perhaps it is too much to ask that the existence of a BQP algorithm for A admits a subroutine having the characteristics of an oracle for A?

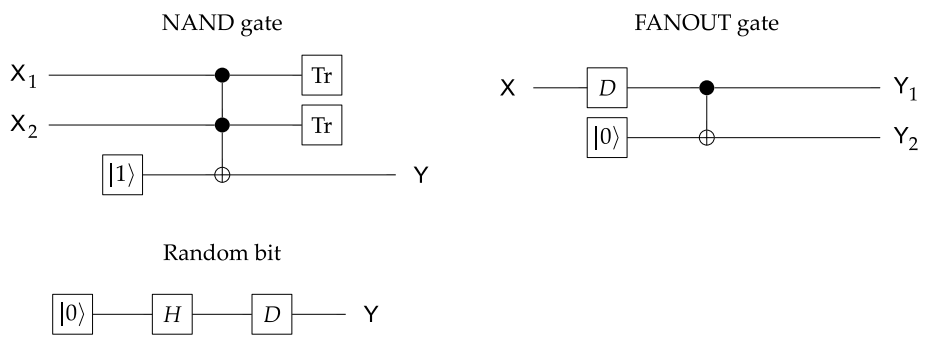
The BQP subroutine theorem establishes that, up to exponentially small error, this is not too much to ask: the existence of an arbitrary BQP algorithm for A implies the existence of a "clean" subroutine for A with the characteristics of an oracle. A precise statement of the theorem follows.

Theorem 4 (BQP subroutine theorem) *Suppose $A = (A_{\text{yes}}, A_{\text{no}})$ is a promise problem in BQP. Then for any choice of a polynomial-bounded function p there exists a polynomial-bounded function q and a polynomial-time generated family of unitary quantum circuits $\{R_n : n \in \mathbb{N}\}$ with the following properties:*

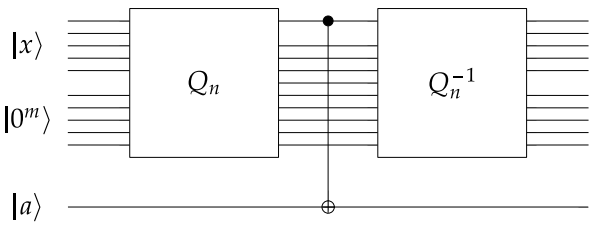
1. *Each circuit R_n implements a unitary operation U_n on $n + q(n) + 1$ qubits.*
2. *For every $x \in A_{\text{yes}}$ and $a \in \Sigma$ it holds that*
$$\langle x, 0^m, a \oplus 1 | U_n | x, 0^m, a \rangle \geq 1 - 2^{-p(n)}$$
for $n = |x|$ and $m = q(n)$.
3. *For every $x \in A_{\text{no}}$ and $a \in \Sigma$ it holds that*
$$\langle x, 0^m, a | U_n | x, 0^m, a \rangle \geq 1 - 2^{-p(n)}$$
for $n = |x|$ and $m = q(n)$.

The proof of this theorem is remarkably simple: given a BQP algorithm for A, one first uses Proposition 3 to obtain a circuit family Q having exponentially small error for A. The circuit illustrated in Fig. 8 then implements a unitary operation with the desired properties. This is essentially a bounded-error quantum adaptation of a classic construction that allows arbitrary deterministic computations to be performed reversibly [26,92].

The following corollary expresses the main implication of the BQP subroutine theorem in complexity-theoretic terms.



Quantum Computational Complexity, Figure 7
Quantum circuit implementations of a NAND gate, a FANOUT gate, and a random bit. The phase-damping gates, denoted by *D*, are only included for aesthetic reasons: they force the purely classical behavior that would be expected of classical gates, but are not required for the quantum simulation of BPP



Quantum Computational Complexity, Figure 8
A unitary quantum circuit approximating an oracle-gate implementation of a BQP computation. Here Q_n is a unitary purification of a circuit having exponentially small error for some problem in BQP

Corollary 5 $\mathrm{BQP}^{\mathrm{BQP}} = \mathrm{BQP}$.

Classical Upper Bounds on BQP

There is no known way to efficiently simulate quantum computers with classical computers—and there would be little point in seeking to build quantum computers if there were. Nevertheless, some insight into the limitations of quantum computers may be gained by establishing containments of BQP in the smallest classical complexity classes where this is possible.

The strongest containment known at this time is given by *counting complexity*. Counting complexity began with Valiant's work [94] on the complexity of computing the permanent, and was further developed and applied in many papers (including [22,47,91], among many others).

The basic notion of counting complexity that is relevant to this article is as follows. Given a polynomial-time nondeterministic Turing machine M and input string $x \in \Sigma^*$, one denotes by $\#M(x)$ the number of *accepting* computation paths of M on x, and by $\#\overline{M}(x)$ the number of *rejecting* computation paths of M on x. A function $f\colon \Sigma^* \to \mathbb{Z}$ is then said to be a *GapP function* if there exists a polynomial-time nondeterministic Turing machine M such that $f(x) = \#M(x) - \#\overline{M}(x)$ for all $x \in \Sigma^*$.

A variety of complexity classes can be specified in terms of GapP functions. For example, a promise problem $A = (A_{\mathrm{yes}}, A_{\mathrm{no}})$ is in PP if and only if there exists a function $f \in \mathrm{GapP}$ such that $f(x) > 0$ for all $x \in A_{\mathrm{yes}}$ and $f(x) \leq 0$ for all $x \in A_{\mathrm{no}}$. The remarkable closure properties of GapP functions allows many interesting facts to be proved about such classes. Fortnow's survey [50] on counting complexity explains many of these properties and gives several applications of the theory of GapP functions. The following closure property is used below.

GapP–multiplication of matrices. Let $p, q\colon \mathbb{N} \to \mathbb{N}$ be polynomial-bounded functions. Suppose that for each $n \in \mathbb{N}$ a sequence of $p(n)$ complex-valued matrices is given

$$A_{n,1},\ A_{n,2},\ \ldots,\ A_{n,p(n)}\ ,$$

each having rows and columns indexed by strings in $\Sigma^{q(n)}$. Suppose further that there exist functions $f, g \in \mathrm{GapP}$ such that

$$f(1^n, 1^k, x, y) = \mathrm{Re}\left(A_{n,k}[x,y]\right)$$
$$g(1^n, 1^k, x, y) = \mathrm{Im}\left(A_{n,k}[x,y]\right)$$

for all $n \in \mathbb{N}$, $k \in \{1, \ldots, p(n)\}$, and $x, y \in \Sigma^{q(n)}$. Then there exist functions $F, G \in \mathrm{GapP}$ such that

$$F(1^n, x, y) = \mathrm{Re}\left((A_{n,p(n)} \cdots A_{n,1})[x,y]\right)$$
$$G(1^n, x, y) = \mathrm{Im}\left((A_{n,p(n)} \cdots A_{n,1})[x,y]\right)$$

for all $n \in \mathbb{N}$ and $x, y \in \Sigma^{q(n)}$. In other words, if there exist two GapP functions describing the real and imaginary parts of the entries of a polynomial sequence of matrices, then the same is true of the product of these matrices.

Now, suppose that $Q = \{Q_n : n \in \mathbb{N}\}$ is a polynomial-time generated family of quantum circuits. For each $n \in \mathbb{N}$, let us assume the quantum circuit Q_n consists of gates $G_{n,1}, \ldots, G_{n,p(n)}$ for some polynomial bounded function p, labeled in an order that respects the topology of the circuit. By tensoring these gates with the identity operation on qubits they do not touch, and using the natural matrix representation of quantum operations, it is possible to obtain matrices $M_{n,1}, \ldots, M_{n,p(n)}$ with the property that

$$M_{n,p(n)} \cdots M_{n,1} \mathrm{vec}(\rho) = \mathrm{vec}(Q_n(\rho))$$

for every possible input density matrix ρ to the circuit. The probability that Q_n accepts a given input x is then

$$\langle 1, 1 | \left(M_{n,p(n)} \cdots M_{n,1}\right) |x, x\rangle\ ,$$

which is a single entry in the product of the matrices.

By padding matrices with rows and columns of zeroes, it may be assumed that each matrix $M_{n,k}$ has rows and columns indexed by strings of length $q(n)$. The assumption that the family Q is polynomial-time generated then allows one to easily conclude that there exist GapP functions f and g so that

$$f(1^n, 1^k, x, y) = \frac{1}{2}\mathrm{Re}\left(A_{n,k}[x,y]\right)$$
$$g(1^n, 1^k, x, y) = \frac{1}{2}\mathrm{Im}\left(A_{n,k}[x,y]\right)$$

for all $n \in \mathbb{N}$, $k \in \{1, \ldots, p(n)\}$, and $x, y \in \Sigma^{q(n)}$. (Note that his fact makes use of the observation that the natural

matrix representation of all of the gates listed in Subsect. "A Finite Universal Gate Set" have entries whose real and imaginary parts come from the set $\{-1, -1/2, 0, 1/2, 1\}$. The numbers $\pm 1/2$ are only needed for the Hadamard gates.) By the property of GapP functions above, it follows that there exists a GapP function F such that

$$\Pr[Q \text{ accepts } x] = \frac{F(x)}{2^{p(|x|)}} .$$

The containment BQP $\subseteq$ PP follows easily. This containment was first proved by Adleman, DeMarrais, and Huang [7] using a different method, and was first argued using counting complexity along similar lines to the above proof by Fortnow and Rogers [51]. There are two known ways that this upper bound can be improved. Using the fact that BQP algorithms have bounded error, Fortnow and Rogers [51] proved that BQP is contained in a (somewhat obscure) counting complexity class called AWPP, which implies the following theorem.

Theorem 6 $\mathrm{PP}^{\mathrm{BQP}} = \mathrm{PP}$.

Another improvement comes from the observation that the above proof that BQP $\subseteq$ PP makes no use of the bounded-error assumption of BQP. It follows that an unbounded error variant of BQP is equal to PP.

PQP Let $A = (A_{\mathrm{yes}}, A_{\mathrm{no}})$ be a promise problem. Then $A \in \mathrm{PQP}$ if and only if there exists a polynomial-time generated family of quantum circuits $Q = \{Q_n : n \in \mathbb{N}\}$, where each circuit Q_n takes n input qubits and produces one output qubit, that satisfies the following properties. If $x \in A_{\mathrm{yes}}$ then $\Pr[Q \text{ accepts } x] > 1/2$; and if $x \in A_{\mathrm{no}}$ then $\Pr[Q \text{ accepts } x] \leq 1/2$.

Theorem 7 $\mathrm{PQP} = \mathrm{PP}$.

Oracle Results Involving BQP

It is difficult to prove separations among quantum complexity classes for apparently the same reasons that this is so for classical complexity classes. For instance, one clearly cannot hope to prove BPP $\neq$ BQP when major collapses such as NC = PP or P = PSPACE are still not disproved. In some cases, however, separations among quantum complexity classes can be proved in relativized settings, meaning that the separation holds in the presence of some cleverly defined oracle. The following oracle results are among several that are known:

1. There exists an oracle A such that $\mathrm{BQP}^A \not\subseteq \mathrm{MA}^A$ [18,30,96]. Such an oracle A intuitively encodes a problem that is solvable using a quantum computer but is not even efficiently verifiable with a classical computer. As the containment BPP $\subseteq$ BQP holds relative to every oracle, this implies that $\mathrm{BPP}^A \subsetneq \mathrm{BQP}^A$ for this particular choice of A.
2. There is an oracle A such that $\mathrm{NP}^A \not\subseteq \mathrm{BQP}^A$ [27]. In less formal terms: a quantum computer cannot find a needle in an exponentially large haystack in polynomial time. This result formalizes a critically important idea, which is that a quantum computer can only solve a given search problem efficiently if it is able to exploit that problem's structure. It is easy to define an oracle search problem represented by A that has no structure whatsoever, which allows the conclusion $\mathrm{NP}^A \not\subseteq \mathrm{BQP}^A$ to be drawn. It is not currently known whether the NP-complete problems, in the absence of an oracle, have enough structure to be solved efficiently by quantum computers; but there is little hope and no indication whatsoever that this should be so. It is therefore a widely believed conjecture that NP $\not\subseteq$ BQP.
3. There is an oracle A such that $\mathrm{SZK}^A \not\subseteq \mathrm{BQP}^A$ [1,5]. This result is similar in spirit to the previous one, but is technically more difficult and rules out the existence of quantum algorithms for unstructured collision detection problems. The graph isomorphism problem and various problems that arise in cryptography are examples of collision detection problems. Once again, it follows that quantum algorithms can only solve such problems if their structure is exploited. It is a major open question in quantum computing whether the graph isomorphism problem is in BQP.

Quantum Proofs

There are many quantum complexity classes of interest beyond BQP. This section concerns one such class, which is a quantum computational analogue of NP. The class is known as QMA, short for *quantum Merlin–Arthur*, and is based on the notion of a *quantum proof*: a quantum state that plays the role of a certificate or witness to a quantum computer that functions as a verification procedure. Interest in both the class QMA and the general notion of quantum proofs is primarily based on the fundamental importance of *efficient verification* in computational complexity. The notion of a quantum proof was first proposed by Knill [69] and consider more formally by Kitaev (presented at a talk in 1999 [65] and later published in [67]).

Definition of QMA

The definition of QMA is inspired by the standard definition of NP included in Subsect. "Computational Complex-

ity" of this article. This definition is of course equivalent to the other well-known definition of NP based on nondeterministic Turing machines, but is much better-suited to consideration in the quantum setting—for nondeterminism is arguably a non-physical notion that does not naturally extend to quantum computing. In the definition of NP from Subsect. "Computational Complexity", the machine M functions as a *verification procedure* that treats each possible string $y \in \Sigma^{p(|x|)}$ as a potential *proof* that $x \in A_{\text{yes}}$. The conditions on M are known as the *completeness* and *soundness* conditions, which derive their names from logic: completeness refers to the condition that true statements have proofs, while soundness refers to the condition that false statements do not.

To define QMA, the set of possible proofs is extended to include quantum states, which of course presumes that the verification procedure is quantum. As quantum computations are inherently probabilistic, a bounded probability of error is allowed in the completeness and soundness conditions. (This is why the class is called QMA rather than QNP, as it is really MA and not NP that is the classical analogue of QMA.)

QMA Let $A = (A_{\text{yes}}, A_{\text{no}})$ be a promise problem, let p be a polynomial-bounded function, and let $a, b\colon \mathbb{N} \to [0,1]$ be functions. Then $A \in \text{QMA}_p(a,b)$ if and only if there exists a polynomial-time generated family of circuits $Q = \{Q_n : n \in \mathbb{N}\}$, where each circuit Q_n takes $n + p(n)$ input qubits and produces one output qubit, with the following properties:

1. *Completeness.* For all $x \in A_{\text{yes}}$, there exists a $p(|x|)$-qubit quantum state ρ such that $\Pr[\text{Q accepts } (x, \rho)] \geq a(|x|)$.
2. *Soundness.* For all $x \in A_{\text{no}}$ and all $p(|x|)$-qubit quantum states ρ it holds that $\Pr[\text{Q accepts } (x, \rho)] \leq b(|x|)$.

Also define $\text{QMA} = \bigcup_p \text{QMA}_p(2/3, 1/3)$, where the union is over all polynomial-bounded functions p.

Problems in QMA

Before discussing the general properties of QMA that are known, it is appropriate to mention some examples of problems in this class. Of course it is clear that thousands of interesting combinatorial problems are in QMA, as the containment $\text{NP} \subseteq \text{QMA}$ is trivial. What is more interesting is the identification of problems in QMA that are not known to be in NP, for these are examples that provide insight into the power of quantum proofs. There are presently just a handful of known problems in QMA that are not known to be in NP, but the list is growing.

The Local Hamiltonian Problem Historically speaking, the first problem identified to be complete for QMA was the *local Hamiltonian problem* [65,67], which can be seen as a quantum analogue of the MAX-k-SAT problem. Its proof of completeness can roughly be seen as a quantum analogue of the proof of the Cook–Levin Theorem [38,72]. The proof has subsequently been strengthened to achieve better parameters [61].

Suppose that M is a Hermitian matrix whose rows and columns are indexed by strings of length n for some integer $n \geq 1$. Then M is said to be *k-local* if and only if it can be expressed as

$$M = P_\pi (A \otimes I) P_\pi^{-1}$$

for an arbitrary matrix A indexed by Σ^k, P_π a permutation matrix defined by

$$P_\pi |x_1 \cdots x_n\rangle = |x_{\pi(1)} \cdots x_{\pi(n)}\rangle$$

for some permutation $\pi \in S_n$, and I denoting the identity matrix indexed by Σ^{n-k}. In less formal terms, M is a matrix that arises from a "gate" on k qubits, but where the gate is described by a $2^k \times 2^k$ Hermitian matrix A rather than a unitary matrix. It is possible to express such a matrix compactly by specifying A along with the bit-positions on which A acts.

Intuitively, a k-local matrix assigns a real number (representing the *energy*) to any quantum state on n qubits. This number depends only on the reduced state of the k qubits where M acts nontrivially, and a limit on this value can be thought of as a local constraint on a given quantum state. Loosely speaking, the k-local Hamiltonian problem asks whether there exists a quantum state that satisfies a collection of such constraints.

The k-local Hamiltonian problem

Input: A collection $H_1, \ldots, H_m$ of k-local Hermitian matrices indexed by strings of length n and satisfying $\| H_j \| \leq 1$ for $j = 1, \ldots, m$.

Yes: There exists an n-qubit quantum state $|\psi\rangle$ such that $\langle\psi|H_1 + \cdots + H_m|\psi\rangle \leq -1$.

No: For every n-qubit quantum state $|\psi\rangle$ it holds that $\langle\psi|H_1 + \cdots + H_m|\psi\rangle \geq 1$.

Theorem 8 *The 2-local Hamiltonian problem is complete for QMA with respect to Karp reductions.*

The completeness of this problem has been shown to be closely related to the universality of the so-called *adiabatic* model of quantum computation [11].

Other Problems Complete for QMA Aside from the local Hamiltonian problem, there are other promise prob-

lems known to be complete for QMA, including the following:

1. Restricted versions of the local Hamiltonian problem. For example, the 2-local Hamiltonian problem remains QMA-complete when the local Hamiltonians are restricted to nearest-neighbor interactions on a two-dimensional array of qubits [81]. The hardness of the local Hamiltonian problem with nearest-neighbor interactions on one-dimensional systems is known to be QMA-complete for 12 dimensional particles in place of qubits [10], but is open for smaller systems including qubits.
2. The *density matrix consistency problem*. Here, the input is a collection of density matrices representing the reduced states of a hypothetical n-qubit state. The input is a yes-instance of the problem if there exists a state of the n-qubit system that is consistent with the given reduced density matrices, and is a no-instance if every state of the n-qubit system has reduced states that have significant disagreement from one or more of the input density matrices. This problem is known to be complete for QMA with respect to Cook reductions [73], but is not known to be complete for Karp reductions.
3. The *quantum clique problem*. Beigi and Shor [23] have proved that a promise version of the following problem is QMA-complete for any constant $k \geq 2$. Given a quantum operation Φ, are there k different inputs to Φ that are perfectly distinguishable after passing through Φ? They name this the quantum clique problem because there is a close connection between computing the zero-error capacity of a classical channel and computing the size of the largest clique in the complement of a graph representing the channel; and this is a quantum variant of that problem.
4. Several problems about quantum circuits. All of the problems mentioned above are proved to be QMA-hard through reductions from the local Hamiltonian problem. Other problems concerning properties of quantum circuits can be proved QMA-complete by more direct means, particularly when the problem itself concerns the existence of a quantum state that causes an input circuit to exhibit a certain behavior. An example in this category is the *non-identity check* problem [59] that asks whether a given unitary circuit implements an operation that is close to some scalar multiple of the identity.

The Group Non-Membership Problem Another example of a problem in QMA that is not known to be in NP is the *group non-membership problem* [17,96]. This problem is quite different from the QMA-complete problems mentioned above in two main respects. First, the problem corresponds to a language, meaning that the promise is vacuous: every possible input can be classified as a yes-instance or no-instance of the problem. (In all of the above problems it is not possible to do this without invalidating the known proofs that the problems are in QMA.) Second, the problem is not known to be complete for QMA; and indeed it would be quite surprising if QMA were shown to have a complete problem having a vacuous promise.

In the group non-membership problem the input is a subgroup H of some finite group G, along with an element $g \in G$. The yes-instances of the problem are those for which $g \notin H$, while the no-instances are those for which $g \in H$.

THE GROUP NON-MEMBERSHIP PROBLEM

Input: Group elements $h_1, \ldots, h_k$ and g from some finite group G. Let $H = \langle h_1, \ldots, h_k \rangle$ be the subgroup generated by $h_1, \ldots, h_k$.

Yes: $g \notin H$.

No: $g \in H$.

Like most group-theoretic computational problems, there are many specific variants of the group non-membership problem that differ in the way that group elements are represented. For example, group elements could be represented by permutations in cycle notation, invertible matrices over a finite field, or any number of other possibilities. The difficulty of the problem clearly varies depending the representation. The framework of *black-box groups*, put forth by Babai and Szemerédi [20], simplifies this issue by assuming that group elements are represented by meaningless labels that can only be multiplied or inverted by means of a *group oracle*. Any algorithm or protocol that solves a problem in this framework can then be adapted to any specific group provided that an efficient implementation of the group operations exists that can replace the group oracle.

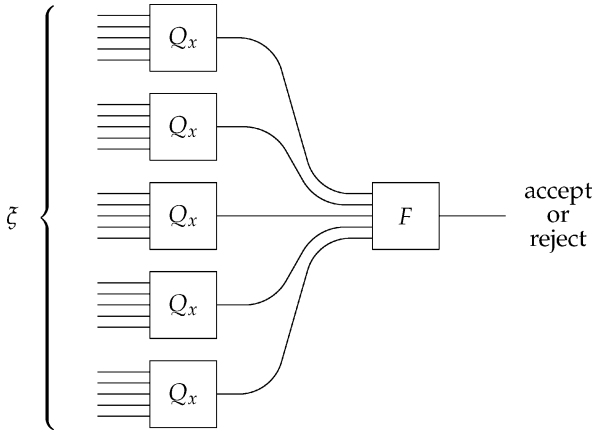
Theorem 9 *The group non-membership problem is in QMA for every choice of a group oracle.*

Error Reduction for QMA

The class QMA is robust with error bounds, which are reflected by the completeness and soundness probabilities, in much the same way that the same is true of BQP. Two specific error reduction methods are presented in this section: *weak error reduction* and *strong error reduction*. The two methods differ with respect to the length of the quantum proof that is needed to obtain a given error bound, which is a unique consideration that arises when working with quantum proofs.

Suppose that $A = (A_{\text{yes}}, A_{\text{no}})$ is a given promise problem for which a QMA-type quantum verification procedure exists. To describe both error reduction procedures, it is convenient to assume that the input to the problem is hard-coded into each circuit in the family representing the verification procedure. The verification procedure therefore takes the form $Q = \{Q_x : x \in \Sigma^*\}$, where each circuit Q_x takes a $p(|x|)$-qubit quantum state as input, for some polynomial-bounded function p representing the quantum proof length, and produces one output qubit. The completeness and soundness probabilities are assumed to be given by polynomial-time computable functions a and b as usual.

The natural approach to error reduction is repetition: for a given input x the verification circuit Q_x is evaluated multiple times, and a decision to accept or reject based on the frequency of accepts and rejects among the outcomes of the repetitions is made. Assuming $a(|x|)$ and $b(|x|)$ are not too close together, this results in a decrease in error that is exponential in the number of repetitions [67]. The problem that arises, however, is that each repetition of the verification procedure apparently destroys the quantum proof it verifies, which necessitates the composite verification procedure receiving $p(|x|)$ qubits for *each* repetition of the original procedure as illustrated in Fig. 9. The form of error reduction implemented by this procedure is called *weak error reduction*, given that the length of the quantum proof must grow as the error decreases. (Of course one may view that it also has a strength, which is that it does not require a significant increase in circuit depth over the original procedure.)



Quantum Computational Complexity, Figure 9
An illustration of the weak error reduction procedure for QMA. The circuits Q_x represent the original verification procedure, and the circuit labeled F outputs acceptance or rejection based on the frequency of accepts among its inputs (which can be adjusted depending on a and b). It cannot be assumed that the input state ξ takes the form of a product state $\rho \otimes \cdots \otimes \rho$; but it is not difficult to prove that there will always be at least one such input state among the states maximizing the acceptance probability of the procedure

The second error reduction procedure for QMA is due to Marriott and Watrous [76]. Like weak error reduction it gives an exponential reduction in error for roughly a linear increase in circuit size, but has the advantage that it does not require any increase in the length of the quantum proof. This form of error reduction is called *strong error reduction* because of this advantage, which turns out to be quite handy in some situations. The procedure is illustrated in Fig. 10. The following theorem follows from an analysis of this procedure.

Theorem 10 *Suppose that $a, b : \mathbb{N} \to [0, 1]$ are polynomial-time computable functions and $q : \mathbb{N} \to \mathbb{N}$ is a polynomial-bounded function such that $a(n) - b(n) \geq 1/q(n)$ for all but finitely many $n \in \mathbb{N}$. Then for every choice of polynomial-bounded functions $p, r : \mathbb{N} \to \mathbb{N}$ such that $r(n) \geq 2$ for all but finitely many n, it holds that* $\text{QMA}_p(a, b) = \text{QMA}_p\left(1 - 2^{-r}, 2^{-r}\right)$.

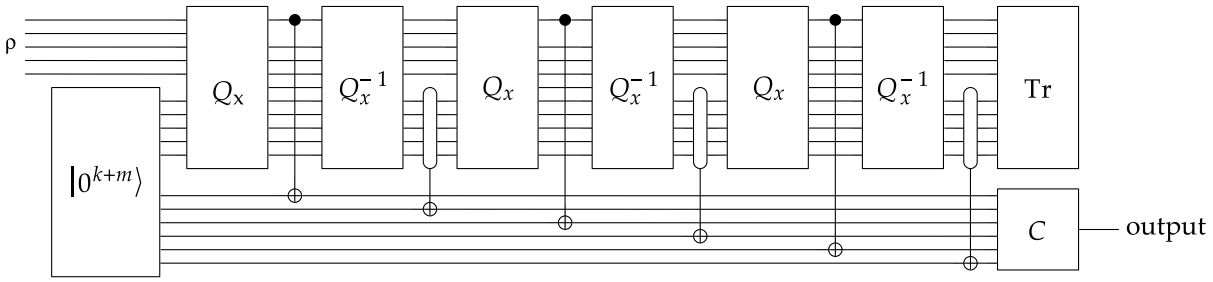
Containment of QMA in PP

By means of strong error reduction for QMA, it can be proved that QMA is contained in PP as follows. Suppose that some promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is contained in QMA. Then by Theorem 10 it holds that

$$A \in \text{QMA}_p\left(1 - 2^{-(p+2)}, 2^{-(p+2)}\right) \tag{2}$$

for some polynomial-bounded function p. What is important here is that the soundness probability is smaller than the reciprocal of the dimension of the space corresponding to the quantum proof, which is where strong error reduction is essential.

Now, consider an algorithm that does not receive any quantum proof, but instead just randomly guesses a quantum proof on p qubits and feeds this proof into a verification procedure having completeness and soundness probabilities consistent with the above inclusion (2). To be more precise, the quantum proof is substituted by the totally mixed state on p qubits. A simple analysis shows that this algorithm accepts every string $x \in A_{\text{yes}}$ with probability at least $2^{-(p(|x|)+1)}$ and accepts every string $x \in A_{\text{no}}$ with probability at most $2^{-(p(|x|)+2)}$. The gap between these two probabilities is enough to establish that $A \in \text{PQP} = \text{PP}$.



Quantum Computational Complexity, Figure 10
Illustration of the strong error reduction procedure for QMA. The circuit Q_x represents a unitary purification of a given QMA verification procedure on an input x, while the circuit C determines whether to accept or reject based on the number of alternations of its input qubits. The quantum proof is denoted by ρ

Classical Proofs for Quantum Verification Procedures

One may also consider quantum verification procedures that receive classical proofs rather than quantum proofs. Aharonov and Naveh [8] defined a complexity class MQA accordingly.

MQA Let $A = (A_{\text{yes}}, A_{\text{no}})$ be a promise problem. Then $A \in \text{MQA}$ if and only if there exists a polynomial-bounded function p and a polynomial-time generated family of circuits $Q = \{Q_n : n \in \mathbb{N}\}$, where each circuit Q_n takes $n + p(n)$ input qubits and produces one output qubit, with the following properties. For all $x \in A_{\text{yes}}$, there exists a string $y \in \Sigma^{p(|x|)}$ such that $\Pr[Q \text{ accepts } (x, y)] \geq 2/3$; and for all $x \in A_{\text{no}}$ and all strings $y \in \Sigma^{p(|x|)}$ it holds that $\Pr[Q \text{ accepts } (x, y)] \leq 1/3$.

(This class was originally named QCMA, and is more commonly known by that name—but it is impossible to resist the urge to give this interesting class a more sensible name.)

When considering the power of quantum proofs, it is the question of whether MQA is properly contained in QMA that arguably cuts to the heart of the issue. Aaronson and Kuperberg [4] studied this question, and based on a reasonable group-theoretic conjecture argued that the group non-membership problem is likely to be in MQA.

Are Two Quantum Proofs Better than One?

An unusual, yet intriguing question about quantum proofs was asked by Kobayashi, Matsumoto, and Yamakami [71]: *are two quantum proofs better than one?* The implicit setting in this question is similar to that of QMA, except that the verification procedure receives *two* quantum proofs that are *guaranteed to be unentangled*. The complexity class QMA(2) is defined to be the class of promise problems having such two-proof systems with completeness and soundness probabilities 2/3 and 1/3, respectively. (It is not known to what extent this type of proof system is robust with respect to error bounds, so it is conceivable that other reasonable choices of error bounds could give distinct complexity classes.)

The restriction on entanglement that is present in the definition of QMA(2) may seem artificial from a physical perspective, as there is no obvious mechanism that would prevent two entities with otherwise unlimited computational power from sharing entanglement. Nevertheless, the question of the power of QMA(2) is interesting in that it turns around the familiar question in quantum computing about the computational power of entanglement. Rather than asking if computational power is limited by a constraint on entanglement, here the question is whether a constraint on entanglement enhances computational power. It is clear that two quantum proofs are no less powerful than one, as a verification procedure may simply ignore one of the proofs—which means that $\text{QMA} \subseteq \text{QMA}(2)$. Good upper bounds on QMA(2), however, are not known: the best upper bound presently known is $\text{QMA}(2) \subseteq \text{NEXP}$, which follows easily from the fact that a nondeterministic exponential-time algorithm can guess explicit descriptions of the quantum proofs and then perform a deterministic exponential-time simulation of the verification procedure.

Quantum Interactive Proof Systems

The notion of efficient proof verification is generalized by means of the *interactive proof system* model, which has fundamental importance in complexity theory and theoretical cryptography. Interactive proof systems, or *interactive proofs* for short, were first introduced by Babai [17,19] and Goldwasser, Micali, and Rackoff [55,56], and have led

to remarkable discoveries in complexity such as the *PCP Theorem* [14,15,42]. In the most commonly studied variant of interactive proof systems, a polynomial-time *verifier* interacts with a computationally unbounded *prover* that tries to convince the verifier of the truth of some statement. The prover is not trustworthy, however, and so the verifier must be specified in such a way that it is not convinced that false statements are true.

Quantum interactive proof systems [66,99] are interactive proof systems in which the prover and verifier may exchange and process quantum information. This ability of both the prover and verifier to exchange and process quantum information endows quantum interactive proofs with interesting properties that distinguish them from classical interactive proofs, and illustrate unique features of quantum information.

Definition of Quantum Interactive Proofs and The Class QIP

A quantum interactive proof system involves an interaction between a prover and verifier as suggested by Fig. 11.

The verifier is described by a polynomial-time generated family

$$V = \left\{V_j^n : n \in \mathbb{N},\ \ j \in \{1, \ldots, p(n)\}\right\}$$

of quantum circuits, for some polynomial-bounded function p. On an input string x having length n, the sequence of circuits

$$V_1^n, \ldots, V_{p(n)}^n$$

determine the actions of the verifier over the course of the interaction, with the value $p(n)$ representing the number of *turns* the verifier takes. For instance, $p(n) = 4$ in the example depicted in Fig. 11. The inputs and outputs of the verifier's circuits are divided into two categories: private *memory* qubits and *message* qubits. The message qubits are sent to, or received from, the prover, while the memory qubits are retained by the verifier as illustrated in the figure. It is always assumed that the verifier *receives* the last message; so if the number of messages is to be $m = m(n)$, then it must be the case that $p(n) = \lfloor m/2 \rfloor + 1$.

The prover is defined in a similar way to the verifier, although no computational assumptions are made: the prover is a family of arbitrary quantum operations

$$P = \{P_j^n : n \in \mathbb{N},\ \ j \in \{1, \ldots, q(n)\}\}$$

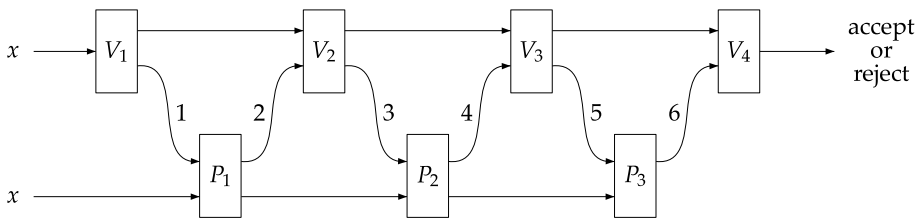
that interface with a given verifier in the natural way. Again, this is as suggested by Fig. 11.

Now, on a given input string x, the prover P and verifier V have an interaction by composing their circuits as described above, after which the verifier measures an output qubit to determine acceptance or rejection. In direct analogy to the classes NP, MA, and QMA, one defines quantum complexity classes based on the completeness and soundness properties of such interactions.

QIP Let $A = (A_{\text{yes}}, A_{\text{no}})$ be a promise problem, let m be a polynomial-bounded function, and let $a, b : \mathbb{N} \to [0, 1]$ be polynomial-time computable functions. Then $A \in \text{QIP}(m, a, b)$ if and only if there exists an m-message quantum verifier V with the following properties:

1. *Completeness.* For all $x \in A_{\text{yes}}$, there exists a quantum prover P that causes V to accept x with probability at least $a(|x|)$.
2. *Soundness.* For all $x \in A_{\text{no}}$, every quantum prover P causes V to accept x with probability at most $b(|x|)$.

Also define $\text{QIP}(m) = \text{QIP}(m, 2/3, 1/3)$ for each polynomial-bounded function m and define $\text{QIP} = \bigcup_m \text{QIP}(m)$, where the union is over all polynomial-bounded functions m.



Quantum Computational Complexity, Figure 11
A quantum interactive proof system. There are six messages in this example, labeled 1, . . . , 6. (There may be polynomially many messages in general.) The arrows each represent a collection of qubits, rather than single qubits as in previous figures. The superscript *n* is omitted in the names of the prover and verifier circuits, which can safely be done when the input length *n* is determined by context

Properties of Quantum Interactive Proofs

Classical interactive proofs can trivially be simulated by quantum interactive proofs, and so the containment PSPACE $\subseteq$ QIP follows directly from IP = PSPACE [75, 87]. Unlike classical interactive proofs, quantum interactive proofs are not known to be simulatable in PSPACE. Simulation is possible in EXP through the use of semidefinite programming [66].

Theorem 11 QIP $\subseteq$ EXP.

Quantum interactive proofs, like ordinary classical interactive proofs, are quite robust with respect to the choice of completeness and soundness probabilities. In particular, the following facts hold [66].

1. Every quantum interactive proof can be transformed into an equivalent quantum interactive proof with *perfect completeness*: the completeness probability is 1 and the soundness probability is bounded away from 1. The precise bound obtained for the soundness probability depends on the completeness and soundness probabilities of the original proof system. This transformation to a perfect-completeness proof system comes at the cost of one additional round (i. e., two messages) of communication.
2. Parallel repetition of quantum interactive proofs with perfect completeness gives an exponential reduction in soundness error. An exponential reduction in completeness and soundness error is also possible for quantum interactive proofs not having perfect completeness, but the procedure is more complicated than for the perfect completeness case.

One of the major differences between quantum and classical interactive proof systems is that quantum interactive proofs can be *parallelized*: every quantum interactive proof system, possibly having polynomially many rounds of communication, can be transformed into an equivalent quantum interactive proof system with *three messages* [66]. This transformation comes at the cost of weakening the error bounds. However, it preserves perfect completeness, and the soundness error can subsequently be reduced by parallel repetition without increasing the number of messages beyond three. If classical interactive proof systems could be parallelized in this way, then it would follow that AM = PSPACE; a collapse that would surprise most complexity theorists and have major implications to the theory.

The following theorem summarizes the implications of the facts just discussed to the complexity classes QIP(m, a, b) defined above.

Theorem 12 *Let $a, b\colon \mathbb{N} \to [0,1]$ be polynomial-time computable functions and let p be a polynomial-bounded function such that $a(n) - b(n) \geq 1/p(n)$ for all but finitely many $n \in \mathbb{N}$. Also let m and r be polynomial-bounded functions. Then*

$$\mathrm{QIP}(m, a, b) \subseteq \mathrm{QIP}(3, 1, 2^{-r}) .$$

For a wide range of completeness and soundness probabilities this leaves just four complexity classes among those defined above: QIP(0) = BQP, QIP(1) = QMA, QIP(2), and QIP(3) = QIP.

The final property of quantum interactive proofs that will be discussed in this section is the existence of an interesting complete promise problem [85]. The problem is to determine whether the operations induced by two quantum circuits are significantly different, or are approximately the same, with respect to the same metric on quantum operations discussed in Subsect. "A Finite Universal Gate Set".

THE QUANTUM CIRCUIT DISTINGUISHABILITY PROBLEM

Input: Quantum circuits Q_0 and Q_1, both taking n input qubits and producing m output qubits.
Yes: $\delta(Q_0, Q_1) \geq 2/3$.
No: $\delta(Q_0, Q_1) \leq 1/3$.

Theorem 13 *The quantum circuit distinguishability problem is QIP-complete with respect to Karp reductions.*

Zero-Knowledge Quantum Interactive Proofs

Interactive proof systems can sometimes be made to have a cryptographically motivated property known as the *zero-knowledge* property [56]. Informally speaking, an interactive proof system is zero-knowledge if a verifier "learns nothing" from an interaction with the prover on an input $x \in A_{\mathrm{yes}}$, beyond the fact that it is indeed the case that $x \in A_{\mathrm{yes}}$. This should hold even if the verifier deviates from the actions prescribed to it by the interactive proof being considered. At first this notion seems paradoxical, but nevertheless there are many interesting examples of such proof systems. For an introductory survey on zero-knowledge, see [93].

Several variants of zero-knowledge are often studied in the classical setting that differ in the particular way that the notion of "learning nothing" is formalized. This article will only consider *statistical* zero-knowledge, which is the variant of zero-knowledge that is most easily adapted to the quantum setting.

Suppose that $A = (A_{\mathrm{yes}}, A_{\mathrm{no}})$ is a promise problem, and (V, P) is a quantum interactive proof system for A.

By this it is meant that V is the *honest verifier* and P is the *honest prover* that behave precisely as the proof system specifies. Whether or not this proof system possesses the zero-knowledge property depends on the characteristics of interactions between a *cheating verifier* V' with the honest prover P on inputs $x \in A_{\text{yes}}$. Figure 12 illustrates such an interaction, wherein it is viewed that the cheating verifier V' is using the interaction with P on input x to compute some quantum operation Φ_x. Informally speaking, the input to this operation represents the verifier's state of knowledge before the protocol is run, while the output represents the verifier's state of knowledge after.

Now, the proof system (V, P) is said to be *quantum statistical zero-knowledge* if, for any choice of a polynomial-time cheating verifier V', the quantum operation Φ_x can be efficiently approximated for all $x \in A_{\text{yes}}$. More precisely, the assumption that V' is described by polynomial-time generated quantum circuits must imply that there exists a polynomial-time generated family $\{Q_x : x \in \Sigma^*\}$ of quantum circuits for which $\delta(\Phi_x, Q_x)$ is negligible for every $x \in A_{\text{yes}}$. Intuitively this definition captures the notion of learning nothing—for anything that the cheating verifier could compute in polynomial time with the help of the prover could equally well have been computed in polynomial time without the prover's help.

The class of promise problems having statistical zero-knowledge quantum interactive proofs is denoted QSZK.

QSZK A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in QSZK if and only if it has a statistical zero-knowledge quantum interactive proof system.

Although it has not been proved, it is reasonable to conjecture that QSZK is properly contained in QIP; for the zero-knowledge property seems to be quite restrictive. Indeed, it was only recently established that there exist nontrivial quantum interactive proof systems that are statistical zero-knowledge [100]. The following facts [97,100] are among those known about QSZK.

1. Statistical zero-knowledge quantum interactive proof systems can be parallelized to two messages. It follows that QSZK $\subseteq$ QIP(2).
2. The class QSZK is closed under complementation: a given promise problem A has a quantum statistical zero-knowledge proof if and only if the same is true for the problem obtained by exchanging the yes- and no-instances of A.
3. Statistical zero-knowledge quantum interactive proof systems can be simulated in polynomial space: QSZK $\subseteq$ PSPACE.

Classical analogues to the first and second facts in this list were shown first [86]. (The classical analogue to the third fact is SZK $\subseteq$ PSPACE, which follows trivially from IP $\subseteq$ PSPACE.) A key step toward proving the above properties (which is also similar to the classical case) is to establish that the following promise problem is complete for QSZK. The problem is a restricted version of the QIP-complete quantum circuit distinguishability problem.

The Quantum State Distinguishability Problem

Input: Quantum circuits Q_0 and Q_1, both taking no input qubits and producing m output qubits. Let ρ_0 and ρ_1 be the density matrices corresponding to the outputs of these circuits.

Yes: $\delta(\rho_0, \rho_1) \geq 2/3$.

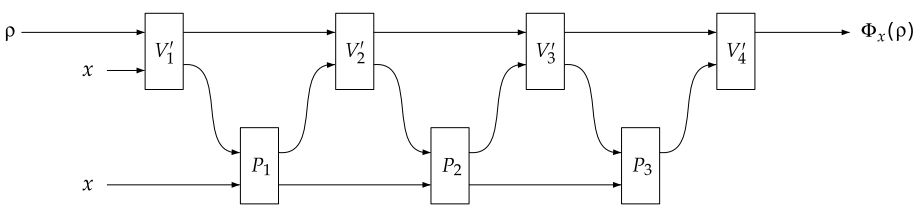
No: $\delta(\rho_0, \rho_1) \leq 1/3$.

Theorem 14 *The quantum state distinguishability problem is QSZK-complete with respect to Karp reductions.*

A quantum analogue of a different SZK-complete problem known as the *entropy difference problem* [53] has recently been shown to be complete for QSZK [16].

Multiple-Prover Quantum Interactive Proofs

Multiple-prover interactive proof systems are variants of interactive proofs where a verifier interacts with two or more provers that are not able to communicate with one another during the course of the interaction. Classical



Quantum Computational Complexity, Figure 12
A cheating verifier V' performs a quantum operation Φ_x with the unintentional help of the honest prover

multiple-prover interactive proofs are extremely powerful: the class of promise problems having multiple-prover interactive proofs (denoted MIP) coincides with NEXP [21]. This is true even for two-prover interactive proofs wherein the verifier exchanges just one round of communication with each prover in parallel [46]. The key to the power of multiple-prover interactive proofs is the inability of the provers to communicate during the execution of the proof system. Similar to a detective interrogating two suspects in separate rooms in a police station, the verifier can ask questions of the provers that require strongly correlated answers, limiting the ability of cheating provers to convince the verifier of a false statement.

The identification of MIP with NEXP assumes a completely classical description of the provers. Quantum information, however, delivers a surprising twist for this model: even when the verifier is classical, an entangled quantum state shared between two provers can allow for non-classical correlations between their answers to the verifier's questions. This phenomenon is better known as a Bell-inequality violation [24] in the quantum physics literature. Indeed, there exist two-prover interactive proof systems that are sound against classical provers, but can be cheated by entangled quantum provers [36].

MIP* A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in MIP* if and only if there exists a multiple-prover interactive proof system for A wherein the verifier is classical and the provers may share an arbitrary entangled state.

One may also consider fully quantum variants of multiple-prover interactive proofs, which were first studied by Kobayashi and Matsumoto [70].

QMIP A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in QMIP if and only if there exists a multiple-prover quantum interactive proof system for A.

Various refinements on these classes have been studied, where parameters such as the number of provers, number of rounds of communication, completeness and soundness probabilities, and bounds on the amount of entanglement shared between provers are taken into account.

The results proved in both [36] and [70] support the claim that it is entanglement shared between provers that is the key issue in multiple-prover quantum interactive proofs. Both models are equivalent in power to MIP when provers are forbidden to share entanglement before the proof system is executed.

At the time of the writing of this article, research into these complexity classes and general properties of multiple-prover quantum interactive proofs is highly active and has led to several interesting results (such as [37,62,63], among others). Despite this effort, little can be said at this time about the relationship among the above classes and other known complexity classes. For instance, only the trivial lower bounds PSPACE $\subseteq$ MIP* and QIP $\subseteq$ QMIP, and no good upper bounds, are known. It has not even been ruled out that non-recursive languages could have multiple-prover quantum interactive proof systems. These difficulties seem to stem from two issues: (i) no bounds are known for the size of entangled states needed for provers to perform optimally, or nearly optimally, in an interactive proof, and (ii) the possible correlations that can be induced with entanglement are not well-understood.

One highly restricted variant of MIP* that has been studied, along with its unentangled counterpart, is as follows.

$\oplus$MIP* A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in $\oplus$MIP* if and only if there exists a one-round two-prover interactive proof system for A wherein the provers each send a single bit to the verifier, and the verifier's decision to accept or reject is determined by the questions asked along with the XOR of these bits. The verifier is classical and the provers are quantum and share an arbitrary entangled state.

$\oplus$MIP This class is similar to $\oplus$MIP*, except that the provers may not share entanglement (and therefore can be assumed to be classical without loss of generality).

It holds that $\oplus$MIP = NEXP for some choices of completeness and soundness probabilities [25,58]. On the other hand, it has been proved [101] that $\oplus$MIP* $\subseteq$ QIP(2) and therefore $\oplus$MIP* $\subseteq$ EXP.

Other Variants of Quantum Interactive Proofs

Other variants of quantum interactive proof systems have been studied, including *public-coin* quantum interactive proofs and quantum interactive proofs with *competing provers*.

Public-Coin Quantum Interactive Proofs Public-coin interactive proof systems are a variant of interactive proofs wherein the verifier performs no computations until all messages with the prover have been exchanged. In place of the verifier's messages are sequences of random bits, visible to both the prover and verifier. Such interactive proof

systems are typically called *Arthur–Merlin games* [17,19], and the verifier and prover are called Arthur and Merlin, respectively, in this setting.

Quantum variants of such proof systems and their corresponding classes are easily defined. For instance, QAM is defined as follows [76].

QAM A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in QAM if and only if it has a quantum interactive proof system of the following restricted form. Arthur uniformly chooses some polynomial-bounded number of classical random bits, and sends a copy of these bits to Merlin. Merlin responds with a quantum state on a polynomial-bounded number of qubits. Arthur then performs a polynomial-time quantum computation on the input, the random bits, and the state sent by Merlin to determine acceptance or rejection.

It is clear that QAM $\subseteq$ QIP(2), but unlike the classical case [54] equality is not known in the quantum setting. It is straightforward to prove the upper bound QAM $\subseteq$ PSPACE, while containment QIP(2) $\subseteq$ PSPACE is not known.

The class QMAM may be defined through a similar analogy with classical Arthur–Merlin games. Here, Merlin sends the first message, Arthur responds with a selection of random bits, and then Merlin sends a second message.

QMAM A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in QMAM if and only if it has a quantum interactive proof system of the following restricted form. Merlin sends a quantum state on a polynomial-bounded number of qubits to Arthur. Without performing any computations, Arthur uniformly chooses some polynomial-bounded number of classical random bits, and sends a copy of these bits to Merlin. Merlin responds with a second quantum state. Arthur then performs a polynomial-time quantum computation on the input, the random bits, and the two states sent by Merlin in order to determine acceptance or rejection.

Even when Arthur is restricted to a single random bit, this class has the full power of QIP [76].

Theorem 15 QMAM = QIP.

Quantum Interactive Proofs with Competing Provers Another variant of interactive proof systems that has been studied is one where a verifier interacts with two *competing* provers, sometimes called the *yes-prover* and the *no-prover*. Unlike ordinary two-prover interactive proof systems, it is now assumed that the provers have conflicting goals: the yes-prover wants to convince the verifier that a given input string is a yes-instance of the problem being considered, while the no-prover wants to convince the verifier that the input is a no-instance. The verifier is sometimes called the *referee* in this setting, given that interactive proof systems of this form are naturally modeled as competitive games between the two provers. Two complexity classes based on interactive proofs with competing provers are the following.

RG A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in RG (short for *refereed games*) if and only if it has a classical interactive proof system with two competing provers. The completeness and soundness conditions for such a proof system are replaced by the following conditions:

1. For every $x \in A_{\text{yes}}$, there exists a yes-prover P_{yes} that convinces the referee to *accept* with probability at least 2/3, regardless of the strategy employed by the no-prover P_{no}.
2. For every $x \in A_{\text{no}}$, there exists a no-prover P_{no} that convinces the referee to *reject* with probability at least 2/3, regardless of the strategy employed by the yes-prover P_{yes}.

QRG A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in QRG (*quantum refereed games*) if and only if it has a quantum interactive proof system with two competing provers. The completeness and soundness conditions for such a proof system are analogous to RG.

Classical refereed games have the computational power of deterministic exponential time [45], and the same is true in the quantum setting [57]: RG = QRG = EXP. The containment EXP $\subseteq$ RG represents an application of the arithmetization technique, while QRG $\subseteq$ EXP exemplifies the power of semidefinite programming.

Other Selected Notions in Quantum Complexity

In this section of this article, a few other topics in quantum computational complexity theory are surveyed that do not fall under the headings of the previous sections. While incomplete, this selection should provide the reader with some sense for the topics that have been studied in quantum complexity.

Quantum Advice

Quantum advice is a formal abstraction that addresses this question: how powerful is quantum software? More precisely, let us suppose that some polynomial-time generated

family of quantum circuits $Q = \{Q_n : n \in \mathbb{N}\}$ is given. Rather than assuming that each circuit Q_n takes n input qubits as for the class BQP, however, it is now assumed that Q_n takes $n + p(n)$ qubits for p some polynomial-bounded function: in addition to a given input $x \in \Sigma^n$, the circuit Q_n will take an *advice state* ρ_n on $p(n)$ qubits. This advice state may be viewed as pre-loaded quantum software for a quantum computer. The advice state may depend on the input length n, but not on the particular input $x \in \Sigma^n$. Similar to a quantum proof, the difficulty of preparing this state is ignored; but unlike a quantum proof the advice state is completely trusted. The quantum complexity class BQP/qpoly is now defined to be the class of promise problems that are solved by polynomial-time quantum algorithms with quantum advice in the natural way. The first published paper on quantum advice was [80], while the definitions and most of the results discussed in this section are due to Aaronson [2].

BQP/qpoly A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in BQP/qpoly if and only if there exists a polynomial-bounded function p, a collection of quantum states $\{\rho_n : n \in \mathbb{N}\}$ where each ρ_n is a $p(n)$-qubit state, and a polynomial-time generated family of quantum circuits $Q = \{Q_n : n \in \mathbb{N}\}$ with the following properties. For all $x \in A_{\text{yes}}$, Q accepts $(x, \rho_{|x|})$ with probability at least 2/3; and for all $x \in A_{\text{no}}$, Q accepts $(x, \rho_{|x|})$ with probability at most 1/3.

Similar to BQP without advice, it is straightforward to shown that the constants 2/3 and 1/3 in this definition can be replaced by a wide range of functions $a, b : \mathbb{N} \to [0, 1]$.

As the notation suggests, BQP/qpoly is a quantum analogue of the class P/poly. This analogy may be used as the basis for several relevant points about BQP/qpoly, quantum advice, and their relationship to classical complexity classes and notions.

1. As is well-known, P/poly may be defined either in a manner similar to the above definition for BQP/qpoly, or more simply as the class of promise problems solvable by polynomial-size Boolean circuit families with no uniformity restrictions. Based on similar ideas, the class BQP/qpoly does not change if the circuit family $\{Q_n : n \in \mathbb{N}\}$ is taken to be an arbitrary polynomial-size circuit family without uniformity constraints.
2. There is a good argument to be made that quantum advice is better viewed as a quantum analogue of *randomized advice* rather than *deterministic advice*. That is, BQP/qpoly can equally well be viewed as a quantum analogue of the (suitably defined) complexity class BPP/rpoly. It happens to be the case, however, that BPP/rpoly = P/poly. (The situation is rather different for *logarithmic-length advice*, where randomized advice is strictly more powerful than ordinary deterministic advice.)
3. Any combination of quantum, randomized, or deterministic advice with quantum, randomized, or deterministic circuits can be considered. This leads to classes such as BQP/rpoly, BQP/poly, P/qpoly, P/rpoly, and so on. (The only reasonable interpretation of BPP/qpoly and P/qpoly is that classical circuits effectively measure quantum states in the standard basis the instant that they touch them.)
   At most three distinct classes among these possibilities arise: BQP/qpoly, BQP/poly, and P/poly. This is because BQP/rpoly = BQP/poly and

$$\begin{aligned}\text{BPP/qpoly} = \text{BPP/rpoly} = \text{BPP/poly} \\ = \text{P/qpoly} = \text{P/rpoly} = \text{P/poly}\,.\end{aligned}$$

   The principle behind these equalities is that nonuniformity is stronger than randomness [6].

The following theorem places an upper-bound on the power of polynomial-time quantum algorithms with quantum advice [2].

Theorem 16 BQP/qpoly $\subseteq$ PP/poly.

Although in all likelihood the class PP/poly is enormous, containing many interesting problems that one can have little hope of being able to solve in practice, the upper-bound represented by this theorem is far from obvious. The most important thing to notice is that the power of quantum advice (to a BQP machine) is simulated by deterministic advice (to a PP machine). This means that no matter how complex, a polynomial-size quantum advice state can never encode more information accessible to a polynomial-time quantum computer than a polynomial-length string can, albeit to an unbounded error probabilistic machine.

Quantum advice has also been considered for other quantum complexity classes such as QMA and QIP. For instance, the following bound is known on the power of QMA with quantum advice [3].

Theorem 17 QMA/qpoly $\subseteq$ PSPACE/poly.

Generally speaking, the study of both quantum and randomized advice is reasonably described as a sticky business when unbounded error machines or interactive protocols are involved. In these cases, complexity classes defined by both quantum and randomized advice can be

highly non-intuitive and dependent on specific interpretations of models. For example, Raz [83] has shown that both QIP/qpoly and IP/rpoly contain all languages, provided that the advice is not accessible to the prover. Likewise, Aaronson [2] has observed that both PP/rpoly and PQP/qpoly contain all languages. These results are quite peculiar, given that significantly more powerful models can become strictly less powerful in the presence of quantum or randomized advice. For instance, even a Turing machine running in exponential space cannot decide all languages with bounded error given randomized advice.

Space-Bounded Quantum Computation

The quantum complexity classes that have been discussed thus far in this article are based on the abstraction that efficient quantum computations are those that can be performed in polynomial time. Quantum complexity classes may also be defined by bounds on space rather than time, but here a different computational model is required to reasonably compare with classical models of space-bounded computation. One simple choice of a suitable model is a hybrid between quantum circuits and classical Turing machines—and although this model is different from the variants of quantum Turing machines that were originally studied in the theory of quantum computing [30,40], the term *quantum Turing machine* (QTM for short) is nevertheless appropriate. There should be no confusion given that no other models of quantum Turing machines are discussed in this article.

Figure 13 illustrates a quantum Turing machine. A quantum Turing machine has a read-only classical input tape, a classical work tape, and a quantum tape consisting of an infinite sequence of qubits each initialized to the zero-state. Three tape heads scan the quantum tape, allowing quantum operations to be performed on the corresponding qubits. (It would be sufficient to have just two, but allowing three tape heads parallels the choice of a universal gate set that allows three-qubit operations.) A single step of a QTM's computation may involve ordinary moves by the classical parts of the machine and quantum operations on the quantum tape: Toffoli gates, Hadamard gates, phase-shift gates, or single-qubit measurements in the computational basis.

The running time of a quantum Turing machine is defined as for ordinary Turing machines. The space used by such a machine is the number of squares on the classical work tape plus the number of qubits on the quantum tape that are ever visited by one of the tape heads. Similar to the classical case, the input tape does not contribute to the space used by the machine because it is a read-only tape.

The following complexity classes are examples of classes that can be defined using this model.

BQL A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in BQL (bounded-error quantum logarithmic space) if and only if there exists a quantum Turing machine M running in polynomial time and logarithmic space that accepts every string $x \in A_{\text{yes}}$ with probability at least 2/3 and accepts every string $x \in A_{\text{no}}$ with probability at most 1/3.

PQL A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in PQL (unbounded-error quantum logarithmic space) if and only if there exists a quantum Turing machine M running in polynomial time and logarithmic space that accepts every string $x \in A_{\text{yes}}$ with probability strictly greater than 1/2 and accepts every string $x \in A_{\text{no}}$ with probability at most 1/2.

BQPSPACE A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in BQPSPACE (bounded-error quantum polynomial space) if and only if there exists a quantum Turing machine M running in polynomial space that accepts every string $x \in A_{\text{yes}}$ with probability at least 2/3 and accepts every string $x \in A_{\text{no}}$ with probability at most 1/3.

PQPSPACE A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in PQPSPACE (unbounded-error quantum polynomial space) if and only if there exists a quantum Turing machine M running in polynomial space that accepts every string $x \in A_{\text{yes}}$ with probability strictly greater than 1/2 and accepts every string $x \in A_{\text{no}}$ with probability at most 1/2.

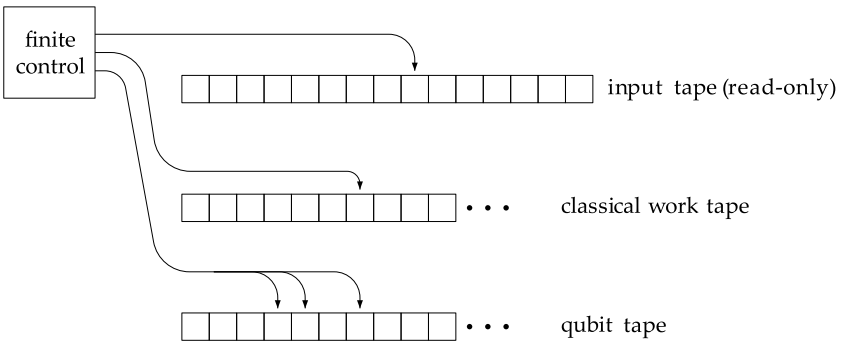
Unlike polynomial-time computations, it is known that quantum information does not give a significant increase in computational power in the space-bounded case [95,98].

Theorem 18 *The following relationships hold.*

1. BQL $\subseteq$ PQL $=$ PL.
2. BQPSPACE $=$ PQPSPACE $=$ PSPACE.

The key relationship in the above theorem, from the perspective of quantum complexity, is PQL $\subseteq$ PL, which can be shown using space-bounded counting complexity. In particular, the proof relies on a theory of GapL functions [12] that parallels the theory of GapP functions, and allows for a variety of matrix computations to be performed in PL.

The above theorem, together with the containment PL $\subseteq$ NC [32], implies that both BQL and PQL are con-



Quantum Computational Complexity, Figure 13
A quantum Turing machine. This variant of quantum Turing machine is classical with the exception of a quantum work tape, each square of which contains a qubit. Quantum operations and measurements can be performed on the qubits scanned by the quantum tape heads

tained in NC. An interpretation of this fact is that logarithmic-space quantum computations can be very efficiently simulated in parallel.

Bounded-Depth Quantum Circuits

The *depth* of a classical or quantum circuit is the maximum number of gates encountered on any path from an input bit or qubit to an output bit or qubit in the circuit. One may reasonably think of circuit depth as the parallel running time, or the number of time units needed to apply a circuit when operations may be parallelized in any way that respects the topology of the circuit.

The following complexity class, first defined by Moore and Nilsson [77], represent bounded-error quantum variants of the class NC.

BQNC A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in BQNC (bounded-error quantum NC) if and only if there exists a logarithmic-space generated family $\{Q_n : n \in \mathbb{N}\}$ of poly-logarithmic depth quantum circuits, where each circuit Q_n takes n input qubits and produces one output qubit, such that $\Pr[Q \text{ accepts } x] \geq 2/3$ for all $x \in A_{\text{yes}}$, and $\Pr[Q \text{ accepts } x] \leq 1/3$ for all $x \in A_{\text{no}}$.

Many other complexity classes based on bounded-depth quantum circuits have been studied as well. The survey of Bera, Green, and Homer [28] discusses several examples.

In the classical case there is a very close relationship between space-bounded and depth-bounded computation [31,32]. This close relationship is based on two main ideas: the first is that space-bounded computations can be simulated efficiently by bounded-depth circuits using parallel algorithms for matrix computations, and the second is that bounded-depth Boolean circuits can be efficiently simulated by space-bounded computations via depth-first traversals of the circuit to be simulated.

For quantum computation this close relationship is not known to exist. One direction indeed holds, as was discussed in the previous subsection: space-bounded quantum computations can be efficiently simulated by depth-bounded circuits. The other direction, which is an efficient space-bounded simulation of bounded-depth quantum circuits, is not known to hold and is arguably quite unlikely. Informally speaking, bounded-depth quantum circuits are computationally powerful, whereas space-bounded quantum Turing machines are not. Three facts that support this claim are as follows.

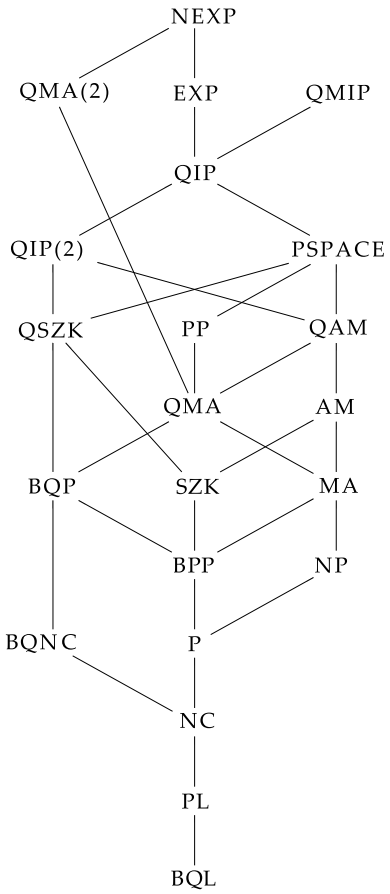
1. Computing the acceptance probability for even constant-depth quantum circuits is as hard as computing acceptance probabilities for arbitrary polynomial-size quantum circuits [48].
2. Shor's factoring algorithm can be implemented by quantum circuits having logarithmic-depth, along with classical pre- and post-processing [35].
3. The quantum circuit distinguishability problem remains complete for QIP when restricted to logarithmic-depth quantum circuits [84].

It is reasonable to conjecture that BQNC is incomparable with BPP and properly contained in BQP.

Future Directions

There are many possible future directions for research in quantum computational complexity theory. The following list suggests just a few of many possibilities.

1. The power of multiple-prover quantum interactive proofs, for both quantum and classical verifiers, is



Quantum Computational Complexity, Figure 14
A diagram of inclusions among most of the complexity classes discussed in this article

very poorly understood. In particular: (i) no interesting upper-bounds are known for either MIP* or QMIP, and (ii) neither the containment NEXP $\subseteq$ MIP* nor NEXP $\subseteq$ QMIP is known to hold.

2. The containment NP $\subseteq$ BQP would be a very powerful incentive to build a quantum computer, to say the least. While there is little reason to hope that this containment holds, there is at the same time little evidence against it aside from the fact that it fails relative to an oracle [27]. A better understanding of the relationship between BQP and NP, including possible consequences of one being contained in the other, is an important direction for further research in quantum complexity.
3. Along similar lines to the previous item, an understanding of the relationship between BQP and the polynomial-time hierarchy has remained elusive. It is not known whether BQP is contained in the polynomial-time hierarchy, whether there is an oracle relative to which this is false, or even whether there is an oracle relative to which BQP is not contained in AM.
4. Interest in complexity classes is ultimately derived from the problems that they contain. An important future direction in quantum complexity theory is to prove the inclusion of interesting computational problems in the quantum complexity classes for which this is possible. Of the many problems that have been considered, one of the most perplexing from the perspective of quantum complexity is the graph isomorphism problem. It is a long-standing open problem whether the graph isomorphism problem is in BQP. A seemingly easier task than showing the inclusion of this problem in BQP is proving it is in co-QMA; but even this problem has remained unresolved.

Bibliography

1. Aaronson S (2002) Quantum lower bound for the collision problem. In: Proceedings of the 35th Annual ACM Symposium on Theory of Computing. ACM Press, New York
2. Aaronson S (2005) Limitations of quantum advice and one-way communication. Theory Comput 1:1–28
3. Aaronson S (2006) QMA/qpoly is contained in PSPACE/poly: de-Merlinizing quantum protocols. In: Proceedings of the 21st Annual IEEE Conference on Computational Complexity. IEEE Ceomputer Society Press, Los Alamitos pp 261–273
4. Aaronson S, Kuperberg G (2007) Quantum versus classical proofs and advice. Theory Comput 3:129–157
5. Aaronson S, Shi Y (2004) Quantum lower bounds for the collision and the element distinctness problems. J ACM 51(4):595–605
6. Adleman L (1978) Two theorems on random polynomial time. In: Proceeding of the 19th Annual IEEE Symposium on Foundations of Computer Science. IEEE Ceomputer Society Press, Los Alamitos pp 75–83
7. Adleman L, DeMarrais J, Huang M (1997) Quantum computability. SIAM J Comput 26(5):1524–1540
8. Aharonov D, Naveh T (2002) Quantum NP – a survey. Available as arXiv.org e-Print quant-ph/0210077
9. Aharonov D, Kitaev A, Nisan N (1998) Quantum circuits with mixed states. In: Proceedings of the 30th Annual ACM Symposium on Theory of Computing. ACM Press, New York, pp 20–30
10. Aharonov D, Gottesman D, Irani S, Kempe J (2007) The power of quantum systems on a line. In: Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science. IEEE Ceomputer Society Press, Los Alamitos pp 373–383
11. Aharonov D, van Dam W, Kempe J, Landau Z, Lloyd S, Regev O (2007) Adiabatic quantum computation is equivalent to standard quantum computation. SIAM J Comput 37(1): 166–194
12. Allender E, Ogihara M (1996) Relationships among PL, #L, and the determinant. RAIRO – Theor Inf Appl 30:1–21
13. Arora S, Barak B (2006) Complexity Theory: A Modern Approach. Web draft available at http://www.cs.princeton.edu/theory/complexity/

14. Arora S, Safra S (1998) Probabilistic checking of proofs: a new characterization of NP. J ACM 45(1):70–122
15. Arora S, Lund C, Motwani R, Sudan M, Szegedy M (1998) Proof verification and the hardness of approximation problems. J ACM 45(3):501–555
16. Aroya A-B, Shma A-T (2007) Quantum expanders and the quantum entropy difference problem. Available as arXiv.org e-print quant-ph/0702129
17. Babai L (1985) Trading group theory for randomness. In: Proceedings of the Seventeenth Annual ACM Symposium on Theory of Computing. ACM Press, New York pp 421–429
18. Babai L (1992) Bounded round interactive proofs in finite groups. SIAM J Discret Math 5(1):88–111
19. Babai L, Moran S (1988) Arthur-Merlin games: a randomized proof system, and a hierarchy of complexity classes. J Comput Syst Sci 36(2):254–276
20. Babai L, Szemerédi E (1984) On the complexity of matrix group problems I. In: Proceedings of the 25th Annual IEEE Symposium on Foundations of Computer Science. IEEE Ceomputer Society Press, Los Alamitos pp 229–240
21. Babai L, Fortnow L, Lund C (1991) Non-deterministic exponential time has two-prover interactive protocols. Comput Complex 1(1):3–40
22. Beigel R, Reingold N, Spielman D (1995) PP is closed under intersection. J Comput Syst Sci 50(2):191–202
23. Beigi S, Shor P (2007) On the complexity of computing zero-error and Holevo capacity of quantum channels. Available as arXiv.org e-Print 0709.2090
24. Bell J (1964) On the Einstein-Podolsky-Rosen paradox. Phys 1(3):195–200
25. Bellare M, Goldreich O, Sudan M (1998) Free bits, PCPs, and non-approximability —towards tight results. SIAM J Comput 27(3):804–915
26. Bennett CH (1973) Logical reversibility of computation. IBM J Res Dev 17:525–532
27. Bennett CH, Bernstein E, Brassard G, Vazirani U (1997) Strengths and weaknesses of quantum computing. SIAM J Comput 26(5):1510–1523
28. Bera D, Green F, Homer S (2007) Small depth quantum circuits. ACM SIGACT News 38(2):35–50
29. Bernstein E, Vazirani U (1993) Quantum complexity theory (preliminary abstract). In: Proceedings of the 25th Annual ACM Symposium on Theory of Computing. ACM Press, New York pp 11–20
30. Bernstein E, Vazirani U (1997) Quantum complexity theory. SIAM J Comput 26(5):1411–1473
31. Borodin A (1977) On relating time and space to size and depth. SIAM J Comput 6:733–744
32. Borodin A, Cook S, Pippenger N (1983) Parallel computation for well-endowed rings and space-bounded probabilistic machines. Inf Control 58:113–136
33. Brassard G (2003) Quantum communication complexity. Found Phys 33(11):1593–1616
34. Cleve R (2000) An introduction to quantum complexity theory. In: Macchiavello C, Palma GM, Zeilinger A (eds) Collected Papers on Quantum Computation and Quantum Information Theory. World Scientific. Singapore pp 103–127
35. Cleve R, Watrous J (2000) Fast parallel circuits for the quantum Fourier transform. In: Proceedings of the 41st Annual IEEE Symposium on Foundations of Computer Science. pp 526–536
36. Cleve R, Høyer P, Toner B, Watrous J (2004) Consequences and limits of nonlocal strategies. In: Proceedings of the 19th Annual IEEE Conference on Computational Complexity. pp 236–249
37. Cleve R, Slofstra W, Unger F, Upadhyay S (2007) Perfect parallel repetition theorem for quantum XOR proof systems. In: Proceedings of the 22nd Annual IEEE Conference on Computational Complexity. pp 109–114
38. Cook S (1972) The complexity of theorem proving procedures. In: Proceedings of the Third Annual ACM Symposium on Theory of Computing. ACM Press, New York pp 151–158
39. de Wolf R (2002) Quantum communication and complexity. Theor Comput Sci 287(1):337–353
40. Deutsch D (1985) Quantum theory, the Church–Turing principle and the universal quantum computer. Proc Roy Soc Lond A 400:97–117
41. Deutsch D (1989) Quantum computational networks. Proc Roy Soc Lond A 425:73–90
42. Dinur I (2007) The PCP theorem by gap amplification. J ACM 54(3)
43. Du D-Z, Ko K-I (2000) Theory of Computational Complexity. Wiley, New York
44. Even S, Selman A, Yacobi Y (1984) The complexity of promise problems with applications to public-key cryptography. Inf Control 61:159–173
45. Feige U, Kilian J (1997) Making games short. In: Proceedings of the 29th Annual ACM Symposium on Theory of Computing. ACM Press, New York pp 506–516
46. Feige U, Lovász L (1992) Two-prover one-round proof systems: their power and their problems. In: Proceedings of the 24th Annual ACM Symposium on Theory of Computing. ACM Press, New York pp 733–744
47. Fenner S, Fortnow L, Kurtz S (1994) Gap-definable counting classes. J Comput Syst Sci 48:116–148
48. Fenner S, Green F Homer S, Zhang Y (2005) Bounds on the power of constant-depth quantum circuits. In: Proceedings of the 15th International Symposium on Fundamentals of Computation Theory. Lect Notes Comput Sci 3623:44–55
49. Feynman R (1983) Simulating physics with computers. Int J Theor Phys 21(6/7):467–488
50. Fortnow L (1997) Counting complexity. In: Hemaspaandra L, Selman A (eds) Complexity Theory Retrospective II. Springer, New York pp 81–107
51. Fortnow L, Rogers J (1999) Complexity limitations on quantum computation. J Comput Syst Sci 59(2):240–252
52. Goldreich O (2005) On promise problems (a survey in memory of Shimon Even [1935–2004]). Electronic Colloquium on Computational Complexity; Report TR05-018
53. Goldreich O, Vadhan S (1999) Comparing entropies in statistical zero-knowledge with applications to the structure of SZK. In: Proceedings of the 14th Annual IEEE Conference on Computational Complexity. pp 54–73
54. Goldwasser S, Sipser M (1989) Private coins versus public coins in interactive proof systems. In: Micali S (ed) Randomness and Computation, vol 5 of Advances in Computing Research. JAI Press, Greenwich, Conn pp 73–90
55. Goldwasser S, Micali S, Rackoff C (1985) The knowledge complexity of interactive proof systems. In: Proceedings of the Seventeenth Annual ACM Symposium on Theory of Computing. ACM Press, New York pp 291–304
56. Goldwasser S, Micali S, Rackoff C (1989) The knowledge

complexity of interactive proof systems. SIAM J Comput 18(1):186–208
57. Gutoski G, Watrous J (2007) Toward a general theory of quantum games. In: Proceedings of the 39th ACM Symposium on Theory of Computing. ACM Press, New York pp 565–574
58. Håstad J (2001) Some optimal inapproximability results. J ACM 48(4):798–859
59. Janzing D, Wocjan P, Beth T (2005) Non-identity-check is QMA-complete. Int J Quantum Inf 3(2):463–473
60. Kaye P, Laflamme R, Mosca M (2007) An introduction to quantum computing. Oxford University Press, Oxford
61. Kempe J, Kitaev A, Regev O (2006) The complexity of the local Hamiltonian problem. SIAM J Comput 35(5):1070–1097
62. Kempe J, Kobayashi H, Matsumoto K, Toner B, Vidick T (2007) Entangled games are hard to approximate. Proceedings ot the 49th Annual IEEE Symposium on Foundations of Computer Science, IEEE Computer Society Press, Los Alamitos 2008
63. Kempe J, Regev O, Toner B (2007) The unique games conjecture with entangled provers is false. Proceedings ot the 49th Annual IEEE Symposium on Foundations of Computer Science, IEEE Computer Society Press, Los Alamitos 2008
64. Kitaev A (1997) Quantum computations: algorithms and error correction. Russ Math Surv 52(6):1191–1249
65. Kitaev A (1999) Quantum NP. Talk at AQIP'99: Second Workshop on Algorithms in Quantum Information Processing. DePaul University, Chicago
66. Kitaev A, Watrous J (2000) Parallelization, amplification, and exponential time simulation of quantum interactive proof system. In: Proceedings of the 32nd ACM Symposium on Theory of Computing. ACM Press, New York pp 608–617
67. Kitaev A, Shen A, Vyalyi M (2002) Classical and Quantum Computation. Graduate Studies in Mathematics, vol 47. American Mathematical Society, Providence
68. Knill E (1995) Approximation by quantum circuits. Technical Report LAUR-95-2225, Los Alamos National Laboratory. Available as arXiv.org e-Print quant-ph/9508006
69. Knill E (1996) Quantum randomness and nondeterminism. Technical Report LAUR-96-2186, Los Alamos National Laboratory. Available as arXiv.org e-Print quant-ph/9610012
70. Kobayashi H, Matsumoto K (2003) Quantum multi-prover interactive proof systems with limited prior entanglement. J Comput Syst Sci 66(3):429–450
71. Kobayashi H, Matsumoto K, Yamakami T (2003) Quantum Merlin–Arthur proof systems: Are multiple Merlins more helpful to Arthur? In: Proceedings of the 14th Annual International Symposium on Algorithms and Computation. Lecture Notes in Computer Science, vol 2906. Springer, Berlin
72. Levin L (1973) Universal search problems. Probl Inf Transm 9(3):265–266 (English translation)
73. Liu Y-K (2006) Consistency of local density matrices is QMA-complete. In: Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques. Springer, Berlin; Lect Notes Comput Sci 4110:438–449
74. Lloyd S (2000) Ultimate physical limits to computation. Nature 406:1047–1054
75. Lund C, Fortnow L, Karloff H, Nisan N (1992) Algebraic methods for interactive proof systems. J ACM 39(4):859–868
76. Marriott C, Watrous J (2005) Quantum Arthur-Merlin games. Comput Complex 14(2):122–152
77. Moore C, Nilsson M (2002) Parallel quantum computation and quantum codes. SIAM J Comput 31(3):799–815
78. Moore G (1965) Cramming more components onto integrated circuits. Electron 38(8):82–85
79. Nielsen MA, Chuang IL (2000) Quantum Computation and Quantum Information. Cambridge University Press, Cambride
80. Nishimura H, Yamakami T (2004) Polynomial time quantum computation with advice. Inf Process Lett 90(4):195–204
81. Oliveira R, Terhal B (2005) The complexity of quantum spin systems on a two-dimensional square lattice. Available as arXiv.org e-Print quant-ph/0504050
82. Papadimitriou C (1994) Computational Complexity. Addison-Wesley, Readig, Mass
83. Raz R (2005) Quantum information and the PCP theorem. In: 46th Annual IEEE Symposium on Foundations of Computer Science. pp 459–468
84. Rosgen B (2008) Distinguishing short quantum computations. Proceedings of the 25th Annual Symposium on theoretical Aspects of Computer Science, IBFI, Schloss Dagstuhl, pp 597–608
85. Rosgen B, Watrous J (2005) On the hardness of distinguishing mixed-state quantum computations. In: Proceedings of the 20th Annual Conference on Computational. pp 344–354
86. Sahai A, Vadhan S (2003) A complete promise problem for statistical zero-knowledge. J ACM 50(2):196–249
87. Shamir A (1992) IP = PSPACE. J ACM 39(4):869–877
88. Shor P (1994) Algorithms for quantum computation: discrete logarithms ande factoring. In: Proceedings of the 35th Annual IEEE Symposium on Foundations of Computer Science. pp 124–134
89. Shor P (1997) Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. SIAM J Comput 26(5):1484–1509
90. Stinespring WF (1955) Positive functions on C^*-algebras. Proc Am Math Soc 6(2):211–216
91. Toda S (1991) PP is as hard as the polynomial-time hierarchy. SIAM J Comput 20(5):865–887
92. Toffoli T (1980) Reversible computing. Technical Report MIT/LCS/TM-151, Laboratory for Computer Science, Massachusetts Institute of Technology
93. Vadhan S (2007) The complexity of zero knowledge. In: 27th International Conference on Foundations of Software Technology and Theoretical Computer Science. Lect Notes Comput Sci 4855:52–70
94. Valiant L (1979) The complexity of computing the permanent. Theor Comput Sci 8:189–201
95. Watrous J (1999) Space-bounded quantum complexity. J Comput Syst Sci 59(2):281–326
96. Watrous J (2000) Succinct quantum proofs for properties of finite groups. In: Proceedings of the 41st Annual IEEE Symposium on Foundations of Computer Science. pp 537–546
97. Watrous J (2002) Limits on the power of quantum statistical zero-knowledge. In: Proceedings of the 43rd Annual IEEE Symposium on Foundations of Computer Science. pp 459–468
98. Watrous J (2003) On the complexity of simulating space-bounded quantum computations. Comput Complex 12: 48–84
99. Watrous J (2003) PSPACE has constant-round quantum interactive proof systems. Theor Comput Sci 292(3):575–588
100. Watrous J (2006) Zero-knowledge against quantum attacks.

In: Proceedings of the 38th ACM Symposium on Theory of Computing. ACM Press, New York pp 296–305
101. Wehner S (2006) Entanglement in interactive proof systems with binary answers. In: Proceedings of the 23rd Annual Symposium on Theoretical Aspects of Computer Science. Lect Notes Comput Sci 3884:162–171

Quantum Computing

Viv Kendon
School of Physics and Astronomy, University of Leeds, Leeds, UK

Article Outline

Glossary

Bit A two state classical system used to represent a binary digit, zero or one.

Bose–Einstein particles Integer spin quantum particles like to be together: any number can occupy the same quantum state.

Classical computing What we can compute within the laws of classical physics.

Error correction In a quantum context, fixing errors without disturbing the quantum superposition.

Entanglement Quantum states can be more highly correlated than classical systems: the extra correlations are known as entanglement.

Fermi–Dirac particles Half-integer spin quantum particles like to be alone: only one such particle can occupy each quantum state.

Quantum communications Using quantum mechanics can gain an advantage when transmitting information.

Quantum dense coding Classical bits can be encoded two for one into qubits for communications purposes.

Quantum key distribution Quantum mechanics allows for secure key distribution in the presence of eavesdroppers and noisy environments. The keys can then be used for encrypted communication.

Quantum teleportation A method to transmit an unknown quantum state using only classical communications plus shared entanglement.

Quantum computing Computation based on the laws of quantum mechanics for the allowed logical operations.

Qubit A two state quantum system such as the spin of an electron or the polarization of a photon. More complex quantum particles (such as atoms) can be used as qubits if just two of their available states are chosen to represent the qubit.

Qubus A quantum version of a computer bus, the fast communications linking memory and processing registers. Can be implemented using a coherent light source (such as a laser).

Scalable A computer architecture designed from modular units that can be efficiently expanded to an arbitrary size.

Squeezing With a pair of complementary quantum observables, making one uncertain so that the other can be measured more precisely.

Threshold result In the context of quantum computation, this results says that error correction can work if the error rate is low enough.

Tunneling Quantum particles can get through barriers that classical particles remain stuck behind. If the barrier is not infinitely high, there is a some probability for the quantum particle to be the other side of it even though it doesn't on average have enough energy to jump over the top.

Unitary operations How to control quantum systems while preserving quantum properties. Quantum systems evolving without any influence from environmental disturbance follow unitary dynamical evolution.

Definition of the Subject

Quantum computing is computing that follows the logic of quantum mechanics. The first part of this subject is well-specified after the hundred-odd years of development of quantum theory. Definitions of computing are fuzzier, and argued over by philosophers: for the purposes of this article we will leave such nuances aside in favor of exploring what is possible when the constraints of classical logic are put aside, but the limitations of the physical world are kept firmly in hand.

Quantum computing is not simply computing that involves quantum effects, that would be too broad to be useful. Since transitors and lasers exploit quantum properties of matter and light, most classical computers would thus be included. Yet it is worth remembering that definitions are never as clearcut as we would like. Figuring out which quantum systems can be simulated efficiently by classical computers is an active area of current research, and pin-

ning down the boundary between quantum and classical is the object of ongoing experimental investigation.

Introduction

This article on *quantum computing* is in the Unconventional Computing section of the encyclopedia. It is thus a somewhat distinctive view of quantum computing, adapted to the context in which you are likely to be reading it. It is self-contained, in that it introduces the concepts you will need to understand the contents, but those desiring more details and a broader perspective on the subject can refer to the introductory article ► Quantum Information Science, Introduction to and referenced articles therein.

To many, quantum computing is already unconventional computing, depending for its conception on the elusive properties of matter at the scale of atoms rather than the simple everyday experience of counting and deduction that underpins classical computation. Yet it is already an established field in which there is a standard approach, and the possibility to be unconventional within a quantum computing context. This article is thus divided into two parts, the first forming a lightening introduction to the standard digital version of quantum computing, and the second expanding into less well-charted territory beside and beyond.

In order to write about such an interdisciplinary subject for a broad audience, I have tried to keep everything at a level a non-specialist can follow, while not oversimplifying to the point of inaccuracy. Please be forgiving about the parts you could have written better yourself, and inquisitive about the parts that may have something new for you.

Digital Quantum Computing

Digital quantum computing developed a more or less parallel structure to classical digital computing, with a quantum version of each component. There are good reasons for this: many of the optimal features of classical digital computing carry over to quantum computing with little or no change. In order to explain why there is a fundamental advantage to using quantum logic for computation, this section will explain the components step by step and then provide examples of how the parts combine into the whole.

What Is a Qubit? (Quantum Mechanics for Dummies)

Quantum computing exploits the logic of quantum mechanics, which is notoriously tricky to understand. The mathematical structure is simple and linear, so treating it as "maths we happen to be able to build (we think)", is one option for getting to grips with it. Here I've chosen the pictorial intuition approach, with some cartoon qubits to help explain how they behave. Those already familiar with quantum mechanics can safely skip over this section. Those still bemused by the end of it may like to sample from the wide range of available textbooks to find an approach which suits them best.

Classical computers use *bits* as their basic unit, any classical system with two states can represent a single bit. Tossing a coin, with the resulting "heads" or "tails" is an example of a two state classical system. In computers, the two states are usually labeled "0" (zero) and "1" (one). Quantum computers use quantum bits, usually called *qubits*, as their basic unit, which are two state quantum systems, and the two states are also usually labeled zero and one. Examples of physical systems with two quantum states include electron spin, photon polarization and phosphorus nuclear spin. These have all been used in experiments aimed towards building a quantum computer. Those who know a bit of quantum mechanics should note that in a quantum computer we usually localize or otherwise separate individual qubits so they are distinguishable from each other, and there are thus no Fermi or Bose statistics to complicate the picture. We will revisit this aspect of quantum mechanics briefly in Subsect. "Topological Quantum Computing".

Here are a couple of cartoon qubits to help explain their quantum properties:  . They are wriggly, fidgety little things, they move so fast you can't tell which way up they are just by looking. If you reach down to grab hold of one the only way you can get a grip is by an arm or a leg. First we have to choose our basis states, labeled $|0\rangle$  and $|1\rangle$  . We write the zero and one in those funny brackets ("kets") to remind us they are quantum states rather than classical bits. To simplify notation, we will sometimes write the state of several qubits together inside one ket thus, $|0\rangle|0\rangle \equiv |00\rangle$. In terms of the cartoon qubits, this means we decided to label "grabbed by the arm" as the zero state and "grabbed by the leg" as the one state. The key quantum property is that qubits can fidget around anywhere in between the zero and one states, in what is called a superposition state, for example, $(|0\rangle + |1\rangle)/\sqrt{2}$  or $(|0\rangle - |1\rangle)/\sqrt{2}$  . If you reach down and grab a qubit in one of these superposition states, you are equally like to grab an arm or a leg, indicat-

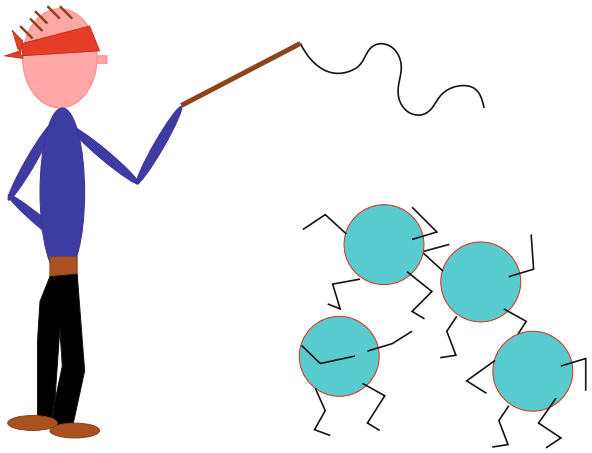
ing zero or one in a basis state. So you can't tell the qubit was in a superposition state by measuring it, the outcome is always a basis state. What's more, grabbing it by the arm hauls it onto its feet, and grabbing it by the leg shoves it into a handstand. So after measuring, what you found is what you now have, regardless of what it was before. Luckily, it is possible to do other things to qubits besides measuring them.

The most general state of one qubit can be written $\alpha|0\rangle + \beta|1\rangle$ with $|\alpha|^2 + |\beta|^2 = 1$, and α, β are complex numbers. The probability of measuring zero is given by $|\alpha|^2$ and the probability of measuring one is given by $|\beta|^2$. (Why complex numbers? Because it works conveniently. If you don't like complex numbers, there are ways to formulate quantum mechanics that avoid using them). Cartoon qubits with limbs don't quite capture all the subtle quantum effects, but they are a useful image to keep in mind if you are new to these concepts. This account also ignores all of the more fundamental issues to do with the interpretation of quantum mechanics, especially those generally referred to as "the quantum measurement problem".

Now consider two qubits (traditionally belonging to Alice and Bob) in the state $(|0\rangle_A|0\rangle_B + |1\rangle_A|1\rangle_B)/\sqrt{2}$. This is an example of an entangled state: the two qubits are completely correlated. Suppose Alice measures her qubit: if she finds $|0\rangle_A$ then she knows Bob will find $|0\rangle_B$ when he measures his qubit. If she finds $|1\rangle_A$ then she knows Bob will find $|1\rangle_B$. Alice can't communicate anything to Bob this way because she can't control whether she will find 0 or 1 when she measures. After they have both measured their qubits, they then share one random bit, which is a resource they can use for other tasks.

The purely quantum feature of entanglement that cannot be reproduced with suitably correlated classical systems is that Alice and Bob can use any basis states they like and the perfect correlation still appears. For example, $(|0\rangle + |1\rangle)/\sqrt{2}$ and $(|0\rangle - |1\rangle)/\sqrt{2}$ are another possible pair of basis states: the rule is they must be orthogonal to each other. This is the equivalent of changing the orientation of your classical coordinate system. However, Alice and Bob do both have to use the same basis: for a thorough discussion of the consequences of this and other reference frames in quantum mechanics, see Bartlett et al. [1].

Now that we have qubits to make a quantum register for our quantum computer, and measurements to read out the state of the quantum register, we need some quantum gates to perform our calculation with. There are two ways to change a quantum state: measurements, which are what happen when you (or something else) observes the state of a quantum system, and unitary dynamics, which is what quantum systems do when no-one is looking.



Quantum Computing, Figure 1
Unitary evolution: controlling qubits without looking at them

Happily, it is possible to steer a quantum system without looking at it (see Fig. 1), so we can make it evolve the way we choose through applying *unitary operations*. Unitary operations are reversible, the simplest example, acting on a single qubit, is the Hadamard gate H. We can define what it does by the effect on basis states:

$$\begin{aligned} H|0\rangle &= (|0\rangle + |1\rangle)/\sqrt{2} \\ H|1\rangle &= (|0\rangle - |1\rangle)/\sqrt{2}\,, \end{aligned} \tag{1}$$

and of course it can also act on superposition states,

$$H(\alpha|0\rangle + \beta|1\rangle) = \frac{1}{\sqrt{2}}\{(\alpha+\beta)|0\rangle + (\alpha-\beta)|1\rangle\}\,. \tag{2}$$

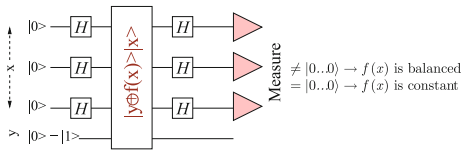
The Hadamard gate rotates the qubit through an angle of $\pi/2$. Notice also that applying the Hadamard gate twice gets you back where you started, $H(H|\psi\rangle) = |\psi\rangle$, the Hadamard gate is its own inverse.

Rotation gates can rotate through any other angle, for example,

$$\begin{aligned} R_\theta|0\rangle &= \cos(\theta/2)|0\rangle + \sin(\theta/2)|1\rangle \\ R_\theta|1\rangle &= \sin(\theta/2)|0\rangle - \cos(\theta/2)|1\rangle. \end{aligned} \tag{3}$$

To carry out computations, we also need gates that act on two qubits at once, for example, the familiar CNOT gate:

$$\begin{aligned} C|0\rangle_c|0\rangle_t &= |0\rangle_c|0\rangle_t \\ C|0\rangle_c|1\rangle_t &= |0\rangle_c|1\rangle_t \\ C|1\rangle_c|0\rangle_t &= |1\rangle_c|1\rangle_t \\ C|1\rangle_c|1\rangle_t &= |1\rangle_c|0\rangle_t \end{aligned} \tag{4}$$



Quantum Computing, Figure 2
Circuit diagram for the Deutsch–Jozsa algorithm for $N < 8$. See text for details. Normalization factors have been omitted from the qubit states

where the first qubit (labeled c) is the control and remains unchanged, while the second qubit (labeled t) is the target and is flipped if the control qubit is in state $|1\rangle$.

Just as with classical computation, small sets of universal quantum gates exist from which any computation can be constructed efficiently [2].

A particular set is usually chosen to suit the physical components of the quantum computer, there are many possible choices, the gates described here are just a few of the one and two qubit gates available.

How Do Qubits Give Us Cool Computing?

Now that we have qubits to form a quantum register, quantum gates to perform our computation, and measurements to read out the result, the next step is to put them all together. We will begin with an example, which will illuminate the basic idea much more clearly than a general description. The Deutsch–Jozsa Algorithm was one of the earliest quantum algorithms, it solves the following promise problem. We are promised that the function $f(x) : x \mapsto \{0, 1\}$ is either

- balanced i. e., the number of times $f(x)$ outputs zero is equal to the number of times $f(x)$ outputs one, compared over all possible input values x (with $0 < x \leq N$), or,
- $f(x)$ is constant, i. e., the output is always either zero or one (but we don't know which) for any input x.

The cost of this computation is measured in "queries" or evaluations of the function $f(x)$. The circuit diagram in Fig. 2 shows how to solve this problem with a quantum computer.

There is one register of $n = \lceil \log_2 N \rceil$ qubits to hold the input x, and one extra single qubit $|y\rangle$. We start with all the qubits in the $|0\rangle$ state. The first trick is the set of Hadamards applied to the $|x\rangle$ register: $H|0\ldots0\rangle = (|0\ldots0\rangle + |0\ldots1\rangle + |0\ldots10\rangle + \cdots + |1\ldots1\rangle)/2^{-n/2}$ which is a superposition of all numbers $0\ldots2^n - 1$. We then make our one query to the oracle that computes $f(x)$ for us, using the superposition of all possible inputs now stored in $|x\rangle$. We get back a superposition of all the answers, which we store by adding it to the $|y\rangle$ qubit.

Remember we cannot detect a superposition by measurement alone, so the superposition of all the possible answers doesn't help us yet. The second trick allows us to detect what sort of superposition of answers we got back. Before adding the answer to $|y\rangle$, we prepare this qubit in the state $(|0\rangle - |1\rangle)/\sqrt{2}$. This can be done by flipping it from $0\rangle$ to $|1\rangle$ then applying a Hadamard gate. Then, when we add the answers to $|y\rangle$, there are three possibilities:

1. the answers are all $|0\rangle$: $|y\rangle|x\rangle = (|0\rangle - |1\rangle)|x\rangle/\sqrt{2} =$ unchanged.
2. the answers are all $|1\rangle$: $|y\rangle|x\rangle = (|1\rangle - |0\rangle)|x\rangle/\sqrt{2} =$ minus what it was before.
3. the answers are half and half: $|y\rangle|x\rangle =$ something else.

The second sequence of Hadamard gates then attempts to undo the first set and convert $|x\rangle$ back to the all zero state. In the first two cases, this succeeds (we can't detect the minus sign when we measure), but in the third case the Hadamards don't get us back where we started so there will be some ones in the $|x\rangle$ qubits which we find when we measure them. So, if we measure all zeros for $|x\rangle$ then $f(x)$ is constant, while if we measure any ones then $f(x)$ is balanced.

To solve this problem classically you have to examine the value of $f(x)$ for one value of x at a time and compare them. As soon as you accumulate both a zero and a one in the set of answers, you know (because of the promise) that $f(x)$ is balanced. But you can't be sure it is constant until you accumulate more than half the answers and see that they are all the same. So the best possible classical algorithm must take at least two queries, and could require up to $N/2 + 1$. Thus the quantum algorithm is fundamentally more efficient, requiring only the one evaluation of $f(x)$.

How Quantum Computing Got Started (A Little Bit of History)

The idea that quantum mechanics could give us fundamentally faster computers first occurred to Feynman [3] and independently to Deutsch [4]. They both observed that quantum systems are like a parallel computer (c.f. Feynman paths or many worlds respectively). It remained a cute idea for the next decade, until a key result showing *error correction* is theoretically possible implied it might be feasible to build one that really works [5,6,7]. Around the same time, Shor [8] found an algorithm for factoring that could in theory break crypto schemes based on products of large primes being hard to factorize. At this point the

funders sat up and took notice. Not just because factoring could become easier with a quantum computer, there are public key crypto schemes based on other "one-way" functions that are hard to compute given the public key, but easy to compute from the private key. But by shifting the boundary between what's easy and what's hard, quantum computing threatens to break other classical one-way functions too.

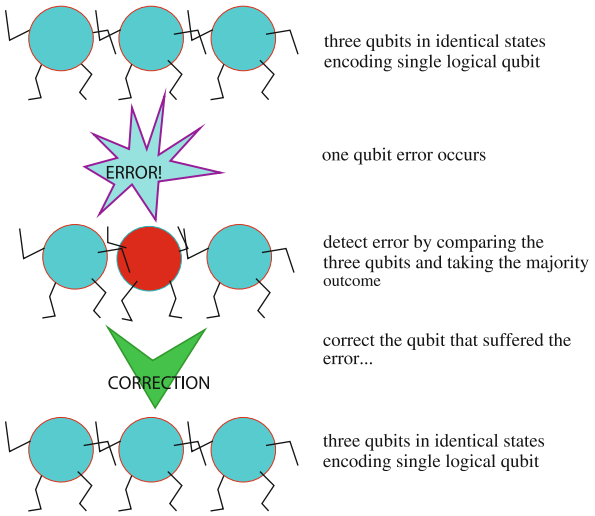
At the same time, *quantum communications* protocols were being developed, in particular, *quantum key distribution* solves the crypto problem created by Shor's algorithm [9]. Quantum resources can double the channel capacity, a technique known as *quantum dense coding* [10], and *quantum teleportation* can transmit a quantum state through a classical channel [11] with the help of some shared entanglement. Commercial quantum communications devices are available now for key distribution. They work over commercial fibre optics, they work well enough to reach from earth to satellites, and handheld devices under development [12,13]. It remains a niche market, but the successful technology in a closely related quantum information field has helped to keep the funding and optimism for quantum computers buoyant.

Error Correction for Quantum Computers

The most important result that allowed quantum computing to move from being an obscure piece of theory to a future technology is that error correction can work to protect the delicate quantum coherences (those signs, or phases, between different components of the superposition). Usually, the ubiquitous environmental disturbances randomize them very rapidly. Quantum systems are very sensitive to noise, which is why we inhabit a largely classical world.

The basic idea is simple: correct the errors faster than they accumulate. The main method for doing this is also conceptually simple: logical qubits are encoded in a larger physical quantum system, and the extra degrees of freedom used for error correction. A very simple example: for every qubit in your computation, use three identical qubits. If a single qubit error occurs, the remaining two will give you the correct answer. This is illustrated in Fig. 3. Two errors hitting the three qubits and the majority outcome will be wrong, though, so this code is only good if the error rates are very low.

What is not at all obvious is that this has any chance of working. Adding extra degrees of freedom (more qubits, for example) adds more places where errors can occur. Adding yet more qubits to correct these extra errors brings in yet more errors, and so on. Nonetheless, if the error rate



Quantum Computing, Figure 3
Error correction using three qubits to encode one

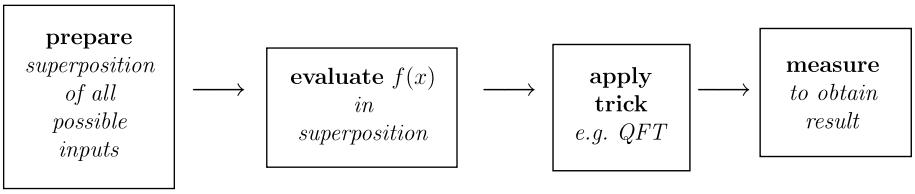
is low enough, even the errors arising from the error correction can be corrected fast enough [5,6]. This is known as the *threshold result*. The details of exactly how to accomplish this are considerable, and employ many techniques from classical error correcting codes [7,8]. For those wanting to know more, there are lots of tutorials available, e. g., [14].

The crucial piece of information is thus what the actual threshold error rate is. It can only be estimated, and current estimates suggest it is around 10^{-3} to 10^{-4} errors per qubit per quantum operation. This is rather smaller than the error rates in most physical systems that have been proposed for quantum computers, but not so small it is clearly unattainable as technology is refined. Various additional methods for avoiding the effects of decoherence have since been developed, such as decoherence-free subspaces [15,16], and using the quantum Zeno effect [17,18] to keep the system evolving in the direction you want.

Programming a Quantum Computer

The general structure of many quantum algorithms follows the same pattern we already saw in the Deutsch–Jozsa algorithm. This is illustrated in Fig. 4.

Each algorithm needs to have its own trick to make it work, so progress on expanding the number of quantum algorithms has been slow. Sometimes the same trick will work for a family of similar problems, but the number of distinct types remains limited. We have already seen an example of a promise problem in the Deutsch–Jozsa algorithm. Shor's algorithm and the many variants on this



Quantum Computing, Figure 4
General structure of many quantum algorithms

method have essentially similar structure, while Grover's search and quantum walks have a distinctly different way of iterating towards the desired result. Quantum simulation, the earliest suggested application for quantum computation, is in a different category altogether and we will return to it near the end, once we have all the tools necessary to understand how it works. There are also quantum versions of almost anything else computational you can think of, such as quantum game theory, and quantum neural networks. Most of these other types of quantum information processing fall into the category of communication protocols, where the advantage is gained through quantum entanglement shared between more than one person or location. To consolidate our understanding of the functioning of quantum algorithms, we will briefly outline Shor's algorithm and then give an overview of quantum walks.

Shor's Algorithm

The task is to find a factor of a number $N = pq$ where p and q are both prime. First choose a co-prime number $a < N$. If a turns out not to be co-prime then the task is done. This can be checked efficiently using Euclid's algorithm, but only happens in a very small number of cases. Then run the quantum computation shown in Fig. 5 to find r, the periodicity of $a^x \pmod N$. The first half of the computation creates a superposition of $a^x \pmod N$ for all possible inputs x up to N^2. This upper limit is not essential, it determines the precision of the output and thus the number of times the computation has to be repeated on average to succeed. Shor's first insight was that this modular exponentiation step can be done efficiently. The answer is then contained in the entanglement between the qubits in the upper and lower registers [19]. Shor's second insight was that a quantum version of the familiar discrete Fourier transform can be used to rearrange the state so that measuring the upper register allows one to calculate the value of r with high probability. Once we have found r, then $a^{r/2} \pm 1$ gives a factor of N (with high probability). If the first attempt doesn't find a factor, repeating the computation a few times with different values of a rapidly increases the chances of success.

The classical difficulty of factoring is not known, but the best classical algorithms we have are sub-exponential (exponential to some slower than linear function of N), whereas Shor's algorithm is polynomial in N. This is roughly speaking an exponential improvement, and it promises to bring factoring into the set of "easy to solve" problems (i. e. polynomial resources) from the "hard to solve" set (exponential resources). There is a whole family of problems using same method, all based around identifying a hidden subgroup of an Abelian group. Some extensions have been made beyond Abelian Groups but this is in general much harder (for a review, see [20]).

Quantum Walks

Classical random walks underpin many of the best classical algorithms, so finding a faster quantum version of a random walk is a good bet for a new type of quantum algorithm. Quantum random walks that are faster than classical random walks have been found in a variety of guises, but turning them into algorithms is harder. They have been applied most successfully to problems related to searching, such as subset finding [21] and element distinctness [22].

A simple recipe for a quantum walk on a line goes as follows:

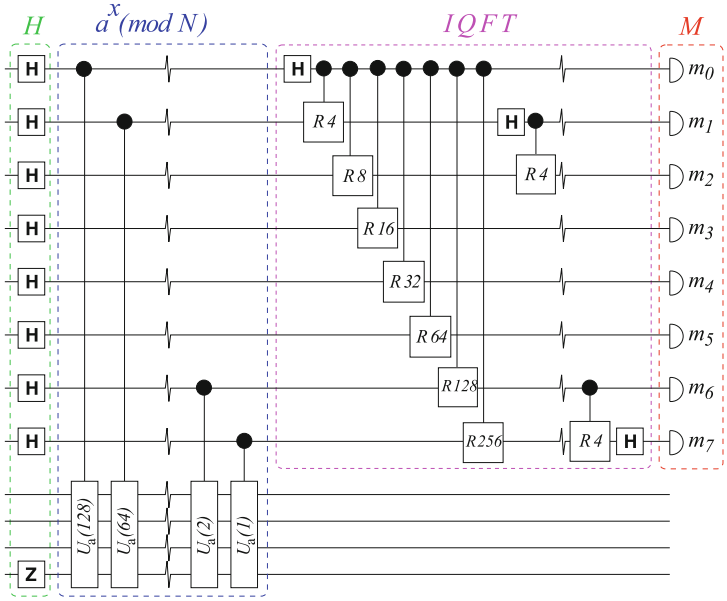
1. Start at the origin
2. Toss a qubit (quantum coin)

$$\mathbf{H}|0\rangle \longrightarrow (|0\rangle + |1\rangle)/\sqrt{2}$$
$$\mathbf{H}|1\rangle \longrightarrow (|0\rangle - |1\rangle)/\sqrt{2}$$

3. Move *left* and *right* according to qubit state

$$\mathbf{S}|x, 0\rangle \longrightarrow |x - 1, 0\rangle$$
$$\mathbf{S}|x, 1\rangle \longrightarrow |x + 1, 1\rangle$$

4. Repeat steps 2. and 3. T times
5. measure position of walker, $-T \leq x \leq T$



Quantum Computing, Figure 5

Circuit diagram for Shor's algorithm for factoring shown for the example of factoring 15. *H* is a Hadamard gate, *R*n is a rotation by π/n, and the *U* gates perform the modular exponentiation (see text). IQFT is inverse quantum Fourier transform, and *M* is the final measurement

If you repeat steps 1. to 5. many times, you get a probability distribution $P(x, T)$, which turns out to spread quadratically faster than a classical random walk.

If we add a little decoherence (measure the quantum walker with prob p at each step) then we get a top hat distribution for just the right amount of noise [23], see Fig. 6. For related results of mixing times on cycles and torii, see Refs. [24,25].

Quantum walks can be used to find a marked item in an unsorted database, the equivalent of finding a person's name from their phone number by searching the telephone directory. Classical solutions have to check on average $N/2$ entries, and in the worst case need to check them all. This problem was investigated by Grover [26,27], who found a quantum algorithm that is quadratically faster than classical searching. His algorithm doesn't use quantum walks, but Shenvi et al. [28] showed that quantum walks can perform the task just as well. The recipe for this is:

1. Start in a uniform distribution over graph with N nodes.
2. Use Grover coin everywhere except the marked node.
3. Run for approx $\frac{\pi}{2}\sqrt{N/2}$ steps.
4. Particle will now be at the marked node with high probability.

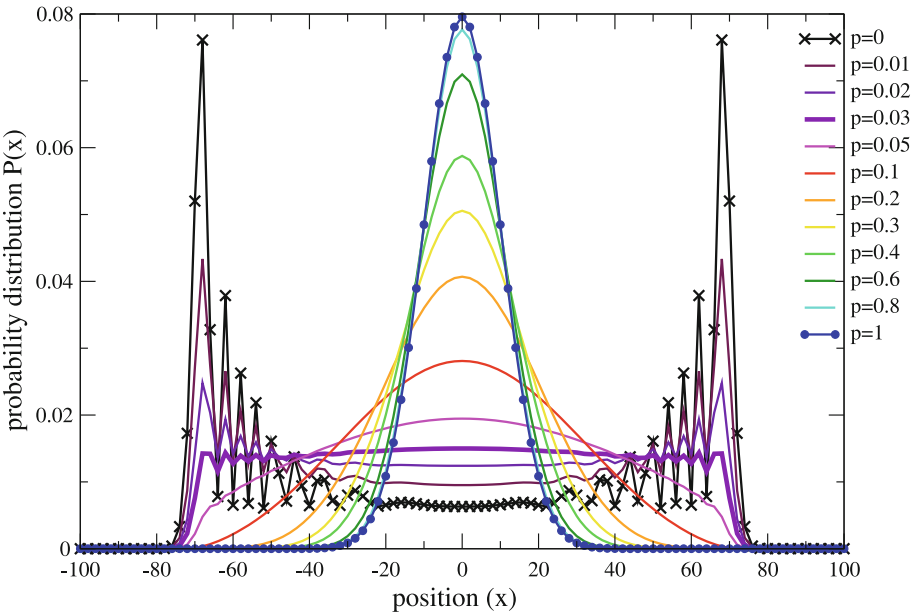
This is the inverse of starting at the origin and trying to get uniform (top hat) distribution. A quadratic speed up is the best that can be achieved by any algorithm tackling this problem [29].

The quantum walk version of Grover's search has been generalized to find more than one item. Magniez et al. [30,31] show how to detect triangles in graphs, Ambainis [22] applies quantum walks to deciding element distinctness, and Childs and Eisenberg [21] generalize to finding subsets. These generally obtain a polynomial improvement over classical algorithms, for example, reducing the running time to $O(N^{2/3})$ compared to $O(N)$ for classical methods.

There is one known problem for which a quantum walk can provide a more impressive speed up. The "glued trees" problem [32] is also about finding a marked state, but this time the starting point is also specified, and an oracle is available to give information about the possible paths between the start and finish. This achieves an exponential speed up, but has not been generalized to other problems. For a short review of quantum walk algorithms, see [33].

How Powerful Is Quantum Computing?

What we've seen so far is quantum computers providing a speed up over classical computation, i. e., new *complexity classes* to add to the zoo [34], but nothing outside of



Quantum Computing, Figure 6
Comparison of a quantum walk of 100 steps with the corresponding classical random walk. Progressively increasing the decoherence produces a "top-hat" distribution for the right choice of decoherence rate

the classical computability limits. The mainstream view is that quantum computing achieves the same *computability* as classical computing, and the quantum advantage may be characterized completely by complexity comparisons. The intuitive way to see that this is likely to be correct is to note that, given the mathematical definition of quantum mechanics, we can always simulate quantum dynamics using classical digital computers, but we generally can't do it efficiently. The chink in this argument is that quantum mechanics is not digital. Qubits can be in an arbitrary superposition $\alpha|0\rangle + \beta|1\rangle$, where $|\alpha|$ can take any real value between zero and one. Thus we can only simulate the quantum dynamics to within some precision dictated by the amount of digital resources we employ. However, we cannot measure α, we can only infer an approximate value from the various measurements we do make. And classical computers also have continuous values for quantities such as voltages that we choose to interpret as binary zeros and ones. So it isn't clear there could be any fundamental difference between classical and quantum in this respect that could separate their computational ability.

So will quantum computers give us a practical advantage over classical computers? How many qubits do we need to make a *useful* quantum computer? This depends on what we want to do:

Simulating a quantum system: for example, $N \times$ two-state particles $\longrightarrow 2^N$ possible different states. The state of the system could be in superposition of all of these 2^N possible states. Classical simulations thus require one complex number per state: $2^{N+1} \times$ size-of-double $\longrightarrow$ 1 Gbyte can store the state of just $N = 26$ 2-state systems. The current record is $N = 36$ in 1 Terabyte of storage [35] – each additional particle *doubles* the memory required. Thus more than 40 or so qubits is beyond current classical resources.

In the cases where we don't need to keep track of all the possible superpositions, e. g., if only nearest neighbor interactions are involved, larger classical simulations can be performed, for example, see [36].

Shor's factoring algorithm: the best classical factoring to date is 200 digit numbers (RSA-200) which is approximately 665 bits. Shor's quantum algorithm needs $2n$ qubits in the QFT register plus $5n$ qubits for the modular exponentiation (lower register plus ancilla qubits), a total of $7n$ logical qubits. A 665 bit number therefore needs 4655 logical qubits. We now need to take account of error correction. Again, this depends on the physical quantum computer and the error rates that have to be corrected. If the error rate is close to the threshold of 10^{-3} to 10^{-4}, then more error correction is needed. For low error rates, maybe 20–200 physical qubits per logical qubit

are required. For high error rates the numbers blow up quickly to maybe 10^5 physical qubits per logical qubit. This suggest that factoring won't be the first useful application of quantum computing, despite being the most famous, we may need Tera-qubit quantum computers to produce useful results here. Although the scaling favors quantum computing, the crossover point is very high.

Ultimate Physical Limits to Computation

This is a subject that is revisited regularly. The version by Lloyd [37] is the best known of these calculations that has an explicitly quantum flavor. Physical limits can be deduced from these fundamental constants:

The speed of light: $c = 3 \times 10^8 \mathrm{m\,s^{-1}}$
Boltzmann's constant: $k = 1.38 \times 10^{-23} \mathrm{J\,K^{-1}}$
The gravitational constant:
$G = 6.67 \times 10^{-11} \mathrm{m^3 kg^{-1}\,s^{-2}}$
Planck's constant: $\hbar = 1.055 \times 10^{-34} \mathrm{J\,s}$

Consider 1 kg of matter occupying volume of 1 liter $= 10^{-3}\,\mathrm{m}^3$ – a little smaller than today's laptops. How fast can this matter compute? Quantum mechanics tells us that the speed of computation is limited by its average energy E to $2E/\pi\hbar = 5 \times 10^{50}$ operations per second with $E = mc^2$ given by the mass of 1 kg. This is because the time-energy Heisenberg uncertainty principle, $\Delta E \Delta t \geq \hbar$, says that a state with a spread in energy of ΔE takes a time Δt to evolve into a different state. This was extended by Margolus and Levitin [38,39] to show that a state with average energy E takes a time $\Delta t \geq \pi\hbar/2E$ to evolve to a different state.

The number of bits it can store is limited by its entropy, $S = k \ln W$ where W is the number of states the system can be in. For m bits, $W = 2^m$, so $S = km$, and the system can store at most m bits. Our estimate of m is thus dependent on what our laptop is made of. It also depends on the energy E, since not all of the W possible states have the same energy, so in practice we don't get the maximum possible $W = 2^m$ number of possible states, since we decided our laptop had a mass-energy of 1 kg. In practice this restriction doesn't change the numbers very much. There are various ways to estimate the number of states, one way is to assume the laptop is made up of stuff much like the early universe: high energy photons (the leftover ones being the cosmic microwave background radiation). This gives around 2×10^{31} bits as the maximum possible memory.

Having noted that entropy S is a function of energy E, we can deduce that the number of operations per second per bit is given by $2Ek/\pi\hbar S \propto kT/\hbar$, where $T = \partial S/\partial E$ is the effective temperature of the laptop. For our 1 kg laptop this means the number of operations per bit is about 10^{19}, and the temperature is about 5×10^8 degrees, a very hot plasma. This also implies the computer must have a parallel type of operation since the bit flip time is far smaller than the time it takes for light to travel from one side to the other. If you compress the computer into a smaller volume to reduce the parallelism it reaches serial operation at the black hole density! The evaporation rate for 1 kg black hole is 10^{-19} seconds, after 10^{32} ops on 10^{16} bits. This is very fast, but similar in size to conventional computers. Prototype quantum computers already operate at these limits, only with most of their energy locked up as mass.

The salutary thing about these calculations is that once you scale back to a computer in which the mass isn't counted as part of the available energy, we are actually within some sort of spitting distance of the physical limits. Computational power has been increasing exponentially for so long we have become used to it doubling every few years, and easily forget that it also requires increasing resources to give us this increased computational capacity. Based on current physics, there isn't some vast reservoir of untapped computational power out there waiting for us to harness as our technology advances, and most of us will see a transition to a significantly different regime in our lifetimes. Which, one may argue, highlights the importance of unconventional computation, where further increases in speed and efficiency may still be available after the standard model has run out of steam.

Can We Build a Quantum Computer?

The short answer to this question is "yes", and people already have built small ones, but we don't know if we can build one big enough to be useful for something we can't already calculate more easily with our classical computers. In 2000, DiVincenzo [40] provided a checklist of five criteria to assess whether a given architecture was capable of being used to build a scalable quantum computer, i. e., one that can be made large enough to be useful:

1. *scalable* qubits – to allow quantum computers to be built large enough to solve real problems
2. prepare initial state – e. g. all qubits in state $|0\rangle$
3. decoherence times far longer than gate times – to allow many quantum operations before the quantum coherence is degraded
4. universal set of quantum gates – a suitable two qubit gate is sufficient
5. measurement of single qubits to read out result – can be hardest part and can be much slower than other operations.

The notion of *scalability* is crucial in the quest to build a useful quantum computer. We need designs that can be tested at small sizes, registers of a few qubits, then scaled up without fundamentally changing the operations we tested at small scales. The most feasible architecture for achieving this is using small, repeatable, connectable units [41].

One of the most important models for scalable quantum computers is the combination of stationary and flying qubits. Flying qubits are usually photons while stationary qubits can be almost any other type that can interact with photons, such as atoms or quantum dots. The motivation for this model is that it can be hard to do everything in one type of physical system, for example, one-qubit gates are easy to apply to photons while two-qubit gates are hard. Also, stationary qubits are limited to nearest neighbor interactions, which then involves many swap operations to allow qubits further apart to be gated together, but this can be overcome using flying intermediaries to connect distant qubits. Atoms in traps combined with photons are currently seen as the best scalable architecture, but this is by no means the only player in the ring.

What Have We Got so Far?

Single qubits – with enough control to initialize them in a chosen state and apply single qubit gates to them – have been demonstrated in a wide variety of systems: this is only a partial list:

- photons – using polarization or path to represent the qubit
- atoms and ions in linear traps – qubit registers
- atoms loaded into optical lattices one per site
- quantum dots – with *single* spins (electrons or holes)
- electrons floating on liquid helium – electron spin
- phosphorus nuclei embedded in silicon – Kane model
- superconducting devices – using either currents or magnetic flux
- nuclear magnetic resonance (NMR) – using nuclear spins as qubits.

Some of these systems have also been demonstrated performing two qubit gates, sometimes even more than one gate at a time maintaining coherence. Progress is slow but steady and this list will likely be out of date by the time you read it, though not likely by more than a few small numbers. For more details of the main types of architectures, [42] is a good starting point. The current capabilities are limited to

- a few (less than 10) qubits entangled to order
- a few (less than 20) quantum gate operations
- NMR can factor 15 (using 7 qubits)
- NMR qubit record is 7 (running an algorithm), and 12 (benchmarking)
- atoms: 6 in controlled superposition.

The vision is driving beautiful experiments: to find out up to date information on current progress, visit the websites of the major experimental collaborations, for example,

- NIST: http://qubit.nist.gov/
- MIT/Quanta: http://www.media.mit.edu/quanta/
- University Vienna: http://www.quantum.at/research/quantum-computation.html
- EU Road map: http://qist.ect.it/Reports/reports.htm

and look for recent articles in Nature and Science.

Unconventional Extensions

Quantum computing as just described appears to mimic classical digital computing quite closely in terms of bits (qubits), gates, error correction, scalable architectures and so on. In many ways this is just because this part of the picture is easiest to relate to those familiar with classical computing. In reality, neither the historical development, nor the creativity with which theorists and experimentalists are tackling the difficult task of trying to build a working quantum computer justify this impression. In this section I will attempt to set the record straight, and hint at some of the wilder territory beyond.

What About Analogue?

Classical analogue computation was once commonplace, yet was swept aside by the advance of digital computation. To understand why this happened we need to review what it means to binary encode our data in a digital computer. Binary encoding has profound implications for the resources required for computation, as this table explains.

Unary vs. Binary Coding		
Number	Unary	Binary
0		0
1	•	1
2	• •	10
3	• • •	11
4	• • • •	100
...	...	...
N	$N \times \bullet$	$\log_2 N$ bits
Read out:	distinguish between measurements with *N* outcomes	$\log_2 N$ measurements with 2 outcomes each
Accuracy:	errors scale linear in *N*	errors scale $\propto \log N$

Binary encoding provides an exponential advantage (reduction) in the amount of memory required to store the data compared to unary encoding. It also means the precision costs are exponentially smaller. To double the accuracy of a unary representation requires double the resources, while a binary representation achieves this with a single extra bit. This is the main reason why digital computing works so well. It is so obvious it is easy to forget that there was once analogue computing in which the data is represented directly as the magnitude of a voltage or water level displacement, or, in the slide rule, with the logarithms etched onto the stick. Those of you too young to have owned a slide rule should now play with Derek's virtual slide rules here: http://www.antiquark.com/sliderule/sim/ to get a feel for what one version of analogue computation is like.

It does not, of course, have to be binary encoding, numbers in base three or base ten, for example, gain the exponential advantage just as well. One of the first to discuss the implications of binary encoding in a quantum computing context was Jozsa [43], and this was refined in Blume–Kohout et al. [44] and Greentree et al. [45]. Binary encoding matters for quantum computing too in the standard digital model. It gains the same exponential advantage from encoding the data into qubits as classical digital computers gain from using bits.

Continuous Variable Quantum Computing

Although many quantum systems naturally have a fixed number of discrete states (polarization of a photon or electronic spin for example), there are also variables like position x and momentum p that can take any real value. These can be used to represent data in the same way as voltages or other continuous quantities are used in classical analogue computers. This quantum version of analogue computation is usually known as continuous variable, or CV computation.

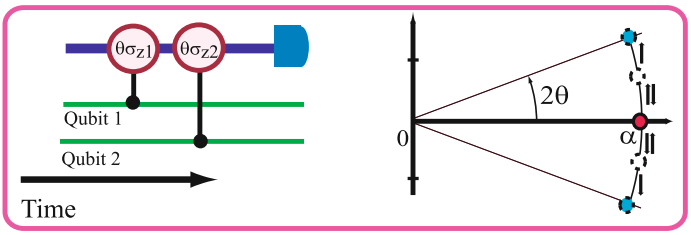
The important difference between quantum and classical continuous variables is that a conjugate pair, such as the position and momentum, are not independent of each other. Instead, x and p are related by the commutator, $[x, p] = i\hbar$ Knowing x exactly implies p is unknowable and vice versa, this is called *squeezing*. The most commonly described continuous variable quantum system uses electromagnetic fields, but rather than specializing to the case of single photons as one does for digital quantum computing, single modes of the field are employed. Braunstein and van Loock [46] provide a comprehensive review of CV quantum information processing for those interested in further details.

Lloyd and Braunstein [47] showed that CV computation is universal, i. e., it can compute anything that a digital quantum computer can compute. They did this by showing that it is possible to obtain any polynomial in x and p efficiently using operations that are known to be available in physical systems. The universal set of operations has to include a nonlinear operation applied to one of the variables, which means in this context higher than quadratic in x and p. For electromagnetic fields, an example of a suitable nonlinear operation is the Kerr effect, which produces a quartic term in x and p. This has the effect of an intensity dependent refractive index. Unfortunately, although it is observable, in real physical systems the Kerr effect is too weak to be of practical use in CV computation. Without the nonlinear term, the operations available are Gaussian preserving, which means they can be simulated efficiently with a classical computer [48]. Some operations are easier than in digital quantum computers, the Fourier transform is a single operation, rather than a long sequence of controlled rotations and Hadamard gates.

Continuous variable information processing is especially suitable for quantum communications because of the practicality of working with bright beams of light instead of single photons. Most quantum communications tasks only require Gaussian operations, which are easily achieved with standard optical elements. Quantum repeaters [49] to extend quantum communications channels over arbitrary distances are a good example. However, a useful CV quantum computer isn't a practical option using current experimental capabilities. The precision that can be obtained in a single mode is limited by the degree of squeezing that can be performed on the mode, currently about 10 dB, which is equivalent to 3.7 bits. An extra bit of precision doubles the required squeezing so CV quantum computation suffers from the same precision problems as classical analogue computing.

Hybrid Quantum Computing Architectures

Having noted above that continuous variable quantum systems are especially suited to quantum communications tasks, the obvious way to use them for quantum computing is to make a high speed quantum "bus" (qubus) to communicate between stationary qubit registers. Although light and matter interact in simple easy ways, creating a suitable gate to transmit the quantum information between qubits is a far from trivial task. Coherent light as used for continuous variable quantum information is also often employed to measure the final state of the matter qubits at the end of the computation.

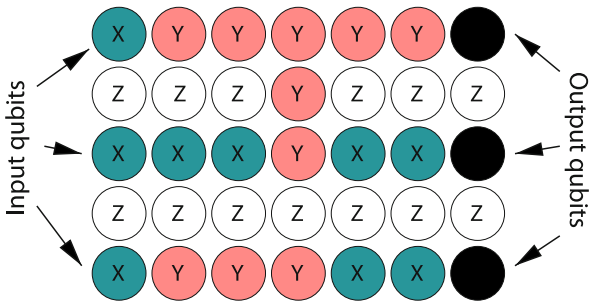


Quantum Computing, Figure 7
Qubus parity gate operation

Unless the required operations are done in a way that reveals only trivial information about the quantum state while the computation is in process, the quantum superpositions will instead be destroyed. The solution is built around an elegant implementation of a parity gate, illustrated in Fig. 7. The left half shows a coherent light beam (blue) interacting with two qubits in turn (green) by means of an interaction that causes a phase shift in the light that depends on the state of the qubit. The right half shows the phase of the light after the interactions. If the two qubits are in the same state, the light will have a phase in one of the two blue positions. If the two qubits are in different states, the two phase shifts will cancel leaving the phase unaltered (red). Provided the angle of the shift, θ is large enough, the two cases can be distinguished by measuring only the coordinate of the horizontal axis, not the vertical distance away from it. For a detailed exposition illustrated in a superconducting qubit setting, see [50].

Cluster State Quantum Computing

First proposed by Raussendorf and Briegel [51], and called a "one way quantum computer", cluster state quantum computing has no direct classical equivalent, yet it has become so popular it is currently almost the preferred architecture for development of quantum computers. Originally designed for atoms in optical lattices, the idea is to entangle a whole array of qubits in global operations, then perform the computation by measuring the qubits in sequence. This architecture trades space resources for a temporal speed up, requiring roughly n^2 physical qubits where the gate model uses n. The measurement sequence can be compressed to allow parallel operation for all but a few measurements that are determined by previous results, and require the outcomes to be fed forward as the computation proceeds. One obvious prerequisite for this architecture is qubits that can be measured quickly and easily. Measurement is often slower than other gate operations so this does limit in practice which types of qubits can be used.



Quantum Computing, Figure 8
Fragment of a cluster state quantum computation. The computation can be thought of as proceeding from left to right, with the qubit register arranged vertically, though all qubits in these gates can be measured at the same time. White qubits do not contribute directly to the computation and are "removed" by measuring in the *Z* basis. Pink qubits are measured in the Y basis and form the gates, at the top a CNOT and at the bottom a Hadamard. Blue qubits are measured in the *X* basis and transmit the logical qubit through the cluster state. The black qubits contain the output state (not yet measured, since their measurements are determined by the next operations). Details in [52]

Since the measurements proceed from one side of the lattice to the other, see Fig. 8, instead of making the whole array in one go, the qubits can be added in rows as they are needed. This also allows the use of a probabilistic method for preparing the qubits, which can be repeated until it succeeds in a separate process to the main computation [53]. This idea grew out of schemes for linear optical quantum computation in which gates are constructed probabilistically and teleported into the main computation once they succeed [54]. Adding rows of qubits "just in time" does remove one of the main advantages of the original proposal, i. e., making the cluster state in a few simple global operations. However, it gains in another respect in that the data is regularly being moved into fresh qubits, so the whole system doesn't have to be kept free of decoherence throughout the computation.

Cluster state quantum computation has been important on a theoretical level also because it made people think

harder about the role of measurement in quantum computing [55]. Instead of striving to obtain perfect unitary quantum evolution throughout the computation, cleverly designed measurements can also be used to push the quantum system along the desired trajectory.

Topological Quantum Computing

In Sect. "Digital Quantum Computing" we learned about qubits: distinguishable quantum two state systems. But there are more exotic quantum species, indeed, quantum particles are indistinguishable when more than one of them is around in the same place. This means if you swap two of them, you can't tell anything happened. And there are precisely two ways to do this for particles living in our familiar three spatial dimensions:

$$|\psi_A\rangle_1|\psi_B\rangle_2 \longrightarrow |\psi_B\rangle_1|\psi_A\rangle_2 \tag{5}$$

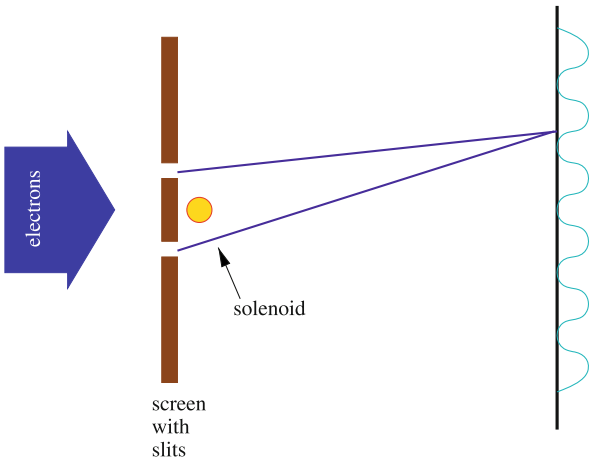
corresponding to *Bose–Einstein particles* (photons, Helium atoms), or

$$|\psi_A\rangle_1|\psi_B\rangle_2 \longrightarrow -|\psi_B\rangle_1|\psi_A\rangle_2 \tag{6}$$

corresponding to *Fermi–Dirac particles* (electrons, protons, neutrons). That extra minus sign can't be measured directly, but it does have observable effects, so we do know the world really works this way. One important consequence is that Fermi–Dirac particles cannot occupy the same quantum state, whereas Bose–Einstein particles can, and frequently do. From this we get, among many other things, lasers (photons – Bose–Einstein particles) and semiconductors (electrons – Fermi–Dirac particles), without which modern technology would be very different indeed.

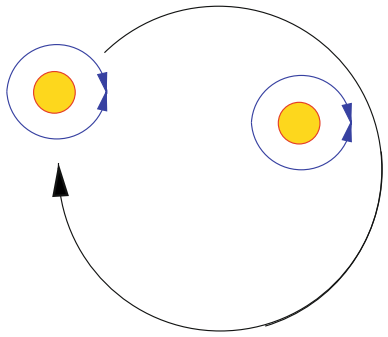
A simple change of sign when swapping particles won't make a quantum computer, but something in between one and minus one can. For this we have to restrict our particles to two spatial dimensions. Then it is possible to have particles, known as *anyons*, that acquire more complicated phase factors when exchanged. Restricting quantum particles to less than three spatial dimensions is not as hard as it sounds. A thin layer of conducting material sandwiched between insulating layers can confine electrons to two-dimensional motion, for example.

The first proposal for a quantum computer using anyons was from Kitaev [56], using an array of interacting spins on a planar surface. The anyons are formed from excitations of the spins – think of this as similar to the way electronic displays such as those found in trains and stations often work by flipping pixels between black and yellow to form characters. This opened up the field to a host of connections with group theory, gauge theory, and braid (knot) theory, and produced a new quantum algorithm to approximate the Jones polynomial [57], a fundamental structure in braid theory. If you already know something about any of the above mathematical subjects, then you can jump into the many accounts of topological quantum computation (also going by the names geometric or holonomic quantum computing). A recent introduction can be found in [58].



Quantum Computing, Figure 9
Aharonov Bohm affect

Here is a simple example of how quantum phases work, to further illuminate the idea. The Aharonov–Bohm effect [59] was discovered in 1959 and has been observed experimentally. It involves two ingredients. First, electrons can be projected through a double slit much like photons can, and they produce an interference pattern when detected on the far side of the slits. If you aren't familiar with Young's double slit experiment [60], note that you need to think about it in terms of single electrons going through both slits in superposition. This is an example of the quantum behavior of particles that combines both wave and particle properties. The interference pattern arises because the electron can't land on the detector in places where the path length difference is half a wavelength, and prefers to land near where the path difference is a whole number of wavelengths. The second ingredient is a solenoid, a coil of wire with a current running through it. It generates a magnetic field inside the coil that does not leak outside. If you put the solenoid between the two slits, see Fig. 9, the interference pattern shifts, even though there is no magnetic field where the electrons are traveling (they can't pass through the solenoid).



Quantum Computing, Figure 10
Anyons: charge (*blue*) encircling magnetic field (flux, *red+yellow*)

From this we deduce two things: interference patterns are a way to detect the extra quantum phases that are generated in topological quantum effects, and, even though there was no interaction between the electrons and the magnetic field, i. e., no energy exchanged, the magnetic field still affected the quantum properties of the electrons just by being inside the two paths the electrons take through the slits.

Now imagine shrinking that experiment down to a magnetic field with a charge (electron) encircling it. This is one way to think of anyons. Confined to two dimensions, the anyons can be moved round each other, and the charge part of one going round the magnetic field part of the other affects the quantum phase. When you want to measure them, you manipulate them so that they form an interference pattern. The rest of the details of how to make them perform quantum computation comes down to the mathematics of group theory.

Apart from delighting the more mathematically-minded quantum computing theorists, this type of quantum computing may have some practical advantages. The effects are generated just by moving the anyons about, and the exact path isn't critical, only that they go around the right set of other anyons. So there is less chance of errors and disturbance spoiling the quantum computation. However, constructing a physical system in which anyonic excitations can be manipulated to order is not so easy. In fact, it has not yet been demonstrated experimentally at all, though there is no shortage of suggestions for how to do it, see for example, [61].

Adiabatic Quantum Computing

Adiabatic quantum computation is the quantum equivalent of simulated annealing. Simulated annealing is a computational method that finds the minimum solution to a problem by starting from an initial state and evolving towards lower values until a minimum is found. In order to avoid being stuck in a local minimum that is higher than the desired solution, some random moves that increase the value are necessary. This can be thought of as starting in a high temperature distribution and slowly cooling the simulation until the lowest energy state is found. The slowness of the cooling is to allow it to get out of local minima on the way.

Quantum mechanics provides a more effective tool for this problem, thanks to the quantum adiabatic theorem. Quantum states don't get stuck in local minima, they can "tunnel" through barriers to lower-lying states. So there is no need to start in a random initial state, instead, the evolution starts in a minimum energy state that is easy to prepare and transforms from there into the minimum energy state we are interested in. The quantum adiabatic theorem tells you how long it will take to evolve to the desired state without ending up in some higher energy states by mistake.

Suppose you can prepare a system in the grounds state $|\sigma_0\rangle$ of the Hamiltonian H_{simple}, and you want to find the ground state $|\phi_0\rangle$ of the Hamiltonian H_{hard}. Then, provided you can evolve the systems under both Hamiltonians together, you applies the following Hamiltonian $H(t)$ to $|\sigma_0\rangle$,

$$H(t) = \{1 - \alpha(t)\}H_{\text{simple}} + \alpha(t)H_{\text{hard}} \tag{7}$$

where $\alpha(0) = 0$, $\alpha(T) = 1$ and $0 \leq \alpha(t) \leq 1$. The total time the simulation will take is T, and $\alpha(t)$ controls how fast the Hamiltonians switch over. The key parameter is the difference, or gap Δ, between the ground and first excited energy states. It varies as the state evolves from $|\sigma_0\rangle$ to $|\phi_0\rangle$ and the smaller it is the slower the state must evolve to stay in the ground state. Let Δ_{min} be the smallest gap that occurs during the adiabatic evolution. Then the quantum adiabatic theorem says that provided $T \gg \varepsilon/\Delta^2_{\text{min}}$, the final state will be $|\phi_0\rangle$ with high probability. Here ε is a quantity that relates the rate of change of the Hamiltonian to the ground state and first excited state. It is usually about the same size as the energy of the system. The important parameter is thus Δ_{min}. For hard problems, Δ_{min} will become small and the required time T will be large.

Adiabatic quantum computing was introduced by Farhi et al. [62] as a natural way to solve SAT problems. SAT stands for satisfiability and a typical problem is expressed as a Boolean expression, for example,

$$\begin{aligned} &B(x_1, x_2, x_3, x_4) \\ &= (x_1 \vee x_3 \vee \tilde{x}_4) \wedge (\tilde{x}_2 \vee x_3 \vee x_4) \wedge (\tilde{x}_1 \vee x_2 \vee \tilde{x}_3) \end{aligned} \tag{8}$$

a set of clauses of three variables each combined with "or" that are then combined together with "and". The tilde over a variable indicates the negation of that variable. The problem is to determine whether there is an assignment of the variables x_1, x_2, x_3, x_4 that makes makes the expression true (in the above simple example there are several such assignments, such as $x_1 = 0, x_2 = 0, x_3 = 0, x_4 = 0$).

The Boolean expression is turned into a Hamiltonian by constructing a cost function for each clause. If the clause is satisfied, the cost is zero, if not, the cost is one unit of energy. The Hamiltonian is the sum of all these cost functions. If the ground state of the Hamiltonian is zero the expression can be satisfied, but if it is non-zero then some clauses remain unsatisfied whatever the values of the variables.

The adiabatic quantum computation only works efficiently if it runs in a reasonably short amount of time, which in turn requires the gap, $\Delta_{\min}$ to remain large enough throughout the evolution. Figuring out whether this is so is not easy in general, and has led to some controversy about the power of quantum computation. The 3SAT problem (clauses with three variables as above) is in the general case a hard problem that requires exponential resources to solve classically. Clearly adiabatic quantum computing can also solve it, but the question is how long it takes. The best indications are that the gap will become exponentially small for hard problems, and thus the adiabatic quantum computation will require exponential time to solve it. For dissenting views see, for example, [63] discussing the traveling salesman problem, another standard "hard" problem. One of the difficulties in analyzing this issue is that adiabatic quantum computation uses the continuous nature of quantum mechanics, and care must be taken when assessing the necessary resources to correctly account for the precision requirements.

On the other hand, we do know that adiabatic quantum computing is at least equivalent to digital computation (i. e., it can solve all the same problems using equivalent resources) [57,64,65]. There are also indications that it can be more resilient to some types of errors [66]. It has even been implemented in a toy system using nuclear magnetic resonance (NMR) quantum computation [67].

Quantum Simulation

We have already noted that quantum simulation was the original application that inspired Feynman [3] and Deutsch [4] to propose the idea of quantum computation. We have also explained why quantum systems cannot generally be simulated efficiently by classical digital computers, because of the need to keep track of the enormous number of superpositions involved. In 1996 Lloyd [68] proved that a quantum system can simulate another quantum system efficiently, making the suggestions from Feynman and Deutsch concrete.

The first step in quantum simulation is trivially simple, the quantum system to be simulated is mapped onto the quantum simulator by setting up a one-to-one correspondence between their state spaces (which are called Hilbert spaces for quantum systems). The second step is non-trivial, Lloyd proved that the unitary evolution of the quantum system being studied can be efficiently approximated in the quantum simulator by breaking it down into a sequence of small steps that use standard operations. Mathematically this can be written

$$\exp\{iHt\} \simeq \left(\exp\{iH_1 t/n\}\exp\{iH_2 t/n\} \ldots \exp\{iH_j t/n\}\right)^n + O(t^2/n)[H_j, H_k]$$

where $\exp\{iHt\}$ is the unitary evolution of the quantum system (H is called the Hamiltonian) and $H_1 \ldots H_j$ are a sequence of standard Hamiltonians available for the quantum simulator. The last term says that the errors are small provided n, the step size, is chosen to be small enough. There are variations and improvements on this that use higher order corrections to this approximation, see for example, [69]. Quantum simulation has even been demonstrated experimentally [70] in a small test system using an nuclear magnetic resonance (NMR) quantum computer.

Just as with the quantum algorithms discussed in the first section, evolving the quantum simulation is only half of the job. Extracting the pertinent information from the simulation requires some tricks, depending on what it is you want to find out. A common quantity of interest is the spectral gap, the energy difference between the ground and first excited states, the quantity Δ discussed in Subsect. "Adiabatic Quantum Computing". One way to obtain this is to create a superposition of the two energy states and evolve it for a period of time. The phase difference between the ground and excited state will now be proportional to the spectral gap. To measure the phase difference requires a Fourier transform, either quantum or classical. Either way, the simulation has to be repeated or extended to provide enough data points to obtain the spectral gap to the desired precision. Even creating the initial state in superposition of ground and excited states is non-trivial. It has been shown that this can be done for some systems of interest by using quasi-adiabatic evolution starting from the ground state. By running the adiabatic evolution a bit too fast, the state of the system doesn't remain perfectly in

the ground state and the required superposition in generated [71,72].

However, because there is no binary encoding, unlike classical simulation on a digital computer, accuracy is a problem. In particular, accuracy does not scale efficiently with time needed to run the simulation [73]. Is this going to be a problem in practice? Poor scaling for accuracy didn't stop people building useful analogue computers from electrical circuits or water pipes in the days before digital computers were widely available. As we noted, a quantum simulator becomes useful at around 40 qubits in size, which means it may well be the first real application of quantum computing to calculate something we didn't know already. In the end, the accuracy required depends on the problem we are solving, so we won't know whether it works until we try.

Future Directions

The wide range of creative approaches to quantum computing are largely a consequence of the early stage of development of the field. There is as yet no proven working method to build a useful quantum computer, so new ideas stand a fair chance of being the winning combination. We do not even have any concrete evidence that we will be able to build a quantum computer that actually outperforms classical computers. But we also have no evidence to the contrary, that quantum computers this powerful are not possible for some fundamental physical reason, so the open question, plus the open questions such a quantum computer may be able to solve, keep the field vibrant and optimistic.

Beyond the precision quantum engineering required to construct any of the designs described thus far, there is plenty of discussion of quantum computational processes in natural systems. While all natural systems are made up of particles that obey quantum mechanics at a fundamental level, the extent to which their behavior requires quantum logic to describe it is less ubiquitous. Quantum coherences are more usually dissipated into the surrounding environment than marshaled into co-ordinated activity.

It has been suggested that the brain may exhibit quantum computational capabilities on two different levels. Firstly, Hameroff and Penrose [74] have argued that brain cells amplify quantum effects that feed into the way the brain functions. This is disputable on physical grounds when the time-scales and energies involved are considered in detail. Secondly, a quantum-like logic for brain processes has been proposed, for example, see [75]. This does not require actual quantum systems to provide it, and is thus not dependent on individual brain cells amplifying quantum effects.

Whether any biological computing is exploiting quantum logic is an open question. Since typical biological temperatures and time scales are generally not a hospitable regime for maintaining quantum coherences, we might expect any such quantum effects to be rare at the level of complex computations. Clearly quantum effects are biologically important for of basic processes such as photosynthesis, for example, and transport properties such as the conductance and coherence of single electrons through biological molecules are the subject of much current study. These are perhaps more likely to feature as problems solved by one of the first quantum simulators, than as examples of natural quantum computers.

Bibliography

Primary Literature

1. Bartlett SD, Rudolph T, Spekkens RW (2006) Reference frames, superselection rules, and quantum information. Rev Mod Phys 79:555; ArXiv:quant-ph/0610030
2. Dawson CM, Nielsen MA (2006) The Solovay-Kitaev algorithm. Quantum Inf Comp 6:81–95; The Solovay-Kitaev theorem dates from 1995, but was only partially published in pieces – this Ref. gives a more comprehensive review; ArXiv:quant-ph/0505030
3. Feynman RP (1982) Simulating physics with computers. Int J Theor Phys 21:467
4. Deutsch D (1985) Quantum-theory, the church-turing principle and the universal quantum computer. Proc R Soc Lond A 400(1818):97–117
5. Knill E, Laflamme R, Zurek W (1996) Threshold accuracy for quantum computation. ArXiv:quant-ph/9610011
6. Aharonov D, Ben-Or M (1996) Fault tolerant quantum computation with constant error. In: Proc 29 th ACM STOC. ACM, NY 176–188; ArXiv:quant-ph/9611025
7. Steane A (1996) Multiple particle interference and quantum error correction. Proc Roy Soc Lond A 452:2551; ArXiv:quant-ph/9601029
8. Shor PW (1997) Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. SIAM J Sci Statist Comput 26:1484
9. Bennett CH, Brassard G (1984) Quantum cryptography: public-key distribution and coin tossing. In: IEEE International Conference on Computers, Systems and Signal Processing. IEEE Computer Society Press, Los Alamitos, pp 175–179
10. Bennett CH, Wiesner SJ (1992) Communication via one – and two-particle operators on Einstein–Podolsky–Rosen states. Phys Rev Lett 69(20):2881–2884
11. Bennett CH, Brassard G, Crépeau C, Jozsa R, Peres A, Wootters WK (1993) Teleporting an unknown quantum state via dual classical and Einstein–Podolsky–Rosen channels. Phys Rev Lett 70:1895–1899
12. Graham-Rowe D (2007) Quantum ATM rules out fraudulent web purchases. New Sci 2629:30–31

13. Duligall JL, Godfrey MS, Harrison KA, Munro WJ, Rarity JG (2006) Low cost and compact quantum cryptography. New J Phys 8:249; ArXiv:quant-ph/0608213v2
14. Steane A (2001) Quantum computing and error correction. In: Gonis, Turchi (eds) Decoherence and its implications in quantum computation and information transfer. IOS Press, Amsterdam, pp 284–298
15. Lidar DA, Chuang IL, Whaley KB (1998) Decoherence free subspaces for quantum computation. Phys Rev Lett 81:2594–2598; ArXiv:quant-ph/9807004v2
16. Lidar DA, Whaley KB (2003) Decoherence-free subspaces and subsystems. In: Benatti F, Floreanini R (eds) Irreversible Quantum Dynamics. Lecture Notes in Physics, vol 622. Springer, Berlin, pp 83–120; ArXiv:quant-ph/0301032
17. Misra B, Sudarshan ECG (1977) The Zeno's paradox in quantum theory. J Math Phys 18:756
18. Beige A, Braun D, Tregenna B, Knight PL (2000) Quantum computing using dissipation to remain in a decoherence-free subspace. Phys Rev Lett 85:762–1766; ArXiv:quant-ph/0004043v3
19. Nielsen M, Chuang I (1996) Talk at KITP Workshop: Quantum Coherence and Decoherence, organized by DiVencenzo DP and Zurek W. http://www.kitp.ucsb.edu/activities/conferences/past/. Accessed 2 Sep 2008
20. Lomont C (2004) The hidden subgroup problem-review and open problems. ArXiv:quant-ph/0411037
21. Childs A, Eisenberg JM (2005) Quantum algorithms for subset finding. Quantum Inf Comput 5:593–604; ArXiv:quant-ph/0311038
22. Ambainis A (2004) Quantum walk algorithms for element distinctness. In: 45th Annual IEEE Symposium on Foundations of Computer Science. IEEE Computer Society Press, Los Alamitos, pp 22–31
23. Kendon V, Tregenna B (2003) Decoherence can be useful in quantum walks. Phys Rev A 67:042315; ArXiv:quant-ph/0209005
24. Richter P (2007) Almost uniform sampling in quantum walks. New J Phys 9:72; ArXiv:quant-ph/0606202
25. Richter P (2007) Quantum speedup of classical mixing processes. Phys Rev A 76:042306; ArXiv:quant-ph/0609204
26. Grover LK (1996) A fast quantum mechanical algorithm for database search. In: Proc 28th Annual ACM STOC. ACM, NY, p 212; ArXiv:quant-ph/9605043
27. Grover LK (1997) Quantum mechanics helps in searching for a needle in a haystack. Phys Rev Lett 79:325; ArXiv:quant-ph/9706033
28. Shenvi N, Kempe J, Whaley BK (2003) A quantum random walk search algorithm. Phys Rev A 67:052307; ArXiv:quant-ph/0210064
29. Bennett CH, Bernstein E, Brassard G, Vazirani U (1997) Strengths and weaknesses of quantum computing. SIAM J Comput 26(5):151–152
30. Magniez F, Santha M, Szegedy M (2003) An $o(n^{1.3})$ quantum algorithm for the triangle problem. ArXiv:quant-ph/0310134
31. Magniez F, Santha M, Szegedy M (2005) Quantum algorithms for the triangle problem. In: Proceedings of 16th ACM-SIAM Symposium on Discrete Algorithms. Society for Industrial and Applied Mathematics, Philadelphia, pp 1109–1117
32. Childs AM, Cleve R, Deotto E, Farhi E, Gutmann S, Spielman DA (2003) Exponential algorithmic speedup by a quantum walk. In: Proc 35th Annual ACM STOC. ACM, NY, pp 59–68; ArXiv:quant-ph/0209131
33. Ambainis A (2003) Quantum walks and their algorithmic applications. Int J Quantum Information 1(4):507–518; ArXiv:quant-ph/0403120
34. Scott Aaronson (2005) The complexity zoo. http://qwiki.caltech.edu/wiki/Complexity_Zoo. Accessed 2 Sep 2008
35. De Raedt K, Michielsen K, De Raedt H, Trieu B, Arnold G, Richter M, Lippert Th, Watanabe H, Ito N (2007) Massive parallel quantum computer simulator. Comp Phys Comm 176:127–136
36. Cirac JI, Verstraete F, Porras D (2004) Density matrix renormalization group and periodic boundary conditions: A quantum information perspective. Phys Rev Lett 93:227205
37. Lloyd S (2000) Ultimate physical limits to computation. Nature 406:1047–1054; ArXiv:quant-ph/9908043
38. Margolus N, Levitin LB (1996) The maximum speed of dynamical evolution. In: Toffoli T, Biafore M, Liao J (eds) Physcomp96. NECSI, Boston
39. Margolus N, Levitin LB (1998) The maximum speed of dynamical evolution. Physica D 120:188–195; ArXiv:quant-ph/9710043v2
40. DiVincenzo DP (2000) The physical implementation of quantum computation. Fortschritte der Physik 48(9–11):771–783
41. Metodi TS, Thaker DD, Cross AW, Chong FT, Chuang IL (2005) A quantum logic array microarchitecture: Scalable quantum data movement and computation. In: 38th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO'05). IEEE Computer Society Press, Los Alamitos, pp 305–318; ArXiv:quant-ph/0509051v1
42. Spiller TP, Munro WJ, Barrett SD, Kok P (2005) An introduction to quantum information processing: applications and realisations. Comptemporary Physics 46:407
43. Jozsa R (1998) Entanglement and quantum computation. In: Huggett SA, Mason LJ, Tod KP, Tsou S, Woodhouse NMJ (eds) The Geometric Universe, Geometry, and the Work of Roger Penrose. Oxford University Press, Oxford, pp 369–379
44. Blume-Kohout R, Caves CM, Deutsch IH (2002) Climbing mount scalable: Physical resource requirements for a scalable quantum computer. Found Phys 32(11):1641–1670; ArXiv:quant-ph/0204157
45. Greentree AD, Schirmer SG, Green F, Lloyd Hollenberg CL, Hamilton AR, Clark RG (2004) Maximizing the hilbert space for a finite number of distinguishable quantum states. Phys Rev Lett 92:097901; ArXiv:quant-ph/0304050
46. Braunstein SL, van Loock P (2005) Quantum information with continuous variables. Rev Mod Phys 77:513–578
47. Lloyd S, Braunstein SL (1999) Quantum computation over continuous variables. Phys Rev Lett 82:1784; ArXiv:quant-ph/9810082v1
48. Bartlett S, Sanders B, Braunstein SL, Nemoto K (2002) Efficient classical simulation of continuous variable quantum information processes. Phys Rev Lett 88:097904; ArXiv:quant-ph/0109047
49. Ladd TD, van Loock P, Nemoto K, Munro WJ, Yamamoto Y (2006) Hybrid quantum repeater based on dispersive cqed interactions between matter qubits and bright coherent light. New J Phys 8:164; ArXiv:quant-ph/0610154v1 doi:10.1088/1367-2630/8/9/184
50. Spiller TP, Nemoto K, Braunstein SL, Munro WJ, van Loock P, Milburn GJ (2006) Quantum computation by communication. New J Phys 8:30; ArXiv:quant-ph/0509202v3
51. Raussendorf R, Briegel HJ (2001) A one-way quantum computer. Phys Rev Lett 86:5188–5191

52. Raussendorf R, Browne DE, Briegel HJ (2003) Measurement-based quantum computation with cluster states. Phys Rev A 68:022312; ArXiv:quant-ph/0301052v2
53. Nielsen MA (2004) Optical quantum computation using cluster states. Phys Rev Lett 93:040503; ArXiv:quant-ph/0402005
54. Yoran N, Reznik B (2003) Deterministic linear optics quantum computation utilizing linked photon circuits. Phys Rev Lett 91:037903; ArXiv:quant-ph/0303008
55. Jozsa R (2005) An introduction to measurement based quantum computation. ArXiv:quant-ph/0508124
56. Kitaev YA (2003) Fault-tolerant quantum computation by anyons. Annals Phys 303:2–30; ArXiv:quant-ph/9707021v1
57. Aharonov D, van Dam W, Kempe J, Landau Z, Lloyd S, Regev O (2004) Adiabatic quantum computation is equivalent to standard quantum computation. ArXiv:quant-ph/0405098
58. Brennen GK, Pachos JK (2007) Why should anyone care about computing with anyons? Proc Roy Soc Lond A 464(2089):1–24; ArXiv:0704.2241v2
59. Aharonov Y, Bohm D (1959) Significance of electromagnetic potentials in quantum theory. Phys Rev 115:485–491
60. Young T (1804) Experimental demonstration of the general law of the interference of light. Phil Trans Royal Soc Lon, London, p 94
61. Jiang L, Brennen GK, Gorshkov AV, Hammerer K, Hafezi M, Demler E, Lukin MD, Zoller P (2007) Anyonic interferometry and protected memories in atomic spin lattices. ArXiv:0711.1365v1
62. Farhi E, Goldstone J, Gutmann S, Sipser M (2000) Quantum computation by adiabatic evolution. ArXiv:quant-ph/0001106
63. Kieu TD (2006) Quantum adiabatic computation and the travelling salesman problem. ArXiv:quant-ph/0601151v2
64. Kempe, Kitaev, Regev (2004) The complexity of the local hamiltonian problem. In: Proc 24th FSTTCS. pp 372–383; ArXiv:quant-ph/0406180
65. Kempe, Kitaev, Regev (2006) The complexity of the local hamiltonian problem. SIAM J Comput 35(5):1070–1097
66. Childs AM, Farhi E, Preskill J (2002) Robustness of adiabatic quantum computation. Phys Rev A 65:012322; ArXiv:quant-ph/0108048
67. van Dam S, Hogg, Breyta, Chuang (2003) Experimental implementation of an adiabatic quantum optimization algorithm. Phys Rev Lett 90(6):067903; ArXiv:quant-ph/0302057
68. Lloyd S (1996) Universal quantum simulators. Science 273:1073–1078
69. Berry DW, Ahokas G, Cleve R, Sanders BC (2007) Efficient quantum algorithms for simulating sparse hamiltonians. Commun Math Phys 270:359; ArXiv:quant-ph/0508139v2
70. Somaroo SS, Tseng CH, Havel TF, Laflamme R, Cory DG (1999) Quantum simulations on a quantum computer. Phys Rev Lett 82:5381–5384; ArXiv:quant-ph/9905045
71. Abrams DS, Lloyd S (1997) Simulation of many-body fermi systems on a universal quantum computer. Phys Rev Lett 79:2586–2589; ArXiv:quant-ph/9703054v1
72. Wu LA, Byrd MS, Lidar DA (2002) Polynomial-time simulation of pairing models on a quantum computer. Phys Rev Lett 89:057904. Due to a proofs mix up there is also an Erratum: 89:139901; ArXiv:quant-ph/0108110v2
73. Brown KR, Clark RJ, Chuang IL (2006) Limitations of quantum simulation examined by simulating a pairing hamiltonian using nuclear magnetic resonance. Phys Rev Lett 97:050504; ArXiv:quant-ph/0601021
74. Hameroff SR, Penrose R (1996) Conscious events as orchestrated spacetime selections. J Conscious Stud 3(1):36–53
75. Khrennikov A (2006) Brain as quantum-like computer. BioSystems 84:225–241; ArXiv:quant-ph/0205092v8

Further Reading

For those who seriously want to learn the quantitative details of quantum computing, this is still the best textbook: Nielsen MA, Chuang IL (2000) Quantum Computation and Quantum Information. CUP, Cambs

For lighter browsing but still with all the technical details, there are several quantum wikis developed by the scientists doing the research:

Quantiki http://www.quantiki.org/wiki/index.php/Main_Page specifically quantum information

Qwiki http://qwiki.stanford.edu/wiki/Main_Page covers wider quantum theory and experiments

For those still struggling with the concepts (which probably means most people without a physics degree or other formal study of quantum theory), there are plenty of popular science books and articles. Please dive in, it's the way the world we all live in works, and there is no reason not dig in deep enough to marvel at the way it fits together and puzzle with the best of us about the bits we can't yet fathom.

Quantum Computing with Trapped Ions

Wolfgang Lange
University of Sussex, Brighton, UK

Article Outline

Glossary

Ion trap Device used for confining charged particles to a small volume of space. There are two types of ion

traps: in a *Paul trap*, the confinement is achieved by using an electric quadrupole radio-frequency field, while a Penning trap uses a static electric quadrupole field and a static magnetic field. For the right field parameters, the charged particles can be stored in the trapping volume permanently.

Lamb–Dicke regime Describes the conditions for strong confinement of ions. In the Lamb–Dicke regime, ions are confined to a region smaller than the wavelength of the optical transition of the ion. Another property of ions in this parameter range is that the ion's recoil energy is less than the energy of a single quantum of vibration in the trap. This suppresses spontaneous sideband transitions. The Lamb–Dicke regime is essential for the reliable operation of laser-induced quantum gates in an ion-trap quantum computer.

Qubit The unit of quantum information processing. A contraction of quantum bit, the term qubit designates quantum system with two quasi-stable states carrying the quantum equivalent of binary information ("0" or "1"). In an ion-trap quantum computer, two long-lived electronic states are employed as quantum memory. Their wavefunctions are represented here by the symbols $|g\rangle$ and $|e\rangle$. The qubit states are manipulated by means of laser pulses.

Quantum register A collection of qubits forms a quantum register. The power of a quantum computer increases exponentially with the size N of the quantum register, since in this case 2^N numbers can be processed in parallel. In ion traps, the largest quantum register implemented so far has size $N = 8$.

Normal modes The motion of ions in a linear string is strongly coupled due to their Coulomb repulsion. The system can still be described by analogy with a set of independent harmonic oscillators, when collective modes of motion are considered, in which all ions oscillate in phase and at the same frequency. These are called normal modes. A linear string of N ions has N normal modes of axial vibration. Each mode has a characteristic distribution of motional amplitudes for the ions. In the lowest frequency mode, all ions oscillate at the same amplitude, so that the string moves like a rigid body.

Phonon bus Since all ions participate in the collective motion, it can be used to transfer quantum information between ions in the string. To this end, each normal mode is considered as a quantum mechanical oscillator, in which quanta of vibrational motion (phonons) can be excited or de-excited. By coupling their excitation to transitions of the ion-qubit, the phonons serve as a quantum data bus to other ions. Reliable transfer requires that the phonons are restricted to a binary system. The state with no vibrational excitation encodes the qubit $|0\rangle$, one quantum of vibration corresponds to $|1\rangle$.

Rabi-oscillations The term describes the change of state of a two-level atom, induced by the coherent excitation with a laser beam close to resonance. After half a period (phase π), the population is completely transferred, after another half period it is returned to the initial distribution again. By choosing suitable phases, amplitudes, detunings and duration of the pulses, the individual qubits may be manipulated in a fully controlled way and quantum gates between different ions may be induced (via the phonon bus).

Coherent superposition According to the laws of quantum mechanics, a qubit can be in an arbitrary superposition of the two basis states with complex amplitudes α and β, written as $\alpha|g\rangle + \beta|e\rangle$. The relative phase is important, for example, $|g\rangle - |e\rangle$ and $|g\rangle + |e\rangle$ are distinct superposition states. As long as a well-defined phase is preserved, the qubit is in a coherent superposition. This is a precondition for quantum information processing.

Decoherence Any loss of coherence of a quantum state is called decoherence. There are many sources of decoherence. In an ion-trap quantum processor, they range from spontaneous decay of the qubit-levels to phase shifts in fluctuating ambient fields. A quantum computation can only proceed reliably while decoherence is negligible. Quantum error correction is a way to counteract decoherence.

Entanglement Entanglement is one of the most important properties of quantum systems and has no classical analogue. A composite system, for example two qubits, is entangled if the states of the individual particles cannot be separated and regarded as independent. Instead, there are strong non-local links between the components, resulting in quantum correlations and state changes of one partner upon measurement of the other. Entangled states have many applications, from spectroscopy to quantum networking. Up to eight ions have been entangled in a deterministic way.

Quantum network A system of either local or distant nodes performing quantum calculations and exchanging results via transport of ions or via photon links. Setting up the latter requires a controlled exchange of quantum data between ions and photons, providing flying qubits. The goal is distributed entanglement, for example to transfer qubits over long distances via teleportation and eventually to perform distributed quantum computation.

Definition of the Subject

In quantum computation, the classical bit as carrier of information is replaced by a two-state quantum system, the quantum bit (*qubit*). The enhanced computing power of qubits is based on two facts: (1) they can simultaneously store arbitrary superpositions of two values 0 and 1 and (2) superpositions in different qubits may be strongly correlated (entangled), even in the absence of any direct physical link. Quantum algorithms exploit these features to efficiently solve problems whose complexity makes any classical method unfeasible.

As a new paradigm of computer science, the concept of quantum computation has generated important results in complexity theory. However, from a practical point of view, the power of quantum information processing can only be unleashed if the necessary quantum hardware is available. What is required is a set of individually accessible binary quantum systems to serve as a quantum register. They must have strong externally controllable interactions between them, but at the same time no coupling to the environment. While there is no system in which these conditions are ideally fulfilled, trapped atomic ions are the closest realization. At present, they provide the most advanced implementation of quantum information processing.

This article summarizes the state-of-the-art of quantum computing with trapped ions. As shown below, all necessary components of a trapped-ion quantum computer have been demonstrated, from quantum memory and fundamental quantum logic gates to simple quantum algorithms. Current experimental efforts are directed towards scaling up the small systems investigated so far and enhancing the fidelity of operations to a level where error correction can be applied efficiently. The first task at which a quantum computer is expected to outperform a classical one is the efficient simulation of quantum systems too complex for classical treatment.

Introduction

The idea of using atomic-scale systems for information processing was first suggested by R. P. Feynman in his lecture *There's plenty of room at the bottom* [1]. In the early 1980s, Feynman and P. Benioff found that by exploiting the dynamics of quantum systems, computations could be performed far more efficiently than using a classical computer [2,3]. They argued that the difficulty of efficiently simulating quantum systems on a classical computer implied the superior power of quantum computing.

The radically novel idea was to not merely use quantum mechanics as a framework for predicting the macroscopic behavior of a system, but to manipulate individual quantum objects directly. While Feynman didn't propose a particular physical implementation, the spectacular progress achieved in atomic and optical physics laboratories around the world in preparing and manipulating single atoms and ions makes these particles an obvious choice as a qubit. Binary quantum information is typically stored in two internal electronic levels.

Owing to their electric charge, ions can be readily trapped by radio-frequency electric fields. However, Coulomb repulsion keeps the trapped ions so far apart, that any direct interaction involving their internal states is negligible. This seems to preclude logic gates for two or more qubits. However, in 1995, I. Cirac and P. Zoller found a way around this problem in a seminal paper which initiated experimental ion-trap quantum computation as an active research field [4]. They proposed to use the long-distance Coulomb repulsion between ions to couple different qubits in a linear string by exchanging single quanta of their vibration.

Trapped ions have other advantages that make them strong contenders as quantum bits. Their internal and external (motional) states can be manipulated using laser light. In atoms or ions, long-lived excited states exist, allowing for the storage and retrieval of quantum superposition states. Since the first experiment demonstrating a simple quantum gate with a single ion, impressive progress has been made in developing or adapting ion trap technologies for quantum information processing.

In this article, the range of techniques used in ion-trap quantum computing is presented, starting with the technological foundations of ion trapping (Sect. "Ion Trap Technology"). The subsequent discussion follows a list of general requirements for a system to serve as a practical device for quantum computation. It was compiled as a general guideline by D. DiVincenzo [5].

(1) The quantum system must provide well-characterized qubits forming a scalable quantum register (Sect. "Ions as Carriers of Quantum Information").
(2) It must be possible to initialize the qubits to a known state (Sect. "Laser Cooling and State Initialization").
(3) A method to efficiently measure individual qubits must exist (Sect. "State Detection of Ionic Qubits").
(4) A universal set of quantum gates is required (Sects. "Single-Ion Operations" and "Two-Qubit Interaction and Quantum Gates").
(5) The time over which quantum states lose coherence should be much longer than gate operation times (Sect. "Decoherence").

At present, trapped ion systems are the only technology, for which all five criteria have been successfully demon-

strated. Major achievements are two-qubit ion gates with a fidelity around 96%, the entanglement of up to 8 ions and a number of simple algorithms (Sect. "Quantum Algorithms").

In his article, DiVincenzo has added two more requirements in case local quantum processors are to exchange quantum information over some distance.

(6) It must be possible to interconvert the stationary qubits of the quantum computer to "flying qubits", suitable for long-distance transmission.
(7) The flying qubits must be faithfully transmitted between specified locations.

These requirements become more important, as schemes for distributed quantum computation are being developed. The obvious choice for flying qubits are photons and there are first results regarding the coupling of ions and photons (Sect. "Distributed Quantum Information with Trapped Ions").

Ion Trap Technology

Ions are particularly well suited for quantum information processing, since due to their charge, they can be confined by electromagnetic fields without affecting their internal electronic levels. Presently, all realizations of quantum information processing with ions use variants of the Paul-trap, in which the ponderomotive force of a time-dependent inhomogeneous electric field leads to stable confinement [6,7]. For a quantum register containing multiple ions, a linear trap geometry is generally chosen.

Linear Paul Trap

The linear Paul trap has its origins in the electric quadrupole mass filter [8], in which transverse confinement of ions with a certain charge to mass ratio is achieved by a radiofrequency quadrupole potential in a plane perpendicular to the axis of the device, assumed to be oriented in the z-direction:

$$\Phi(x, y, t) = (U - V\cos\Omega t)\ \frac{x^2 - y^2}{2r_0^2}\,, \tag{1}$$

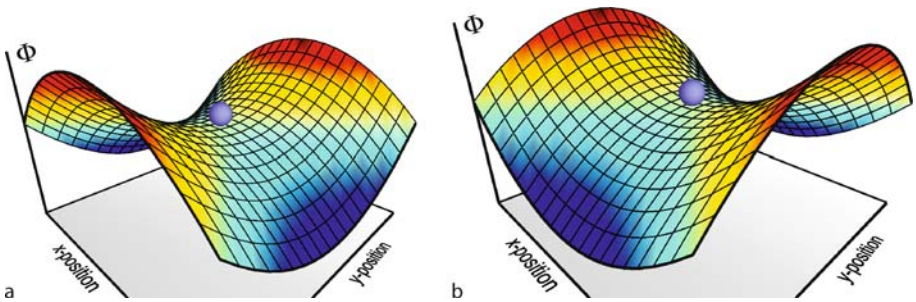
where $\pm U/2$ are the dc-voltages and $\pm V/2\cos\Omega t$ the radiofrequency-voltages applied to the quadrupole electrodes at a distance r_0 from the trap center. Equation (1) represents a saddle potential which at any particular time provides confinement in one direction only, as shown in Fig. 1. By using alternating voltages and averaging over a period of the radiofrequency, stable trapping may be achieved in both transverse directions. Figure 2 indicates how the trap parameters V and U must be chosen to obtain confinement. In practice, $U = 0$ and $eV \ll m\Omega^2 r_0^2/2$ are chosen, where m and e denote mass and charge of the ion.

To generate the potential (1), four hyperbolically shaped electrodes are required. Usually these are approximated by simpler structures like rods with circular or triangular cross section, since close to the trap axis the resulting anharmonicity is negligible.

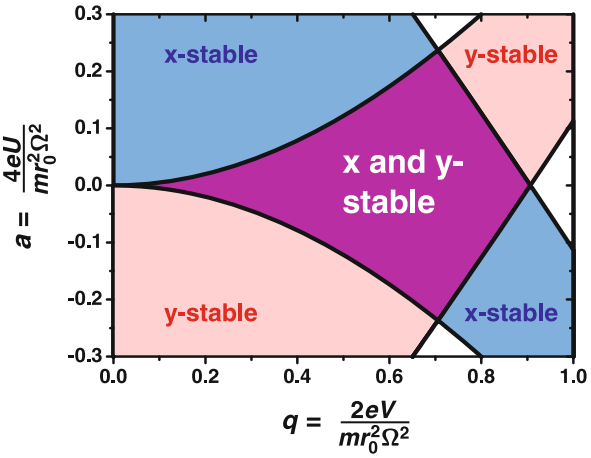
On a timescale long compared to the period of the radiofrequency, the ions move as if they are radially confined in a parabolic pseudopotential given by $\Psi = e^2|\nabla\Phi|^2/4m\Omega^2$. The motion in the radial directions is therefore harmonic, with the radial *secular frequency*

$$\omega_r \approx \frac{eV}{\sqrt{2}m\Omega r_0^2}\,. \tag{2}$$

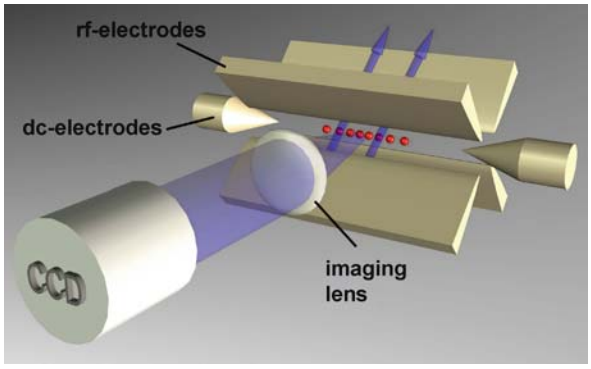
The pseudopotential Ψ provides confinement in the radial direction only, while the motion along the z-axis is not restricted. For three-dimensional confinement, an additional static potential must be applied in the z-direction



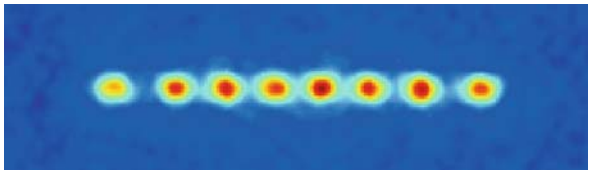
Quantum Computing with Trapped Ions, Figure 1
Two-dimensional saddle potential for a charged particle in the linear Paul trap according to Eq. (1). Due to the quadrupole shape, at any given time there is confinement only along one axis. a Potential at $t = 0$, providing y-confinement only. b Potential at $t = \pi/\Omega$, providing x-confinement only. Stable confinement in both directions is achieved by rapidly alternating the sign of the potential



Quantum Computing with Trapped Ions, Figure 2
Stability diagram for the linear Paul trap as a function of radiofrequency amplitude *V* and dc-amplitude *U*, scaled to dimensionless parameters *q* and *a*, respectively. Confinement in both *x*- and *y*-direction is achieved in the purple region at the center, defining the operating range of the trap



Quantum Computing with Trapped Ions, Figure 3
Schematic drawing of a linear ion trap, as used for quantum information processing. The ions are arranged in a linear string along the trap axis, held by rf- and dc electric fields. The state of the ions is detected by imaging them to a CCD camera. © R. Blatt, University of Innsbruck, Austria



Quantum Computing with Trapped Ions, Figure 4
String of eight ions in a linear Paul trap [10]. Color indicates the intensity of the fluorescent light detected with a CCD camera. The average distance between adjacent ions is about 10 µm. See Refs. [9,11,12] for other ion crystal structures. © R. Blatt, University of Innsbruck, Austria

which is either achieved using additional electrodes at either end of the trap or by segmenting the linear rod electrodes and applying a positive dc-voltage U_z to the outer segments. This provides a static harmonic well along the z-axis which is characterized by the longitudinal trap frequency

$$\omega_z = \sqrt{\frac{2\kappa e U_z}{m z_0^2}} \,. \tag{3}$$

Here, z_0 is half the length between the axially confining electrodes and κ is a geometric factor accounting for the specific electrode configuration. The presence of the axially confining field weakens the radial confinement which is reduced to

$$\omega_r' = \sqrt{\omega_r^2 - \frac{1}{2}\omega_z^2} \,. \tag{4}$$

A sketch of a linear ion trap is shown in Fig. 3. Values for $\omega_r/2\pi$ in ion traps used for quantum information processing range from 3 MHz to 10 MHz, while typical values for $\omega_z/2\pi$ are 1 MHz to 4 MHz.

If multiple ions are confined and cooled to a sufficiently low temperature in a Paul trap (see Sect. "Vibrational Cooling of a Single Ion"), they form ordered structures [9,10,11]. If the radial confinement is strong enough ($\omega_r \gg \omega_z$), ions arrange themselves in a linear pattern along the trap axis at distances determined by the equilibrium of their mutual Coulomb repulsion and the potential providing axial confinement. Figure 4 shows an example of a string of eight $^{40}Ca^+$ ions in a linear trap. The equilibrium spacing of the ions is not uniform and must be determined numerically [12].

The number of ions N that fit on the axis of a linear ion trap is limited by the ratio of radial to axial confinement to approximately $N < 1.82\,(\omega_r/\omega_z)^{1.13}$ [14]. For larger ion numbers, the equilibrium positions no longer coincide with the z-axis of the trap. This must be avoided, since off the axis, the ions undergo *micromotion*, an oscillatory motion around their equilibrium position, driven by the trapping field at frequency Ω. Only on the trap axis, the ions are protected from micromotion, since ideally it corresponds to a node of the confining radiofrequency-field. However, micromotion can still occur if stray electric fields shift the ions off the nodal line of the radiofrequency field. Since this may result in rf-heating of the ions and precludes controlled quantum operations, stray fields must be carefully compensated [15,16].

Spherical Paul Trap

If only a single ion is to be stored, the spherical variant of the Paul trap may be used [17]. Here, the average ponderomotive force of an alternating quadrupole potential is used to provide three-dimensional confinement. The corresponding potential is

$$\Phi(x, y, z, t) = (U - V\cos\Omega t)\,\frac{x^2 + y^2 - 2z^2}{2r_0^2}\,. \quad (5)$$

The ideal electrodes are two hyperbolic endcaps along the z-axis and a hyperbolic ring in the $x - y$-plane, but simplified configurations with only two cylindrical endcaps or only one ring-electrode [18,19,20,21] are possible.

In a spherical Paul trap, the micromotion vanishes completely only at the origin, which is why it is suited only for the storage of a single ion. It was used in early studies of quantum information processing in an ion trap [22,23].

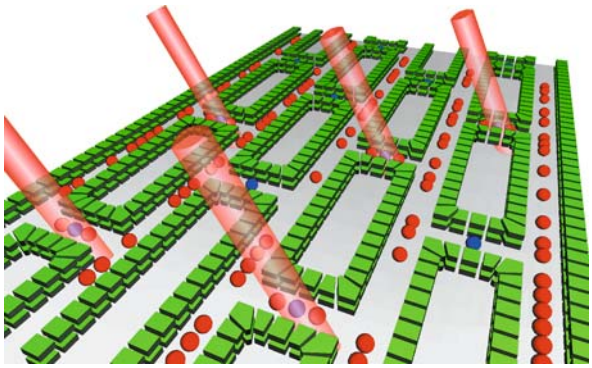
In the future, spherical Paul traps may again play an important role as single-ion microtraps, arranged in large, scalable arrays [24]. Coupling of adjacent ions would be accomplished through phase shifts, mediated by state-dependent Coulomb interaction rather than the exchange of vibrational quanta, as will be discussed in Sect. "Vibrational Coupling" for quantum information processing in linear ion strings.

Large-Scale Ion Traps

There is no fundamental limit to the length of a string of trapped ions and hence to the size of a multi-ion quantum register. However, the technical challenges of manipulating many ions in a single trapping zone are substantial. In particular, the large number of vibrational modes of a long ion crystal makes reliable transfer of quantum information difficult (see Sect. "Vibrational Coupling"). So far, quantum operations have been performed with up to eight ions in a single trap.

In order to circumvent the problems associated with long ion chains, trap architectures made up of interconnected individual trap segments have been proposed [25]. In these schemes, processing of quantum information is performed locally in special regions of the trap containing only a small number of ions, which are retrieved from and returned to memory regions. Ion transfer is achieved by changing the axially confining fields with dc-electrodes placed along the trapping zones. This type of trap is also known as a *quantum charge-coupled device* (QCCD). A sketch of a possible electrode layout is shown in Fig. 5.

An important issue for quantum computation in segmented traps is to maintain coherence of internal states during transport of the qubits. This has been demonstrated over a straight distance of 1.2 mm after separating two ions originally held in a single trap [26]. In order to extract arbitrary ions from a quantum register, junctions have to be implemented, for example, a T-junction joining two linear trap sections at an angle of 90°. A trap of this type was realized at the University of Michigan and used to swap the positions of two ions [27].



Quantum Computing with Trapped Ions, Figure 5
Sketch of an ion trap with multiple segments, providing memory regions for storing qubits (*red*) and interaction regions for logic operations, including ions to sympathetically cool the vibrational motion (*blue*). Shuttling between different locations is accomplished by suitable dc-voltages applied to the electrodes shown in *green*

Microtraps and Surface-Electrode Traps

The number of electrodes required for multi-zone traps quickly grows as more processing or memory segments are included. These devices require integrated fabrication technologies to replace the manual assembly and alignment of electrodes used in simpler traps. Two approaches have been pursued to make ion traps amenable to microfabrication.

(1) *Three-dimensional microtraps*, fabricated from a monolithic multilayer microchip [28,29]. The trapping geometry is similar to that in a macroscopic trap but minimum features are smaller and no assembly is required. A chip-trap has been fabricated from a doped gallium-arsenide heterostructure [30].
(2) *Surface-electrode traps* further simplify the fabrication of trapping structures. They are derived from three-dimensional Paul traps, but have all their electrodes moved to a single plane [31,32]. The ions are again trapped in a minimum of the pseudopotential, which is typically located several tens of μm above the sur-

face. A planar trap for single ions has been demonstrated at NIST [31].

Using microfabrication techniques for ion traps makes them suitable for miniaturization and scaling. As an additional benefit of chip-traps, microelectronics, e. g., for electrode potential control can be placed directly on the chip [33].

Penning Traps

An alternative way to trap ions in three dimensions is the Penning trap [6]. In contrast to the Paul-trap, there are no alternating fields. Stable confinement of ions is accomplished with a static magnetic field, providing a radial force in the xy-plane and a static electric field exerting a restoring force in the z-direction. While the z-motion in the Penning trap is harmonic, in the radial plane the ions orbit the center in a combination of cyclotron and magnetron motion. This makes a fixed radial arrangement of ions difficult. However, by using a rotating electric field technique, large radial ion crystals rotating at a controlled rate were produced in a Penning trap [34]. Such a collection of ions is equivalent to a quantum hard disk and has been proposed for quantum information processing [35].

Another scheme for quantum computing with Penning traps uses an array of miniature Penning traps each containing only a single ion [36]. Ions could be shuttled between individual traps using suitable arrangements of electric fields. To perform two-bit quantum gates, two ions would be combined in a single trap, forming an axial crystal, which can be manipulated by analogy with to the equivalent structure in a Paul trap. Axial two-ion Coulomb crystals have recently been observed in experiment [37].

An array of Penning traps holding single *electrons* has been proposed as a scalable quantum information processor [38]. Here, gates are implemented by controlling the Coulomb interaction with radio-frequency and microwave techniques. In the remainder of this article, Penning traps will not be further discussed, since to date, all realizations of quantum information processing have been performed in Paul traps.

Ions as Carriers of Quantum Information

In an ion-trap quantum computer, information is encoded in two internal (electronic) states of the ion. Among the large number of levels, only those with a long natural lifetime are suitable for storing quantum information. This excludes levels with electric dipole transitions to lower lying states. For easy manipulation of the qubit-transition, the levels should be coupled either by a two-photon Raman transition, an electric quadrupole or a magnetic dipole transition. Ideally, the states should be insensitive to fluctuations of external magnetic or electric fields to minimize decoherence.

Two classes of states have been used for quantum information processing with ions so far: two hyperfine levels of the ground state of an ion with nuclear spin and a combination of an electronic ground state and a low-lying metastable state.

Hyperfine State Qubits

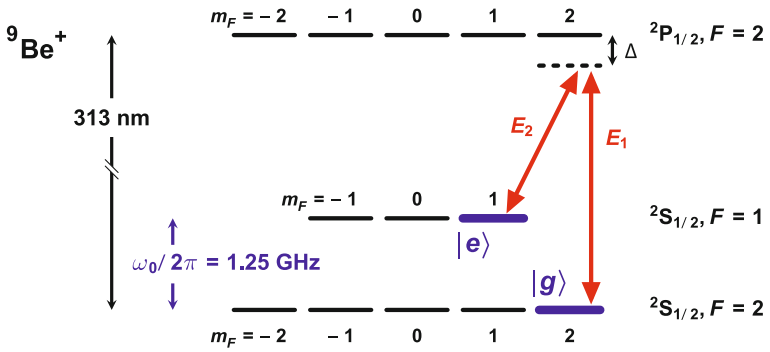
In any isotope with non-vanishing nuclear spin, the electronic ground state energy levels are split by the interaction of the electrons' angular momentum with the nuclear magnetic moment (and, if applicable, the electric quadrupole moment). These so-called hyperfine states are coupled by magnetic dipole transitions and hence have an extremely long lifetime. Therefore, they are ideally suited for reliably storing quantum information. An example is the $^9Be^+$ ion, which was used in the first implementations of quantum information processing with ions [22]. The two hyperfine levels are distinguished by the quantum numbers $F = 1$ and $F = 2$, differing in the relative orientation of electronic angular momentum and nuclear spin.

The relevant level structure of $^9Be^+$ is shown in Fig. 6. The two basis states of the quantum bit are represented by a magnetic sublevel of each hyperfine state. A common choice is $|F = 2, m_F = 2\rangle$ and $|F = 1, m_F = 1\rangle$. While the outermost levels with $m_F = \pm F$ are most conveniently initialized, they are subject to random shifts by fluctuating magnetic fields, leading to decoherence. For a quantum memory with longer lifetime, a first-order magnetic-field independent transition at a finite magnetic bias field should be used [39].

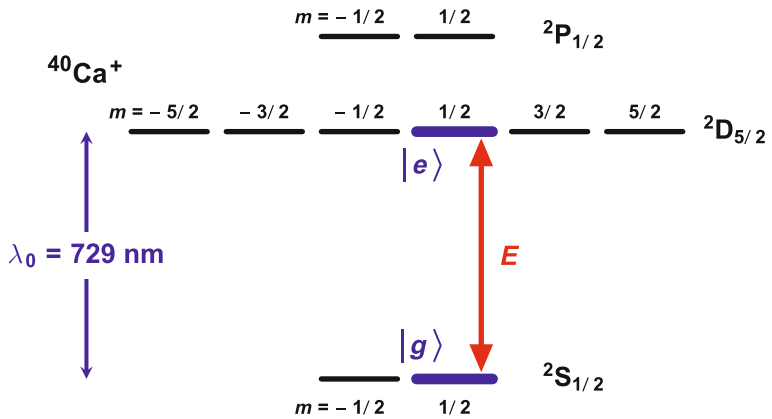
Metastable State Qubits

Long lifetimes for ionic qubits may also be achieved by employing excited electronic states which are connected to lower lying levels only by higher order electromagnetic multipole transitions. This is the case for ions with a D-state with lower energy than the first excited P-state. It decays to the ground state via an electric quadrupole transition with a rate around $1s^{-1}$. An example is $^{40}Ca^+$ [40], which is used for ion-trap quantum information processing in a number of laboratories. Here, the states $|^2S_{1/2}, m_j = 1/2\rangle$ and $|^2D_{5/2}, m_j = 1/2\rangle$ are usually chosen as basis states. The level structure of $^{40}Ca^+$ is shown in Fig. 7.

All ions investigated for quantum information processing fall in either of these two categories. An overview



Quantum Computing with Trapped Ions, Figure 6
Qubit-implementation in $^9Be^+$. As indicated, the quantum information is stored in the two magnetic sublevels $|F = 2, m_F = 2\rangle$ and $|F = 1, m_F = 1\rangle$. In order to change the state of the qubit, Raman transitions driven by laser fields E_1 and E_2, off-resonant from the level $P_{1/2}$, are employed



Quantum Computing with Trapped Ions, Figure 7
Qubit-implementation in $^{40}Ca^+$. Quantum information is stored in a magnetic sublevel of the ground state $|^2S_{1/2}, m_j = 1/2\rangle$ and of the metastable state $|^2D_{5/2}, m_j = 1/2\rangle$. The state of the qubit is changed by resonantly driving the quadrupole transition between S and D with a laser field E at 729 nm

of isotopes that have been used is given in Table 1. In the remainder of this article, the two physical basis states of a qubit will be designated by the symbols $|g\rangle$ for the lower (ground-) state and $|e\rangle$ for the higher (excited) state, as indicated in Fig. 6 and 7. Note that in the literature, alternative notations are used, e. g., $|\downarrow\rangle$ and $|\uparrow\rangle$ for hyperfine qubits, $|S\rangle$ and $|D\rangle$ symbolizing the orbital states involved, or the generic $|0\rangle$ and $|1\rangle$. The 2^N basis states of a quantum register with N ions are represented by analogy with single qubit-states by specifying N entries: $|x_1x_2x_3 \ldots x_N\rangle$, where x_i is either g or e.

Quantum Computing with Trapped Ions, Table 1
Isotopes used in investigations of quantum information processing with trapped ions

Scheme	Ion Species
Hyperfine qubit	$^9Be^+$ [22,39], $^{43}Ca^+$ [41,42], $^{111}Cd^+$ [43], $^{171}Yb^+$ [44,45], $^{25}Mg^+$ [46]
Metastable qubit	$^{40}Ca^+$ [40,41], $^{88}Sr^+$ [47,48,49,50]

Laser Cooling and State Initialization

Ionization

Before a quantum register can be initialized, the ion trap must be loaded with the desired isotope. This is most efficiently achieved by photo-ionizing an atomic beam [51,52,53]. If one of the excitation steps in multi-photon ionization is resonant, a specific isotope may be loaded, distinguished by its isotope shift [51,54]. If no suitable laser source for photo-ionization is available, electron-impact ionization may be used instead. It is not state-

selective and hence only suited for loading the most abundant species among an ion's isotopes. Due to the long trap lifetimes of many hours to days, the trap has to be loaded only infrequently.

Initialization of the Quantum Register

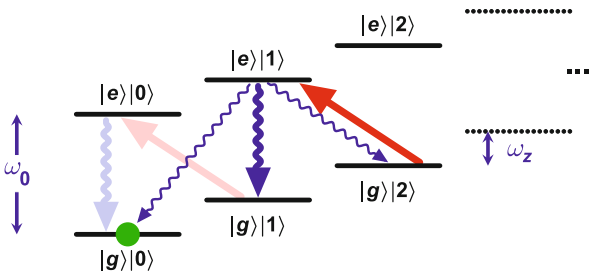
Any quantum computation must start from a well-defined, i. e. pure state of the quantum register. Therefore, initially all ions must be transferred to a specific internal state, typically the lower basis state, so that the register is initialized to $|ggg\ldots g\rangle$. In ions, this is achieved by the well-established method of optical pumping [55]. A laser with a suitable polarization repeatedly excites the ions to a set of states from which they eventually decay to the desired state $|g\rangle$. This state is either not coupled to the exciting laser or driven to a state whose only decay channel is back to $|g\rangle$ (cycling transition). In any case, all ions reach their initial state after a few excitation-emission cycles. In contrast to the quantum gates discussed below, the initialization of the register is irreversible.

Not all internal states are amenable to initialization by optical pumping with high fidelity. In some cases, it may be necessary to pump to an auxiliary state, which is then coherently transferred to the desired initial state by suitable laser pulses [39].

Vibrational Cooling of a Single Ion

In an ion-trap quantum processor, different ions are coupled by the Coulomb interaction via their motional degrees of freedom (see Sect. "Vibrational Coupling"). In order to transfer quantum information in this way, the ions' harmonic oscillation in the trapping potential must be cooled close to its quantum mechanical ground state. This is achieved by different stages of laser cooling, exploiting the momentum transfer between ions and light.

Initial cooling of the ions is provided by Doppler cooling [56,57,58], using an atomic transition with a natural line width Γ larger than the vibrational frequencies (ω_r or ω_z) of the ions in the trap. If the cooling laser is slightly red-detuned from resonance, the Doppler-shift from the ions' motion ensures that photons are preferably absorbed when an ion is moving towards the laser. The resulting momentum transfer, averaged over many scattering events, reduces the kinetic energy and thus the temperature of the ion. The final temperature which can be reached by this technique is given by the Doppler temperature $T_D = \hbar\Gamma/2k_B$ [59] where k_B denotes Boltzman's constant. T_D is on the order of a few millikelvin, which for typical trap frequencies corresponds to a residual excitation of one to tens of vibrational quanta.



Quantum Computing with Trapped Ions, Figure 8
Resolved sideband cooling in a two-level system (states $|g\rangle$, $|e\rangle$). Associated with each level is a ladder of quantized vibrational states $|n\rangle$, labeled by their vibrational quantum number. Red sideband excitation (*red arrow*), followed predominantly by spontaneous emission on the carrier, reduces the vibrational excitation by one. After a few cycles, the vibrational ground state (*green circle*) is reached, which is not coupled to the excitation

In order to reach the vibrational ground state, an additional cooling stage must be applied. To this end, a transition is used whose Γ is smaller than the frequency of the vibration to be cooled, e. g., $\Gamma \ll \omega_z$. In this case, the ion's absorption spectrum is composed of a line at the transition frequency ω_0 and a series of sidebands at $(\omega_0 \pm n\omega_z)$ with integer n. The strength of these sidebands is given by the vibrational excitation.

Very efficient cooling is obtained by tuning a laser to the first red sideband at $\omega_0 - \omega_z$ (see Fig. 8). If the ion is well localized (Lamb–Dicke regime), subsequent spontaneous emission occurs predominantly at the carrier frequency ω_0, so that there is a net reduction of the ion's kinetic energy by $\hbar\omega_z$. After a few excitation-emission cycles, the ground state of vibration is reached with high probability.

The technique is known as sideband cooling. The ion's mean vibrational quantum number in the trapping potential is reduced to $\langle n\rangle = (\Gamma/2\omega_z)^2 \ll 1$ [60,61]. The technical challenge is to find a transition narrower than the longitudinal trap frequency ω_z. In ions with calcium-like level schemes, this is possible on the $S_{1/2} - D_{5/2}$ quadrupole transition [62]. For hyperfine-qubits, it is necessary to use a two-photon stimulated Raman transition for sideband-cooling [61].

Vibrational Cooling of Ion Strings

In an ion string, the motion of ions is strongly coupled due to their Coulomb repulsion. It is therefore more natural to consider N normal modes of motion of the string in the z-direction, rather than N individual ions. The lowest normal mode has frequency ω_z, the same as that of a sin-

gle ion. Normal modes of ion strings will be discussed in Sect. "Normal Modes".

The cooling of a normal mode proceeds by analogy with the cooling of a single ion. It is not required that all the ions in a string are interacting with the laser. Momentum transfer from the mode can already be achieved through a single ion in the string. This fact provides the possibility of using different ions for cooling and data operations. In this case, one ion species is used for sideband-cooling the collective motion of the entire string, while the remaining ions are used in the subsequent quantum processing. This method is called sympathetic cooling [63,64,65,66]. While only the one vibrational mode to be used as a quantum data-bus must be cooled to the ground state, the remaining $N-1$ vibrational modes should be cold enough not to interfere with quantum operations, ideally also residing in their ground state of motion.

When all ions in the string are in a known internal state and the vibrational modes cooled to (or close to) their ground states of motion, the quantum register is ready for quantum information processing [67].

Single-Ion Operations

After initialization, the state of each qubit may be modified by means of suitable laser pulses applied on the qubit transition. Figure 10 shows the geometry for laser excitation. An important precondition is that the ions may be addressed individually. This requires the laser beam to be focused tighter than the minimum distance between ions, given by [12]

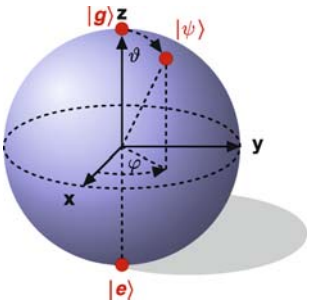
$$s_{\min} \approx \left(\frac{2e^2}{\pi\epsilon_0 m\omega_z^2}\right)^{1/3} N^{-0.56} \qquad (6)$$

for a string of N ions with mass m. For typical experimental parameters, $s_{\min}$ ranges from 4 to 10 µm. Focusing must be considerably better to minimize residual laser intensity at the position of adjacent ions and scattered light, which would lead to unwanted excitation of other ions in the string. The laser beam is switched between ions by means of an electro-optic deflector.

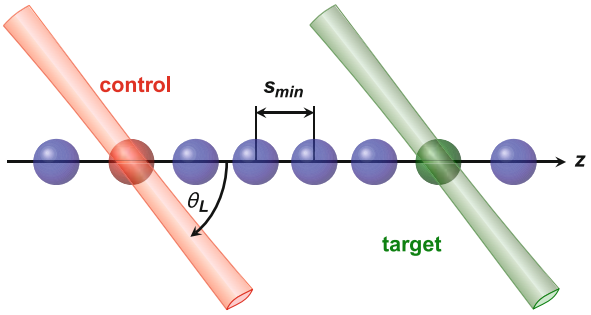
Single-qubit operations are best visualized in the Bloch picture. Here, a general pure state of a qubit is parametrized by two angles, a polar angle ϑ $(0 \le \vartheta \le \pi)$ and an azimuthal angle φ $(0 \le \varphi \le 2\pi)$ such that

$$|\psi\rangle = \cos\frac{\vartheta}{2}\,|g\rangle + \sin\frac{\vartheta}{2}\,e^{i\varphi}\,|e\rangle\ . \qquad (7)$$

Each qubit state can be represented by a vector pointing in the direction given by (ϑ, φ) (the Bloch-vector), ending on the surface of a sphere of unit radius, the so-called



Quantum Computing with Trapped Ions, Figure 9
Bloch-sphere representation of the state of a qubit. Each point on the surface of the sphere represents a quantum state $|\psi\rangle$ of the two-level system, with the poles representing the basis states $|g\rangle$ and $|e\rangle$. A suitable laser pulse rotates the state through a polar angle θ and an azimuthal angle ϕ, so that arbitrary points on the sphere can be reached



Quantum Computing with Trapped Ions, Figure 10
Geometry of selective excitation of ions in a quantum register. Shown are the two positions of a laser beam, consecutively addressing control- and a target-ion in a two-qubit quantum gate

Bloch-sphere shown in Fig. 9. The poles represent the basis states $|g\rangle$ and $|e\rangle$, while equal weight superpositions of these two states lie on the equator. Note that the phase associated with $|g\rangle$ is arbitrarily chosen to be zero. This is no longer possible in two-qubit operations, when differences between the phases of qubits are important.

In ionic qubits, the states can be manipulated by exciting the ion with electromagnetic pulses. The transition frequencies of qubits encoded in *metastable* atomic states (see Sect. "Metastable State Qubits") are in the optical domain, so that a direct transition driven by a narrowband laser may be used. In the case of $^{40}Ca^{+}$, this is the S-D quadrupole transition (cf. Fig. 7).

As any two-level atom resonantly excited by electromagnetic radiation, the qubit undergoes Rabi-oscillations, i. e., periodically changes between the states $|g\rangle$ and $|e\rangle$. This corresponds to a rotation of the Bloch-vector. The oscillation frequency Ω_0 (Rabi-frequency for resonant exci-

tation) is given by the time-dependent amplitude $E_0(t)$ of the exciting field and the (quadrupole) coupling strength:

$$\Omega_0(t) = E_0(t)\,\frac{\omega_L Q}{2\hbar c}\,. \tag{8}$$

$\boldsymbol{E}(t) = E_0(t)\,\boldsymbol{\epsilon}\,\cos(\boldsymbol{k}\cdot\boldsymbol{r} - \omega_L t + \phi)$ is the electric field strength of the laser with frequency ω_L and Q a component of the electric quadrupole tensor of the transition, determined by the excitation geometry and polarization $\boldsymbol{\epsilon}$ of the field. The phase ϕ of the laser determines the axis in the xy-plane, around which the Bloch-vector is rotated. For example, $\phi = 0$ leads to a rotation around the x-axis, while $\phi = \pi/2$ rotates around the y-axis. The most important parameter is the rotation angle θ. It is given by the pulse area, obtained by integrating the Rabi-frequency over the pulse duration: $\theta = \int \Omega_0(t)\,\mathrm{d}t$.

After excitation by a pulse characterized by the parameters θ and ϕ, the initial state of a qubit $c_g\,|g\rangle + c_e\,|e\rangle$ is therefore rotated according to the following transformation:

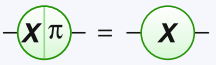
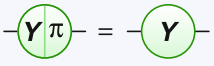
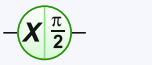
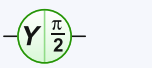
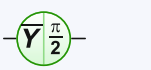
$$\begin{pmatrix} c_g \\ c_e \end{pmatrix} \rightarrow \begin{pmatrix} \cos\theta/2 & -ie^{-i\phi}\,\sin\theta/2 \\ -ie^{i\phi}\,\sin\theta/2 & \cos\theta/2 \end{pmatrix} \cdot \begin{pmatrix} c_g \\ c_e \end{pmatrix}. \tag{9}$$

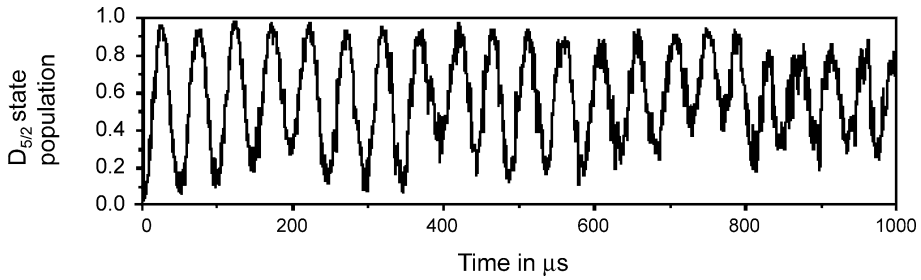
Table 2 gives an overview of important single-qubit operations and the way they are implemented in an ion-trap quantum computer by choosing the pulse area θ and phase ϕ of laser pulses resonant with the qubit transition. Other single-qubit transformations can be implemented by combining rotations around different axes. An example is the Hadamard transform. The transformation matrix and its realization by X- and Y-pulses in symbolic notation is

$$\begin{pmatrix} c_g \\ c_e \end{pmatrix} \rightarrow \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \cdot \begin{pmatrix} c_g \\ c_e \end{pmatrix} \tag{10}$$

Quantum Computing with Trapped Ions, Table 2

Important single qubit operations according to Eq. (9) and their realization with laser pulses. The corresponding transformations on the Bloch-sphere are often referred to as $R(\theta, \phi)$ in the literature. The symbols are used in the quantum circuits discussed below

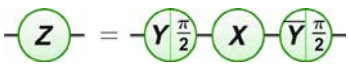
θ	ϕ	Description	Symbol
4π	any	4π-pulse – **identity**, no change of qubit: $\|g\rangle \rightarrow \|g\rangle$ $\quad$ $\|e\rangle \rightarrow \|e\rangle$	—
2π	any	2π-pulse – **sign change** of qubit-state. $\|g\rangle \rightarrow -\|g\rangle$ $\quad$ $\|e\rangle \rightarrow -\|e\rangle$ This has no effect if applied on the qubit -transition since both levels experience the same phase shift. However, it is one of the most important operations in two-ion gates if applied on a transition with selective coupling to the basis states.	—
π	0	π-pulse – **NOT-gate** = qubit flip (omitting common phase factor $-i$): $\|g\rangle \rightarrow \|e\rangle$ $\quad$ $\|e\rangle \rightarrow \|g\rangle$	
π	$\frac{\pi}{2}$	π-pulse, bit-flip and relative sign change: $\|g\rangle \rightarrow \|e\rangle$ $\quad$ $\|e\rangle \rightarrow -\|g\rangle$	
$\frac{\pi}{2}$	0	$\pi/2$-pulse (basis states are rotated around x-axis to equally weighted superposition): $\|g\rangle \rightarrow \frac{\|g\rangle - i\,\|e\rangle}{\sqrt{2}}$ $\quad$ $\|e\rangle \rightarrow \frac{\|e\rangle - i\,\|g\rangle}{\sqrt{2}}$	
$\frac{\pi}{2}$	$\frac{\pi}{2}$	$\pi/2$-pulse (basis states are rotated around y-axis to equally weighted superposition): $\|g\rangle \rightarrow \frac{\|e\rangle + \|g\rangle}{\sqrt{2}}$ $\quad$ $\|e\rangle \rightarrow \frac{\|e\rangle - \|g\rangle}{\sqrt{2}}$	
$\frac{\pi}{2}$	$-\frac{\pi}{2}$	$\pi/2$-pulse around negative y-axis – identical to combination of **Hadamard -gate** and **Z-gate** (sign change of $\|e\rangle$): $\|g\rangle \rightarrow \frac{\|g\rangle - \|e\rangle}{\sqrt{2}}$ $\quad$ $\|e\rangle \rightarrow \frac{\|g\rangle + \|e\rangle}{\sqrt{2}}$	



Quantum Computing with Trapped Ions, Figure 11
Rabi-oscillations of a single $^{40}Ca^{+}$-ion on the blue sideband starting in the motional ground state. The Rabi-frequency is given by Eq. (17) instead of Eq. (8), but the flopping between two states is analogous. The decay of contrast is a measure of decoherence in the system (Sect. "Decoherence"). In the example shown, coherence is maintained for up to 1 ms [62]

Pulses around the z-axis, too, are most conveniently excited by a combination of pulses around the x- and y-axes. An example is the realization of a π-rotation around the z-axis by three pulses:

$$\begin{pmatrix} c_g \\ c_e \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \cdot \begin{pmatrix} c_g \\ c_e \end{pmatrix} \tag{11}$$



For the second class of qubits, *hyperfine qubits*, transition frequencies between the basis states are several GHz and hence require microwave pulses if driven directly [68]. Due to the large wavelength of microwaves, different ions in a quantum register can only be addressed individually if their transition frequencies are split, for example by the Zeeman-effect when using a strong magnetic field gradient.

An alternative, which has been used in most implementations of hyperfine qubits to date, is to drive the qubit using a two-photon stimulated Raman transition involving a third electronic level, connected to the qubit states through an electric dipole transition. In Fig. 6, the intermediate state is $^2P_{1/2}$. By detuning the driving fields from resonance by an amount Δ which is large compared to the inverse lifetime of the auxiliary state, excitation of this state is largely avoided [61]. The advantage is that individual ions may be addressed with the Raman beams.

The effective Rabi-frequency for a Raman-transition, to be used instead of expression (8), is

$$\Omega_0 = \frac{\Omega_1 \Omega_2}{\Delta}, \qquad \Omega_i = E_{0i} \frac{\boldsymbol{\mu}_i \cdot \boldsymbol{\epsilon}_i}{2\hbar} \tag{12}$$

where Ω_i is the Rabi-frequency for transition i and μ_i the corresponding electric dipole moment. The qubit is resonantly excited if the difference frequency $\omega_{L1} - \omega_{L2}$ of the two lasers (which replaces ω_L in the direct driving scheme) is tuned to the qubit frequency ω_0. The role of the $\boldsymbol{k}$-vector in the case of direct driving is played by the difference $\boldsymbol{k}_1 - \boldsymbol{k}_2$ and the relevant phase determining the single-ion dynamics is given by the difference of the phase constants of the two laser beams, $\phi = \phi_1 - \phi_2$. With these modifications, Eq. (9) applies.

The ion dynamics is more complicated if their vibration in the trapping potential is taken into account. This is the basis of two-bit quantum gates in ion traps discussed in Sect. "Two-Qubit Interaction and Quantum Gates". Generally, excitation of a transition on a vibrational sideband (see Fig. 8) leads to Rabi-oscillations with a modified Rabi-frequency.

By measuring the state of the qubit as a function of the length of the Rabi pulses (and hence the pulse area θ), Rabi-oscillations can be observed experimentally. An example is shown in Fig. 11. The decay of the contrast of the oscillations is an indication of the loss of coherence between the qubit levels. This can be used to determine the decoherence-time of quantum memory (see Sect. "Decoherence").

State Detection of Ionic Qubits

At the end of any data processing, a quantum computer must be subjected to a measurement of its register, in order to determine the result of the calculation. It is important for the success of the quantum computation to achieve a high efficiency readout. One of the greatest advantages of ion-trap quantum processors is that the state of each qubit may be determined with almost 100% efficiency. This is in contrast to other carriers of quantum information such as photons or nuclear spins in a bulk sample.

Quantum Jump Detection

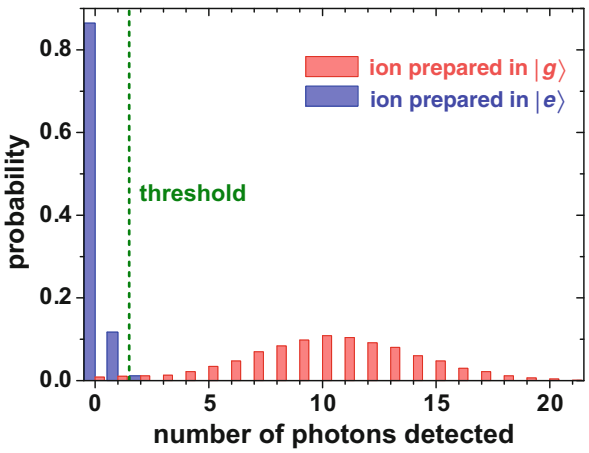
The quantum state of an ionic qubit is measured optically by monitoring the fluorescent light emitted upon selective

laser excitation of one basis state of the qubit on a strong electric dipole transition. The population in the other basis state is either not affected by the probe laser due to selection rules or has previously been transferred (shelved) to another state inaccessible to the probe. The method is therefore also known as *electron shelving* [69].

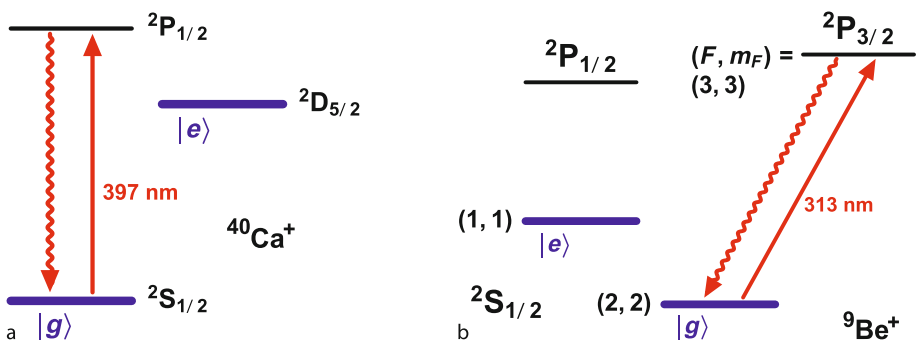
The crucial point of the procedure is that the fluorescence is emitted on a transition back to the original qubit-level (cycling transition), so that the cycle of excitation and emission of fluorescent light can be repeated indefinitely. In practice, transitions to levels outside the cycle occur, though at a small rate. The number of photons extracted from a single qubit can reach $N_{\mathrm{ph}} \approx 10^6$ [70]. Even for a small photon detection efficiency η_{d}, the number of detected photons is large and can be distinguished from the other, non-fluorescing level with almost 100% certainty [71]. The probability for a false interpretation of a zero-photon signal is only $\exp(-\eta_{\mathrm{d}}\, N_{\mathrm{ph}})$. Originally, the method has been used to detect quantum jumps between two atomic levels by observing the intermittent fluorescence on a transition coupled to only one of the levels [69,72], hence the name *quantum jump detection*. It is now employed as the standard readout procedure in quantum information processing with ions.

Figure 12 shows the cycling transitions used in two different qubit systems. Additional repumping lasers may be required to avoid population trapping in uncoupled levels. An example of typical photon counting statistics in the two qubit states is shown in Fig. 13. By exciting the entire ion string simultaneously and imaging the fluorescent light with a CCD camera (cf. Fig. 3), the state of the entire quantum register can be read out simultaneously.

According to the laws of quantum mechanics, a single measurement cannot retrieve the complete information on the quantum state, ruling out single-shot detection of arbitrary superpositions and entanglement between different qubits. Rather, a projection to the basis states is obtained with a probability given by the magnitude squared of the corresponding coefficient. Correlations between qubits are lost in this way. Alternatively, single qubits might be subjected to a state rotation (see Sect. "Single-Ion Operations") prior to measurement, mapping coherent superpositions of a qubit's states onto the basis states, which then are probed as described above. A useful transformation is achieved with $\pi/2$-pulses.



Quantum Computing with Trapped Ions, Figure 13
Histogram of detected photons after an ion is prepared in either the $|g\rangle$ or the $|e\rangle$ state. In correspondence with Fig. 12 $|g\rangle$ designates the fluorescing, while $|e\rangle$ is not coupled to the probe laser and scatters practically no photons. Using the indicated threshold to determine the state of the ion, a high detection fidelity is achieved (97.9% in the case of Yb shown above [45])



Quantum Computing with Trapped Ions, Figure 12
Transitions used in the electron shelving detection of the state of an ion-qubit. Only when the ion is in state $|g\rangle$, excitation occurs (*red arrow*) and strong fluorescence back to the original state (*wavy line*) is observed. a Relevant levels in $^{40}Ca^+$ b Relevant levels in $^9Be^+$

If correlations between different qubits are of interest, e. g., entanglement of two qubits, the two-ion gates described in Sect. "Two-Qubit Interaction and Quantum Gates" may be used to disentangle the ions first and measure the qubits separately. For a more complete characterization of a quantum operation, methods of quantum state tomography must be used [73]. By preparing a certain quantum result repeatedly and averaging over measurements of 3^N different observables (for an N-ion register), the density matrix of the system may be reconstructed, providing full information of the quantum state [74].

Two-Qubit Interaction and Quantum Gates

The most important requirement for any quantum computer is the implementation of gates, logically connecting the quantum states of two or more qubits in a coherent way. Trapped ions in a linear trap are several μm apart and therefore do not interact directly, for example through their dipole moments. Instead, the interaction is established indirectly, based on two principles: (1) Due to the long-range Coulomb repulsion between ions, their motion is strongly coupled, leading to collective vibrations of the ion-string. (2) By driving the qubit transition with a laser tuned to a motional sideband, the internal state of any ion and the collective vibration of the entire string may be coupled.

Using the collective vibration to logically connect different ions requires introducing a new qubit to the system, stored in the motional quantum state of the ion string. Since all ions participate in the vibration, it constitutes a *quantum data bus* for the entire register.

Normal Modes

The description of the collective vibration of a linear string of ions in the trapping potential is based on the concept of normal modes. In a normal mode, all ions oscillate in phase at the same frequency. A single normal mode can provide the bus for the transfer of quantum information between different ions. A linear string of N ions has N normal modes of collective axial vibration. The two with the lowest vibrational frequencies are of particular significance.

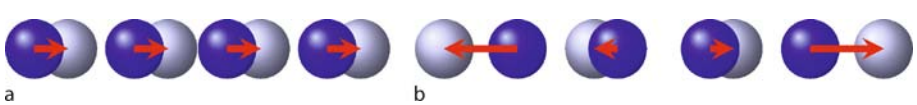
The fundamental mode has a frequency ν equal to the axial frequency of the trap ($\nu = \omega_z$) and is known as the center-of-mass (COM) mode. Here, all ions move synchronously, i. e., in the same direction and by equal amounts, so that the relative positions of the ions stay the same. It has the advantage that all ions couple to the motion with the same strength. The disadvantage is that it also couples strongly to fluctuating ambient electric fields, leading to large heating rates (see Sect. "Decoherence"). Therefore, the next higher normal mode is often preferable. It is known as the stretch- or breathing mode, since the ions on opposite sides of the center move in opposite directions with an amplitude proportional to their distance from the center. Its frequency is $\nu = \sqrt{3}\,\omega_z$. The motion of a string of four ions in the COM and stretch-mode is indicated in Fig. 14.

The higher order modes correspond to a more complicated motion of the ions. Their exact frequencies depend on the number of ions in the string and must be calculated numerically [12]. The vibrational excitation spectrum of an ion string becomes increasingly complex at higher ion number N, since not only the number of normal modes increases, but also oscillation at the sum- and difference-frequencies contributes. This makes quantum operations based on the selective coupling of ions to one particular mode difficult in long strings.

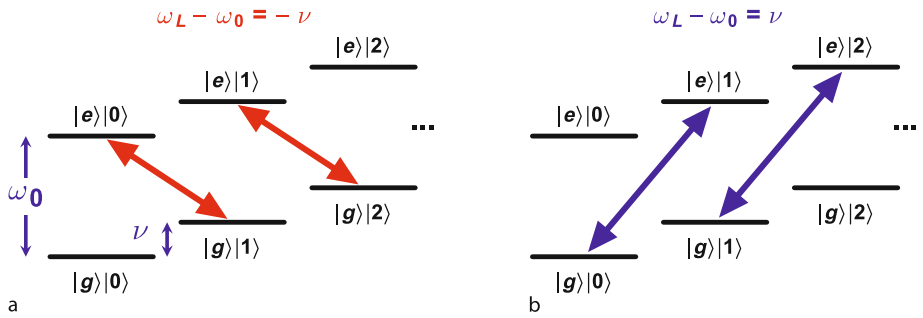
In the limit of low vibrational excitation, each mode must be treated as a quantum mechanical harmonic oscillator, the state of which can only be changed in units of single vibrational quanta of energy $\hbar\nu$ (phonons), where ν is the angular frequency of this particular mode. For quantum computation, one collective normal mode is selected for the exchange of information, while the remaining $N-1$ modes are treated as spectator modes, which should remain unexcited during the calculation.

Vibrational Coupling

In order to use the motion of ions for information exchange, there must be a coupling between the internal state of an ion and the vibration. This is achieved by detuning the exciting laser from the ion's resonance at ω_0 to either the red or the blue vibrational sideband of the desired mode, i. e., to a frequency $\omega_0 \pm \nu$.



Quantum Computing with Trapped Ions, Figure 14
The two lowest normal modes of a string of four ions. The positions of the ions are shown at two different times. Arrows indicate the motion of the ions. a COM-mode at $\nu = \omega_z$; b stretch-mode at $\nu = \sqrt{3}\,\omega_z$



Quantum Computing with Trapped Ions, Figure 15
Vibrational structure of the qubit states and transitions for excitation on a the first red sideband and b the first blue sideband. The three lowest vibrational levels for each qubit state are shown, displaced horizontally according to the number of phonons for clarity

Assuming that the vibrational mode is in a state with phonon number n, red sideband excitation drives the transitions

$$|g\rangle\,|n\rangle \longleftrightarrow |e\rangle\,|n-1\rangle \qquad (n = 1, 2, \dots)\,, \tag{13}$$

while blue sideband excitation drives the transitions

$$|g\rangle\,|n\rangle \longleftrightarrow |e\rangle\,|n+1\rangle \qquad (n = 0, 1, \dots)\,. \tag{14}$$

Here, the first symbol represents the ion's state and the second the state of the vibrational mode, with $|n\rangle$ corresponding to a number state with n phonons in the mode. The transitions corresponding to (13) and (14) are depicted in Fig. 15. In each case, a change of the state of the ionic qubit coincides with a change of the phonon number, i. e., the state of the data bus.

By analogy with the resonant case discussed in Sect. "Single-Ion Operations", the evolution of the system for sideband excitation is characterized by an oscillatory change of populations at a modified Rabi-frequency. The probability of sideband transitions scales with a factor of η, the so-called Lamb–Dicke parameter. It is given by

$$\eta = \frac{2\pi\, a_0}{\lambda}\,, \tag{15}$$

where a_0 is the size of an ion's wavepacket and λ the wavelength of the transition. For the stretch-mode of a two-ion crystal, for example, the Lamb–Dicke parameter in the ground-state is given by

$$\eta = \pm\frac{1}{\sqrt[4]{12}}\,\frac{\omega_L}{c}\,\cos\theta_L\,\sqrt{\frac{\hbar}{2m\omega_z}}\,, \tag{16}$$

where θ_L is the angle between the laser beam and the trap axis (cf. Fig. 10). The Rabi-frequency for the first sideband transitions (13) and (14) depends on η and the number of excited vibrational quanta:

$$\begin{array}{lrcl} \text{Red sideband:} & \Omega^-(n) & = & \Omega_0\,\eta\,\sqrt{n} \\ \text{Blue sideband:} & \Omega^+(n) & = & \Omega_0\,\eta\,\sqrt{n+1}\,, \end{array} \tag{17}$$

where Ω_0 is the Rabi-frequency for resonant excitation from Eq. (8) or Eq. (12).

SWAP Gate

A two-qubit gate that illustrates the use of sideband excitation in a system with one ionic and one vibrational qubit is the SWAP-gate. In order to restrict the infinite ladder of vibrational states to a binary system, the qubit is represented by the two lowest vibrational states $|0\rangle$ and $|1\rangle$. The computational space is then spanned by the four basis states

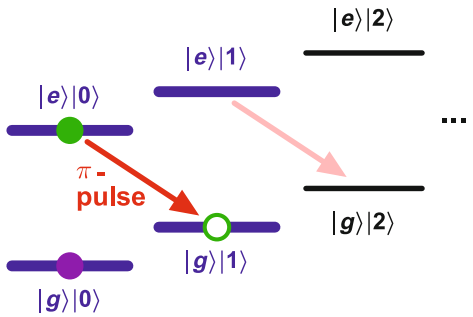
$$|g\rangle\,|0\rangle\,, |e\rangle\,|0\rangle\,, |g\rangle\,|1\rangle \text{ and } |e\rangle\,|1\rangle\,, \tag{18}$$

i. e. the first two pairs of levels in Fig. 15. Assuming that the vibrational qubit is initially in the state $|0\rangle$ (after sideband cooling), an arbitrary superposition state $\alpha\,|g\rangle + \beta\,|e\rangle$ of the ion can be swapped to the vibrational mode by applying a π-pulse on the first red sideband (Fig. 16). With the state $|g\rangle\,|0\rangle$ not coupled to the red sideband, the mapping obtained is

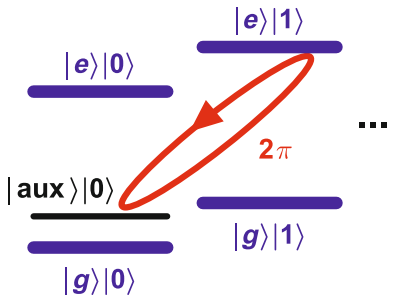
$$(\alpha|g\rangle + \beta|e\rangle) \otimes |0\rangle \longrightarrow |g\rangle \otimes (\alpha|0\rangle + \beta|1\rangle)\,. \tag{19}$$

This operation is the basis for making the quantum state of an ion in a string accessible to any other ion sharing the collective motion. Two-ion gates may be realized by combining the above SWAP operation with a gate for the vibrational mode and a second ion.

The transformation described by Eq. (19) requires that the vibrational mode isn't excited initially. In particular, if



Quantum Computing with Trapped Ions, Figure 16
Mapping of a qubit from an ion to the vibrational mode using a π-pulse on the red sideband



Quantum Computing with Trapped Ions, Figure 17
Controlled Z-gate in the Cirac–Zoller scheme. The laser pulse couples the auxiliary level to the state $|e\rangle\,|1\rangle$, leaving all other levels unaffected. Using a 2π-pulse results in a sign change of the $|e\rangle\,|1\rangle$-component

state $|e\rangle\,|1\rangle$ is exposed to a red sideband pulse, a transition to $|g\rangle\,|2\rangle$ would occur, which is outside the computational subspace. More elaborate pulse sequences are required to keep the ion dynamics within the space spanned by the four states of Eq. (18). A SWAP-gate that preserves the computational basis was realized by pulses of blue sideband excitation [75].

Cirac–Zoller Gate

The coupling of two ions in a string through their collective external motion was first proposed in 1995 in a seminal paper by Cirac and Zoller [4]. At the heart of their proposal is a two-qubit *controlled Z-gate* with one qubit stored in the vibrational state of the COM-mode and the other stored in the internal state of an ion. The controlled Z-gate is the simplest universal quantum gate, from which arbitrary logic gates may be built in conjunction with single qubit operations. The vibrational state controls the transformation of the ion state to which the Z-gate of Eq. (11) is applied if and only if the vibration qubit is in state $|1\rangle$. Alternatively, this can be expressed as the wave function $|\psi\rangle$ of the system acquiring a change of sign if both input qubits are in the excited state ($|1\rangle$ and $|e\rangle$) and is left unchanged in all other cases. It corresponds to the following truth table.

input	output
$\|0\rangle\|g\rangle$	$\|0\rangle\|g\rangle$
$\|0\rangle\|e\rangle$	$\|0\rangle\|e\rangle$
$\|1\rangle\|g\rangle$	$\|1\rangle\|g\rangle$
$\|1\rangle\|e\rangle$	$-\|1\rangle\|e\rangle$

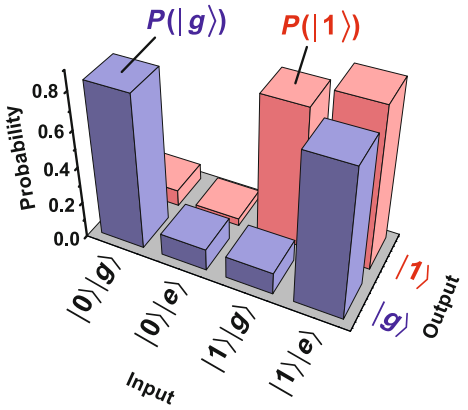
As shown in Tab. 2, a sign change, is achieved by applying a 2π-pulse to the internal states of the ion. In the Cirac–Zoller scheme, the required conditioning on the state $|e\rangle$ of the ion is obtained by using a transition which couples exclusively to the upper internal state of the ion. This requires an auxiliary electronic level, e. g., another Zeeman sublevel of the ground state. Conditioning on the vibrational state $|1\rangle$ is achieved by tuning to the first blue vibrational sideband, which induces a transition to a lower state only if at least one vibrational quantum is present. The scheme of this controlled Z-gate is illustrated in Fig. 17.

Another important gate may be implemented by combining the controlled Z-gate with single qubit rotations. By applying a resonant $\pi/2$-pulse to the ion (changing the electronic states but leaving the vibration unaffected) before the controlled Z-gate and an inverse $\pi/2$-pulse after, a controlled NOT (CNOT) gate is realized, in which the target bit (carried by the electronic state) is flipped depending on the state of the control bit (represented by the vibrational state). Its truth table is

input	output
$\|0\rangle\|g\rangle$	$\|0\rangle\|g\rangle$
$\|0\rangle\|e\rangle$	$\|0\rangle\|e\rangle$
$\|1\rangle\|g\rangle$	$\|1\rangle\|e\rangle$
$\|1\rangle\|e\rangle$	$\|1\rangle\|g\rangle$

The $\pi/2$-pulses effectively change the computational basis in which the controlled Z-gate is performed.

The described CNOT-gate for the internal and the vibrational state of a single ion was demonstrated in 1995 at NIST in Boulder [22] with $^9Be^+$, the first implementation of an ion-trap quantum gate. The auxiliary state $|\text{aux}\rangle = |S_{1/2}, F = 2, m_F = 0\rangle$ was chosen. The experimental results for different combinations of input states are shown in Fig. 18, confirming the realization of the above truth table.



Quantum Computing with Trapped Ions, Figure 18
Experimental verification of the truth table of a CNOT quantum gate, starting from the four combinations of basis states indicated on the input axis [22]. The bars represent the measured probability for being in the lower atomic state $|g\rangle$ (blue) and in the first excited vibrational state $|1\rangle$ (red). The ionic qubit is flipped if the vibration is initially in state $|1\rangle$, while the vibrational state itself is unchanged

In the original scheme of Cirac and Zoller, the qubits are stored in the electronic states of two ions in a string, while the center-of-mass motion of the string is used as a data bus only during the gate operation. Therefore, before applying the controlled Z-gate to the target ion, the state of the first ion (control qubit) must be transferred to the vibrational motion by means of the SWAP operation described in Sect. "SWAP Gate". After completion of the gate, the state of the control ion must be restored by mapping the state of the vibrational mode back to the first ion.

The implementation of the full Cirac–Zoller gate with qubits stored in different ions has been accomplished recently in Innsbruck [76]. In this experiment, a simplified procedure for realizing the controlled Z-gate was used, avoiding the need for an auxiliary level. In principle, a controlled Z-operation can be accomplished by applying a 2π-pulse on the blue sideband of the qubit transition. However, Eq. (17) implies that state $|g\rangle|1\rangle$ has a blue sideband Rabi-frequency $\sqrt{2}$ times higher than $|g\rangle|0\rangle$, so that this state experiences a $2\sqrt{2}\pi$ pulse. A 2π-pulse for all levels (except $|e\rangle|0\rangle$, which does not couple to blue sideband excitation) can be achieved by using a transition composed of four pulses, rotating the state of the target ion with different angles and around different axes [77].

The dynamics of this gate is illustrated in Fig. 19, showing the levels coupled by effective 2π-pulses. Note that there is no excitation of levels outside the computational subspace, since 2π rotations do not change the occupation of the states involved. The experimentally determined truth table is shown in Fig. 20. The corresponding fidelity of the gate operation is between 70% and 80%.

Geometric Gates

The direct coupling of qubits and vibration in the Cirac–Zoller gate requires the bus-mode to be in the quantum mechanical ground state, a condition which is not easy to satisfy for long times in the presence of motional heating. In addition, individual addressing of control and target ion is needed. In order to relax these requirements, a different type of gate was proposed, based on electronic-state-dependent motional displacements [78,79,80,81,82,83]. Being simpler and less error-prone, the technology offers high fidelity gate operation.

In a geometric gate, the state of the harmonic oscillator corresponding to the vibrational bus-mode of the ions is displaced in phase-space (i. e. its position coordinate $\langle z\rangle$ and momentum $\langle p\rangle$ are changed), conditioned on the internal state of the qubits involved. The displacement is applied in such a way that it follows a closed loop in phase space, returning the vibration to its original state. For a detuning $\delta = \omega_L - \omega_0$ between the effective frequency difference ω_L of the Raman excitation and the qubit transition frequency, this occurs after a time τ given by $(\nu - \delta)\tau = 2\pi m$ for integer m. At this time, any entanglement between the bus and the qubits vanishes. Therefore, the only effect of the operation is that the system acquires a phase ϕ (geometric phase), equal to the area in phase space enclosed by the trajectory of the displacement, as shown in Fig. 21 [82]:

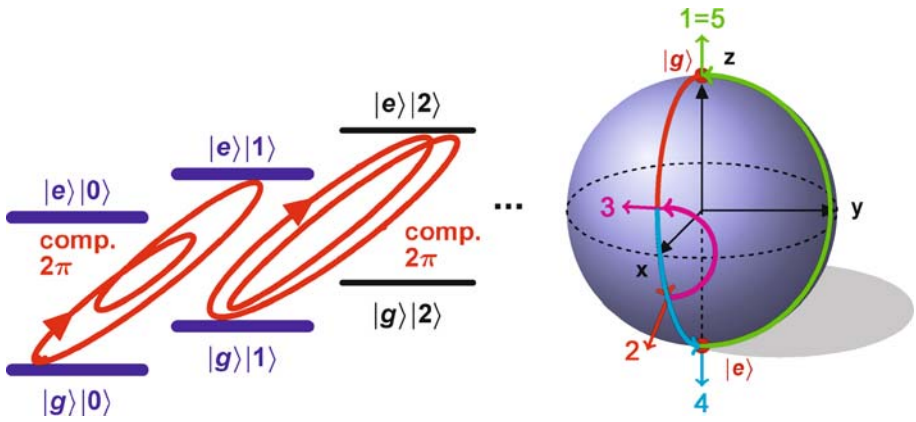
$$\phi = \frac{1}{\hbar}\int_{\text{area}} \mathrm{d}z\ \mathrm{d}p = 2\pi m \left(\frac{\eta\Omega}{\nu - \delta}\right)^2 . \tag{20}$$

The phase ϕ has properties which make it particularly suitable for a quantum gate:

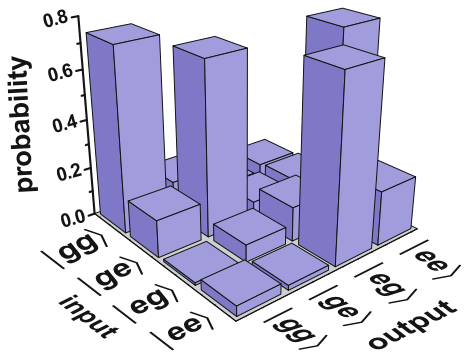
(1) It depends on the internal state of the two qubits involved (e. g., ϕ is non-zero if and only if the two qubits are different);
(2) It is independent of the initial state of the bus-mode, as long as excursions stay within the Lamb–Dicke regime (i. e. are smaller than the wavelengths of the transitions involved). Therefore, the phase of the system is much less susceptible to motional heating.

Insensitivity to the bus-mode excitation could also be reached by making the detuning $\nu - \delta$ from single-photon resonance large [79]. However, this leads to an unfavorably slow evolution of the interaction between the qubits.

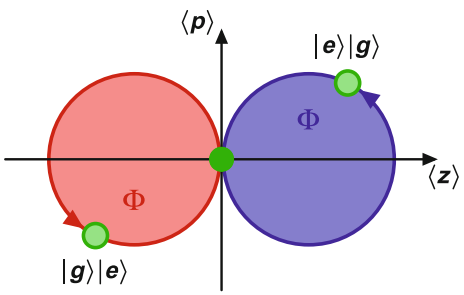
The selective displacement of the bus-mode for certain qubit states is achieved through optical forces exerted by



Quantum Computing with Trapped Ions, Figure 19
Controlled Z-gate with composite pulses. Using a sequence of four blue sideband pulses, all ions in $|g\rangle\,|0\rangle$, $|g\rangle\,|1\rangle$ and $|e\rangle\,|1\rangle$ experience a 2π-rotation, while $|e\rangle\,|0\rangle$ is unchanged. This corresponds to a controlled Z-gate, up to an overall minus sign. The drawing on the right shows the motion of the Bloch vector of a target qubit starting and ending in the state $|g\rangle\,|0\rangle$, picking up a phase shift of π



Quantum Computing with Trapped Ions, Figure 20
Experimentally observed truth table of the Cirac–Zoller CNOT derived from measurement of the joint probabilities of the two qubits [76]



Quantum Computing with Trapped Ions, Figure 21
Phase-space representation of the bus-mode excitation during a geometric gate. The blue and the red circle are the trajectories for internal states $|e\rangle|g\rangle$ and $|g\rangle|e\rangle$, respectively. In each case, a phase ϕ is acquired

suitable Raman pulses. Two different methods have been applied, resulting in different qubit dynamics and therefore different gates.

Mølmer–Sørensen or *xy*-Gate In this case, bichromatic Raman beams are used. They provide two equally strong effective fields, with opposite detunings $\pm\delta$ from the red and the blue vibrational sideband of the qubit-transition. This drives a two-photon transition, corresponding to the two-qubit interaction Hamiltonian

$$H_I = -\frac{\hbar}{2}\frac{(\eta\Omega)^2}{\nu-\delta}(\hat{\sigma}_x \otimes \hat{\sigma}_x)\,, \tag{21}$$

where $\hat{\sigma}_x = (|g\rangle\langle e| + |e\rangle\langle g|)/\sqrt{2}$ is a Pauli spin-matrix. In the Bloch-picture, the Hamiltonian (21) provides a synchronous rotation of the internal states of both qubits around the x-axis (cf. Fig. 9). Alternatively, a rotation around the y-axis could have been chosen, hence the name xy-gate. A gate based on this type of interaction was first implemented in an experiment generating entangled states $(|gg\rangle + i|ee\rangle)/\sqrt{2}$ of two qubits and $(|gggg\rangle - i|eeee\rangle)/\sqrt{2}$ of four qubits, starting from a string of ions initialized in the internal ground state [84,85]. It has recently been used for the entanglement of magnetic-field insensitive qubit-states [86,87].

***z*-Gate** This gate is similar to the xy-gate, except that different basis states are affected by the interaction. Here, the frequency difference ω_L of two Raman beams is chosen to lie close to the vibrational frequency ν. In this case, displacement is provided by the optical dipole force from the

Raman beams, which is different for each qubit-state. If the spacing between the ions, as well as polarization and detuning of the Raman-beams are chosen correctly, there will be a vibrational displacement only if the ions are in different states. Again, for a closed loop of the displacement in phase space, the vibrational state decouples from that of the qubits, and the state of the system acquires a phase shift ϕ. The Hamiltonian corresponding to this interaction is similar to Eq. (21), except that the states of the qubits are rotated around the z-axis of the Bloch-sphere. The transformations of the basis state for the two types of geometric gates discussed are summarized in the following table:

input	z-gate	MS-gate
$\|g\rangle\|g\rangle$	$\|g\rangle\|g\rangle$	$(\|g\rangle\|g\rangle + ie^{-i\phi}\|e\rangle\|e\rangle)/\sqrt{2}$
$\|g\rangle\|e\rangle$	$e^{i\phi}\|g\rangle\|e\rangle$	$(\|g\rangle\|e\rangle + ie^{-i\phi}\|e\rangle\|g\rangle)/\sqrt{2}$
$\|e\rangle\|g\rangle$	$e^{i\phi}\|e\rangle\|g\rangle$	$(\|e\rangle\|g\rangle + ie^{i\phi}\|g\rangle\|e\rangle)/\sqrt{2}$
$\|e\rangle\|e\rangle$	$\|e\rangle\|e\rangle$	$(\|e\rangle\|e\rangle + ie^{i\phi}\|g\rangle\|g\rangle)/\sqrt{2}$

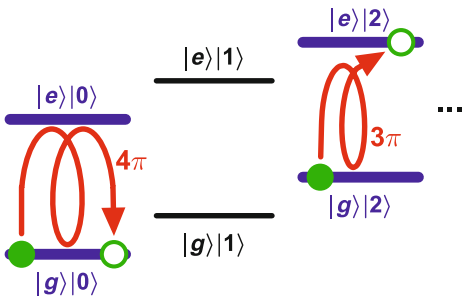
A z-gate has been realized experimentally with two $^9Be^+$-ions [88], interacting via the stretch-mode. The entangled state $(|00\rangle\text{-}i|11\rangle)/\sqrt{2}$ was produced with a fidelity of 97%. The high fidelity confirms the benefit of reduced requirements regarding thermally excited motion and addressing of the ions in a geometric gate. The z-gate was used to generate and investigate entanglement between a single $^{40}Ca^+$-ion and its vibrational motion [89], as well as two $^{40}Ca^+$-ions [90].

CNOT-Wavepacket Gate

A different route to a simplified CNOT-gate was proposed by Monroe and coworkers [91]. In contrast to gates of the Cirac–Zoller type, which use sideband transitions for coupling internal and motional states, the gate of Monroe et al. uses pulses tuned to resonance with the qubit transition. The necessary conditioning on the state of the vibrational bus-mode is obtained by taking into account the higher-order dependence of the Rabi-frequency on the Lamb–Dicke parameter η from Eq. (15) . In fact, if the extension of the wavepacket corresponding to the ion is not small compared with the wavelength of the qubit transition, i. e., $\eta \gtrsim 1$, the system is no longer in the Lamb-Dicke regime and the resonant Rabi-frequency becomes

$$\Omega(n) = \Omega_0 \, e^{-\eta^2/2} \, L_n(\eta^2) \,, \tag{22}$$

where Ω_0 is the Rabi-frequency in the Lamb–Dicke limit given by either Eq. (8) or Eq. (12) and $L_n(x)$ is the Laguerre polynomial of order n.



Quantum Computing with Trapped Ions, Figure 22
CNOT-gate with control qubit stored in the motional ground state (|0⟩) and second excited state (|2⟩) of the bus-mode. A laser pulse drives a 4π or a 3π transition, returning the ion to its initial electronic state or toggling its state when the bus is in state 0 or 2, respectively [92]

By adjusting η, the transition can be driven by a single pulse of a suitable length τ in such a way that, for example, the phase change of the system is $\theta = 4\pi$ if the vibration is in state $|0\rangle$ and $\theta = 3\pi$ if it is in state $|2\rangle$ (note that this requires using the basis states $|0\rangle$ and $|2\rangle$ for the bus, replacing the usual $|0\rangle$ and $|1\rangle$). In the first case, there is no change of the system, while in the latter there is a flip of the qubit state as required for a CNOT-gate. This gate was realized experimentally at NIST [92] using $\eta = 0.359$.

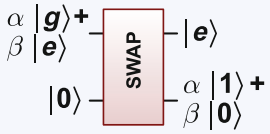
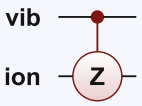
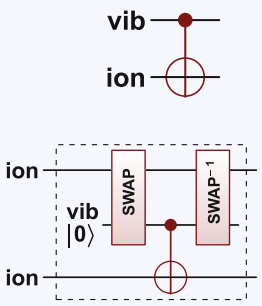
Physically the conditional dynamics of the gate relies on the different size of the ions's wavepacket in different vibrational states. The advantages of this type of gate are that only a single pulse is needed, no auxiliary levels are involved and there are no differential Stark-shifts of the levels during the gate operation.

Fast Gates

The two-ion gates discussed so far use one normal mode of the string to couple the qubits. In this case, the gate speed is limited by the mode frequency ν, or, more generally, the trap frequency ω_z, since on a faster time-scale, the vibrational sidebands cannot be resolved for selective excitation. Recently, geometric gates have been proposed which offer processing speeds beyond the vibrational frequencies [93,94,95,96]. This is possible in long ion strings, if all normal modes of vibration are excited simultaneously. By using suitably shaped laser pulses [93,95] or a fast sequence of precisely timed pulses [93,94,95,96], the time required for a two-ion phase-gate may be reduced below $2\pi/\omega_z$. The reason is that with fast pulses local oscillations of the addressed ions may be excited, in which none of the other ions undergo significant motion. This requires the coherent excitation of a large number of normal modes with a minimum of five laser pulses [96].

Quantum Computing with Trapped Ions, Table 3

Important two-qubit gates employed in ion-trap quantum computation. The table lists the matrices representing the gate operators for the joint basis states, along with the symbols used in quantum circuits. See Tab. 2 for single-ion operations

Gate	Description	Quantum Circuit	Ref.
SWAP	Ion state swapped to bus-mode $\begin{array}{c} \\ g0 \\ g1 \\ e0 \\ e1 \end{array} \begin{array}{c} \begin{array}{cccc} g0 & g1 & e0 & e1 \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \end{array}$ Method: blue sideband composite π-pulse		Sect. "SWAP Gate", [75]
CZ	**Controlled Z-gate:** one combination of basis states acquires minus sign (=π phase shift, hence the name *phase gate*) $\begin{array}{c} \\ 0g \\ 0e \\ 1g \\ 1e \end{array} \begin{array}{c} \begin{array}{cccc} 0g & 0e & 1g & 1e \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} \end{array}$ Methods: • 2π-pulse on auxiliary transition • blue sideband composite 2π-pulses		Sect. "Cirac–Zoller Gate", [22], [76]
CNOT	**Controlled NOT gate:** NOT applied to target ion if control qubit (vibration) is 1 $\begin{array}{c} \\ 0g \\ 0e \\ 1g \\ 1e \end{array} \begin{array}{c} \begin{array}{cccc} 0g & 0e & 1g & 1e \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \end{array}$ Methods: • CZ between $\pi/2$-pulses • wavepacket-gate • two-ion CNOT if combined with SWAP gate		Sect. "Cirac–Zoller Gate" [22], [76], Sect. "CNOT-Wavepacket Gate", [92]

The most important fundamental logic gates realized in ion traps are summarized in Tab. 3, along with their matrix representation and the corresponding notation in quantum circuit diagrams.

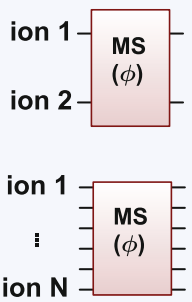
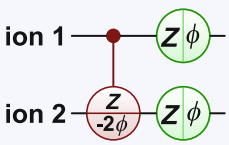
Decoherence

The most important property distinguishing a qubit from a classical bit is that its two states have a well-defined relative phase, i. e., they are coherent. Uncontrollable interactions of a qubit with its environment destroy this coherence, a process known as *decoherence*. This includes the radiative decay of qubit states. Quantum memory as well as quantum gates are affected by decoherence and decoherence rates are important figures of merit for any quantum processing device. Trapped ion systems have been demonstrated to reach very low decoherence rates, making them the most successful implementation of quantum information processing to date. One way to measure decoherence is through the loss of contrast of the coherent Rabi-oscillation either on the carrier or on a sideband, as shown in Fig. 11.

Internal State Decoherence

The coherence of internal states is limited by the radiative decay of the qubit-levels. The excited states are either metastable with lifetimes on the order of seconds or hyperfine ground-states with even smaller decay rates.

Quantum Computing with Trapped Ions, Table 3
continued

Gate	Description	Quantum Circuit	Ref.
MS	**Mølmer–Sørensen gate:** multi-ion entanglement $\frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 0 & 0 & ie^{-i\phi} \\ 0 & 1 & ie^{-i\phi} & 0 \\ 0 & ie^{i\phi} & 1 & 0 \\ ie^{i\phi} & 0 & 0 & 1 \end{pmatrix}$ (gg, ge, eg, ee) Methods: • bichromatic Raman excitation • closed loop in motional phase-space • scalable to *N* ions		Sect. "Mølmer-Sørensen or *xy*-gate", [79], [84], [97]
$Z\phi$	**Geometric Z-phase gate:** phase shift ϕ for ions in different basis states $\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & e^{i\phi} & 0 & 0 \\ 0 & 0 & e^{i\phi} & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$ (gg, ge, eg, ee) Methods: • optical dipole force • phase shift ϕ from closed trajectory in phase space • controlled Z up to single bit rotations		Sect. "*z*-gate", [88], [89]

The dominant contribution to the decoherence of internal states in ions used for quantum information processing are fluctuating ambient magnetic fields [98,99,100]. Through the Zeeman-shift of the levels, they result in random fluctuations of the transition frequency between the qubit states. The sensitivity to magnetic field fluctuations may be reduced by using transitions which are independent of magnetic field to first order. This strategy has been successfully applied in atomic clock systems [101]. An example are states with magnetic quantum number $m = 0$ at zero magnetic field, but similar transitions also exist at finite magnetic field. For a single magnetic field insensitive qubit, memory coherence times greater than 10 s were observed [39], five orders of magnitude larger than in experiments with magnetic field dependent states in $^9Be^+$ [102]. Another source of decoherence are the lasers driving the qubit transition. Via frequency and intensity fluctuations, they contribute to the loss of phase coherence of the qubits.

Coherence is more difficult to maintain as the length of the quantum register increases. Collective dephasing is presently the largest source of decoherence and the limiting factor for the generation of large entangled states (see Sect. "Multiparticle Entanglement"). It is caused, for example, by variations of magnetic, electric field or laser intensity across the size of a typical ion string. Ions in different locations or ions being shuttled through a large-scale trap architecture may be affected by these variations in slightly different ways.

Qubits in Decoherence-free Subspace

Frequently, the environment couples to each qubit of a quantum register in an identical way. In this case, quantum information may be protected by encoding it not in the physical basis states $|g\rangle$ and $|e\rangle$ of the qubits, but in two entangled states of two physical qubits, for example

$$|\Psi_{\pm}\rangle = \frac{1}{\sqrt{2}}\left(|ge\rangle \pm |eg\rangle\right) . \tag{23}$$

These logical states are invariant against common phase shifts of individual basis states. For example, if both ions

simultaneously acquire a phase shift ϕ in state $|e\rangle$, i.e, undergo the transformation $|e\rangle \rightarrow \exp(i\phi)\,|e\rangle$, $|\Psi_{\pm}\rangle$ is not affected, while the individual qubits would lose their phase coherence. The space spanned by $|\Psi_{\pm}\rangle$ is called decoherence-free subspace (DFS) [103,104,105]. It has been demonstrated [85], that a physical qubit can be reversibly encoded in a DFS, achieving coherence times two orders of magnitude longer than for the physical qubit. DFS-encoded states have been used in conjunction with qubits based on field-independent transitions to produce even longer-lived entanglement of the logical qubits [39].

Vibrational State Decoherence

The coupling required for quantum gates in a string of trapped ions is provided by their collective motion. The mode selected as a data-bus must preserve coherent superpositions of the motional ground state $|0\rangle$ and the state $|1\rangle$. Any motional decoherence of the ion chain degrades gate fidelities and must be avoided.

The ions vibrating in a string constitute an oscillating electric dipole and hence their motion couples to fluctuating electric fields in the environment. These may be generated, for example, by patch electric fields on the trap electrodes. The resulting noise in the ions' motion is a prime source of decoherence in ion trap quantum information processing.

The most obvious environmental influence on the vibration of a string is the excitation of vibrational quanta (phonons), corresponding to motional heating of the mode. This has been investigated in a number of experiments [62,65,106]. By measuring the motional excitation as a function of time starting in the ground state, heating rates between 1/ms [106] and 0.005/ms [62] were observed. Factors influencing the heating rate are the distance d between ions and electrodes (scaling approximately as d^{-4} [107]) and coating of trap electrodes with atoms during loading of the trap, resulting in patch-potentials. Therefore, careful shielding of the electrodes from atomic beam exposure is important. Another way to reduce heating rates is cooling the electrodes [107,108].

For a multi-ion string, the rate of motional heating strongly depends on which normal mode is used. The COM-mode has heating rates comparable to that of a single ion, as it can be excited by a homogeneous fluctuating field coupling to each individual ion in an equivalent way. The stretch-mode, on the other hand, can only be excited by a field gradient and is therefore much less sensitive to fluctuating fields emanating from the trap electrodes. Therefore, for high fidelity gate operations, the stretch mode is preferable.

While motional heating sets an upper limit to the coherence time of bus qubits, for reliable quantum operations also dephasing of motional superpositions must be taken into account. This has been investigated at NIST [109], in Innsbruck [110] and in Oxford [111]. Dephasing times found are on the order of 100 ms. Dephasing is faster for larger differences in the quantum numbers of the states involved, making it advantageous to use the lowest vibrational states $|0\rangle$ and $|1\rangle$.

Even though only one normal mode is used for quantum information transfer, heating of the remaining $N-1$ axial vibration modes in an N-ion string (spectator modes) is detrimental, since it affects the Rabi-frequency of the bus-mode. This may be alleviated by cooling all modes of the string, ideally to the quantum mechanical ground state.

With typical gate durations on the order of 1 μs, decoherence times of 1 s would permit 10^6 quantum operations to be performed. Another approach is to consider the probability of error during a gate operation [112]. It should be smaller than 10^{-4} to be able to successfully apply quantum error correction. This corresponds to a fidelity of $F > 0.9999$, which present quantum gates do not yet achieve.

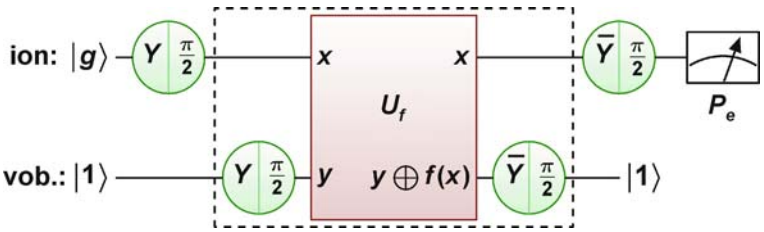
Quantum Algorithms

The two-ion gates described in the previous section are universal, i. e., any logic circuit may be realized by combining them with single qubit operations [113,114]. In this section, examples of algorithms implemented in ion traps are presented. The most important quantum algorithms have been demonstrated in ion-trap systems in their simplest form.

Deutsch–Jozsa Algorithm

The Deutsch–Jozsa algorithm [115] is the simplest example of quantum parallelism. The task it performs is to decide if a function f is constant, i. e., has either always the result $f(x) = 0$ or always $f(x) = 1$, independent of the input x, or if it is balanced, i. e. its output depends on the input ($f(x) = x$ or $f(x) = \mathrm{NOT}x$). With classical means, two function evaluations are necessary, whereas the quantum algorithm performs the task with a single application of the function to a superposition state.

The algorithm has been demonstrated with a single ionic qubit, using its vibrational excitation as an auxiliary qubit [75] (see Fig. 23). The required pulses were applied either resonantly or on the blue sideband, in the latter case using composite pulses in order to avoid excitation of more than one phonon (cf. Sect. "Cirac–Zoller Gate").



Quantum Computing with Trapped Ions, Figure 23
Quantum circuit implementing the Deutsch–Jozsa algorithm with a trapped ion. The upper line represents the ionic qubit $|x\rangle$, the lower line an auxiliary vibrational qubit. In the central box, one of four possible functions $f(x)$ is implemented by acting on y with the identity, a NOT gate and/or a CNOT-gate controlled by x. A final measurement of the ion in state $|e\rangle$ indicates a balanced function after only a single function evaluation [75]

Quantum Teleportation

A more complex task and one of fundamental importance for quantum computation is quantum teleportation, the transfer of an unknown quantum state from one qubit to another in a distant location [116]. Even though no finite number of measurements is sufficient to obtain a full specification of a quantum state, the transfer is possible with the help of non-local correlations provided by an entangled pair of qubits. While probabilistic teleportation has been demonstrated with photonic qubits, two ion-trap experiments have implemented a fully deterministic quantum teleportation protocol. At the University of Innsbruck, $^{40}Ca^{+}$-ions were used [99], while at NIST the experiment was performed with $^{9}Be^{+}$ [98].

The teleportation procedure requires three ionic qubits. The essential steps are summarized in Fig. 24. Initially, ions B and C are prepared in a maximally entangled state using the gates described in Sect. "Cirac–Zoller Gate" (Innsbruck) and Sect. "Geometric Gates" (NIST). Subsequently, the state to be transferred is prepared in ion A by a single-qubit rotation. The central step of the teleportation protocol is a so-called Bell-state measurement of the states of ions A and B. This is achieved by using another gate to entangle qubits A and B and, after subjecting them to a $\pi/2$-pulse, measuring the state of both qubits. Taken separately, none of the four possible measurement outcomes (gg, ge, eg, or ee), nor the remaining ion C carry any information about the original state. Only when one of four unitary state rotations is applied to ion C, chosen depending on the outcome of the measurement, the state of ion C becomes identical to the original state A.

In their implementation of this protocol, the two groups have employed quite different technologies. The qubits are stored differently and different realizations of the phase gate are employed. The Innsbruck group addresses individual ions with tightly focused lasers, protecting the remaining ions from the target ion's fluorescence by hiding them in another internal state. At NIST, ions are selectively moved to separate zones in a segmented trap, where they can be manipulated individually while maintaining entanglement. In spite of these differences, in both experiments the quantum bit was transferred with a fidelity of around 75%. In a quantum computer, teleportation may be applied to transfer quantum information without physically moving qubits.

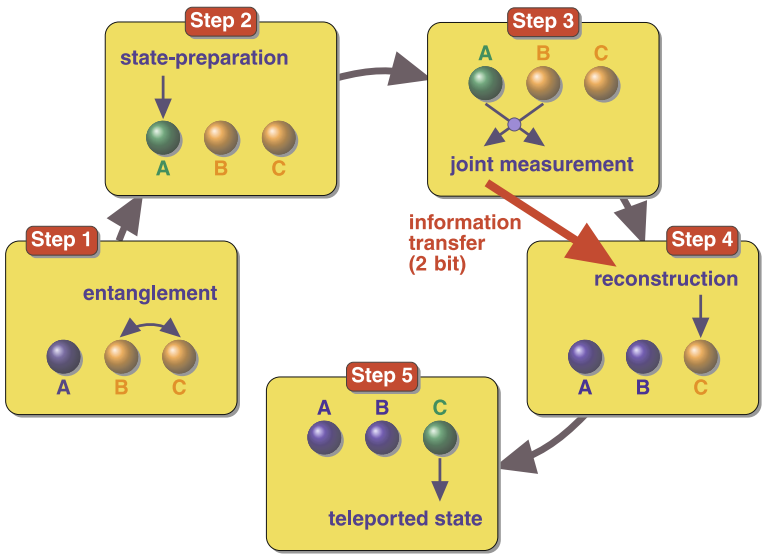
Quantum Fourier Transform

One of the most celebrated quantum algorithms was developed by P. Shor for finding the prime factors of a large number [117]. A central element of this algorithm is the *quantum Fourier transform* of a set of qubits. Here, the amplitudes x_k of a superposition of basis states are Fourier-transformed, resulting in a new state with amplitudes y_j:

$$\sum_{k=0}^{N-1} x_k\,|k\rangle \longrightarrow \sum_{j=0}^{N-1} y_j\,|j\rangle \quad \text{with} \quad y_j = \sum_{k=0}^{N-1} x_k e^{i2\pi jk/N}. \tag{24}$$

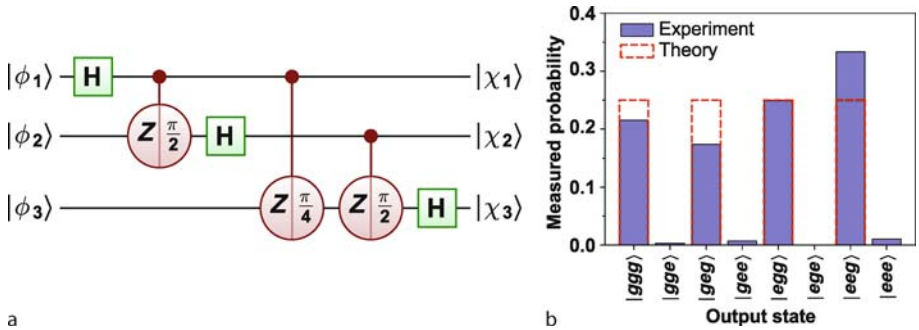
Here, $|k\rangle$ is related to the corresponding ion-state through its binary representation with $0 \equiv g$ and $1 \equiv e$.

The quantum Fourier transform was experimentally demonstrated in a system of three beryllium ions for states with periods 1,2,4,8 and approximately 3 [118]. The sequence of gates that was applied is shown in Fig. 25a. It comprises Hadamard transforms and controlled Z-gates with different phase factors. In the experiment, these are replaced by equivalent sequences of X- and Y-pulses. A simplification of the scheme is possible by performing the read-out of each qubit immediately after the last Hadamard-gate is applied. Measurement of the individual ions is achieved by separating them in a segmented linear trap. The subsequent quantum gates in Fig. 25a must



Quantum Computing with Trapped Ions, Figure 24

Schematic representation of quantum teleportation in a chain of three ions: the state of ion A (green) is transferred to ion C with the help of an entangled pair of ions B and C. Blue arrows indicate the manipulation of ions with laser pulses, while the red arrow corresponds to the transmission of two bit of classical information. This protocol is used in two independent experiments [98,99], which differ only in technical details



Quantum Computing with Trapped Ions, Figure 25

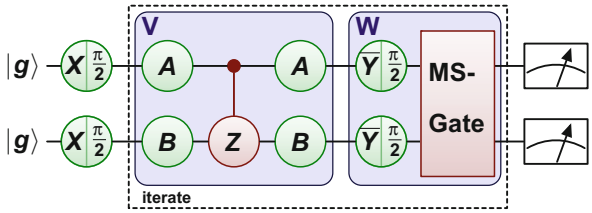
a Circuit diagram for quantum Fourier transform of three qubits, composed of Hadamard transforms and controlled Z-gates (with rotation angles $\theta = \pi/2$ and $\pi/4$). In the actual experiment, these gates were replaced by equivalent rotation sequences around the *x*- and *y*-axis; b experimental output for the quantum Fourier transform of the input state $|gee\rangle + |eee\rangle$ with period 4 [118]

then be replaced by classically controlled phase shifts. Figure 25b shows an example of the outcome of such a measurement. A state with a period 4 in an 8 element space has non-zero Fourier components with a periodicity of 2 (= 8/4). Therefore, only every other state is observed in the output. Depending on the input state, an accuracy between 87% and 99% was reached.

Grover's Quantum Search Algorithm

Grover's search algorithm is another example of a quantum computer outperforming its classical counterpart. It succeeds in finding a marked entry in a database of length n after $\pi\sqrt{n}/4$ queries, rather than n [119]. It requires an operator V (oracle) marking the target state $|a\rangle$ of the search by flipping its sign and an operator W to amplify the contribution of the target state, until, after a few iterations, only the target state has a sizable probability. The position of the target state in the database is then revealed after a measurement. The corresponding quantum circuit is shown in Fig. 26.

In the experiment [120], a four element database was set up with two ionic qubits. Using $\pi/2$-pulses, an equally weighted superposition of all basis states is prepared as



Quantum Computing with Trapped Ions, Figure 26
Scheme of Grover's search algorithm for N qubits encoding $n = 2^N$ entries, illustrated for $N = 2$. Initially, all qubits are prepared in the ground state. $\pi/2$-pulses prepare an equal superposition of all basis states. The oracle V flips the sign of one marked element $|a\rangle$. Which state is marked is determined by the operators A and B, which can be either NOT (X-gate, see Tab. 2) or the identity. Two more $\pi/2$-pulses followed by a Mølmer–Sørensen gate amplify the weight of the marked state. After a number of iterations, the qubits are measured

the input: $|\phi\rangle = \left(|gg\rangle + |ge\rangle + |eg\rangle + |ee\rangle\right)/2$. The oracle V as well as the amplification W were implemented using the Mølmer–Sørensen gate (Sect. "Mølmer-Sørensen or xy-gate") together with single-qubit rotations. The iteration step in Fig. 26 takes the state $|\phi\rangle$ through the following transformation:

$$|\phi\rangle \xrightarrow{V} (|\phi\rangle - |a\rangle) \xrightarrow{W} |a\rangle \ .$$

Therefore, the marked state $|a\rangle$ should be retrieved in only a single query. The measured probability was 60%, surpassing the classical limit of 50%. For a register size larger than $N = 2$, the operators V and W have to be applied iteratively.

Quantum Error Correction

Any quantum computer is subject to noise, resulting in uncontrolled, irreversible changes of the qubits involved. Even using decoherence-free subspaces (see Sect. "Qubits in Decoherence-Free Subspace"), residual noise still reduces the fidelity of quantum memory and quantum gates. The ability to correct these errors is therefore of great importance for the reliable operation of a quantum computer. The challenge is that error correction must be accomplished without gaining any knowledge about the stored information, since this would destroy any quantum superpositions.

An example for an error which might occur during the quantum computation is the flip of a single qubit. The correction of an error of this type has been demonstrated at NIST [100]. To this end, a redundant encoding of a single logical qubit in three physical qubits was applied. A qubit state $\alpha\,|g\rangle + \beta\,|e\rangle$ was encoded as an entangled state $\alpha\,|ggg\rangle + \beta\,|eee\rangle$. At NIST, the two-ion phase gate described in Sect. "Geometric Gates", extended to three ions was used. The logical qubit was then subjected to an artificial bit-flip error of variable rate. After decoding the logical qubit by reversing the three-ion entangling gate, information on which error has occurred was obtained by measuring two of the physical qubits. The outcome determined which correction operation was applied to the remaining qubit to restore the original state. Finally, the qubit was analyzed to determine the efficiency of the method. The protocol is illustrated in Fig. 27.

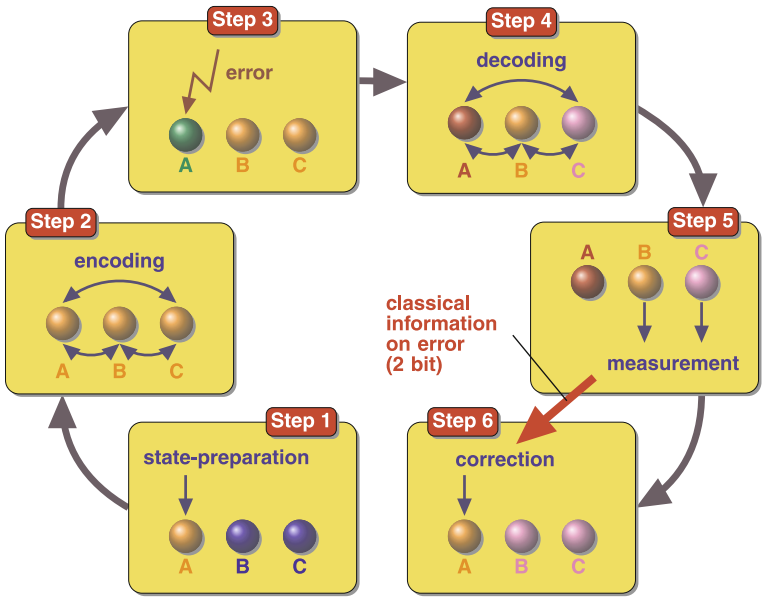
The data showed that quantum error correction improved the fidelity if bit-flips occurred with a probability larger than 25%. For smaller errors, the imperfections of the gates applied during encoding and decoding of the logical qubit outweighed the benefit of correction and lead to a reduced fidelity. Improved gates are necessary before a fault tolerant operation can be achieved. Since quantum error correction schemes require a large number of physical quantum bits, passive protection against errors, for example through decoherence-free subspaces, is important.

Multiparticle Entanglement

As shown in the previous section, entanglement is an essential resource in many quantum algorithms. In the course of a quantum computation, a large number of qubits must be entangled. It is therefore of great relevance to investigate highly entangled quantum states of many qubits. This is not only an important demonstration of the amount of control over a quantum system, but can also serve as a resource for quantum information processing [121].

Entangled states may be generated with the help of two- or multi-bit quantum gates (Sect. "Two-Qubit Interaction and Quantum Gates"). An early achievement was the entanglement of four ions [84] using the Mølmer–Sørensen gate (Sect. "Mølmer-Sørensen or xy-gate"). Since then, the development of new gates and better control over experimental parameters have led to the creation and verification of entangled states of up to eight ions. Each additional particle greatly increases the difficulty of the experiment.

Two different classes of entangled states have been investigated. The first type is called Greenberger–Horne-Zeilinger (GHZ) or *Schrödinger cat* states and consists of equal superpositions of maximally different quantum



Quantum Computing with Trapped Ions, Figure 27
Schematic representation of quantum error correction of flips of a Single logical quantum bit logical quantum bit, encoded in a chain of three ions. This protocol was demonstrated experimentally [100]

states of N ions.

$$|\mathrm{GHZ}_N\rangle = \frac{1}{\sqrt{2}}\left(|\underbrace{ggg\ldots g}_{N}\rangle + e^{i\varphi}|\underbrace{eee\ldots e}_{N}\rangle\right). \quad (25)$$

Three-ion GHZ-states were generated and controlled in 2004 at NIST ($F = 89\%$) [122] and the University of Innsbruck ($F = 72\%$) [123]. In both experiments, the potential of entanglement for practical applications was explored.

The largest GHZ-state so far was generated at NIST, using a generalization of the phase gate described in Sect. "z-gate" for up to $N = 6$ ions [97]. Together with a common single-bit rotation of all the ions before and after the phase gate, the algorithm involved three steps, irrespective of the number of ions to be entangled, making this a very efficient method for multiparticle entanglement generation. The fidelity obtained for the six-ion entangled state was estimated to be better than 51%.

In order to verify the coherence of the entangled states, experimenters have subjected them to another phase gate, identical to the entangling operation, but in a reference system rotated around the z-axis of the Bloch-sphere by an angle ϕ. In this way, the coherence of the state $|\mathrm{GHZ}_N\rangle$ is converted to a population difference, which can be directly measured. For a perfect N-ion GHZ-state, the population difference and hence the fluorescence signal varies as

$$P_{gg\ldots g} = \tfrac{1}{2}\left[1 - \cos(N\phi)\right]. \quad (26)$$

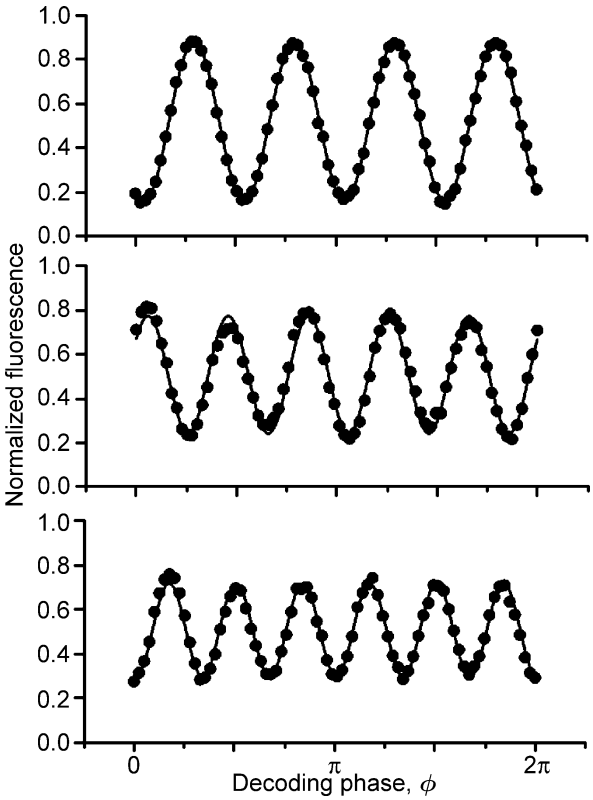
Results of this method for GHZ-states with four, five, and six ions are shown in Fig. 28. The observed contrast provides a lower bound for the coherence and hence the fidelity of the state.

This method of probing coherences is closely related to Ramsey spectroscopy and, indeed, GHZ-states of multiple ions have been applied in spectroscopy. The N-fold increased oscillation frequency observed in Fig. 28 can be used to improve the phase sensitivity of a Ramsey spectrometer by a factor of $\sqrt{N}$ compared to unentangled atoms [122]. Other applications of GHZ-states include quantum error correction (Sect. "Quantum Error Correction") and the deterministic preparation of Bell states.

A different class of entangled states are the N-particle W-states, consisting of a superposition of N states with exactly one particle in state $|g\rangle$ and all others in $|e\rangle$:

$$|\mathrm{W}_N\rangle = \frac{1}{\sqrt{N}}\left(e^{i\varphi_1}|e\ldots eeeg\rangle + e^{i\varphi_2}|e\ldots eege\rangle + e^{i\varphi_3}|e\ldots egee\rangle + \cdots + e^{i\varphi_N}|gee\ldots ee\rangle\right). \quad (27)$$

These states were generated in Innsbruck with $N = 3$ [123] and later up to $N = 8$ [124]. First, all ions are



Quantum Computing with Trapped Ions, Figure 28
Fluorescence measurement of decoded GHZ-states as a function of the decoding phase ϕ for $N = 4$, 5, and 6 ions. The sinusoidal dependence on $N\phi$ is a signature of an N-ion GHZ state [97]. The observed contrast is related to the coherence of the state

initialized in state $|e\rangle$ and the vibrational COM-mode in state $|1\rangle$. N pulses on the blue sideband of the qubit transition, are consecutively applied to each ion in the string, exciting the level $|g\rangle$ of that ion with a probability of $1/N$. The entanglement is verified by measuring all N^2 elements of the density matrix, completely characterizing the quantum state. For the state $|W_8\rangle$, a fidelity of $F = 72\%$ was determined.

The entanglement properties of W-states are very different from those of GHZ-states. If the state of one particle is measured to be $|e\rangle$, the remaining particles are still entangled. In addition, W-states are robust against global dephasing. They might serve as a resource for quantum information processing as well as quantum communication.

The six-ion GHZ- and eight-ion W-states define the state-of-the-art in the generation of multiparticle entanglement. The generation of even larger entangled states is a major challenge, due to the increased sensitivity to decoherence and inhomogeneous effects.

Distributed Quantum Information with Trapped Ions

The quantum systems discussed so far were confined to single ion traps. However, many important applications require the distribution of quantum information among different locations. A prime example is quantum communication, which at present is based on photons as carriers of quantum information. Substantial benefits are expected from linking quantum communication with quantum computing. For example, small systems of trapped ions could act as routers or repeaters of quantum information. Teleportation is another application which is relevant mainly over long distances, i. e., between remote ion traps. Even quantum computation itself could gain from being spread among a number of local processing sites (distributed quantum computation), linked by quantum communication channels [125].

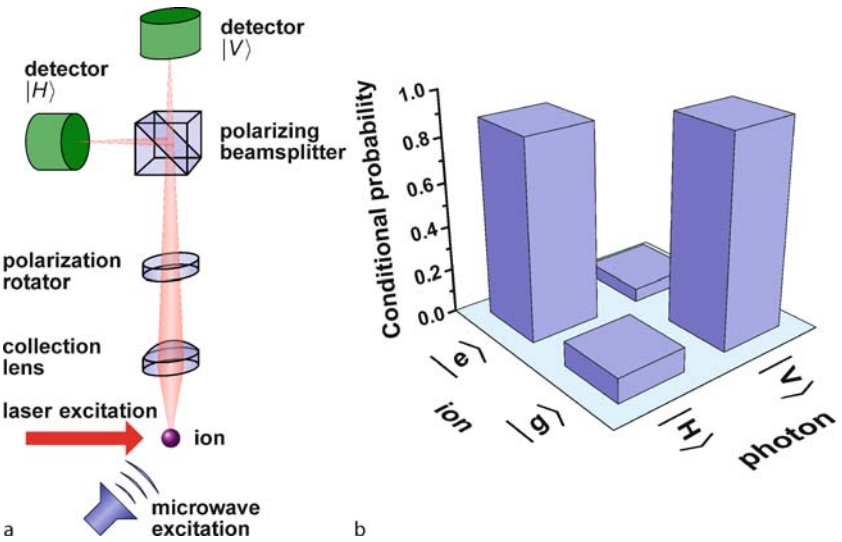
It is therefore an important goal to establish reliable quantum networks involving spatially separated traps. A first step is to create entanglement between ions stored in remote locations. There are three methods to achieve this:

(1) An entangled pair of ions is prepared in one location, and one of the particles is subsequently transported to a distant site.
(2) Entanglement is created between an ionic and a photonic qubit, the latter transferring entanglement deterministically to an ion in a distant trap.
(3) Entanglement is created locally in two distant traps between an ion and a photon on each side. Subjecting the emitted photons to a joint measurement projects the ions left behind to an entangled state. This scheme works probabilistically, as it is dependent on a particular measurement outcome.

The first method is suited for entanglement distribution within an extended trap architecture (Sect. "Large-Scale Ion Traps"), while the latter two methods allow long distance entanglement.

Ion–Photon and Remote Ion–Ion Entanglement

A crucial step towards mapping quantum information from an atomic ion to a photon is to generate entanglement between trapped ions and photons, establishing a long-distance communication channel. This was demonstrated at the University of Michigan using a single $^{111}Cd^+$-ion [126]. In the experiment sketched in Fig. 29a, the ion was excited to a state with two channels of spontaneous decay, leading to different hyperfine states of the



Quantum Computing with Trapped Ions, Figure 29

a Experimental setup for generation of ion-photon entanglement. Photons scattered from the excitation beam are analyzed by a beamsplitter distinguishing V- and H-polarization. The polarization rotator is used to change the measurement basis. The qubit-transition is driven by a microwave field. b Measured conditional probabilities after both atomic and photonic qubits are rotated by a polar angle $\theta = \pi/2$ on the Bloch sphere with equal phase Φ. If the atomic and photonic qubits were not entangled but in a statistically mixed state, all conditional probabilities in the figure would have been 0.5 [126]

ion ($|g\rangle$ or $|e\rangle$). Information on which transition has occurred is contained in the polarization of the emitted photon (horizontal=$|\mathrm{H}\rangle$ or vertical=$|\mathrm{V}\rangle$). Taking into account the relative strength of the transitions, the resulting state is $\sqrt{1/3}\,|\mathrm{H}\rangle\,|e\rangle + \sqrt{2/3}\,|\mathrm{V}\rangle\,|g\rangle$, which is entangled. In the experiment, it was obtained with a fidelity of $F = 97\%$ (see Fig. 29b). Note that this method of generating entanglement is probabilistic, since it relies on the spontaneous emission of a photon.

The generation of entanglement between ion and photon is only an intermediate step towards the goal of entangling the quantum states of ions in different traps. This has been achieved by the same group. Initially, single $^{171}\mathrm{Yb}^{+}$-ions are stored in two identical traps 1 m apart. Using the same principle as described above, an entangled ion-photon state $\left(|e\rangle_i\,|\nu_e\rangle_i - |g\rangle_i\,|\nu_g\rangle_i\right)/\sqrt{2}$ is created in each trap, where the index i identifies the trap. The photon states $|\nu_g\rangle$ and $|\nu_e\rangle$ are distinguished by their frequency. The two photon-states are then brought to interference on a beamsplitter and measured at its two output ports. In case of a coincident detection, the state of the atoms is projected to the entangled state $\left(|e\rangle_1\,|g\rangle_2 - |g\rangle_1\,|e\rangle_2\right)/\sqrt{2}$. An experimental verification yields a fidelity of 63% [127].

Ion-Trap Cavity-QED

One of the drawbacks of entanglement using spontaneously emitted photons is the low success probability, due to the emission of photons into the full solid angle. As a result, only one ion–ion pair was entangled every 8.5 min in the experiment reported in Ref. [127]. A solution to this problem is to place the ion in a high-finesse Fabry–Perot optical cavity resonant with the emitted photons. The cavity enhances the coherent interaction between ion and photon and provides a well-defined mode for the photon, resulting in a predetermined direction of propagation when it eventually escapes through a semi-transparent mirror. Individual $^{40}\mathrm{Ca}^{+}$-ions have already been stored in and coupled to an optical cavity, localized well within the Lamb–Dicke regime [128,129].

Optical cavities have applications beyond enhancing the success rate of probabilistic entanglement schemes. If the cavity mode volume is small enough, the coherent interaction between ions and photons becomes stronger than all spontaneous decay processes and the system dynamics becomes deterministic. This is known as the strong coupling limit of cavity-QED. It has been proposed as a technique to transfer quantum states or distribute entanglement deterministically in a quantum network [130]. According to this scheme, the qubit state of an ion is mapped to a cavity photon with the help of a laser pulse. This photon leaks out of the first cavity and enters a second cavity, in which it is absorbed by another ion. A requirement for this transfer is that the coupling between the second ion and the second cavity is provided by a laser pulse which is time-reversed with respect to the first pulse and

that the photon-wavepacket carrying the quantum bit is time-symmetric.

While the strong coupling limit hasn't been reached with single trapped ions yet, the system has already been used as a source of single photons on demand, emitting photons with a predetermined shape [131]. Cavity-QED techniques have many potential applications in quantum information processing with trapped ions, in particular for the reversible mapping of qubits between photons and ions [125].

Future Directions

As shown in this review, all essential elements of quantum information processing have been demonstrated with trapped ions, making them the most successful technology for quantum computation to date. The biggest challenge now is to scale up the present systems to a size where algorithms of practical interest could be run. At the same time, the fidelity of gate operations must be improved, in order to reach the point where quantum error correction could be successfully applied. The implementation of quantum error correction will further increase the demand for larger quantum registers, due to the memory overhead of logical qubits.

The most promising route to large-scale quantum computation is a segmented trap architecture with a large number of zones, each storing only a few ions [25]. Present activities in laboratories around the world are directed towards building suitable traps using nanofabrication methods. It is still a major challenge to combine all the required techniques in one scalable system. The implementation of a quantum computer with 300 qubits is discussed by A.M. Steane [132].

Before a universal quantum computer will be available, trapped ion systems are expected to find applications in the analog simulation of other quantum systems, as was originally proposed by Feynman [2]. The idea is to use the evolution of a system of trapped ions to mimic the behavior of other, less accessible systems with equivalent Hamiltonians. An example are quantum phase transitions in spin systems, which could be investigated with trapped ions [133]. A recent experiment has simulated the transition from paramagnetic to ferromagnetic order in a quantum system of two ions [134]. A quantum simulation that was successfully performed at NIST demonstrated the increased sensitivity of 2nd- and 3rd-order nonlinear beam splitters [135]. Trapped ions are even predicted to simulate cosmological particle creation in the early universe [136].

Beyond quantum computation or simulation, the techniques developed for ion trap quantum computers have already started to find applications in other fields. One example is frequency standards. Here, the use of N entangled ions has been demonstrated to increase the sensitivity by a factor of $\sqrt{N}$ [122]. Methods from ion-trap quantum information processing have also been applied to spectroscopic investigation of $^{27}Al^{+}$ by using $^{9}Be^{+}$ to cool, initialize and detect the Al-ion, which lacks a suitable level structure [137]. The rapid progress of ion-trap technology not only makes harnessing the power of quantum computation a realistic possibility in the near future, but already provides new tools for fundamental and applied research.

Bibliography

Primary Literature

1. Feynman RP (1960) There's plenty of room at the bottom. Eng Sci 23:22–36
2. Feynman RP (1982) Simulating physics with computers. Int J Theor Phys 21:467–488
3. Benioff P (1982) Quantum-mechanical models of turing-machines that dissipate no energy. Phys Rev Lett 48:1581–1585
4. Cirac JI, Zoller P (1995) Quantum computations with cold trapped ions. Phys Rev Lett 74:4091–4094
5. DiVincenzo DP (2000) The physical implementation of quantum computation. Fortschritte Phys-Prog Phys 48:771–783
6. Ghosh PK (1995) Ion Traps. Clarendon, Oxford
7. Paul W (1990) Electromagnetic traps for charged and neutral particles. Rev Mod Phys 62:531–540
8. Paul W, Steinwedel H (1953) Ein neues Massenspektrometer ohne Magnetfeld. Z Naturforsch 8:448–450
9. Waki I, Kassner S, Birkl G et al (1992) Observation of ordered structures of laser-cooled ions in a quadrupole storage ring. Phys Rev Lett 68:2007–2010
10. Nagerl HC, Bechter W, Eschner J et al (1998) Ion strings for quantum gates. Appl Phys B-Lasers Opt 66:603–608
11. Raizen MG, Gilligan JM, Bergquist JC et al (1992) Ionic-crystals in a linear paul trap. Phys Rev A 45:6493–6501
12. Drewsen M, Brodersen C, Hornekaer L et al (1998) Large ion crystals in a linear Paul trap. Phys Rev Lett 81:2827–2830
13. James DFV (1998) Quantum dynamics of cold trapped ions with application to quantum computation. Appl Phys B-Lasers Opt 66:181–190
14. Hughes RJ, James DFV, Gomez JJ et al (1998) The Los Alamos trapped ion quantum computer experiment. Fortschritte Phys-Prog Phys 46:329–361
15. Berkeland DJ, Miller JD, Bergquist JC et al (1998) Minimization of ion micromotion in a Paul trap. J Appl Phys 83:5025–5033
16. Hoffges JT, Baldauf HW, Eichler T et al (1997) Heterodyne measurement of the fluorescent radiation of a single trapped ion. Opt Commun 133:170–174
17. Dehmelt H (1967) Radiofrequency Spectroscopy of Stored Ions. Adv At Mol Phys 3:53
18. Straubel H (1955) Zum Öltröpfchenversuch von Millikan. Naturwissenschaften 42:506–507
19. Brewer RG, Devoe RG, Kallenbach R (1992) Planar ion microtraps. Phys Rev A 46:R6781–R6784
20. Schrama CA, Peik E, Smith WW et al (1993) Novel miniature ion traps. Opt Commun 101:32–36

21. Jefferts SR, Monroe C, Bell EW et al (1995) Coaxial-resonator-driven rf (paul) trap for strong confinement. Phys Rev A 51:3112–3116
22. Monroe C, Meekhof DM, King BE et al (1995) Demonstration of a fundamental quantum logic gate. Phys Rev Lett 75:4714–4717
23. Turchette QA, Wood CS, King BE et al (1998) Deterministic entanglement of two trapped ions. Phys Rev Lett 81:3631–3634
24. Cirac JL, Zoller P (2000) A scalable quantum computer with ions in an array of microtraps. Nature 404:579–581
25. Kielpinski D, Monroe C, Wineland DJ (2002) Architecture for a large-scale ion-trap quantum computer. Nature 417:709–711
26. Rowe MA, Ben-Kish A, Demarco B et al (2002) Transport of quantum states and separation of ions in a dual RF ion trap. Quantum Inform Comput 2:257–271
27. Hensinger WK, Olmschenk S, Stick D et al (2006) T-junction ion trap array for two-dimensional ion shuttling, storage, and manipulation. Appl Phys Lett 88:034101
28. Madsen MJ, Hensinger WK, Stick D et al (2004) Planar ion trap geometry for microfabrication. Appl Phys B-Lasers Opt 78:639–651
29. Brownnutt M, Wilpers G, Gill P et al (2006) Monolithic microfabricated ion trap chip design for scaleable quantum processors. New J Phys 8:232
30. Stick D, Hensinger WK, Olmschenk S et al (2006) Ion trap in a semiconductor chip. Nat Phys 2:36–39
31. Seidelin S, Chiaverini J, Reichle R et al (2006) Microfabricated surface-electrode ion trap for scalable quantum information processing. Phys Rev Lett 96:253003
32. Chiaverini J, Blakestad RB, Britton J et al (2005) Surface-electrode architecture for ion-trap quantum information processing. Quantum Inform Comput 5:419–439
33. Kim J, Pau S, Ma Z et al (2005) System design for large-scale ion trap quantum information processor. Quantum Inform Comput 5:515–537
34. Mitchell TB, Bollinger JJ, Dubin DHE et al (1998) Direct observations of structural phase transitions in planar crystallized ion plasmas. Science 282:1290–1293
35. Porras D, Cirac JI (2006) Phonon superfluids in sets of trapped ions. Found Phys 36:465–476
36. Castrejon-Pita JR, Ohadi H, Crick DR et al (2007) Novel designs for Penning ion traps. J Mod Opt 54:1581–1594
37. Crick DR, Ohadi H, Bhatti I et al (2008) Two-ion Coulomb crystals of Ca+ in a Penning trap. Opt Express 16:2351–2362
38. Ciaramicoli G, Marzoli I, Tombesi P (2003) Scalable quantum processor with trapped electrons. Phys Rev Lett 91:017901
39. Langer C, Ozeri R, Jost JD et al (2005) Long-lived qubit memory using atomic ions. Phys Rev Lett 95:060502
40. Nagerl HC, Roos C, Rohde H et al (2000) Addressing and cooling of single ions in Paul traps. Fortschritte Phys-Prog Phys 48:623–636
41. Lucas DM, Donald CJS, Home JP et al (2003) Oxford ion-trap quantum computing project. Philos Trans R Soc Lond Ser A-Math Phys Eng Sci 361:1401–1408
42. Benhelm J, Kirchmair G, Rapol U et al (2007) Measurement of the hyperfine structure of the S-1/2-D-5/2 transition in Ca-43^{+}. Phys Rev A 75:032506
43. Deslauriers L, Haijan PC, Lee PJ et al (2004) Zero-point cooling and low heating of trapped Cd-111^{+} ions. Phys Rev A 70:043408
44. Balzer C, Braun A, Hannemann T et al (2006) Electrodynamically trapped Yb+ ions for quantum information processing. Phys Rev A 73:041407
45. Olmschenk S, Younge KC, Moehring DL et al (2007) Manipulation and detection of a trapped Yb+ hyperfine qubit. Phys Rev A 76:052314
46. Schaetz T, Friedenauer A, Schmitz H et al (2007) Towards (scalable) quantum simulations in ion traps. J Mod Opt 54:2317–2325
47. Berkeland DJ (2002) Linear Paul trap for strontium ions. Rev Sci Instrum 73:2856–2860
48. Brown KR, Clark RJ, Labaziewicz J et al (2007) Loading and characterization of a printed-circuit-board atomic ion trap. Phys Rev A 75:015401
49. Brownnutt M, Letchumanan V, Wilpers G et al (2007) Controlled photoionization loading of Sr-88^{+} for precision ion-trap experiments. Appl Phys B-Lasers Opt 87:411–415
50. Letchumanan V, Wilpers G, Brownnutt M et al (2007) Zero-point cooling and heating-rate measurements of a single Sr-88^{+} ion. Phys Rev A 75:063425
51. Kjaergaard N, Hornekaer L, Thommesen AM et al (2000) Isotope selective loading of an ion trap using resonance-enhanced two-photon ionization. Appl Phys B-Lasers Opt 71:207–210
52. Gulde S, Rotter D, Barton P et al (2001) Simple and efficient photo-ionization loading of ions for precision ion-trapping experiments. Appl Phys B-Lasers Opt 73:861–863
53. Deslauriers L, Acton M, Blinov BB et al (2006) Efficient photoionization loading of trapped ions with ultrafast pulses. Phys Rev A 74:063421
54. Lucas DM, Ramos A, Home JP et al (2004) Isotope-selective photoionization for calcium ion trapping. Phys Rev A 69:012711
55. Happer W (1972) Optical-pumping. Rev Mod Phys 44:169
56. Wineland DJ, Drullinger RE, Walls FL (1978) Radiation-pressure cooling of bound resonant absorbers. Phys Rev Lett 40:1639–1642
57. Neuhauser W, Hohenstatt M, Toschek P et al (1978) Optical-sideband cooling of visible atom cloud confined in parabolic well. Phys Rev Lett 41:233–236
58. Wineland DJ, Itano WM (1979) Laser cooling of atoms. Phys Rev A 20:1521–1540
59. Stenholm S (1986) The semiclassical theory of laser cooling. Rev Mod Phys 58:699–739
60. Diedrich F, Bergquist JC, Itano WM et al (1989) Laser cooling to the zero-point energy of motion. Phys Rev Lett 62:403–406
61. Monroe C, Meekhof DM, King BE et al (1995) Resolved-sideband raman cooling of a bound atom to the 3d zero-point energy. Phys Rev Lett 75:4011–4014
62. Roos C, Zeiger T, Rohde H et al (1999) Quantum state engineering on an optical transition and decoherence in a Paul trap. Phys Rev Lett 83:4713–4716
63. Kielpinski D, King BE, Myatt CJ et al (2000) Sympathetic cooling of trapped ions for quantum logic. Phys Rev A 61:032310
64. Schmidt-Kaler F, Roos C, Nagerl HC et al (2000) Ground state cooling, quantum state engineering and study of decoherence of ions in Paul traps. J Mod Opt 47:2573–2582
65. Rohde H, Gulde ST, Roos CF et al (2001) Sympathetic ground-state cooling and coherent manipulation with two-ion crystals. J Opt B-Quantum Semicl Opt 3:S34–S41
66. Barrett MD, DeMarco B, Schaetz T et al (2003) Sympathetic

cooling of Be-9$^+$ and Mg-24$^+$ for quantum logic. Phys Rev A 68:042302
67. King BE, Wood CS, Myatt CJ et al (1998) Cooling the collective motion of trapped ions to initialize a quantum register. Phys Rev Lett 81:1525–1528
68. Mintert F, Wunderlich C (2001) Ion-trap quantum logic using long-wavelength radiation. Phys Rev Lett 87:257904
69. Nagourney W, Sandberg J, Dehmelt H (1986) Shelved optical electron amplifier – observation of quantum jumps. Phys Rev Lett 56:2797–2799
70. Wineland DJ, Bergquist JC, Itano WM et al (1980) Double-resonance and optical-pumping experiments on electromagnetically confined, laser-cooled ions. Opt Lett 5:245–247
71. Blatt R, Zoller P (1988) Quantum jumps in atomic systems. Eur J Phys 9:250–256
72. Bergquist JC, Hulet RG, Itano WM et al (1986) Observation of quantum jumps in a single atom. Phys Rev Lett 57:1699–1702
73. Vogel K, Risken H (1989) Determination of quasiprobability distributions in terms of probability distributions for the rotated quadrature phase. Phys Rev A 40:2847–2849
74. Roos CF, Lancaster GPT, Riebe M et al (2004) Bell states of atoms with ultralong lifetimes and their tomographic state analysis. Phys Rev Lett 92:220402
75. Gulde S, Riebe M, Lancaster GPT et al (2003) Implementation of the Deutsch-Jozsa algorithm on an ion-trap quantum computer. Nature 421:48–50
76. Schmidt-Kaler F, Haffner H, Riebe M et al (2003) Realization of the Cirac-Zoller controlled-NOT quantum gate. Nature 422:408–411
77. Childs AM, Chuang IL (2001) Universal quantum computation with two-level trapped ions. Phys Rev A 6301:012306
78. Milburn GJ (1999) Simulating nonlinear spin models in an ion trap. arXiv:quant-ph/9908037v1
79. Sorensen A, Molmer K (1999) Quantum computation with ions in thermal motion. Phys Rev Lett 82:1971–1974
80. Molmer K, Sorensen A (1999) Multiparticle entanglement of hot trapped ions. Phys Rev Lett 82:1835–1838
81. Solano E, de Matos RL, Zagury N (1999) Deterministic Bell states and measurement of the motional state of two trapped ions. Phys Rev A 59:R2539–R2543
82. Sorensen A, Molmer K (2000) Entanglement and quantum computation with ions in thermal motion. Phys Rev A 6202:022311
83. Milburn GJ, Schneider S, James DFV (2000) Ion trap quantum computing with warm ions. Fortschritte Phys – Prog Phys 48:801–810
84. Sackett CA, Kielpinski D, King BE et al (2000) Experimental entanglement of four particles. Nature 404:256–259
85. Kielpinski D, Meyer V, Rowe MA et al (2001) A decoherence-free quantum memory using trapped ions. Science 291: 1013–1015
86. Haljan PC, Brickman KA, Deslauriers L et al (2005) Spin-dependent forces on trapped ions for phase-stable quantum gates and entangled states of spin and motion. Phys Rev Lett 94:153602
87. Haljan PC, Lee PJ, Brickman KA et al (2005) Entanglement of trapped-ion clock states. Phys Rev A 72:062316
88. Leibfried D, DeMarco B, Meyer V et al (2003) Experimental demonstration of a robust, high-fidelity geometric two ion-qubit phase gate. Nature 422:412–415
89. McDonnell MJ, Home JP, Lucas DM et al (2007) Long-lived mesoscopic entanglement outside the Lamb-Dicke regime. Phys Rev Lett 98:063603
90. Home JP, McDonnell MJ, Lucas DM et al (2006) Deterministic entanglement and tomography of ion-spin qubits. New J Phys 8:188
91. Monroe C, Leibfried D, King BE et al (1997) Simplified quantum logic with trapped ions. Phys Rev A 55:R2489–R2491
92. DeMarco B, Ben-Kish A, Leibfried D et al (2002) Experimental demonstration of a controlled-NOT wave-packet gate. Phys Rev Lett 89:267901
93. Garcia-Ripoll JJ, Zoller P, Cirac JI (2003) Speed optimized two-qubit gates with laser coherent control techniques for ion trap quantum computing. Phys Rev Lett 91:157901
94. Duan LM, Blinov BB, Moehring DL et al (2004) Scalable trapped ion quantum computation with a probabilistic ion-photon mapping. Quantum Inform Comput 4:165–173
95. Garcia-Ripoll JJ, Zoller P, Cirac JI (2005) Quantum information processing with cold atoms and trapped ions. J Phys B – At Mol Opt Phys 38:S567–S578
96. Zhu SL, Monroe C, Duan LM (2006) Trapped ion quantum computation with transverse phonon modes. Phys Rev Lett 97:050505
97. Leibfried D, Knill E, Seidelin S et al (2005) Creation of a six-atom 'Schroedinger cat' state. Nature 438:639–642
98. Barrett MD, Chiaverini J, Schaetz T et al (2004) Deterministic quantum teleportation of atomic qubits. Nature 429:737–739
99. Riebe M, Haffner H, Roos CF et al (2004) Deterministic quantum teleportation with atoms. Nature 429:734–737
100. Chiaverini J, Leibfried D, Schaetz T et al (2004) Realization of quantum error correction. Nature 432:602–605
101. Fisk PTH, Sellars MJ, Lawn MA et al (1995) Very high-q microwave spectroscopy on trapped yb-171$^+$ ions – application as a frequency standard. IEEE Trans Instrum Meas 44:113–116
102. Meekhof DM, Monroe C, King BE et al (1996) Generation of nonclassical motional states of a trapped atom. Phys Rev Lett 76:1796–1799
103. Palma GM, Suominen KA, Ekert AK (1996) Quantum computers and dissipation. Proc R Soc London Ser A-Math Phys Eng Sci 452:567–584
104. Barenco A, Berthiaume A, Deutsch D et al (1997) Stabilization of quantum computations by symmetrization. SIAM J Comput 26:1541–1557
105. Zanardi P, Rasetti M (1997) Noiseless quantum codes. Phys Rev Lett 79:3306–3309
106. Turchette QA, Kielpinski D, King BE et al (2000) Heating of trapped ions from the quantum ground state. Phys Rev A 6106:063418
107. Deslauriers L, Olmschenk S, Stick D et al (2006) Scaling and suppression of anomalous heating in ion traps. Phys Rev Lett 97:103007
108. Labaziewicz J, Richerme P, Brown KR et al (2007) Compact, filtered diode laser system for precision spectroscopy. Opt Lett 32:572–574
109. Turchette QA, Myatt CJ, King BE et al (2000) Decoherence and decay of motional quantum states of a trapped atom coupled to engineered reservoirs. Phys Rev A 62:053807
110. Schmidt-Kaler F, Gulde S, Riebe M et al (2003) The coherence of qubits based on single Ca+ ions. J Phys B-At Mol Opt Phys 36:623–636
111. Lucas DM, Keitch BC, Home JP et al (2007) A long-lived mem-

ory qubit on a low-decoherence quantum bus. arXiv:0710.4421v1 [quant-ph]
112. Steane AM (2002). Overhead and noise threshold of fault-tolerant quantum error correction. quant-ph/0207119
113. Barenco A, Bennett CH, Cleve R et al (1995) Elementary gates for quantum computation. Phys Rev A 52:3457–3467
114. Schmidt-Kaler F, Haffner H, Gulde S et al (2003) How to realize a universal quantum gate with trapped ions. Appl Phys B-Lasers Opt 77:789–796
115. Deutsch D (1985) Quantum-theory, the church-turing principle and the universal quantum computer. Proc R Soc London Ser A-Math Phys Eng Sci 400:97–117
116. Bennett CH, Brassard G, Crepeau C et al (1993) Teleporting an unknown quantum state via dual classical and einstein-podolsky-rosen channels. Phys Rev Lett 70:1895–1899
117. Shor PW (1994) Algorithms for quantum computation – discrete logarithms and factoring. In: Goldwasser S (ed) Proceedings of the 35th Annual Symposium on Foundations of Computer Science, Santa Fe. IEEE Computer Society Press, Washington, pp 124–134
118. Chiaverini J, Britton J, Leibfried D et al (2005) Implementation of the semiclassical quantum Fourier transform in a scalable system. Science 308:997–1000
119. Grover LK (1997) Quantum mechanics helps in searching for a needle in a haystack. Phys Rev Lett 79:325–328
120. Brickman KA, Haljan PC, Lee PJ et al (2005) Implementation of Grover's quantum search algorithm in a scalable system. Phys Rev A 72:050306
121. Bennett CH, DiVincenzo DP (2000) Quantum information and computation. Nature 404:247–255
122. Leibfried D, Barrett MD, Schaetz T et al (2004) Toward Heisenberg-limited spectroscopy with multiparticle entangled states. Science 304:1476–1478
123. Roos CF, Riebe M, Haffner H et al (2004) Control and measurement of three-qubit entangled states. Science 304:1478–1480
124. Haffner H, Hansel W, Roos CF et al (2005) Scalable multiparticle entanglement of trapped ions. Nature 438:643–646
125. Monroe C (2002) Quantum information processing with atoms and photons. Nature 416:238–246
126. Blinov BB, Moehring DL, Duan LM et al (2004) Observation of entanglement between a single trapped atom and a single photon. Nature 428:153–157
127. Moehring DL, Maunz P, Olmschenk S et al (2007) Entanglement of single-atom quantum bits at a distance. Nature 449:68–71
128. Guthöhrlein GR, Keller M, Hayasaka K et al (2001) A single ion as a nanoscopic probe of an optical field. Nature 414:49–51
129. Mundt AB, Kreuter A, Becher C et al (2002) Coupling a single atomic quantum bit to a high finesse optical cavity. Phys Rev Lett 89:103001
130. Cirac JI, Zoller P, Kimble HJ et al (1997) Quantum state transfer and entanglement distribution among distant nodes in a quantum network. Phys Rev Lett 78:3221–3224
131. Keller M, Lange B, Hayasaka K et al (2004) Continuous generation of single photons with controlled waveform in an ion-trap cavity system. Nature 431:1075–1078
132. Steane AM (2007) How to build a 300 bit, 1 giga-operation quantum computer. Quantum Inform Comput 7:171–183
133. Porras D, Cirac JI (2004) Effective quantum spin systems with trapped ions. Phys Rev Lett 92:207901
134. Friedenauer A, Schmitz H, Glueckert JT et al (2008) Simulating a quantum magnet with trapped ions. Nat Phys 4:757-761
135. Leibfried D, DeMarco B, Meyer V et al (2002) Trapped-ion quantum simulator: Experimental application to nonlinear interferometers. Phys Rev Lett 89:247901
136. Schutzhold R, Uhlmann M (2007) Analogue of cosmological particle creation in an ion trap. Phys Rev Lett 99:201301
137. Schmidt PO, Rosenband T, Langer C et al (2005) Spectroscopy using quantum logic. Science 309:749–752

Books and Reviews

Wineland DJ, Monroe C, Itano WM et al (1998) Trapped-ion quantum simulator. Phys Scr T76:147–151
Ekert A, Jozsa R (1996) Quantum computation and Shor's factoring algorithm. Rev Mod Phys 68:733–753
Steane A (1997) The ion trap quantum information processor. Appl Phys B – Lasers Opt 64:623–642
Wineland DJ, Barrett M, Britton J et al (2003) Quantum information processing with trapped ions. Philos Trans R Soc Lond Ser A-Math Phys Eng Sci 361:1349–1361
Leibfried D, Blatt R, Monroe C et al (2003) Quantum dynamics of single trapped ions. Rev Mod Phys 75:281–324
Steane A (1998) Quantum computing. Rep Prog Phys 61:117–173
Wineland DJ, Monroe C, Itano WM et al (1998) Experimental issues in coherent quantum-state manipulation of trapped atomic ions. J Res Natl Inst Stand Technol 103:259–328
Blatt R, Wineland D (2008) Entangled states of trapped atomic ions. Nature 453:1008–1015
Häffner H, Roos CF, Blatt R (2008) Quantum computing with trapped ions. Phys Rep 469:155–204

Quantum Computing Using Optics

Gerard J. Milburn, Andrew G. White
Centre for Quantum Computer Technology,
The University of Queensland, Brisbane, Australia

Article Outline

Glossary

Avalanche photodiode (APD) A device for counting photons that absorbs a single photon and, with some probability, produces a single electrical signal.

Beam splitter A linear optical device that partially transmits, and partially reflects, an incoming beam of light.

Bell inequality The results of local measurements of dichotomic observables on each component of two correlated classical systems satisfy a correlation function that is less than or equal to a universal bound. This bound can be exceeded by correlated quantum systems.

Bell states Four orthogonal, maximally entangled states of two qubits that violate a Bell inequality.

Cluster State A highly entangled state of many qubits that enables quantum computation by sequences of conditional single qubit measurements, each conditioned on the results of previous measurements.

C-SIGN A two-qubit gate that leaves all logical states unchanged unless both qubits take the value one, in which case the state suffers a π phase shift.

Entangled state The state of a multi-component quantum system which cannot be expressed as a convex combination of tensor product states of each subsystem. Entangled states cannot be prepared by local operations on each subsystem, even when supplemented by classical communication.

HOM interference Hong, Ou and Mandel discovered that when indistinguishable single photon pulses arrive simultaneously at each of the two input ports of a fifty-fifty beam splitter, the probability for detecting two photons, co-incidently, at each of the two output ports drops to zero. In other words, the photons are either both reflected or both transmitted.

Mixed state A quantum state that is not completely known and thus has non zero-entropy.

Photon A single quantum excitation of an orthonormal optical mode. A field in such a state has definite intensity and completely random phase, so that the average field amplitude is zero.

Pure state A quantum state that is completely known and thus has zero entropy.

Quantum computation The ability to process information in a physical device using unitary evolution of superpositions of the physical states that encode the logical states.

Qubit The fundamental unit of quantum information in which two orthogonal states encode one bit of information. Unlike a classical bit, the physical system that forms the qubit can be in a superposition of the two logical states simultaneously.

Quantum teleportation A quantum communication protocol based on measurement, feed-forward and shared entanglement.

Shor's algorithm An algorithm for finding the prime factors of large integers by unitarily processing information in a quantum computer.

Tomography A measurement scheme for experimentally determining a quantum state in which a large sequence of physical systems, prepared in the same state, are subjected to measurements of a carefully chosen set of physical observables.

Unitary A transformation of a quantum state that is physically, and thus logically, reversible. Unitary transformations take pure states to pure quantum states.

Definition of the Subject

Quantum computation is a new approach to information processing based on physical devices that operate according to the quantum principles of superposition and unitary evolution [5,9]. This enables more efficient algorithms than are available for a computer operating according to classical principles, the most significant of which is Shor's efficient algorithm for finding the prime factors of a large integer [51]. There is no known efficient algorithm for this task on conventional computing hardware.

Optical implementations of quantum computing have largely focused on encoding quantum information using single photon states of light. For example, a single photon could be excited to one of two carefully defined orthogonal mode functions of the field with different momentum directions. However, as optical photons do not interact with each other directly, physical devices that enable one encoded bit of information to unitarily change another are hard to implement. In principle it can be done using a Kerr nonlinearity as was noted long ago [34,61], but Kerr nonlinear phase shifts are too small to be useful. Knill et al. [24] discovered another way in which the state of one photon could be made to act conditionally on the state of another using a measurement based scheme. We discuss this approach in some detail here as it has led to experiments that have already demonstrated many of the key elements required for quantum computation with optics.

Introduction

The optical fiber communication network is the largest artificial complex system on the planet, enabling the internet, the growth of economies and a massive connectivity of minds presaging a but dimly seen revolution. Information, encoded in pulses of light, courses through the system at more than 10 terabits per second. It is hard to believe that it is barely 20 years since the first transoceanic optical fiber was installed, yet the system is still growing at an astonish-

ing pace. The insatiable appetite for *bandwidth* in modern economies does not look like abating any time soon.

The entire system is held together by thin fibers of glass guiding pulses of light. The huge increase in bandwidth that modern communication networks have enabled follows directly from the very high carrier frequency of optical pulses. A wealth of modulation techniques have been developed to exploit the potential of this bandwidth: it is currently limited only by the speed of the switches in the network required to control the flow of information. Early networks required the switches to be largely electronic and this meant first converting the light pulses to electronic pulses. However increasingly these switches are being replaced with all-optical devices.

The routers and switches in the optical fiber network are essentially small computers processing packets of information at the fastest possible speeds. In the 1980s there was a research program to build computers entirely from optical switches processing information encoded optically. The astonishing progress in silicon technology meant that optical computers could never compete with the shrinking scale of semiconductors. The size of an optical switch is largely determined by the wavelength of light. This is beginning to change with the rise of plasmonics and nano photonics. However this constraint does not pose a problem for an optical communication system that spans the entire planet and much of the early work on all-optical switches found its way into the optical fiber system.

Optical fiber networks operate by the principles of classical physics; the computers they connect operate by the principles of classical logic. The semiconductors used as light sources and detectors function by quantum principles, but these do not influence network operation or logic. This is the physics and logic of Newton, Faraday and Maxwell. The motion of charges in semiconductor circuits is governed by the classical understanding of electric and magnetic fields, while the light pulses coursing down a glass fiber are perfectly well described by Maxwell's theory of electromagnetic radiation. However this situation will inevitably change and the first indications are already upon us. We have known since the early part of the last century that the deepest description of the physical world is not classical but quantum. It is now clear that the quantum world enables new computation and communication tasks that are difficult, if not impossible, in a classical world.

Over the same period that the optical communication system has been built, a small group of visionaries have speculated on the ultimate limits to conventional computation. There are two strands to this question. The first strand takes us down a road of ever decreasing dimensions. The astonishing growth in semiconductor technology is the direct result of acquiring the technical ability to make transistors smaller and smaller so that billions can fit onto a single chip. A very natural question is: how small can it get? Ultimately the answer to this question is the domain of quantum mechanics.

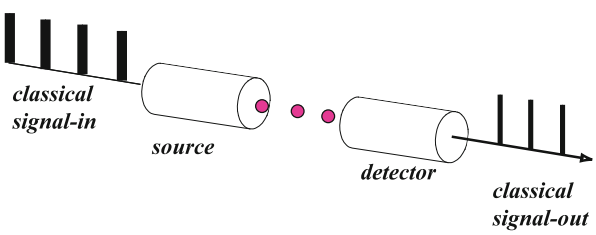
The second strand joins a more abstract path that began with von Neumann and Turing, and took a quantum twist in the mid 1980s when Feynman [9] asked if a physical computer operating by quantum principles would be a more efficient computational machine than a conventional classical computer. We now know that the answer is yes.

The quantum description of light began with Einstein in 1905 with his explanation of the photoelectric effect [7]:

> when a light ray starting from a point is propagated, the energy is not continuously distributed over an ever increasing volume, but it consists of a finite number of energy quanta, localized in space, which move without being divided and which can be absorbed or emitted only as a whole.

In Einstein's explanation, the energy of each quanta, or *photon* to use the modern word, is proportional to its frequency while the intensity of the light determines the number of photons per second passing through some given area. Einstein's insight is now routinely confirmed in a semiconductor device known as the avalanche photodiode (APD). This is a device that produces a current pulse when a single photon is absorbed. If we turn the intensity of the optical source down to very low levels and connect the APD to an audio amplifier we can hear the individual clicks as the photons arrive on the surface of the detector. In this operational sense, a single photon is a detection event, Fig. 1

The weak pulses of light used in conventional optical communication systems contain a huge number of pho-



Quantum Computing Using Optics, Figure 1
A source generates a sequence of optical pulses conditional on some classical input, e. g. electrical pulses. A detector registers a sequence of classical electrical pulses that we interpret as due to the propagation of individual photons from source to detector. We might call this the *ballistic* picture of a photon

tons, and furthermore, the number of photons per pulse is not fixed but fluctuates from pulse to pulse. This is a direct consequence of the kind of light source that is used, the *laser*.

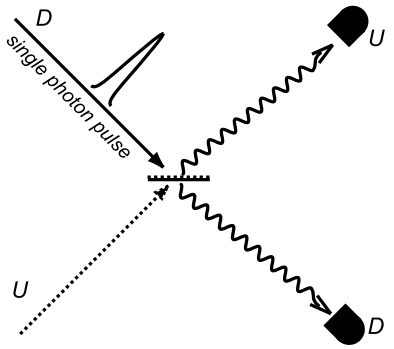
The laser is a true quantum device but it necessarily produces pulses of light with an indeterminate number of photons. The reason for this is intimately connected with the coherence properties of laser light. In a laser light source both the amplitude of the light and its phase fluctuate as little as quantum theory will permit. However these fluctuations are related by an uncertainty principle: if we try and produce a sequence of pulses with a well defined number of photons we necessarily make the phase random from pulse to pulse (we shall make this idea more precise in what follows). The random phase would destroy the key feature of first order optical coherence that makes a laser so useful.

Are there any light sources that controllably produce pulses of light each with one and only one photon per pulse? Until very recently, the answer was no. However the discovery of optical communication schemes (quantum key distribution) and computation schemes (as described below) based on quantum principles have given a strong incentive for building such sources. We will now explain how single photons enable quantum computation and postpone our discussion of exactly what a single photon source is to later.

Quantum Computation with Single Photons

Let us now consider the experiment depicted in Fig. 2. A source (labeled D for downward) is producing a sequence of single photons which are directed towards a 50/50 beam splitter. This is a classical optical device which in the wave theory would be described as simply dividing the wave intensity equally between a reflected beam and a transmitted beam. If we had a source of very high intensity, each of the photodetectors in the reflected path and the transmitted path would record on average an equal number of counts per second. As we reduce the intensity to the single photon level however, we see an irreducible uncertainty in which detector will register the photon in each trial. On average it is still the case that over many trials, one half are recorded at the D-detector and one half at the U-detector. If we were to persist in our *ballistic* view of a single photon, we might be tempted to say that each photon is either reflected or transmitted at random; it is a simple coin toss. This picture would adequately capture the experimental facts, *for this experiment*.

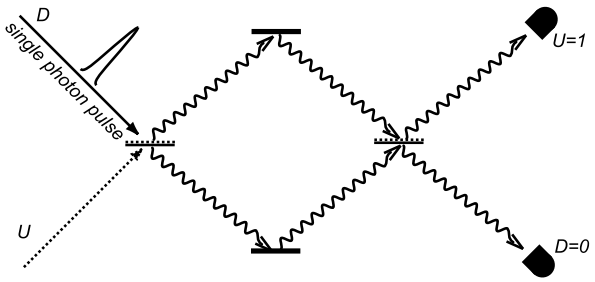
The beam splitter experiment has a binary outcome: the photon is either detected at U or D. To encode the result at the output we need a single bit of information. Indeed if we allow another single photon source into the upward going input to the beam splitter, we will need a single binary number to encode the input state as well. The kind of encoding we have just described is called *dual rail*. In more precise terms we have encoded a single bit into one of two perfectly distinguishable momentum states of a single photon.



Quantum Computing Using Optics, Figure 2
A single photon pulse incident on a beam splitter in the downward (D) direction. It can be detected in the upward going direction (U) or the downward going direction (D)

At first sight, it would appear that the experiment we have just described is a fanciful coin toss. That this is not the case can be seen if we 'toss the coin again', that is to say, we take the photon after the beam splitter and, instead of running it into a photodetector, we reflect it back onto an identical beam splitter and then ask for the probability for it to be reflected or transmitted (see Fig. 3). It is well known that in this case we can adjust things so that the probability for the photon to be detected at U is certain. Irreducible uncertainty has been turned into certainty simply by changing the experimental conditions for detecting the photon, not by changing the state of the single photon source.

Note that the price we pay for certainty is complete loss of knowledge of what happened to the photon at the left beam splitter. A photon can be detected at either detector in Fig. 3 in two indistinguishable ways corresponding to the two (unknowable) outcomes of reflection or transmission at the left beam splitter. Returning to the simple experiment in Fig. 1, can we say that it is a simple coin toss before the detectors register the photon? Does the output of the beam splitter really encode one bit of information? No. We capture the dual quantum/classical nature of the state of a single photon after a beam splitter by saying that it is a *superposition* of the two mutually exclusive alternatives. In such a case the output is said to encode a single



Quantum Computing Using Optics, Figure 3
A single photon pulse incident on a beam splitter in the downward (D) direction. It can be detected in the upward going direction (U) or the downward going direction (D)

quantum bit, or *qubit*. Quantum computation and communication runs on qubits not bits.

The quantum description of the state of the photon after a single beam splitter requires us to give a list of the *probability amplitudes* for the two mutually exclusive possibilities for detection in either the downward or the upward direction. The detection probabilities are then given by the square of the probability amplitudes. For example the input state in Fig. 1 is certain to be in the downward direction, so the ordered pair of probability amplitudes $(1, 0)$ in which the first (second) entry corresponds to detection in the D (U) direction. After a general (not 50/50) beam splitter the state of the photon is depicted by the ordered pair (r, t) where $|r|^2 + |t|^2 = 1$. So in fact the beam splitter has enacted the linear transformation $(1, 0) \rightarrow (r, t)$.

One might think that the input state $(0, 1)$ is transformed in exactly the same way. However these two input states are physically distinguishable: in one case the photon is going down and in the other it is going up with certainty. Mathematically this is captured by the fact that the input vectors are orthogonal. The beam splitter is a perfectly reversible device and does not destroy information by making two distinguishable alternatives indistinguishable. The problem is avoided when we note that a full quantum theory of beam splitters will in fact show that $(0, 1) \rightarrow (-r, t)$. There is a central lesson here: operations on qubits must not take two physically distinguishable states and make them indistinguishable. We call this sort of transformation *unitary*.

We now make a change to more conventional notation. The input states $(1, 0)$ and $(0, 1)$ are written as $|1, 0\rangle$ and $|0, 1\rangle$ (ordering is preserved) so that the beam splitter transformations are then written as

$$|1, 0\rangle \rightarrow r|1, 0\rangle + t|0, 1\rangle \tag{1}$$

$$|0, 1\rangle \rightarrow -r|1, 0\rangle + t|0, 1\rangle \, . \tag{2}$$

We now formalize the concept of a qubit by making a distinction between the *logical* state of a qubit and the physical states of the system used to encode it. In the example used here the relation ship is

$$|0\rangle = |1, 0\rangle \tag{3}$$

$$|1\rangle = |0, 1\rangle \, . \tag{4}$$

We then say that we have a qubit code that uses one photon and two optical modes per qubit. The beam splitter transformation on the logical code acts like

$$|0\rangle \rightarrow r|0\rangle + t|1\rangle \tag{5}$$

$$|1\rangle \rightarrow -r|0\rangle + t|1\rangle \, . \tag{6}$$

This transformation on logical qubits is called a *one-qubit gate*.

If we want to encode two qubits in this dual rail scheme we will need two single photon pulses and four modes, which may be achieved by two independent beam splitters. In a notation that keeps track of the ordering of the beam splitters by an ordering of states from left to right, the logical state $|1, 0\rangle$ would then be represented physically by

$$|1\rangle \otimes |0\rangle = |0, 1\rangle \otimes |1, 0\rangle \, . \tag{7}$$

Note that the number of mutually distinguishable output states from N beam splitters (and N photons) increases exponentially as 2^N. This simply means that for the logical code of N qubits there are 2^N possible logical states.

If we continue in this fashion we are not going to have a very compact notation. We can make things easier by using the fact that our qubits are ordered corresponding to a physical order of the underlying beam splitters. We can then represent a state, for example $|1\rangle \otimes |0\rangle \otimes |1\rangle \otimes |0\rangle$, by regarding it as the binary code for an integer, in this case $|10\rangle$ as $10 = 1 \times 2^3 + 0 \times 2^2 + 1 \times 2^1 + 0 \times 2^0$.

It is then clear that with N physical resources we can code 2^N binary numbers (or their corresponding integers). If we could somehow act on all these numbers simultaneously we might have a very efficient processor. This is precisely the hope of quantum computation. For example, if we pass N photons downward through N independent beam splitters, the state at the output is superposition state

$$|0\rangle \rightarrow 2^{-N/2} \sum_{x=0}^{2^N-1} |x\rangle \, , \tag{8}$$

an equal superposition of all integers from zero up to $2^N - 1$. If we could somehow continue to postpone mea-

surement and act unitarily on all these numbers simultaneously, we might gain a considerable advantage in computational efficiency. This was the idea first captured in precise form by David Deutsch [5].

Suppose for example we set aside some number, say N, of photons to encode the input register and some number, say K, to encode an output register. We first prepare the N photons so as to encode the state produced in Eq. (8) and the K to encode the state $|0\rangle$. Then the unitary processing through some, as yet unspecified, physical device could implement the transformation

$$\sum_x |x\rangle|0\rangle \rightarrow \sum_x |x\rangle|f(x)\rangle \,, \tag{9}$$

for some function f (we drop the normalization for simplicity). Of course now when we read out the output register we only get one value of the function with a random distribution. This may not sound very useful if you are in fact interested in a particular value of the function, but in that case why do all computations of the function at once? The primary reason why you might want to evaluate a function on all inputs is when you are not so much interested in any particular value but rather some global property of the function itself – for example is it constant? It is precisely that kind of computation that Deutsch [3] showed could be done more efficiently on a quantum computer.

A far more interesting example was described by Shor [51] who gave an efficient quantum algorithm for finding the period of a modular function. The particular modular function he used is part of an algorithm for finding the prime factors of large integers, so that in effect he gave an efficient quantum algorithm for finding such prime factors. The assumed intractability of this problem for a classical computer is why it is used as a method for public-key encryption. If someone had a quantum computer all these crypto systems would be vulnerable to attack. We describe below a simple optical experiment which captures the kind of methods by which a quantum computer could implement Shor's algorithm.

Of course no one yet has a quantum computer capable of posing a threat to public-key crypto systems, although it begins to look like it might be possible. The problem is to figure out what kind of physical systems would enable the arrow in Eq. (9). Various suggestions have been made including single photon optical systems, which we discuss below. At the level of logical qubits we need two types of elementary transformations. The first we have already seen: It is the single qubit transformation in Eq. (5). In addition all we need is some form of interaction that correlates the state of at least two qubits. One example is the controlled-SIGN or c-SIGN, gate,

$$|x\rangle|y\rangle \rightarrow (-1)^{x\cdot y}|x\rangle|y\rangle \,. \tag{10}$$

For single photon, dual rail encoding this presents a major problem.

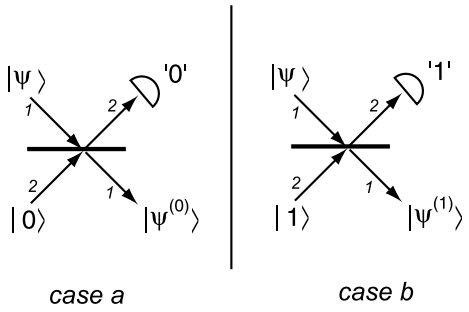
If we look at the c-SIGN gate at the level of physical qubits it will require some kind of intensity dependent phase change so that only a two photon component acquires a π phase change. This is possible with a Kerr nonlinearity as was noted long ago [34,61]. Nonlinear optical materials with an intensity dependent refractive index do indeed exist and are generically referred to as Kerr nonlinear materials. The problem is that Kerr nonlinear phase shifts are so small that to expect an intensity as low as two photons to give a π phase change in an optical material of practical size is to expect too much. Fortunately there is another way.

Conditional Optical Two-Qubit Gates

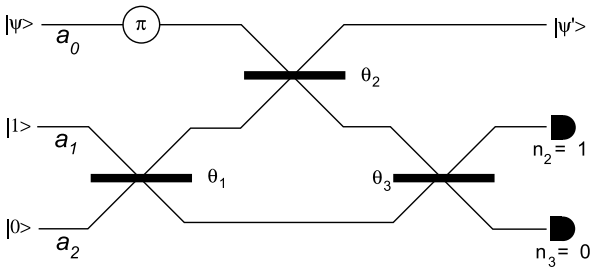
Measurements play quite a different role in quantum mechanics than they do in classical mechanics. In the latter, it is the objective of accurate measurement to reveal the preexisting values of dynamical variables. Of course in reality both the measurement and the preparation of the initial state are subject to noise. However it is assumed that both may be sufficiently refined to reveal the true dynamical state of a system. If we know all there is to know about a classical system, all measurements are dispersion free in principle.

Quantum mechanics is an irreducibly statistical theory. This means that even if a system has been prepared in state about which we have maximal knowledge, there will be at least one measurement the results of which are uncertain. Fortunately there is also at least one physical quantity that may be measured for which the results are certain. If we make a perfect measurement of this quantity we have in effect prepared the system in a pure state of which we have maximal knowledge. Measurement thus plays an essential role in state preparation, where knowledge of the prepared state is heralded by the measurement result.

Can we use the unique role that measurement plays in quantum mechanics to conditionally implement a required state transformation? Consider the situation shown in Fig. 4. In a dual rail scenario, two modes are mixed on a beam splitter. One mode is assumed to be in the vacuum state (a) or a one photon state (b), while the other mode is arbitrary. A single photon counter is placed in the output port of mode-2. What is the conditional state of mode-1 given a count of n photons?



Quantum Computing Using Optics, Figure 4
A conditional state transformation conditioned on photon counting measurements



Quantum Computing Using Optics, Figure 5
A conditional state transformation on three optical modes, conditioned on photon counting measurements on the ancilla modes a_1, a_2. The signal mode, a_0 is subjected to a π phase shift

The conditional state of mode a_1 is given by (unnormalized),

$$|\tilde{\psi}^{(i)}\rangle_1 = \hat{\Upsilon}(i)|\psi\rangle_1 \,, \tag{11}$$

where

$$\hat{\Upsilon}(i) = {}_2\langle i|U(\theta)|i\rangle_2 \tag{12}$$

with $i = 1, 0$. Here $U(\theta)$ is the unitary transformation implemented by the beam splitter as we described in the introduction. The probability for the specified photo count event is given by,

$$P(i) = {}_1\langle\psi|\Upsilon^\dagger(i)\hat{\Upsilon}(i)|\psi\rangle_1 \,, \tag{13}$$

which fixes the normalization of the state,

$$|\psi^{(i)}\rangle_1 = \frac{1}{\sqrt{P(i)}}|\tilde{\psi}^{(i)}\rangle_1 \,. \tag{14}$$

It is then possible to show that

$$\hat{\Upsilon}(0) = \sum_{n=0}^{\infty} \frac{(\cos\theta - 1)^n}{n!}(a_1^\dagger)^n a_1^n$$

$$\hat{\Upsilon}(1) = \cos\hat{\Upsilon}^{(0)} - \sin^2\theta\, a_1^\dagger \hat{\Upsilon}^{(0)} a_1 \,.$$

In order to see how we can use these kind of transformations to effect a c-SIGN gate, consider the situation shown in Fig. 5. Three optical modes are mixed on a sequence of three beam splitters with beam splitter parameters θ_i, with corresponding reflection and transmission amplitudes, $r_i = \cos\theta_i$, $t_i = \sin\theta_i$. The *ancilla* modes, a_1, a_2 are restricted to be in the single photon states $|1\rangle_2, |0\rangle_3$ respectively.

We will assume that the *signal* mode, a_0, is restricted to have *at most* two photons, thus

$$|\psi\rangle = \alpha|0\rangle_0 + \beta|1\rangle_0 + \gamma|2\rangle_0 \,. \tag{15}$$

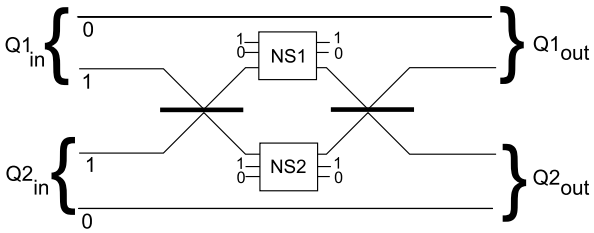
The reason we are only interested in two photon states is that in dual rail encoding a general two qubit state can have at most two photons. We can now chose the beam splitter parameters so that, conditional on the detectors both registering a one, the signal state is transformed as

$$|\psi\rangle \to |\psi'\rangle = \alpha|0\rangle + \beta|1\rangle - \gamma|2\rangle \,, \tag{16}$$

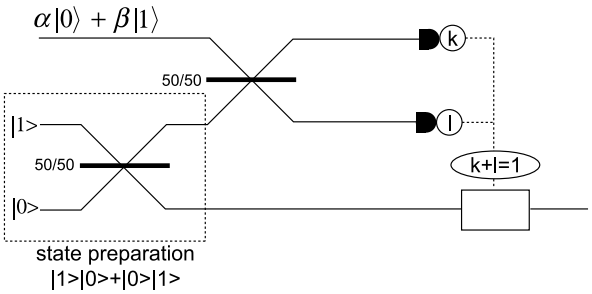
with a probability that is *independent* of the input state $|\psi\rangle$. This last condition is essential as in a quantum computation, the input state to a general two qubit gate is completely unknown. We will call this transformation the NS (for nonlinear sign shift) gate. This can be achieved using [24]: $\theta_1 = -\theta_3 = 22.5\,\text{deg}$ and $\theta_2 = 65.53\,\text{deg}$. The probability of the conditioning event $(n_2 = 1, n_3 = 0)$ is 1/4. Note that we can't be sure in a given trial if the correct transformation will be implemented. Such a gate is called a *nondeterministic* gate. However the key point is that success is heralded by the results on the photon counters (assuming ideal operation).

We can now proceed to a c-SIGN gate in the dual rail basis. Consider the situation depicted in Fig. 6. We first take two dual rail qubits encoding for $|1\rangle|1\rangle$. The single photon components of each qubit are directed towards a 50/50 beam splitter where they overlap perfectly in space and time and produces a state of the form $|0\rangle_2|2\rangle_3 + |2\rangle_2|0\rangle_3$, a effect known as Hong–Ou–Mandel (HOM) interference [15]. We then insert an NS gate into each output arm of the HOM interference. When the conditional gates in each arm work, which occurs with probability 1/16, the state is multiplied by an overall minus sign. Finally we direct these modes towards another HOM interference. The output state is thus seen to be $-|1\rangle|1\rangle$. One easily checks the three other cases for the input logical states to see that this device implements the c-SIGN gate with a probability of 1/16 and successful operation is heralded.

Clearly a sequence of nondeterministic gates is not going to be much use: the probability of success after



Quantum Computing Using Optics, Figure 6
A conditional state transformation conditioned on photon counting measurements. A c-SIGN gate that works with probability of 1/16. It uses HOM interference and two NS gates



Quantum Computing Using Optics, Figure 7
A partial teleportation system for single photons states using a linear optics

a few steps will be exponentially small. The key idea in using nondeterministic gates for quantum computation is based on the idea of gate teleportation of Gottesmann and Chuang [12]. In quantum teleportation an unknown quantum state can be transferred from A to B provided A and B first share an entangled state. Gottesmann and Chuang realized that it is possible to simultaneously teleport a two qubit quantum state and implement a two qubit gate in the process by first applying the gate to the entangled state that A and B share prior to teleportation.

We use a non deterministic NS gate to prepare the required entangled state, and only complete the teleportation when the this stage is known to work. The teleportation step itself is non deterministic but, as we see below, by using the appropriate entangled resource the teleportation step can be made near deterministic. The near deterministic teleportation protocol requires only photon counting and fast feed-forward. We do not need to make measurements in a Bell basis.

A nondeterministic teleportation measurement is shown in Fig. 7. The client state is a one photon state in mode-0 $\alpha|0\rangle_0 + \beta|1\rangle_0$ and we prepare the entangled ancilla state

$$|t_1\rangle = |01\rangle_{12} + |10\rangle_{12} \tag{17}$$

where mode-1 is held by the sender, A, and mode-2 is held by the receiver, B. For simplicity we omit normalization constants wherever possible. This an ancilla state is easily generated from $|01\rangle_{12}$ by means of a beam splitter.

If the total count is $n_0 + n_1 = 0$ or $n_0 + n_1 = 2$, an effective measurement has been made on the client state and the teleportation has failed. However if $n_0 + n_1 = 1$, which occurs with probability 0.5, teleportation succeeds with the two possible conditional states being

$$\alpha|0\rangle_2 + \beta|1\rangle_2 \quad \text{if } n_0 = 1, n_1 = 0 \tag{18}$$

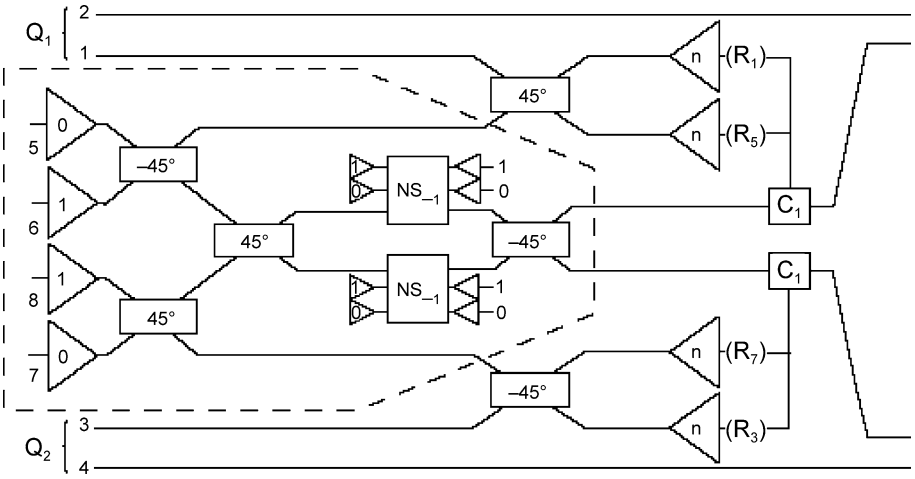
$$\alpha|0\rangle_2 - \beta|1\rangle_2 \quad \text{if } n_0 = 0, n_1 = 1\,. \tag{19}$$

When successful, this procedure implements a joint measurement on modes 0 and 1. In the conventional deterministic teleportation protocol the joint measurement is a simultaneous measurement of the commuting operators XX and ZZ where X, Z are respectively the Pauli-x and Pauli-z operators. This is a Bell measurement. In the teleportation protocol considered here, we only have a partial Bell measurement. We will refer to it as a non-deterministic teleportation protocol, $\mathbf{T}_{1/2}$. Note that teleportation failure is detected and corresponds to a photon number measurement of the state of the client qubit. Detected number measurements are a very special kind of error and can be easily corrected by a suitable error correction protocol. For further details see [24] and the review [26].

The next step is to use $\mathbf{T}_{1/2}$ to effect a conditional sign flip c-SIGN$_{1/4}$ which succeeds with probability 1/4. Note that to implement c-SIGN on two bosonic qubits in modes 1, 2 and 3, 4 respectively, we can first teleport the first modes of each qubit to two new modes (labeled 6 and 8) and then apply c-SIGN to the new modes. When using $\mathbf{T}_{1/2}$, we may need to apply a sign correction. Since this commutes with c-SIGN, there is nothing preventing us from applying c-SIGN to the prepared state before performing the measurements. The implementation is shown in Fig. 8 and now consists of first trying to prepare two copies of $|t_1\rangle$ with c-SIGN already applied, and then performing two partial Bell measurements. Given the prepared state, the probability of success is $(1/2)^2$. The state can be prepared using c-SIGN$_{1/16}$, which means that the preparation has to be retried an average of 16 times before it is possible to proceed.

The probability of successful teleportation can be boosted to $1 - 1/(n+1)$ using more entangled resource state of the kind

$$|t_n\rangle_{1\ldots 2n} = \sum_{j=0}^{n} |1\rangle^j |0\rangle^{n-j} |0\rangle^j |1\rangle^{n-j}\,. \tag{20}$$



Quantum Computing Using Optics, Figure 8

A c-SIGN two qubit gate with teleportation to increase success probability to 1/4. When using the basic teleportation protocol (T1), we may need to apply a sign correction. Since this commutes with c-SIGN, it is possible to apply c-SIGN to the prepared state before performing the measurements, reducing the implementation of c-SIGN to a state-preparation (outlined) and two teleportations. The two teleportation measurements each succeed with probability 1/2, giving a net success probability of 1/4. The correction operations C1 consist of applying the phase shifter when required by the measurement outcomes

The notation $|a\rangle^j$ means $|a\rangle|a\rangle \ldots$, j times. The modes are labeled from 1 to $2n$, left to right. Note that the state exists in the space of n bosonic qubits, where the kth qubit is encoded in modes $n+k$ and k (in this order).

We can teleport the state $\alpha|0\rangle_0 + \alpha|1\rangle_0$ using $|t_n\rangle_{1\ldots 2n}$. We first couple the client mode to half of the ancilla modes by applying an $n+1$ point Fourier transform on modes 0 to n. This is defined by the mode transformation

$$a_k \to \frac{1}{\sqrt{n+1}} \sum_{l=0}^{n} \omega^{kl} a_l \,, \tag{21}$$

where $\omega = e^{i2\pi/(n+1)}$. This transformation does not change the total photon number and is implementable with passive linear optics. After applying the Fourier transform, we measure the number of photons in each of the modes 0 to n. If the measurement detects k bosons altogether, it is possible to show [24] that if $0 < k < n+1$, then the teleported state appears in mode $n+k$ and only needs to be corrected by applying a phase shift. The modes $2n-l$ are in state 1 for $0 \le l < (n-k)$ and can be reused in future preparations requiring single bosons. The modes are in state 0 for $n-k < l < n$. If $k = 0$ or $k = n+1$ an effective measurement of the client is made, and the teleportation fails. The probability of these two events is $1/(n+1)$, regardless of the input. Note that again failure is detected and corresponds to measurements in the basis $|0\rangle, |1\rangle$ with the outcome known. Note that both the necessary correction and the receiving mode are unknown until after the measurement.

Cluster State Methods

About the same time it was realized that measurements would provide a path to optical single photon computing, Raussendorf and Briegel [48] gave an independent and remarkably novel method by which measurement alone could be used to do quantum information processing. In their approach to quantum computation, an array of qubits is initially prepared in a special entangled state called a cluster state. The computation then proceeds by making a sequence of single qubit measurements. Each measurement is made in a basis that depends on prior measurement outcomes; in other words, the results of past measurements are fed forward to determine the basis for future measurements. Subsequently Popescu showed that the linear optical measurement based scheme of Knill et al. can be interpreted as a Raussendorf and Briegel measurement based quantum computation [44].

Nielsen [37] realized that the LOQC model of [24] could be used to efficiently assemble the cluster using the nondeterministic teleportation t_n. As we saw the failure mode of this gate constituted an accidental measurement of the qubit in the computational basis. The key point is that such an error does not destroy the entire assembled cluster but merely detaches one qubit from the cluster. This enables a protocol to be devised that produces a clus-

ter that grows on average. The LOQC cluster state method dramatically reduces the number of optical elements required to implement the original LOQC scheme. Of course if large single photon Kerr nonlinearities were available, the optical cluster state method could be made deterministic [16].

In its simplest form the Raussendorf and Breigel scheme begins with a two dimensional array of qubits. Each qubit is prepared in a superposition of the computational basis states, $|0\rangle + |1\rangle$. Then an entangling operation is performed between nearest neighbor qubits in the lattice using the two-qubit controlled sign operation

$$|x\rangle|y\rangle \mapsto (-1)^{xy}|x\rangle|y\rangle \,. \tag{22}$$

In the next step one or more qubits are measured in a particular basis and depending on the results of that measurement another basis is chosen for subsequent qubit measurements. Any circuit model of a quantum algorithm can be mapped onto the two dimensional lattice through a sequence of conditional measurements on subsets of qubits.

The key difficulty in doing this with a dual rail optical scheme is that, as we have noted, the transformation in Eq. (22) is very difficult to implement using a deterministic unitary transformation as optical nonlinearities are too small. However a conditional scheme of the kind discussed above might work by conditionally entangling sequences of qubits, provided failure at any point did not destroy the entire developing lattice of entanglement. This is indeed the case because of a key feature of the LOQC model of [24] scheme. If a teleportation gate failure is heralded, it corresponds to an effective measurement of one of the qubits at input. This feature was used by Knill et al. to establish the scalability of the scheme as detected measurements errors can easily be protected by a suitable code.

The importance of this feature for a conditional cluster state assembly is that an accidental measurement of one of the qubits in a developing cluster simply detaches it from the cluster without destroying the remaining entanglement. The picture then emerges of a kind of random cluster assembly in which the cluster grows when a gate succeeds and gets pruned if a gate fails. As long as the probability for gate success is greater than 0.5 there is a chance that the cluster overall will grow. As this does not require a very large teleportation resource, $|t_n\rangle_{1\ldots 2n}$, it can be achieved with quite modest overheads of linear optics and single photons. For this reason optical cluster state methods are preferred over the original LOQC scheme of Knill et al. A number of schemes have been proposed to efficiently assemble a cluster via this probabilistic growth. A recent application of percolation theory is a good example of the kind of optimization that is required [20].

Experimental Implementations

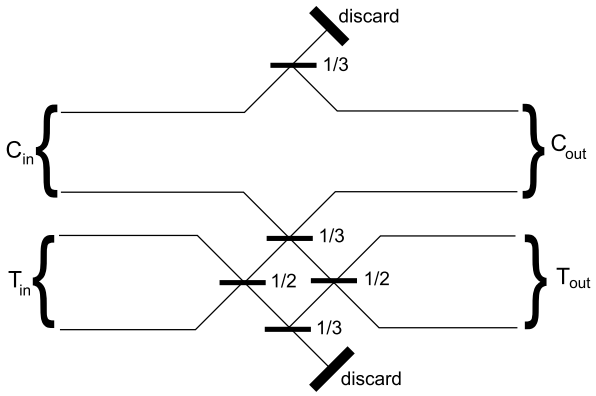
Two Qubit Gates

Considerable progress has been made on demonstrating conditional two-qubit gates with single photon states. After the initial proposal was made by Knill, Laflamme and Milburn [24], there was a flurry of theoretical work to propose linear-optical two-qubit gates that would be easier to realize experimentally. These can be divided into two categories, *internal-ancilla* gates [22,41,42,46], where the ancilla photons are intrinsic to the operation of the logic gates, and *external-ancilla* gates [14,47], where ancilla photons can be used to verify correct gate operation by performing a quantum non-demolition measurement [25] on the gate outputs. There are a variety of configurations of internal-ancilla gates, including simplified [46] and efficient [22] versions of the KLM gate; and gates that gain in efficiency using entangled ancilla photons [41,42].

An entangled internal-ancilla gate was soon realized in Johns Hopkins, where entangling operation was suggested by a $61.5 \pm 7.4\%$ visibility fringe [43]. An unambiguous demonstration of entangling gate operation was performed at the University of Queensland with an external-ancilla gate [39]: all four entangled Bell states were produced as a function of only the logical values of the input qubits, for a single operating condition of the gate. Both these gates filtered on photon-number, i. e. they required the four or two input photons to be detected to signal successful gate operation. An important requirement for LOQC is that it must be possible to detect successful gate operation by measurement of the ancilla photons and then feed-forward this information to the logic photons: this was realized with an external-ancilla gate at the University of Vienna [10].

Linear-optic gates, both internal and external ancilla, have achieved a wide range of firsts and proof-of-principle demonstrations: the first full characterization of a quantum-logic gate, in *any* architecture [38]; production of cluster [58] and graph [31] states, and their use for Grover's algorithm [58]; production of the highest entanglement [27] and fastest gate operation [45] of any physical architecture; and the first demonstration of Shor's algorithm exhibiting the core processes, coherent control, and resultant entangled states required in a full-scale implementation [29,30].

Measuring gate operation is non-trivial. Further, it is desirable to diagnose, and if possible correct, error behaviors introduced by a real gate, such as phase or bit-flip errors, which can induce the wrong amount or type of entanglement, or decoherence. To date, there have been



Quantum Computing Using Optics, Figure 9
An external-ancilla CNOT gate. When the control is in the logical-one state, the control and target photons interfere non-classically at the central 1/3 beam splitter which causes a π phase-shift in the upper arm of the central interferometer and the target state qubit is flipped. The qubit value of the control is unchanged. Correct operation has probability 1/9 and occurs when a single photon is detected in each output; this can be done with quantum non-demolition (QND) measurements using external ancilla photons [25] or by filtering on photon number in the final detection. Experimentally, the two modes of each qubit are distinguished by orthogonal polarizations. These may be split into separate spatial modes and sent through normal beamsplitters [38,39] or left in one spatial mode and sent through partially-polarizing beamsplitters [21,27,40]

a wide variety of measures used to gauge the quality of two-qubit gates. A comprehensive comparison of the various measures, and an architecture-independent measurement standard for two-qubit gates, is given in [60]. The key is quantum process tomography, which allows reconstruction of the quantum state transfer function of the gate. Tomography requires sampling the statistics of a fixed number of measurement outcomes, at least 256 for a two-qubit gate. With these statistics data inversion can be devised to reconstruct the gate process. From this we can calculate its overlap, or fidelity, with respect to the ideal quantum-logic gate, e. g. in recent experiments at the University of Queensland we obtained a process fidelity of $F_p = 98.2 \pm 0.3\%$ for a controlled-$\pi/4$ gate [28]. The process fidelity is itself not a metric [11], but forms the basis of several, such as the average gate fidelity, i. e. the fidelity of every gate output state with the ideal, averaged over every possible input state. This is given simply by,

$$\overline{F} = \frac{dF_p + 1}{d + 1}\,, \tag{23}$$

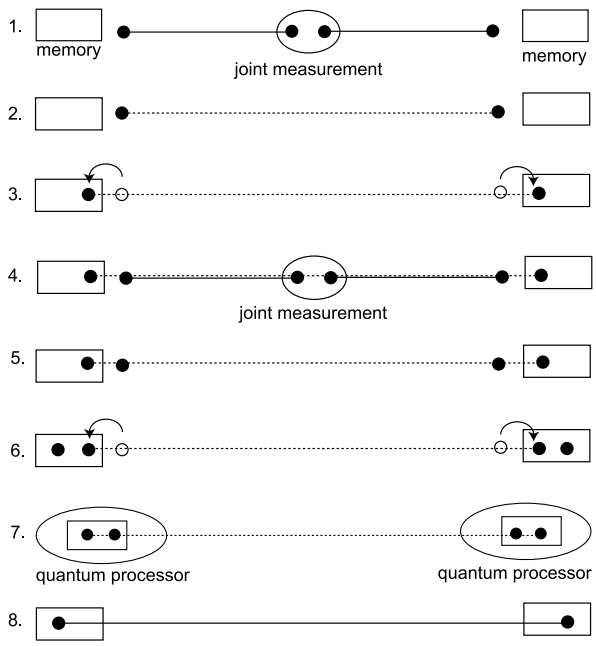
where d is the dimension of the process matrix, e. g. $d = 16$ for two-qubit gates.

These measures are useful for comparing gate performances, but do not provide an error probability-per-gate to allow direct comparison with fault-tolerance thresholds. A recent publication introduces a semidefinite programming technique to do exactly this, using the measured process matrix [59]. Such fault-tolerant benchmarking identifies the magnitude, but not the source of errors – critical in identifying the critical technology path for improving gate operation. This requires a comprehensive theoretical model of the quantum-logic gate, and its errors. Using such a model, Ref. [59] identifies small amounts of multi-photon emission as the dominant, yet previously unrecognized, source of error in linear-optical quantum computing – higher-order terms of 0.3—0.8% lead to gate errors of ∼16%! Removing multi-photon emission puts photonic quantum computing within striking distance of a recently predicted fault-tolerance threshold [23].

Quantum Optical Algorithms

The simplest useful algorithm for quantum optical information processing is the *quantum repeater*. This is more conventionally referred to as quantum communication protocol rather than an algorithm, but realistic systems will require few qubit quantum computers to do entanglement distillation and purification giving it an algorithmic flavor. Furthermore quantum repeaters could enable distributed quantum computation. We also mention it here as the various quantum optical repeater protocols that have been suggested [6,50,52] will have immediate application in long distance quantum key distribution and thus are pivotal to developing quantum optical networks across the global optical fiber network. This raises many new questions in the area of communication complexity and a great deal of work remains to be done to understand such systems [18,55].

The general idea of a quantum repeater is sketched in Fig. 10. There are three key elements at each node of the network: (i) a pair of entangled qubits, (ii) quantum memories, (iii) a few qubit quantum computer for purification distillation. In the first step two entangled pairs are produced at the origin, one of each pair is held there, while the other member of each pair is sent to distant locations. A measurement is made on the two qubits kept at the origin at step 1, leaving partially entangled qubits at the remote locations, step 2. The remote qubits are then loaded into quantum memory, step 3. The process is repeated, steps 4 and 5 until two qubits are held at the remote locations. A purification algorithm is then run at each remote location so that finally a maximally entangled pair is held at the remote locations separated by twice the distance that



Quantum Computing Using Optics, Figure 10
A simple schematic for a quantum repeater

separated the pairs in step 1. In the quantum optical realization both the generation of entanglement and purification can be done by heralded non deterministic processes.

The previous scheme of course requires a good quantum memory and a great deal of research is underway to develop such systems for photons. One promising approach is to use polarized atomic ensembles [35]. The quantum memory itself may need to have error correction algorithms running to maintain the coherence of the entangled pairs until they required for some future quantum information processing task.

Many current cryptographic protocols rely on the computational difficulty of finding the prime factors of a large number: a small increase in the size of the number leads to an exponential increase in computational resources. Shor's quantum algorithm for factoring composite numbers faces no such limitation [51], and its realization represents a major challenge in quantum computation. Only one step of Shor's algorithm to find the factors of a number N requires a quantum routine. Given a randomly chosen co-prime C (where $1 < C < N$ and the greatest common divisor of C and N is 1), the quantum routine finds the *order* of C modulo N, defined to be the minimum integer r that satisfies $C^r \mod N = 1$. It is straightforward to find the factors from the order. Consider $N = 15$: if we choose $C = 2$, the quantum routine finds $r = 4$, and the prime factors are given by the non-trivial greatest common divisor of $C^{r/2} \pm 1$ and N, i. e. 3 and 5; similarly if we choose the next possible co-prime, $C = 4$, we find the order $r = 2$, yielding the same factors.

The quantum routine in Shor's algorithm can factor a k-bit number using $72k^3$ elementary quantum gates, e. g. factoring the smallest meaningful number, 15, requires 4608 gates operating on 21 qubits [1], where the gates are one-, two-, or three-qubit logic gates. Recognizing this is well beyond the reach of current technology, Ref. [1] introduced a compiling technique which exploits properties of the number to be factored, allowing exploration of Shor's algorithm with a vastly reduced number of resources. The compiled algorithms are not scalable in themselves, but do allow the characterization of core processes required in a full-scale implementation of Shor's algorithm – including the ability to generate entanglement between qubits by coherent application of a series of quantum gates. In the only previous demonstration of Shor's algorithm, a compiled set of gate operations were implemented in a liquid NMR architecture [56]. Unfortunately, in such a system the qubits are at all times in a highly mixed state [2], and the dynamics can be fully modeled classically [32], so that neither the entanglement nor the coherent control at the core of Shor's algorithm can be implemented or verified.

The quantum order-finding routine consists of three distinct steps: i) *register initialization*, $|0\rangle^{\otimes n}|0\rangle^{\otimes m} \rightarrow (|0\rangle + |1\rangle)^{\otimes n}|0\rangle^{\otimes m-1}|1\rangle = \sum_{x=0}^{2^n-1}|x\rangle|0\rangle^{\otimes m-1}|1\rangle$, where the argument-register is prepared in an equal coherent superposition of all possible arguments (normalization omitted by convention); ii) *modular exponentiation*, which by controlled application of the order-finding function produces the entangled state $\sum_{x=0}^{2^n-1}|x\rangle|C^x \mod N\rangle$; iii) the *inverse Quantum Fourier Transform* (QFT) followed by measurement of the argument-register in the logical basis, which with high probability extracts the order r after further classical processing. If the routine is standalone, the inverse QFT can be performed without two-qubit gates using an approach based on local measurement and feed-forward.

In recent experiments, Shor's algorithm was performed for $N = 15$ using internal-ancilla logic gates [29, 30]. Co-primes of $C = 4$ and $C = 2$ were realized using circuits of two controlled-SIGN gates, with three and four qubit inputs, respectively. To no-ones surprise, the prime factors of 3 and 5 were found. What was surprising is that was near-ideal algorithm performance – far better than expected given the known errors inherent in the logic gates. This result highlights a subtle point with respect to benchmarking quantum algorithms: *algorithm performance* is not necessarily an accurate indicator of

the underlying *circuit performance*. This is particularly the case in Shor's algorithm, where ideally the algorithm produces mixed states. From algorithm performance alone it is not possible to distinguish between the desired mixture arising from entanglement with the function-register, and the undesired mixture due to environmental decoherence. In [29] this was quantified by using quantum state tomography [17] to measure the *joint* state of the argument and function registers after modular exponentiation. The joint state was both mixed and entangled, only partially overlapping with the expected states, a sure indication of environmental decoherence. Such accurate knowledge of circuit performance is crucial if the circuits are to be incorporated as sub-routines in larger algorithms.

Reprise: Single Photon States

The two key requirements for conditional quantum computing with single photons are; (i) single photon sources and, (ii) highly efficient single photon detectors capable of resolving zero, one or two photon detection events. Of course we will need a lot more than just this. In order to make the scheme scalable some kind of photonic memory, or fast switching, or both, would be desirable. Once a particular entangled multi photon state resource has been prepared it might be necessary to store it for some time until it is required for use. We first discuss in some detail what is required for a single photon source.

It is now time to become much more precise about what is meant by a single photon state. The quantum electromagnetic field is described by an electric field operator [57],

$$\vec{E}(\vec{x},t) = i\sum_{n,\nu}\sqrt{\frac{\hbar\omega_n}{2\epsilon_o V}} \vec{e}_{l,\nu}\left[\mathrm{e}^{i(\vec{k}_n.\vec{x}-\omega_n t)}a_{n,\nu} - \mathrm{e}^{-i(\vec{k}_n.\vec{x}-\omega_n t)}a^\dagger_{n,\nu}\right], \quad (24)$$

where $\vec{e}_{n,\nu}$ are two orthonormal polarization vectors ($\nu = 1,2$) which satisfy $\vec{k}_n.\vec{e}_{n,\nu} = 0$, as required for a transverse field, and the frequency is given by the dispersion relation $\omega_n = c|\vec{k}_n|$. The positive and negative frequency amplitude operators are respectively $a_{n,\nu}$ and $a^\dagger_{m,\nu}$, with bosonic commutation relations

$$[a_{n,\nu}, a_{n',\nu'}] = \delta_{\nu\nu'}\delta_{nn'}\,, \quad (25)$$

with all other commutations relations zero.

Typically we are interested in sources defined by optical cavity modes so that emission is directional and defined by the cavity spatio-temporal mode structure. Much of the difficulty in building single photon sources is in designing the optical cavity to ensure emission into a preferred set of modes. We will assume that the only modes that are excited have the same plane polarization and are all propagating in the same direction, which we take to be the x-direction. The positive frequency components of the quantum electric field for these modes are then

$$E^{(+)}(x,t) = i\sum_{n=0}^{\infty}\left(\frac{\hbar\omega_n}{2\epsilon_0 V}\right)^{1/2} a_n \mathrm{e}^{-i\omega_n(t-x/c)}\,. \quad (26)$$

In ignoring all the other modes, we are implicitly assuming that all our measurements do not respond to the vacuum state, an assumption which is justified by the theory of photon-electron detectors [57]. Let us further assume that all excited modes of this form have frequencies centered on *carrier frequency* of $\Omega \gg 1$. Then we can approximate the positive frequency components by

$$E^{(+)}(x,t) = i\left(\frac{\hbar\Omega_n}{2\epsilon_0 Ac}\right)^{1/2}\sqrt{\frac{c}{L}}\sum_{n=0}^{\infty} a_n \mathrm{e}^{-i\omega_n(t-x/c)} \quad (27)$$

where A is a characteristic transverse area. This operator has dimensions of electric field. In order to simplify the dimensions we now define a field operator that has dimensions of $\mathrm{s}^{-1/2}$. Taking the continuum limit we thus define the positive frequency operator

$$a(x,t) = \mathrm{e}^{-i\Omega(t-x/c)}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\mathrm{d}\omega' a(\omega')\mathrm{e}^{-i\omega'(t-x/c)}\,, \quad (28)$$

where we have made a change of variable $\omega \mapsto \Omega + \omega'$ and used the fact that $\Omega \gg 1$ to set the lower limit of integration to minus infinity, and

$$[a(\omega_1), a^\dagger(\omega_2)] = \delta(\omega_1 - \omega_2)\,. \quad (29)$$

In this form the moment $n(x,t) = \langle a^\dagger(x,t)a(x,t)\rangle$ has units of s^{-1}. This moment determines the probability per unit time (the count rate) to count a photon at space-time point (x,t) [57]. The field operators $a(t)$ and $a^\dagger(t)$ can be taken to describe the field emitted from the end of an optical cavity, which selects the directionality.

We will contrast single photon states with multimode coherent states defined by a multimode displacement operator acting on the vacuum $D|0\rangle$, defined implicitly by

$$D^\dagger a(\omega)D = a(\omega) + \alpha(\omega)\,, \quad (30)$$

where consistent with proceeding assumptions, $\alpha(\omega)$ is peaked at $\omega = 0$ which corresponds to a carrier frequency

$\Omega \gg 1$. The average field amplitude for this state is

$$\langle a(x,t)\rangle = \mathrm{e}^{-i\Omega(t-x/c)}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\alpha(\omega)\mathrm{e}^{-i\omega(t-x/c)} \equiv \alpha(x,t)\mathrm{e}^{-i\Omega(t-x/c)}\,, \tag{31}$$

which implicitly defines the average complex amplitude of the field as the Fourier transform of the frequency dependent displacements $\alpha(\omega)$ We can also calculate the probability per unit time to detect a photon in this state at space-time point (x,t). This is given by $n(x,t) = |\alpha(x,t)|^2$. Note that in this case the second order moment $\langle a^\dagger(x,t)a(x,t)\rangle$ factories, a result characteristic of fields with first order coherence. A coherent state is is closest to our intuitive idea of a classical electromagnetic field.

The multimode single photon state is defined by

$$|1\rangle = \int_{-\infty}^{\infty}\nu(\omega)a^\dagger(\omega)|0\rangle\,. \tag{32}$$

Normalization requires that

$$\int_{-\infty}^{\infty}\mathrm{d}\omega|\nu(\omega)|^2 = 1\,. \tag{33}$$

This last condition implies that the total number of photons, integrated over all modes, is unity,

$$\int_{-\infty}^{\infty}\mathrm{d}\omega\langle a^\dagger(\omega)a(\omega)\rangle = 1\,. \tag{34}$$

This state has zero average field amplitude but

$$n(x,t) = |\nu(t-x/c)|^2\,, \tag{35}$$

where $\nu(t)$ is the Fourier transform of $\nu(\omega)$. So while the state has zero average field amplitude there is apparently some sense in which the coherence implicit in the superposition of Eq. (32) is manifest. In fact comparing this to the case of a coherent state, Eq. (31), we see that the expression for $n(t)$ is also determined by the Fourier transform of a coherent amplitude. For this state the function $\nu(\phi)$ is periodic in the phase $\phi = t - x/c$ and it is not difficult to choose a form with a well defined pulse sequence. However care should be exercised in interpreting these pulses. They do not represent a sequence of pulses each with one photon rather they represent a single photon coherently superposed over all pulses. Once a photon is counted in a particular pulse, the field is returned to the vacuum state. A review of current efforts to produce such a state may be found in [53]. More recent results may be found in [4,8,13,19].

In the absence of true single photon pulse sources, most of the experimental work we have described has been done with a conditional source based on photon pair production in spontaneous parametric down conversion. We now give a model of the kind of state such sources produce.

In the case of a continuous wave pump, the entangled two photon produced by SPDC may be well approximated by

$$|\psi\rangle = \int_0^{\infty}\mathrm{d}\omega_1\int_0^{\infty}\mathrm{d}\omega_2\,\rho(\omega_1,\omega_2)a^\dagger(\omega_1)b^\dagger(\omega_2)|0\rangle\,, \tag{36}$$

where a, b designate distinguishable modes, for example, distinguished either by polarization or wave vector [54]. More precisely, this is the conditional state given that a down conversion process has in fact taken place at all. The probability per unit time for this event is the down conversion efficiency. For spontaneous parametric down conversion with a continuous pump field at frequency 2Ω we can approximate

$$\rho(\omega_1,\omega_2) = \delta(\omega_1+\omega_2-2\Omega)\alpha(\omega_1)\,. \tag{37}$$

We now make the change of variable $\epsilon = \Omega - \omega_1$ and assume that the bandwidth, B, over which $\alpha(\omega_1)$ is significantly different from zero is such that $\Omega \gg B$, then we can write

$$|\psi\rangle = \int_{-\infty}^{\infty}\mathrm{d}\epsilon\beta(\epsilon)a^\dagger(-\epsilon)b^\dagger(\epsilon)|0\rangle\,, \tag{38}$$

where we have defined $\beta(\epsilon) = \alpha(\Omega-\epsilon)$ and $a(-\epsilon) \equiv a(\Omega-\epsilon), b(\epsilon) \equiv b(\Omega+\epsilon)$. We will also assume that $\beta(-\epsilon) = \beta(\epsilon)$. Normalization of $|\psi\rangle$ requires that

$$\int_{-\infty}^{\infty}|\beta(\epsilon)|^2 = 1\,. \tag{39}$$

The probability per unit time to detect a photon from this field with a unit efficiency detector is in fact unity, $n_a(t) = \langle a^\dagger(t)a(t)\rangle = 1$ The probability per unit time to detect a photon from mode-a is thus independent of time. This simply means that photons will be counted at randomly distributed times from each of the fields a, b. This is a reflection of the fact that the state Eq. (38) is invariant under time translations in a Lorentzian frame.

On the other hand let us now compute the coincidence rate,

$$C(t,t') = \langle\psi|a^\dagger(x,t')a(x,t')b^\dagger(x,t)b(x,t)|\psi\rangle\,. \tag{40}$$

This is given by

$$C(t,t') = \left|\int_{-\infty}^{\infty}\beta(\epsilon)\mathrm{e}^{-i\epsilon(t'-t)}\right|^2 \tag{41}$$

$$= |\tilde{\beta}(\tau)|^2 \tag{42}$$

$$\equiv C(\tau)\,, \tag{43}$$

where $\tau = t' - t$. The coincidence rate is thus symmetrical about $t'-t=0$ and is peaked at $t'=t$. Even though

photons are detected at random, independently from each beam, they are highly correlated in time. In the case of spontaneous SPDC the distributions function $\beta(\epsilon)$ is given approximately [49]

$$\beta(\epsilon) \propto \frac{1}{\kappa^2 + \omega^2} \,. \tag{44}$$

We can now consider a heralded single photon source made by detecting one of the photon pairs and then ask for the kind of single photon state conditionally produced in the other mode. In the case of a CW pump, we first must provide a temporal filter on the detected mode. One can think of this as a time dependent detector that is switched on an off over some time interval. In frequency domain this is simply a filter. If such a detector is placed at the b mode, the conditional state of the a mode is given by Eq. (32) with

$$\nu(\omega) \propto \mathrm{e}^{-\omega^2/\kappa^2} \,, \tag{45}$$

which corresponds to a Gaussian temporal pulse.

If we drive SPDC with a pulsed pump, the pump itself provides a natural temporal filter. In this case the frequency distribution function in Eq. (37) is not delta correlated but takes the form [54]

$$\rho(\omega_1, \omega_2) = \exp\left[-(\omega_1 + \omega_2 - 2\Omega)^2/\sigma_\mathrm{p}^2\right] \alpha(\omega_1) \,, \tag{46}$$

where Ω is the pump carrier frequency and σ_p is the bandwidth of the pump pulse. It is still the case however that no photon down conversion event may take place within the pump pulse window. Thus the source is not a deterministic single photon source. Migdall [33] has proposed a way to over come this by multiplexing many heralded conditional SPDC sources with a conditioning detection on one mode of each of the multiplexed pairs. Another approach has been implemented by the Polzik group [36]. They used a cavity to enhance the parametric down conversion to implement a frequency tunable source of heralded single photons with a narrow bandwidth of 8 MHz. This approach is particularly important as frequency tunability makes the source compatible with atomic quantum memories.

Future Directions

Optical systems are certain to be used for future quantum communication protocols. Indeed the first steps have already been taken with quantum key distribution. In this article we have seen that it is also possible to process quantum information optically using heralded non deterministic schemes of various kinds and simple examples have been implemented experimentally. This greatly enhances the practicability of quantum communication schemes that require some quantum processing, such as quantum repeaters. While this has not yet been realized in practice, we expect that the first demonstrations are not far away.

In the effort to produce single photon sources we are learning new ways to encode and process information in optical pulses. We have seen that coherent communication can be done using single photons despite the fact that the average field amplitude for such states is zero. If information can be encoded and decoded on photon number states, this would represent a major step beyond quantum communication protocols like quantum key distribution.

It is far from clear however if optical schemes will be viable for large scale quantum computation. Currently ion trap schemes and schemes based on super-conducting devices offer the most likely way forward for quantum computation per se. However a number of investigators are turning to the concept of a hybrid quantum computer which combines optical and matter based qubits. Optical qubits with heralded non deterministic processing combined with matter based quantum memories is poised to make significant achievements.

Hybrid schemes offer a path to distributed quantum computation between many nodes each made up of a few hundred qubits. The nodes do not need to be far apart: they could simply be different parts of a single quantum computation device. However, if you will permit us some license, it is not too difficult to imagine a single quantum computation spanning the entire planet, with matter based nodes connected by quantum optical communication channels. Such a system would hold in its web massively entangled quantum states and exhibit a complexity that would make our current optical communication system look rather simple by comparison.

Bibliography

1. Beckman D, Chari AN, Devabhaktuni S, Preskill J (1996) Phys Rev A 54:1034
2. Braunstein SL et al (1999) Phys Rev Lett 83:1054
3. Cleve R, Ekert A, Henderson L, Macchiavello C, Mosca M (1999) On quantum algorithms. LANL quant-ph/9903061
4. Darquie B, Jones MPA, Dingjan J, Beugnon J, Bergamini S, Sortais Y, Messin G, Browaeys A, Grangier P (2005) Science 309:454
5. Deutsch D (1985) Quantum-theory, the Church–Turing principle and the universal quantum computer. Proc R Soc Lond A 400:97–117
6. Duan L-M, Lukin MD, Cirac JI, Zoller P (2001) Nature 414:413
7. Einstein A (1905) On a Heuristic Point of View about the Creation and Conversion of Light? Ann Physik 17:132
8. Eisaman MD, Fleischhauer M, Lukin MD, Zibrov AS (2006) To-

ward quantum control of single photons. Opt Photonics News 17:22–27
9. Feynman RP (1982) Simulating physics with computers. Int J Theor Phys 21:467
10. Gasparoni S et al (2004) Phys Rev Lett 93:020504
11. Gilchrist A, Langford NK, Nielsen MA (2005) Distance measures to compare real and ideal quantum processes. Phys Rev A 71:062310
12. Gottesman D, Chuang IL (1999) Demonstrating the viability of universal quantum computation using teleportation and single-qubit operations. Nature 402:390–393
13. Hennrich M, Legero T, Kuhn A, Rempe G (2004) Photon statistics of a non-stationary periodically driven single-photon source. New J Phys 6:86
14. Hofmann H, Takeuchi S (2002) Phys Rev A 66:024308
15. Hong CK, Ou ZY, Mandel L (1987) Measurement of subpicosecond time intervals between two photons by interference. Phys Rev Lett 59:2044
16. Hutchinson GD, Milburn GJ (2004) Nonlinear quantum optical computing via measurement. J Modern Opt 51:1211–1222
17. James DFV, Kwiat PG, Munro WJ, White AG (2001) Measurement of qubits. Phys Rev A 64:052312
18. Jiang L, Taylor JM, Khaneja N, Lukin MD (2007) Optimal approach to quantum communication using dynamic programming. arXiv:quant-ph/0710.5808
19. Keller M, Lange B, Hayasaka K, Lange W, Walther H (2003) Continuous generation of single photons with controlled waveform in an ion-trap cavity system. Nature 431:1075
20. Kieling K, Rudolph T, Eisert J (2007) Percolation, renormalization, and quantum computing with nondeterministic gates. Phys Rev Lett 99:130501
21. Kiesel N, Schmid C, Weber U, Ursin R, Weinfurter H (2007) Phys Rev Lett 99:250505
22. Knill E (2002) Phys Rev A 66:052306
23. Knill E (2005) Quantum computing with realistically noisy devices. Nature 434:39–44
24. Knill E, Laflamme R, Milburn GJ (2001) Efficient linear optical quantum computation. Nature 409:46
25. Kok P, Lee H, Dowling JP (2002) Phys Rev A 66:063814
26. Kok P, Munro WJ, Nemoto K, Ralph TC, Dowling JP, Milburn GJ (2007) Linear optical quantum computing. Rev Mod Phys 79:135
27. Langford NK, Weinhold TJ, Prevedel R, Gilchrist A, O'Brien JL, Pryde GJ, White AG (2005) Phys Rev Lett 95:210504
28. Lanyon BP, Barbieri M, Almeida MP, Jennewein T, Ralph TC, Resch KJ, Pryde G, O'Brien JL, Gilchrist A, White AG (2007) Quantum computing using shortcuts through higher dimensions. appear in Nat Phys
29. Lanyon BP, Weinhold TJ, Langford NK, Barbieri M, James DFV, Gilchrist A, White AG (2007) Phys Rev Lett 99:250505
30. Lu C-Y, Browne DE, Yang T, Pan J-W (2007) Phys Rev Lett 99:250504
31. Lu C-Y, Zhou X-Q, Gühne O, Gao W-B, Zhang J, Yuan Z-S, Goebel A, Yang T, Pan J-W (2007) Experimental entanglement of six photons in graph states. Nature Phys 3:91
32. Menicucci NC et al (2002) Phys Rev Lett 88:167901
33. Migdall A et al (2002) Phys Rev A 66:053805
34. Milburn GJ (1989) A quantum optical fredkin gate. Phys Rev Lett 62:2124–2127
35. Mishina OS, Kupriyanov DV, Muller JH, Polzik ES (2007) Spectral theory of quantum memory and entanglement via Raman scattering of light by an atomic ensemble. Phys Rev A 75:042326
36. Neergaard-Nielsen JS, Melholt Nielsen B, Takahashi H, Vistnes AI, Polzik ES (2007) High purity bright single photon source. Opt Express 15:7940
37. Nielsen MA (2004) Optical quantum computation using cluster states. Phys Rev Lett 93:040503
38. O'Brien JL et al (2004) Phys Rev Lett 93:080502
39. O'Brien JL, Pryde GJ, White AG, Ralph TC Branning D (2003) Demonstration of an all-optical quantum controlled-NOT gate. Nature 426:264
40. Okamoto R, Hofmann HF, Takeuchi S, Sasaki K (2007) Phys Rev Lett 99:250506
41. Pittman TB, Jacobs BC, Franson JD (2001) Probabilistic quantum logic operations using polarizing beam splitters. Phys Rev A 64:062311
42. Pittman TB, Jacobs BC, Franson JD (2002) Phys Rev Lett 88:257902
43. Pittman TB, Fitch MJ, Jacobs BC, Franson JD (2003) Experimental controlled-NOT logic gate for single photons in the coincidence basis. Phys Rev A 68:032316
44. Popescu S (2006) KLM quantum computation as a measurement based computation. arXiv:quant-ph/0610025
45. Prevedel R et al (2007) Nature 445:65
46. Ralph TC, White AG, Munro WJ, Milburn GJ (2001) Simple scheme for efficient linear optics quantum gates. Phys Rev A 65:012314
47. Ralph TC, Langford NK, Bell TB, White AG (2002) Linear optical controlled-NOT gate in the coincidence basis. Phys Rev A 65:062324
48. Raussendorf R, Briegel HJ (2001) A one-way quantum computer. Phys Rev Lett 86:5188
49. Rhode PP, Ralph TC (2005) Phys Rev A 71:032320
50. Sangouard N, Simon C, Minar J, Zbinden H, de Riedmatten H, Gisin N (2007) Long-distance entanglement distribution with single-photon sources. arXiv:quant-ph/0706.1924v1
51. Shor P (1994) Algorithms for quantum computation: Discrete logarithms and factoring. Proc 35th annual symposium on foundations of computer science. See also LANL preprint quant-ph/9508027
52. Simon C et al (2007) Phys Rev Lett 98:190503
53. Special Issue on Single photon Sources (2004) Single photon Sources 51(9–10)
54. U'Ren AB, Mukamel E, Banaszek K, Walmsley IA (2003) Phil Trans Roy Soc A 361:1471
55. Van Meter R, Ladd TD, Munro WJ, Nemoto K (2007) System Design for a Long-Line Quantum Repeater. arXiv:quant-ph/0705.4128v1
56. Vandersypen LMK et al (2001) Nature 414:883
57. Walls DF, Milburn GJ (2008) Quantum Optics, 2nd edn. Springer, Berlin
58. Walther P, Resch KJ, Rudolph T, Schenck E, Weinfurter H, Vedral V, Aspelmeyer M, Zeilinger A (2005) Experimental one-way quantum computing. Nature 434:169
59. Weinhold TJ, Gilchrist A, Resch KJ, Doherty AC, O'Brien JL, Pryde GJ, White AG (2008) Understanding photonic quantum-logic gates: The road to fault tolerance. arxiv 0808.0794
60. White AG et al (2007) Measuring two-qubit gates. J Opt Soc Am B 24:172–183
61. Yamamoto Y, Kitagawa M, Igeta K (1988) Proc 3rd Asia-Pacific phys conf 779. World Scientific, Singapore

Quantum Cryptography

Hoi-Kwong Lo, Yi Zhao
Center for Quantum Information and Quantum Control, Department of Physics and Department of Electrical and Computer Engineering, University of Toronto, Toronto, Canada

Article Outline

Glossary

One-time pad One-time pad is a classical encryption algorithm invented by Gilbert Vernam in 1917. In one-time pad algorithm, the legitimate users share a random key (e. g. a random binary string) that is not known to anyone else. The message is combined with this random key ("pad") which is as long as the message. The key is used only once ("one-time"). The most typical usage is in binary case, where an XOR operation is applied between the message and the key to achieve the ciphertext. Claudé Shannon proved that the one-time pad provides perfect secrecy in 1949. The perfect secrecy is defined that the ciphertext does not give any additional information on the message.

Key distribution problem The key distribution problem originates from the one-time pad encryption. The one-time pad encryption requires that the two parties share a secret random key before the communication. This key is usually generated by one party. The key distribution problem is how to distribute this random key from one party to the other party securely. This problem is non-solvable classically, but is solvable via quantum key distribution.

Quantum key distribution Quantum key distribution (QKD) is a method to distribute a random key between two parties securely. The main idea is to encode the bit value on the quantum state of certain particle (usually photon) and send the particle to the receiver. The quantum no-cloning theorem guaranteed that any eavesdropper cannot duplicate the encoded quantum information perfectly.

BB84 BB84 is the first and so far the most popular quantum cryptography protocol. It was proposed by C. H. Bennet and G. Brassard in 1984 [1]. In the original BB84 proposal, the quantum information is encoded on the polarizations of photons. Later BB84 was extended to the phase coding. Detailed description of BB84 protocol can be found in Sect. "Introduction".

B92 B92 is a quantum cryptography protocol proposed by C. H. Bennet in 1992 [2]. It uses two non-orthogonal states (e. g. the horizontally – and 45° polarized photons) to denote "0" and "1". It is simpler than BB84 protocol in implementation.

E91 E91 is a quantum cryptography protocol proposed by A. Ekert in 1991 [3]. It is based on entangled photon pairs. E91 protocol is often used in free-space quantum key distribution.

Uni-directional QKD Uni-directional QKD is the QKD scheme in which Alice (sender) generates the photon, encodes the quantum information on it, and sends it to Bob (receiver).

Bi-directional QKD, or "Plug & play" QKD Bi-directional QKD, or "Plug & play" QKD is the QKD scheme in which Bob generates strong laser pulses and sends them to Alice. Alice encodes her quantum information on the pulse and attenuates the pulse to single-photon level, and sends it back to Bob through the same channel. This design can automatically compensate the phase and the polarization drifting in the channel.

Single photon source Single photon source is the light source that can generate a *single* photon on demand. A perfect single photon source should have zero probability to generate multi photons once triggered. Single photon source is required in the original BB84 protocol. However, it is no longer under absolute demand due to the discovery and implementation of decoy state QKD.

Fainted laser source Fainted laser source is the light source that has a standard pulsed laser source and a heavy attenuator. The average output photon number is usually set to ~ 0.1 photon per pulse. This low average photon number is to suppress the production of multi photon signals. However, due to the poisson nature of laser source, the probability of multi photon production can never reach zero unless the laser is turned off.

Single photon detector Single photon detector is sensitive to the weakest light signals – signals with single photons. Most single photon detectors are threshold by means that they can only detect the arrival of one or more photons, but cannot count the number of photons within one signal.

Dark count Dark count is the event that the detector reports a detection while no photon actually hits it. It is a key parameter for single photon detectors. Dark count becomes important when the channel loss between the sender and the receiver is high (i. e., when very few photons can reach the receiver).

Qubit Qubit (or quantum bit) is the fundamental unit of quantum information. Whereas a classical bit can take value of either "0" or "1", a qubit can take a value in any superposition of two distinguishable (i. e., orthogonal) states commonly labeled by $|0\rangle$ and $|1\rangle$. In other words, a (pure) qubit state can be written in the form $a|0\rangle + b|1\rangle$ where a and b are complex numbers. The normalization constraint is that $|a|^2 + |b|^2 = 1$. Physically, a qubit can be encoded in any two-level quantum system, such as the two polarization of a single photon or two atomic levels of an atom.

Bit-flip Bit-flip is a typical noise in both classical and quantum communication. In quantum cryptography, a bit-flip in the channel will transform an initial state $|i\rangle = a|0\rangle + b|1\rangle$ into the final state $|f\rangle = a|1\rangle + b|0\rangle$.

Phase-flip Phase-flip is a typical noise that is unique in quantum communication. A phase-flip in the quantum channel will transform an initial state $|i\rangle = a|0\rangle + b|1\rangle$ into the final state $|f\rangle = a|0\rangle - b|1\rangle$.

Definition of the Subject

Quantum cryptography is the synthesis of quantum mechanics with the art of code-making (cryptography). The idea was first conceived in an unpublished manuscript written by Stephen Wiesner around 1970 [4]. However, the subject received little attention until its resurrection by a classic paper published by Bennett and Brassard in 1984 [1]. The goal of quantum cryptography is to perform tasks that are impossible or intractable with conventional cryptography. Quantum cryptography makes use of the subtle properties of quantum mechanics such as the quantum no-cloning theorem and the Heisenberg uncertainty principle. Unlike conventional cryptography, whose security is often based on unproven computational assumptions, quantum cryptography has an important advantage in that its security is often based on the laws of physics. Thus far, proposed applications of quantum cryptography include quantum key distribution (abbreviated QKD), quantum bit commitment and quantum coin tossing. These applications have varying degrees of success. The most successful and important application – QKD – has been proven to be unconditionally secure. Moreover, experimental QKD has now been performed over hundreds of kilometers over both standard commercial telecom optical fibers and open-air. In fact, commercial QKD systems are currently available on the market.

On a wider context, quantum cryptography is a branch of quantum information processing, which includes quantum computing, quantum measurements, and quantum teleportation. Among all branches, quantum cryptography is the branch that is closest to real-life applications. Therefore, it can be a concrete avenue for the demonstrations of concepts in quantum information processing. On a more fundamental level, quantum cryptography is deeply related to the foundations of quantum mechanics, particularly the testing of Bell-inequalities and the detection efficiency loophole. On a technological level, quantum cryptography is related to technologies such as single-photon measurements and detection and single-photon sources.

Introduction

The best-known application of quantum cryptography is quantum key distribution (QKD). The goal of QKD is to allow two distant participants, traditionally called Alice and Bob, to share a long random string of secret (commonly called the key) in the presence of an eavesdropper, traditionally called Eve. The key can subsequently be used to achieve a) perfectly secure communication (via one-time-pad, see below) and b) perfectly secure authentication (via Wigman–Carter authentication scheme), thus achieving two key goals in cryptography.

The best-known protocol for QKD is the Bennett and Brassard protocol (BB84) published in 1984 [1]. The procedure of BB84 is as follows (also shown in Table 1).

1. Quantum communication phase
   (a) In BB84, Alice sends Bob a sequence of photons, each independently chosen from one of the four polarizations – vertical, horizontal, 45-degrees and 135-degrees.
   (b) For each photon, Bob randomly chooses one of the two measurement bases (rectilinear and diagonal) to perform a measurement.
   (c) Bob records his measurement bases and results. Bob publicly acknowledges his receipt of signals.
2. Public discussion phase
   (a) Alice broadcasts her bases of measurements. Bob broadcasts his bases of measurements.

Quantum Cryptography, Table 1
Procedure of BB84 protocol

Alice's bit sequence	0	1	1	1	0	1	0	0	0	1
Alice's basis	+	×	+	+	×	+	×	×	+	×
Alice's photon polarization	↔	↘	↕	↕	↗	↕	↗	↗	↔	↘
Bob's basis	+	+	×	+	+	×	×	+	+	×
Bob's measured polarization	↔	↕	↘	↕	↔	↗	↗	↕	↔	↘
Bob's sifted measured polarization	↔			↕			↗		↔	↘
Bob's data sequence	0			1			0		0	1

(b) Alice and Bob discard all events where they use different bases for a signal.
(c) To test for tampering, Alice randomly chooses a fraction, p, of all remaining events as test events. For those test events, she publicly broadcasts their positions and polarizations.
(d) Bob broadcasts the polarizations of the test events.
(e) Alice and Bob compute the error rate of the test events (i. e., the fraction of data for which their value disagree). If the computed error rate is larger than some prescribed threshold value, say 11%, they abort. Otherwise, they proceed to the next step.
(f) Alice and Bob each convert the polarization data of all remaining data into a binary string called a raw key (by, for example, mapping a vertical or 45-degrees photon to "0" and a horizontal or 135-degrees photon to "1"). They can perform classical post-processing such as error correction and privacy amplification to generate a final key.

Notice that it is important for the classical communication channel between Alice and Bob to be authenticated. Otherwise, Eve can easily launch a man-in-the-middle attack by disguising as Alice to Bob and as Bob to Alice. Fortunately, authentication of an m-bit classical message requires only logarithmic in m-bit of an authentication key. Therefore, QKD provides an efficient way to expand a short initial authentication key into a long key. By repeating QKD many times, one can get an arbitrarily long secure key.

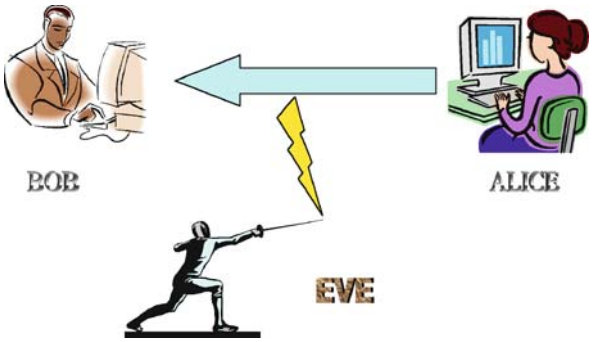
This article is organized as follows. In Sect. "Quantum Key Distribution: Motivation and Introduction", we will discuss the importance and foundations of QKD; in Section "Security Proofs", we will discuss the principles of different approaches to prove the unconditional security of QKD; in Section "Experimental Fundamentals", we will introduce the history and some fundamental components of QKD implementations; in Section "Experimental Implementation of BB84 Protocol", we will discuss the implementation of BB84 protocol in detail; in Section "Other Quantum Key Distribution Protocols", we will discuss the proposals and implementations of other QKD protocols; in Section "Quantum Hacking", we will introduce a very fresh and exciting area – quantum hacking – in both theory and experiments. In particular, we provide a catalog of existing eavesdropping attacks; in Section "Beyond Quantum Key Distribution", we will discuss some topics other than QKD, including quantum bit commitment, quantum coin tossing, etc.; in Section "Future Directions", we will wrap up this article with prospectives of quantum cryptography in the future.

Quantum Key Distribution: Motivation and Introduction

Cryptography – the art of code-marking – has a long and distinguished history of military and diplomatic applications, dating back to ancient civilizations in Mesopotamia, Egypt, India and China. Moreover, in recent years cryptography has widespread applications in civilian applications such as electronics commerce and electronics businesses. Each time we go on-line to access our banking or credit card data, we should be deeply concerned with our data security.

Key Distribution Problem and One-Time-Pad

Secure communication is the best-known application of cryptography. The goal of secure communication is to allow two distant participants, traditionally called Alice and Bob, to communicate securely in the presence of an eavesdropper, Eve. See Fig. 1. A simple example of an encryption scheme is the Caesar's cipher. Alice simply shifts each letter in a message alphabetically by three letters. For instance, the word NOW is mapped to QRZ, because N → O → P → Q, O → P → Q → R and W → X → Y → Z. According to legends, Julius Caesar used Caesar's cipher to communicate with his generals. An encryption by an alphabetical shift of a fixed but arbitrary number of po-



Quantum Cryptography, Figure 1
Communication in presence of an eavesdropper

sitions is also called a Caesar's cipher. Note that Caesar's cipher is not that secure because an eavesdropper can simply exhaustively try all 26 possible combinations of the key to recover the original message.

In conventional cryptography, an unbreakable code does exist. It is called the one-time-pad and was invented by Gilbert Vernam in 1918 [5]. In the one-time-pad method, a message (traditionally called the plain text) is first converted by Alice into a binary form (a string consisting of "0"s and "1"s) by a publicly known method. A key is a binary string of the same length as the message. By combining each bit of the message with the respective bit of the key using XOR (i. e. addition modulo two), Alice converts the plain text into an encrypted form (called the cipher text). i. e. for each bit $c_i = m_i + k_i \mod 2$. Alice then transmits the cipher text to Bob via a broadcast channel. Anyone including an eavesdropper can get a copy of the cipher text. However, without the knowledge of the key, the cipher text is totally random and gives no information whatsoever about the plain text. For decryption, Bob, who shares the same key with Alice, can perform another XOR (i. e. addition modulo two) between each bit of the cipher text with the respective bit of the key to recover the plain text. This is because $c_i + k_i \mod 2 = m_i + 2k_i \mod 2 = m_i \mod 2$.

Notice that it is important not to re-use a key in a one-time-pad scheme. Suppose the same key, k, is used for the encryption of two messages, m_1 and m_2, then the cipher texts are $c_1 = m_1 + k \mod 2$ and $c_2 = m_2 + k \mod 2$. Then, Eve can simply take the XOR of the two cipher texts to obtain $c_1 + c_2 \mod 2 = m_1 + m_2 + 2k \mod 2 = m_1 + m_2 \mod 2$, thus learning non-trivial information, namely the parity of the two messages.

The one-time-pad method is commonly used in top-secret communication. The one-time-pad method is unbreakable, but it has a serious drawback: it supposes that Alice and Bob initially share a random string of secret that is as long as the message. Therefore, the one-time-pad simply shifts the problem of secure communication to the problem of key distribution. This is the key distribution problem. In top-secret communication, the key distribution problem can be solved by trusted couriers. Unfortunately, trusted couriers can be bribed or compromised. Indeed, in conventional cryptography, a key is a classical string consisting of "0" and "1"s. In classical physics, there is no fundamental physical principle that can prevent an eavesdropper from copying a key during the key distribution process.

A possible solution to the key distribution problem is public key cryptography. However, the security of public key cryptography is based on unproven computational assumptions. For example, the security of standard RSA crypto-system invented by Rivest–Shamir–Adleman (RSA) is based on the presumed difficulty of factoring large integers. Therefore, public key distribution is vulnerable to unanticipated advances in hardware and algorithms. In fact, quantum computers – computers that operate on the principles of quantum mechanics – can break standard RSA crypto-system via the celebrated Shor's quantum algorithm for efficient factoring [6].

Quantum No-Cloning Theorem and Quantum Key Distribution (QKD)

Quantum mechanics can provide a solution to the key distribution problem. In quantum key distribution, an encryption key is generated randomly between Alice and Bob by using non-orthogonal quantum states. In contrast to classical physics, in quantum mechanics there is a quantum no-cloning theorem (see below), which states that it is fundamentally impossible for anyone including an eavesdropper to make an additional copy of an unknown quantum state.

A big advantage of quantum cryptography is *forward security*. In conventional cryptography, an eavesdropper Eve has a transcript of all communications. Therefore, she can simply save it for many years and wait for breakthroughs such as the discovery of a new algorithm or new hardware. Indeed, if Eve can factor large integers in 2100, she can decrypt communications sent in 2008. We remark that Canadian census data are kept secret for 92 years on average. Therefore, factoring in the year 2100 may violate the security requirement of our government today! And, no one in his/her sane mind can guarantee the impossibility of efficient factoring in 2100 (except for the fact that he/she may not live that long). In contrast, quantum cryptography guarantees forward security. Thanks to the quan-

tum no-cloning theorem, an eavesdropper does *not* have a transcript of all quantum signals sent by Alice to Bob.

For completeness, we include the statement and the proof of the quantum no-cloning theorem below.

Quantum No-Cloning Theorem *An unknown quantum state cannot be copied.*

(a) The case without ancilla: Given an unknown state $|\alpha\rangle$, show that a quantum copying machine that can map $|\alpha\rangle|0\rangle \to |\alpha\rangle|\alpha\rangle$ does not exist.

(b) The general case: Given an unknown state $|\alpha\rangle$, show that a quantum copying machine that can map $|\alpha\rangle|0\rangle|0\rangle \to |\alpha\rangle|\alpha\rangle|u_\alpha\rangle$ does not exist.

Proof

(a) Suppose the contrary. Then, a quantum cloning machine exists. Consider two orthogonal input states $|0\rangle$ and $|1\rangle$ respectively. We have

$$|0\rangle|0\rangle \to |0\rangle|0\rangle$$

and

$$|1\rangle|0\rangle \to |1\rangle|1\rangle\,.$$

Consider a general input $|\alpha\rangle = a|0\rangle + b|1\rangle$. Since a unitary transformation is linear, by linearity, we have

$$\begin{aligned} |\alpha\rangle|0\rangle &= (a|0\rangle + b|1\rangle)|0\rangle \\ &\to a|0\rangle|0\rangle + b|1\rangle|1\rangle\,. \end{aligned} \tag{1}$$

In contrast, for quantum cloning, we need:

$$\begin{aligned} |\alpha\rangle|0\rangle &\to (a|0\rangle + b|1\rangle)(a|0\rangle + b|1\rangle) \\ &= a^2|0\rangle|0\rangle + ab|0\rangle|1\rangle + ab|1\rangle|0\rangle + b^2|1\rangle|1\rangle\,. \end{aligned} \tag{2}$$

Clearly, if $ab \neq 0$, the two results shown in Eqs. (1) and (2)) are different. Therefore, quantum cloning (without ancilla) is impossible.

(b) similar. □

More generally, for general quantum states, information gain implies disturbance.

Theorem (Information Gain implies disturbance) *Given one state chosen from one of the two distinct non-orthogonal states, $|u\rangle$ and $|v\rangle$ (i.e. $|\langle u|v\rangle| \neq 0$ or 1), any operation that can learn any information about its identity necessarily disturbs the state.*

Proof Given a system initially in state either $|u\rangle$ and $|v\rangle$. Suppose an experimentalist applies some operation on the system. The most general thing that she can try to do is to prepare some ancilla in some standard state $|0\rangle$ and couple it to the system. Therefore, we have:

$$|u\rangle|0\rangle \to |u\rangle|\phi_u\rangle \tag{3}$$

and

$$|v\rangle|0\rangle \to |v\rangle|\phi_v\rangle \tag{4}$$

for some states $|\phi_u\rangle$ and $|\phi_v\rangle$.

In the end, the experimentalist lets go of the system and keeps the ancilla. He/she may then perform a measurement on the ancilla to learn about the initial state of the system.

Recall that quantum evolution is unitary and as such it preserves the inner product. Now, taking the inner product between Eqs. (3) and (4), we get:

$$\begin{aligned} \langle u|v\rangle\langle 0|0\rangle &= \langle u|v\rangle\langle\phi_u|\phi_v\rangle \\ \langle u|v\rangle &= \langle u|v\rangle\langle\phi_u|\phi_v\rangle \\ \langle u|v\rangle(1 - \langle\phi_u|\phi_v\rangle) &= 0 \\ (1 - \langle\phi_u|\phi_v\rangle) &= 0 \\ |\phi_u\rangle &= |\phi_v\rangle\,, \end{aligned} \tag{5}$$

where in the fourth line, we have used the fact that $|\langle u|v\rangle| \neq 0$.

Now, the condition that $|\phi_u\rangle = |\phi_v\rangle$ means that the final state of the ancilla is independent of the initial state of the system. Therefore, a measurement on the ancilla will tell the experimentalist nothing about the initial state of the system. □

Therefore, any attempt by an eavesdropper to learn information about a key in a QKD process will lead to disturbance, which can be detected by Alice and Bob who can, for example, check the bit error rate of a random sample of the raw transmission data.

The standard BB84 protocol for QKD was discussed in Sect. "Introduction". In the BB84 protocol, Alice prepares a sequence of photons each randomly chosen in one of the four polarizations – vertical, horizontal, 45-degrees and 135-degrees. For each photon, Bob chooses one of the two polarization bases (rectilinear or diagonal) to perform a measurement. Intuitively, the security comes from the fact that the two polarization bases, rectilinear and diagonal, are conjugate observables. Just like position and momentum are conjugate observables in the standard Heisenberg uncertainty principle, no measurement by an eavesdropper Eve can determine the value of both observables simultaneously. In mathematics, two conjugate observables are represented by two non-commuting Hermitian matrices. Therefore, they cannot be simultaneously

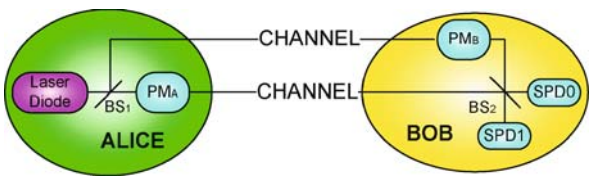
diagonalized. This impossibility of simultaneous diagonalization implies the impossibility of simultaneous measurements of two conjugate observables.

Example of a Simple Eavesdropper Attack: Intercept-Resend Attack

To illustrate the security of quantum cryptography, let us consider the simple example of an intercept-resend attack by an eavesdropper Eve, who measures each photon in a randomly chosen basis and then resends the resulting state to Bob. For instance, if Eve performs a rectilinear measurement, photons prepared by Alice in the diagonal bases will be disturbed by Eve's measurement and give random answers. When Eve resends rectilinear photons to Bob, if Bob performs a diagonal measurement, then he will get random answers. Since the two bases are chosen randomly by each party, such an intercept-resend attack will give a bit error rate of $0.5 \times 0.5 + 0.5 \times 0 = 25\%$, which is readily detectable by Alice and Bob. Sophisticated attacks against QKD do exist. Fortunately, the security of QKD has now been proven. This subject will be discussed further in Sect. "Security Proofs".

Equivalence Between Phase and Polarization Encoding

Notice that the BB84 protocol can be implemented with any two-level quantum system (qubits). In Sect. "Introduction" and the above discussion, we have described the BB84 protocol in terms of polarization encoding. This is just one of the many possible types of encodings. Indeed, it should be noted that other encoding method, particularly, phase encoding also exists. In phase encoding, a signal consists of a superposition of two time-separated pulses, known as the reference pulse and the signal pulse. See Fig. 2 for an illustration of the phase encoding scheme. The information is encoded in the relative phase between two pulses. i. e., the four possible states used by Alice are $1/\sqrt{2}(|R\rangle + |S\rangle)$, $1/\sqrt{2}(|R\rangle - |S\rangle)$, $1/\sqrt{2}(|R\rangle + i|S\rangle)$, $1/\sqrt{2}(|R\rangle - i|S\rangle)$.



Quantum Cryptography, Figure 2
Conceptual schematic for phase-coding BB84 QKD system. PM: Phase Modulator; BS: Beam Splitter; SPD: Single Photon Detector

Notice that, mathematically the phase encoding scheme is equivalent to the polarization encoding scheme. They are simply different embodiments of the same BB84 protocol.

Security Proofs

"The most important question in quantum cryptography is to determine how secure it really is." (Brassard and Crépeau [7]) Security proofs are very important because a) they provide the foundation of security to a QKD protocol, b) they provide a formula for the key generation rate of a QKD protocol and c) they may even provide a construction for the classical post-processing protocol (for error correction and privacy amplification) that is necessary for the generation of the final key. Without security proofs, a real-life QKD system is incomplete because we can never be sure about how to generate a secure key and how secure the final key really is.

Classification of Eavesdropping Attacks

Before we discuss security proofs, let us first consider eavesdropping attacks. Notice that there are infinitely many eavesdropping strategies that an eavesdropper, Eve, can perform against a QKD protocol. They can be classified as follows:

Individual attacks In an individual attack, Eve performs an attack on each signal independently. The intercept-resend attack discussed in Subsect. "Example of a Simple Eavesdropper Attack: Intercept-Resend Attack" is an example of an individual attack.

Collective attacks A more general class of attacks is collective attack where for each signal, Eve independently couples it with an ancillary quantum system, commonly called an ancilla, and evolves the combined signal/ancilla unitarily. She can send the resulting signals to Bob, but keep all ancillas herself. Unlike the case of individual attacks, Eve postpones her choice of measurement. Only after hearing the public discussion between Alice and Bob, does Eve decide on what measurement to perform on her ancilla to extract information about the final key.

Joint attacks The most general class of attacks is joint attack. In a joint attack, instead of interacting with each signal independently, Eve treats all the signals as a single quantum system. She then couples the signal system with her ancilla and evolves the combined signal and ancilla system unitarily. She hears the public discussion between Alice and Bob before deciding on which measurement to perform on her ancilla.

Proving the security of QKD against the most general attack was a very hard problem. It took more than 10 years, but the unconditional security of QKD was finally established in several papers in the 1990s. One approach by Mayers [8] was to prove the security of the BB84 directly. A simpler approach by Lo and Chau [9], made use of the idea of entanglement distillation by Bennett, DiVincenzo, Smolin and Wootters (BDSW) [10] and quantum privacy amplification by Deutsch et al. [11] to solve the security of an entanglement-based QKD protocol. The two approaches have been unified by the work of Shor and Preskill [12], who provided a simple proof of security of BB84 using entanglement distillation idea. Other early security proofs of QKD include Biham, Boyer, Boykin, Mor, and Roychowdhury [13], and Ben-Or [14].

Approaches to Security Proofs

There are several approaches to security proof. We will discuss them one by one.

1. Entanglement distillation
   Entanglement distillation protocol (EDP) provides a simple approach to security proof [9,11,12]. The basic insight is that entanglement is a sufficient (but not necessary) condition for a secure key. Consider the noiseless case first. Suppose two distant parties, Alice and Bob, share a maximally entangled state of the form $|\phi\rangle_{AB} = 1/\sqrt{2}(|00\rangle_{AB} + |11\rangle_{AB})$. If each of Alice and Bob measure their systems, then they will both get "0"s or "1"s, which is a shared random key. Moreover, if we consider the combined system of the three parties – Alice, Bob and an eavesdropper, Eve, we can use a pure-state description (the "Church of Larger Hilbert space") and consider a pure state $|\psi\rangle_{ABE}$. In this case, the von Neumann entropy [15] of Eve $S(\rho_E) = S(\rho_{AB}) = 0$. This means that Eve has absolutely no information on the final key. This is the consequence of the standard Holevo's theorem. See, like, [16].
   In the noisy case, Alice and Bob may share N pairs of qubits, which are a noisy version of N maximally entangled states. Now, using the idea of entanglement distillation protocol (EDP) discussed in BDSW [10], Alice and Bob may apply local operations and classical communications (LOCCs) to distill from the N noisy pairs a smaller number, say M almost perfect pairs i. e., a state close to $|\phi\rangle_{AB}^{M}$. Once such a EDP has been performed, Alice and Bob can measure their respective system to generate an M-bit final key.
   One may ask: how can Alice and Bob be sure that their EDP will be successful? Whether an EDP will be successful or not depends on the initial state shared by Alice and Bob. In the above, we have skipped the discussion about the verification step. In practice, Alice and Bob can never be sure what initial state they possess. Therefore, it is useful for them to add a verification step. By, for example, randomly testing a fraction of their pairs, they have a pretty good idea about the properties (e. g., the bit-flip and phase error rates) of their remaining pairs and are pretty confident that their EDP will be successful.
   The above description of EDP is for a quantum-computing protocol where we assume that Alice and Bob can perform local quantum computations. In practice, Alice and Bob do not have large-scale quantum computers at their disposal. Shor and Preskill made the important observation that the security proof of the standard BB84 protocol can be reduced to that of an EDP-based QKD protocol [9,11]. The Shor–Preskill proof [12] makes use of the Calderbank–Shor–Steane (CSS) code, which has the advantage of decoupling the quantum error correction procedure into two parts: bit-flip and phase error correction. They can go on to show that bit-flip error correction corresponds to standard error correction and phase error correction corresponds to privacy amplification (by random hashing).
2. Communication complexity/quantum memory
   The communication complexity/quantum memory approach to security proof was proposed by Ben-Or [14] and subsequently by Renner and Koenig [17]. See also [18]. They provide a formula for secure key generation rate in terms of an eavesdropper's quantum knowledge on the raw key: Let Z be a random variable with range $\mathcal{Z}$, let $\boldsymbol{\rho}$ be a random state, and let F be a two-universal function on $\mathcal{Z}$ with range $S = \{0,1\}^s$ which is independent of Z and $\boldsymbol{\rho}$. Then [17]

$$d(F(Z)|\{F\} \otimes \boldsymbol{\rho}) \leq \tfrac{1}{2} 2^{-\frac{1}{2}(S_2([\{Z\}\otimes\boldsymbol{\rho}]) - S_0([\boldsymbol{\rho}]) - s)} .$$

   Incidentally, the quantum de Finnetti's theorem [19] is often useful for simplifying security proofs of this type.
3. Twisted state approach
   What is a necessary and sufficient condition for secure key generation? From the entanglement distillation approach, we know that entanglement distillation a sufficient condition for secure key generation. For some time, it was hoped that entanglement distillation is also a necessary condition for secure key generation. However, such an idea was proven to be wrong in [20,21], where it was found that a necessary and sufficient condition is the distillation of a private state, rather than a maximally entangled state. A private state is a "twisted" version of a maximally entangled state.

They proved the following theorem in [20]: a state is private in the above sense iff it is of the following form

$$\gamma_m = U|\psi_{2^m}^+\rangle_{AB}\langle\psi_{2^m}^+| \otimes \varrho_{A'B'} U^\dagger \tag{6}$$

where $|\psi_d\rangle = \sum_{i=1}^d |ii\rangle$ and $\varrho_{A'B'}$ is an arbitrary state on A', B'. U is an arbitrary unitary controlled in the computational basis

$$U = \sum_{i,j=1}^{2^m} |ij\rangle_{AB}\langle ij| \otimes U_{ij}^{A'B'} . \tag{7}$$

The operation (7) will be called "twisting" (note that only $U_{ii}^{A'B'}$ matter here, yet it will be useful to consider general twisting later).

Proof. (copied from [20]) The authors of [20] proved for $m = 1$ (for higher m, the proof is analogous). Start with an arbitrary state held by Alice and Bob, $\rho_{AA'BB'}$, and include its purification to write the total state in the decomposition

$$\begin{aligned}\Psi_{ABA'B',E} = {} & a|00\rangle_{AB}|\Psi_{00}\rangle_{A'B'E} + b|01\rangle_{AB}|\Psi_{01}\rangle_{A'B'E} \\ & + c|10\rangle_{AB}|\Psi_{10}\rangle_{A'B'E} + d|11\rangle_{AB}|\Psi_{11}\rangle_{A'B'E}\end{aligned} \tag{8}$$

with the states $|ij\rangle$ on AB and Ψ_{ij} on $A'B'E$. Since the key is unbiased and perfectly correlated, we must have $b = c = 0$ and $|a|^2 = |d|^2 = 1/2$. Depending on whether the key is $|00\rangle$ or $|11\rangle$, Eve will hold the states

$$\varrho_0 = Tr_{A'B'}|\Psi_{00}\rangle\langle\Psi_{00}|, \quad \varrho_1 = Tr_{A'B'}|\Psi_{11}\rangle\langle\Psi_{11}|. \tag{9}$$

Perfect security requires $\varrho_0 = \varrho_1$. Thus there exists unitaries U_{00} and U_{11} on $A'B'$ such that

$$\begin{aligned}|\Psi_{00}\rangle &= \sum_i \sqrt{p_i}|U_0\phi_i^{A'B'}\rangle|\varphi_i^E\rangle \\ |\Psi_{11}\rangle &= \sum_i \sqrt{p_i}|U_1\phi_i^{A'B'}\rangle|\varphi_i^E\rangle .\end{aligned} \tag{10}$$

After tracing out E, we will thus get a state of the form Eq. (6), where $\varrho_{A'B'} = \sum_i p_i|\phi_i\rangle\langle\phi_i| \square$.

The main new ingredient of the above theorem is the introduction of a "shield" part to Alice and Bob's system. That is, in addition to the systems A and B used by Alice and Bob for key generation, we assume that Alice and Bob also hold some ancillary systems, A' and B', often called the shield part. Since we assume that Eve has no access to the shield part, Eve is further limited in her ability to eavesdrop. Therefore, Alice and Bob can derive a higher key generation rate than the case when Eve does have access to the shield part.

An upshot is that even a bound entangled state can give a secure key. A bound state is one whose formation (via local operations and classical communications, LOCCs) requires entanglement, but which does not give any distillable entanglement. In other words, even though no entanglement can be distilled from a bound entangled state, private states (a twisted version of entangled states) *can* be distilled from a bound entangled state.

In summary, secure key generation is a more general theory than entanglement distillation.

4. Complementary principle

   Another approach to security proof is to use the complementary principle of quantum mechanics. Such an approach is interesting because it shows the deep connection between the foundations of quantum mechanics and the security of QKD. In fact, both Mayers' proof [8] and Biham, Boyer, Boykin, Mor, and Roychowdhury's proof [13] make use of this complementary principle. A clear and rigorous discussion of the complementary principle approach to security proof has recently been achieved by Koashi [22].

   The key insight of Koashi's proof is that Alice and Bob's ability to generate a random secure key in the Z-basis (by a measurement of the Pauli spin matrix σ_Z) is equivalent to the ability for Bob to help Alice prepare an eigenstate in the complementary, i. e., X-basis (σ_X), with their help of the shield. The intuition is that an X-basis eigenstate, for example, $|+\rangle_A = 1/\sqrt{2}(|0\rangle_A + |1\rangle_A)$, when measured along the Z-basis, gives a random answer.

5. Other ideas for security proofs

   Here we discuss two other ideas for security proofs, namely, a) device-independent security proofs and b) security from the causality constraint. Unfortunately, these ideas are still very much under development and so far a complete version of a proof of unconditional security of QKD based on these ideas with a finite key rate is still missing.

   Let us start with a) device-independent security proofs So far we have assumed that Alice and Bob know what their devices are doing exactly. In practice, Alice and Bob may not know their devices for sure. Recently, there has been much interest in the idea of device-independent security proofs. In other words, how to prove security when Alice and Bob's devices cannot be trusted. See, for example, [23]. The idea is to look only at the input and output variables. A handwaving argument goes as follows. Using their probability distribution, if one can demonstrate the violation of some Bell inequalities, then one cannot explain the data by a separable system. How to develop such a handwaving argument into a full proof of unconditional security is an important question.

The second idea b) security from the causality constraint is even more ambitious. The question that it tries to address is the following. How can one prove security when even quantum mechanics is wrong? In [24] and references cited therein, it was suggested that perhaps a more general physical principle such as the no-signaling requirement for space-like observables could be used to prove the security of QKD.

Classical Post-processing Protocols

As noted in Sect. "Introduction", after the quantum communication phase, Alice and Bob then proceed with the classical communication phase. In order to generate a secure key, Alice and Bob have to know what classical post-processing protocol to apply to the raw quantum data. This is a highly non-trivial question. Indeed, a priori, given a particular procedure for classical post-processing, it is very hard to know whether it will give a secure key or what secure key will be generated. In fact, it is sometimes said that in QKD, the optical part is easy, the electronics part is harder, but the hardest part is the classical post-processing protocol. Fortunately, security proofs often give Alice and Bob clear ideas on what classical post-processing protocol to use. This highlights the importance for QKD practitioners to study the security proofs of QKD.

Briefly stated, the classical post-processing protocol often consists of a) test for tampering and b) key generation. In a) test for tampering, Alice and Bob may randomly choose a fraction of the signals for testing. For example, by broadcasting the polarizations of those signals, they can work out the bit error rate of the test signals. Since the test signals are randomly chosen, they have high confidence on the bit error rate of the remaining signals. If the bit error rate of the tested signal is higher than a prescribed threshold value, Alice and Bob abort. On the other hand, if the bit error rate is lower than or equal to the prescribed value, they proceed with the key generation step with the remaining signals. They first convert their polarization data into binary strings, the raw keys, in a prescribed manner. For example, they can map a vertical or 45-degrees photon to "0" and a horizontal or 135-degrees photon to "1". As a result, Alice has a binary string x and Bob has a binary string y. However, two problems remain. First, Alice's string may differ from Bob's string. Second, since the bit error rate is non-zero, Eve has some information about Alice's and Bob's string. The key generation step may be divided into the following stages:

1. Classical pre-processing
   This is an optional step. Classical pre-processing has the advantage of achieving a higher key generation rate and tolerating a higher bit error rate [25,26,27].
   Alice and Bob may pre-process their data by either a) some type of error detection algorithm or b) some random process. An example of an error detection algorithm is a B-step [25], where Alice randomly permutes all her bits and broadcasts the parity of each adjacent pair. In other words, starting from $\vec{x} = (x_1, x_2, \ldots, x_{2N-1}, x_{2N})$, Alice broadcasts a string $\vec{x_1} = (x_{\sigma(1)} + x_{\sigma(2)} \mod 2, x_{\sigma(3)} + x_{\sigma(4)} \mod 2, \ldots, x_{\sigma(2N-1)} + x_{\sigma(2N)} \mod 2)$, where σ is a random permutation chosen by Alice. Moreover, Alice informs Bob which random permutation, σ, she has chosen. Similarly, starting from $\vec{y} = (y_1, y_2, \ldots, y_{2N-1}, y_{2N})$, Bob randomly permutes all his bits using the same σ and broadcasts the parity bit of all adjacent pairs. I. e. Bob broadcasts $\vec{y_1} = (y_{\sigma(1)} + y_{\sigma(2)} \mod 2, y_{\sigma(3)} + y_{\sigma(4)} \mod 2, \ldots, y_{\sigma(2N-1)} + y_{\sigma(2N)} \mod 2)$. For each pair of bits, Alice and Bob keep the first bit iff their parities of the pair agree. For instance, if $x_{\sigma(2k-1)} + x_{\sigma(2k)} \mod 2 = y_{\sigma(2k-1)} + y_{\sigma(2k)} \mod 2$, then Alice keeps $x_{\sigma(2k-1)}$ and Bob keeps $y_{\sigma(2k-1)}$ as their new key bit. Otherwise, they drop the pair $(x_{\sigma(2k-1)}, x_{\sigma(2k)})$ and $(y_{\sigma(2k-1)} x_{\sigma(2k)})$ completely. Notice that the above protocol is an error detection protocol. To see this, let us regard the case where $x_i \neq y_i$ as an error during the quantum transmission stage. Suppose that for each bit, i, the event $x_i \neq y_i$ occurs with an independent probability p. For each k, the B-step throws away the cases where a single error has occurred for the two locations $\sigma(2k-1)$ and $\sigma(2k)$ and keeps the cases when either no error or two errors has occurred. As a result, the error probability after the B-step is reduced from $O(p)$ to $O(p^2)$. The random permutation of all the bit locations ensures that the error model can be well described by an independent identical distribution (i.i.d.).
   An example of a random process is an adding noise protocol [26] where, for each bit x_i, Alice randomly and independently chooses to keep it unchanged or flip it with probabilities, $1 - q$ and q respectively, where the probability q is publicly known.
2. Error correction
   Owing to noises in the quantum channel, Alice and Bob's raw keys, x and y, may be different. Therefore, it is necessary for them to reconcile their keys. One simple way of key reconciliation is forward key reconciliation, whose goal is for Alice to keep the same key x and Bob to change his key from y to x. Forward key reconciliation can be done by either standard error correcting codes such as low-density-parity-check (LDPC) codes

or specialized (one-way or interactive) protocols such as Cascade protocol [28].

3. Privacy amplification
   To remove any residual information Eve may have about the key, Alice and Bob may apply some algorithm to compress their partially secure key into a shorter one that is almost perfectly secure. This is called privacy amplification. Random hashing and a class of two-universal hash functions are often suitable for privacy amplification. See for example [29] and [18] for discussion.

Composability

A key generated in QKD is seldom used in isolation. Indeed, one may concatenate a QKD process many times, using a small part of the key for authentication each time and the remaining key for other purposes such as encryption. It is important to show that using QKD as a sub-routine in a complicated cryptographic process does not create new security problems. This issue is called the composability of QKD and, fortunately, has been solved in [30].

Composability of QKD is not only of academic interest. It allows us to refine our definition of security [30,31] and directly impacts on the parameters used in the classical post-processing protocol.

Security Proofs of Practical QKD Systems

As will be discussed in Sect. "Experimental Fundamentals", practical QKD systems suffer from real-life imperfections. Proving the security of QKD with practical systems is a hard problem. Fortunately, this has been done with semi-realistic models by Inamori, Lütkenhaus and Mayers [32] and in a more general setting by Gottesman, Lo, Lütkenhaus, and Preskill [33].

Experimental Fundamentals

Quantum cryptography can ensure the secure communication between two or more legitimate parties. It is more than a beautiful idea. Conceptually, it is of great importance in the understandings of both information and quantum mechanics. Practically, it can provide an ultimate solution for confidential communications, thus making everyone's life easier.

By implementing the quantum crypto-system in the real life, we can test it, analyze it, understand it, verify it, and even try to break it. Experimental QKD has been performed since about 1989 and great progress has been made. Now, you can even buy QKD systems on the market.

A typical QKD set-up includes three standard parts: a source (Alice), a channel, and a detection system (Bob).

A Brief History

The First Experiment The proposal of BB84 [1] protocol seemed to be simple. However, it took another five years before it was first experimentally demonstrated by Bennett, Bessette, Brassard, Salvail, and Smolin in 1989 [34]. This first demonstration was based on polarization coding. Heavily attenuated laser pulses instead of single photons were used as quantum signals, which were transmitted over 30 cm open air at a repetition rate of 10 Hz.

From Centimeter to Kilometer 30 cm is not that appealing for practical communications. This short distance is largely due to the difficulty of optical alignment in free space. Switching the channel from open air to optical fiber is a natural choice. In 1993, Townsend, Rarity, and Tapster demonstrated the feasibility of phase-coding fiber-based QKD over 10 km telecom fiber [35] and Muller, Breguet, and Gisin demonstrated the feasibility of polarization-coding fiber-based QKD over 1.1 km telecom fiber [36]. (Also, Jacobs and Franson demonstrated both free-space [37] and fiber-based QKD [38].) These are both feasibility demonstrations by means that neither of them applied random basis choosing at Bob's side. Townsend's demonstration seemed to be more promising than Muller's due to the following reasons.

1. The polarization dispersion in fibers is highly unpredictable and unstable. Therefore polarization coding requires much more controlling in the fiber than phase coding. In fiber-based QKD implementations, phase coding is in general more preferred than polarization coding.
2. Townsend et al. used 1310 ns laser as the source, while Muller et al. used 800 ns laser as the source. 1310 nm is the second window wavelength of telecom fibers (the first window wavelength is 1550 nm). The absorption coefficient of standard telecom fiber at 1310 nm is 0.35 dB/km, comparing to 3 dB/km at 800 nm. Therefore the fiber is more transparent to Townsend et al.'s set-up.

P. D. Townsend demonstrated QKD with Bob's random basis selection in 1994 [39]. It was phase-coding and was over 10 km fiber. The source repetition rate was 105 MHz (which is quite high even by today's standard) but the phase modulation rate was 1.05 MHz. This mismatch brought a question mark on its security.

Getting Out of the Lab It is crucial to test QKD technique in the field deployed fiber. Muller, Zbinden, and Gisin successfully demonstrated the first QKD experiment outside the labs with polarization coding in 1995 [40,41]. This demonstration was performed over 23 km installed optical fiber under Lake Geneva. (Being under water, quantum communication in the optical fiber suffered less noise.)

There is less control over the field deployed fiber than fiber in the labs. Therefore its stabilization becomes challenging. To solve this problem, A. Muller et al. designed the "plug & play" structure in 1997 [42]. A first experiment of this scheme was demonstrated by H. Zbinden et al. in the same year [43]. Stucki, Gisin, Guinnard, Robordy, and Zbinden later demonstrated a simplified version of the "plug & play" scheme under Lake Geneva over 67 km telecom fiber in 2002 [44].

With a Coherent Laser Source The original BB84 [1] proposal required a single photon source. However, most QKD implementations are based on faint lasers due to the great challenge to build the perfect single photon sources. In 2000, the security of coherent laser based QKD systems was analyzed first against individual attacks [45]. Finally, the unconditional security of coherent laser based QKD systems was proven in 2001 [32] and in a more general setting in 2002 [33]. Gobby, Yuan, and Shields demonstrated an experiment based on [45] in 2005 [46] (Note that this work was claimed to be unconditionally secure. However, due to the limit of [45], this is only true against individual attacks rather than the most general attack).

The security analysis in [32,33] will severely limit the performance of unconditionally secure QKD systems. Fortunately, since 2003 the decoy state method has been proposed [47,48,49,50,51,52] by Hwang and extensively analyzed by our group at the University of Toronto and by Wang. The first experimental demonstration of decoy state QKD was reported by us in 2006 [53] over 15 km telecom fiber and later over 60 km telecom fiber [54]. Subsequently, decoy state QKD was further demonstrated by several other groups [55,56,57,58,59]. The readers may refer to Subsect. "Decoy State Protocols" for details of decoy state protocols.

Sources

Single Photon Sources are demanded by the original BB84 [1] proposal. Suggested by its name, the single photon sources are expected to generate exactly one photon on demand. The bottom line for a single photon source is that no more than one photon can be generated at one time. It is very hard to build a perfect single photon source (i. e., no multi-photon production). Despite tremendous effort made by many groups, perfect single photon source is still far from practical. Fortunately, the proposal and implementation of decoy state QKD (see Subsect. "Decoy State Protocols") make it unnecessary to use single photon sources in QKD.

Parametric Down-Conversion (PDC) Sources are often used as the entanglement source. Its principle is that a high energy (~ 400 nm) photon propagates through a highly non-linear crystal (usually BBO), producing two entangled photons with frequency halved. PDC sources are usually used for entanglement-based QKD systems (e. g. Ekert [3] protocol).

PDC sources are also used as "triggered single photon sources", in which Alice possesses a PDC source and monitor one arm of its outputs. In case that Alice sees a detection, she knows that there is one photon emitted from the other arm. Experimental demonstration of QKD with PDC sources is reported in [60].

Attenuated Laser Sources are the most commonly used sources in QKD experiments. They are essentially the same as the laser sources used in classical optical communication except for that heavy attenuation is applied on them. They are simple and reliable, and they can reach Gigahertz with little challenge. In BB84 system and differential-phase-shift-keying (DPSK) system (to be discussed in Subsect. "Differential-Phase-Shift-Keying (DPSK) Protocols"), the laser source is usually attenuated to below 1 photon per pulse. In Gaussian-modulated coherent-state (GMCS) system (to be discussed in Subsect. "Gaussian-Modulated Coherent State (GMCS) Protocol"), the laser source is usually attenuated to around 100 photons per pulse.

Attenuated laser sources used to be considered to be non-ideal for BB84 systems as they always have non-negligible probability of emitting multi-photon pulses regardless how heavily they are attenuated. However, the discovery and implementation of decoy method [47,48,49,50,51,52,53,54,55,56,57,58] made coherent laser source much more appealing. With decoy method, it is possible to make BB84 system with laser source secure without significant losses on the performance.

Channels

Standard Optical Single-Mode Fiber (SMF) is the most popular choice for now. It can connect two arbitrary points, and can easily be extended to networks.

Moreover, it is deployed in most developed urban areas.

SMF has two "window wavelengths": one is 1310 nm and the other is 1550 nm. The absorptions at these two wavelengths are particularly low ($\sim$ 0.35 dB/km at 1310 nm, and $\sim$ 0.21 dB/km at 1550 nm). Nowadays most fiber-based QKD implementations use 1550 nm photons as information carriers.

The main disadvantage of optical fiber is its birefringence. The strong polarization dispersion made it hard to implement polarization-coding system. Also it has strong spectral dispersion, which affects the high speed (10+ GHz) QKD systems heavily [61] as the pulses are broadened and overlap with each other. For this reason, the loss in fibers (0.21 dB/km at 1550 nm) puts an limit on the longest distance that a fiber-based QKD system can reach.

Free Space is receiving more and more attention recently. It is ideal for the polarization coding. There is negligible dispersion on the polarization and the frequency. However, the alignment of optical beams can be challenging for long distances, particularly due to the atmospheric turbulence. Notice that open-air QKD requires a direct line of sight between Alice and Bob (unless some forms of mirrors are used). Buildings and mountains are serious obstacles for open-air QKD systems.

The greatest motivation for open-air QKD scheme is the hope for ground-to-satellite [62] and satellite-to-satellite quantum communication. As there is negligible optical absorption in the outer space, we may be able to achieve inter-continental quantum communication with free-space QKD.

Detection Systems

InGaAs-APD Single Photon Detectors are the most popular type of single-photon detectors in fiber-based QKD and they are commercially available. InGaAs-APD Single Photon Detectors utilize the avalanche effect of semiconductor diodes. A strong biased voltage is applied on the InGaAs diode. The incident photon will trigger the avalanche effect, generating a detectable voltage pulse. The narrow band gap of InGaAs made it possible to detect photons at telecom wavelengths (1550 nm or 1310 nm).

InGaAs APD based single photon detectors have simple structure and commercially packaged. They are easy to calibrate and operate. The reliability of InGaAs APD is relatively high. They normally work at $-50°C$ to $-110°C$ to lower the dark count rate. This temperature can be easily achieved by thermal-electric coolers. The detection efficiency of InGaAs-APD based single photon detectors is usually $\sim$ 10% [63].

In single photon detectors, a key parameter (besides detection efficiency) is the dark count rate. The dark count is the event that the detector generates a detection click while no actual photon hits it (i. e. "false alarm"). The dark count rate of InGaAs single photon detector is relatively high (10^{-5} per gate; The concept of gating will be introduced below.) even if it is cooled. The after-pulse effect is that the dark count rate of the detector increases for a time period after a successful detection. This effect is serious for InGaAs single photon detectors. Therefore the blank circuit is often introduced to reduce this effect. The mechanism of the blank circuit is that the detector is set to be deactivated for a time period, which is called the "dead time", after a detection event. The dead time should be set to long enough so that when the detector is re-activated, the after-pulse effect is negligible. The dead time for InGaAs single photon detector is typically in the order of microseconds [63]. The long after pulse effect, together with the large timing jitter limits the InGaAs-APD based single photon detectors to work no faster than several megahertz. Moreover, the blank circuit reduces the detection efficiency of InGaAs-APD based single photon detectors.

An additional method to reduce the dark count rate is to apply the gating mode, i. e., the detectors are only activated when the photons are expected to hit them. Gating mode reduces the dark count rate by several orders and is thus used in most InGaAs-APD single photon detectors. However, it may open up a security loophole [64,65].

There is a trade-off between the detection efficiency and the dark count rate. As the biased voltage on an APD increases, both the detection efficiency and the dark count rate increase.

Recently, it has been reported that, by gating an InGaAs detector in a sinusoidal manner, it is possible to reduce the dead time and operate a QKD system at 500 MHz, see [66]. This result seems to be an important development which could make InGaAs detectors competitive with newer single photon detector technologies such as SSPDs (to be introduced below).

Si-APD Single Photon Detectors are ideal for detection of visible photons (say 800 nm). They have negligible dark count rate and can work at room temperature. They are very compact in size. More importantly, they have high detection efficiency ($>$ 60%) and can work at gigahertz. These detectors are ideal for free space QKD

systems. However, the band gap of silicon is too large to detect photons at telecom wavelength (1550 nm or 1310 nm), and the strong attenuation of telecom fiber on visible wavelengths makes it impractical to use visible photons in long distance fiber-based QKD systems.

Parametric Up-Conversion Single Photon Detectors try to use Si-APD to detect telecom wavelength photons. It uses periodically poled lithium niobate (PPLN) waveguide and a pumping light to up-convert the incoming telecom frequency photons into visible frequency, and uses Si-APD to detect these visible photons. The high speed and low timing-jitter of Si-APDs make it possible to perform GHz QKD on fiber-based system with up-conversion single photon detectors [67,68].

The efficiency of up-conversion detectors is similar to that of InGaAs APD single photon detectors. There is also a trade-off between the detection efficiency and the dark count rate. When increasing the power of the pumping light, the conversion efficiency will increase, improving the detection efficiency. Meanwhile, more pumping photons and up-converted pumping photons (with frequency doubled) will pass through the filter and enter the Si-APD, thus increasing the dark count rate [68].

Transiting-Edge Sensor (TES) is based on critical state superconductor rather than semiconductor APDs. It uses squared superconductor (typically tungsten) thin film as "calorimeter" to measure the electron temperature. A biased voltage is applied on the thin film to keep it in critical state. Once one or more photons are absorbed by the sensor, the electron temperature will change, leading to a change of the current. This current change can be detected by a superconductive quantum-interference device (SQUID) array [69].

The TES single photon detectors can achieve very high detection efficiency (up to 89%) at telecom wavelength. The dark count rate is negligible. Moreover, TES detectors can resolve photon numbers. This is because the electron temperature change is proportional to the number of photons that have been absorbed.

The thermal nature of TES detectors limits their counting rates. Once some photons were absorbed by the sensor, it would take a few microseconds before the heat is dissipated to the substrate. This long relaxation time limits the counting rate of TES detector to no more than a few megahertz [69]. This is a major drawback of TES.

The bandwidth of TES detector is extremely wide. The detector is sensitive to all the wavelengths. Even the black body radiation from the fiber or the environment can trigger the detection event, thus increasing the dark count rate. To reduce the dark counts caused by other wavelengths, a spectral filter is necessary. However, this will increase the internal loss and thus reduce the detection efficiency.

One of the greatest disadvantage of TES detector is its working temperature: 100 mK. This temperature probably requires complicated cooling devices [69].

Superconductive Single Photon Detectors (SSPDs) also use superconductor thin film to detect incoming photons. However, instead of using a piece of plain thin film, a pattern of zigzag superconductor (typically NbN) wire is formed. The superconductive wire is set to critical state by applying critical current through it. Once a photon hits the wire, it heats a spot on the wire and makes the spot over-critical (i. e., non-superconductive). As the current is the same as before, the current density in the areas around this hot-spot increases, thus making these areas non-superconductive. As a result, a section of the wire becomes non-superconductive, and a voltage spike can be observed as the current is kept constant [61].

The SSPD can achieve very high (up to 10 GHz) counting rate. This is because the superconductor wire used in SSPD can dissipate the heat in tens of picoseconds. It also has very low dark count rate (around 10 Hz) due to the superconductive nature. The SSPDs should be able to resolve the incident photon number in principle. However, photon number resolving SSPDs have never been reported yet.

The efficiency of SSPD is lower than that of TES. This is because only part ($\sim$ 50%) of the sensing area is covered by the wire. The fabrication of such complicated zigzag superconductor wire with smooth edge is also very challenging.

The working temperature of SSPD ($\sim$3K) is significantly higher than that of TES. SSPDs can work in closed-cycle refrigerator. Moreover, this relatively high working temperature significantly reduces the relaxation time [61].

Homodyne Detectors are used to count the *photon number* of a very weak pulse ($\sim$ 100 photons). The principle is to use a very strong pulse (often called the local oscillator) to interfere with the weak pulse. Then use two photo diodes to convert the two resulting optical pulses into electrical signals, and make a substraction between the two electrical signals.

The homodyne detectors are in general very efficient as there is always some detection output given some input signals. However, the noise of the detectors as well as in the electronics is very significant. Moreover,

the two photo diodes in the homodyne detector have to be identical, which is hard to meet in practice.
The homodyne detectors are commonly used in GMCS QKD systems, and they are so far the only choice for GMCS QKD systems. Recently, homodyne detectors have also been used to implement the BB84 protocol.

Truly Quantum Random Number Generators

An important but often under-appreciated requirement for QKD is a high data-rate truly quantum random number generator (RNG). An RNG is needed because most QKD protocols (with the exception of a passive choice of bases in an entanglement-based QKD protocol) require Alice to choose actively random bases/signals. Given the high repetition rate of QKD, such a RNG must have a high data-rate. To achieve unconditional security, a standard software-based pseudo-random number generator cannot be used because it is actually deterministic. So, a high data rate quantum RNG is a natural choice. Incidentally, some firms such as id Quantique do offer commercial quantum RNGs. Unfortunately, it is very hard to generate RNG by quantum means at high-speed. In practice, some imperfections/bias in the numbers generated by a quantum RNG are inevitable. The theoretical foundation of QKD is at risk because existing security proofs all assume the existence of perfect RNGs and do not apply to imperfect RNGs.

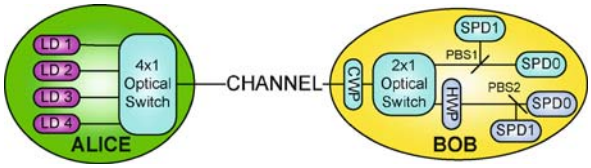
Experimental Implementation of BB84 Protocol

In this section, we will focus mainly on the optical layer. We will skip several important layers. In practice, the control/electronics layer is equally important. Moreover, it is extremely challenging to implement the classical post-processing layer in real-time, if one chooses block sizes of codes to be long enough to achieve unconditional security.

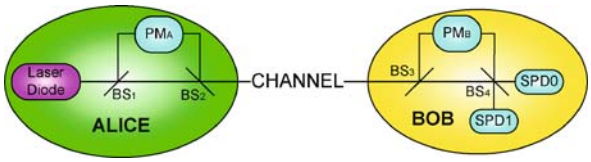
Polarization Coding

Polarization coding usually uses four laser sources generating the four polarization states of BB84 [1] protocol. A conceptual schematic is shown in Fig. 3. Note that due to polarization dispersion of a fiber, usually people need some compensating like the waveplates.

The polarization compensation should be implemented dynamically as the polarization dispersion in the fiber changes frequently. This is solved by introducing the electrical polarization controller in [58].



Quantum Cryptography, Figure 3
Conceptual schematic for polarization-coding BB84 QKD system. LD: Laser Diode; CWP: Compensating Wave Plate; HWP: Half Wave Plate; PBS: Polarizing Beam Splitter; SPD: Single Photon Detector



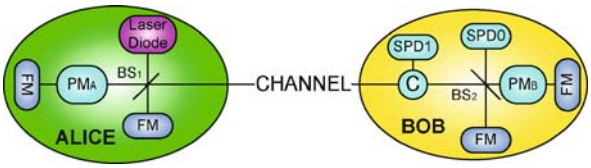
Quantum Cryptography, Figure 4
Conceptual schematic for double Mach–Zehnder interferometer phase-coding BB84 QKD system. PM: Phase Modulator; BS: Beam Splitter; SPD: Single Photon Detector

Phase Coding

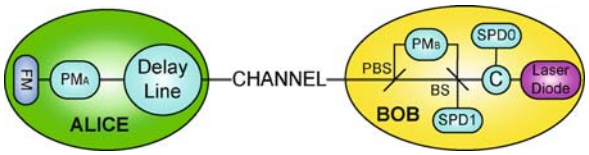
Original Scheme is basically a big interferometer as shown in Fig. 2. However it is not practical as the stability of such a huge interferometer is extremely poor.

Double Mach–Zehnder Interferometer Scheme is an improved version of the original proposal. It has two interferometers and there is only one channel connecting Alice and Bob (comparing to the two channels in the original proposal). A conceptual set-up is shown in Fig. 4. We can see that the two signals travel through the same channel. They only propagate through different paths locally in the two Mach–Zehnder interferometers. Therefore people only need to compensate the phase drift of the local interferometers (the polarization drift in the channel still needs to be compensated). This is a great improvement over the original proposal. However, the local compensation has to be implemented in real time. This is quite challenging. An example set-up that implemented the real time compensation of both polarization and phase drifting is reported in [70].

Faraday–Michelson Scheme is an improved version of the double Mach–Zehnder interferometer scheme. It still has two Mach–Zehnder interferometers but each interferometer has only one beam splitter. The light propagates through the same section of fiber twice due to the Faraday mirror. The schematic is shown in Fig. 5. We can see that the polarization drift is self-compensated. This is a great advance in uni-direc-



Quantum Cryptography, Figure 5
Conceptual schematic for Faraday–Michelson phase-coding BB84 QKD system. FM: Faraday Mirror, PM: Phase Modulator; BS: Beam Splitter; C: Circulator; SPD: Single Photon Detector



Quantum Cryptography, Figure 6
Conceptual schematic for "Plug & Play" phase-coding BB84 QKD system. FM: Faraday Mirror, PM: Phase Modulator; BS: Beam Splitter; PBS: Polarizing Beam Splitter; C: Circulator; SPD: Single Photon Detector

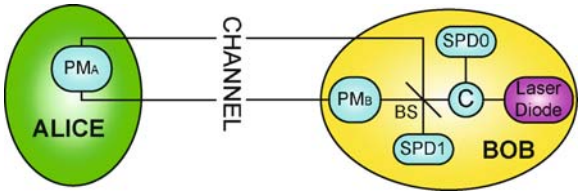
tional QKD implementation and is first proposed and implemented in [71].

Nonetheless, the phase drift of local interferometers still needs compensation. Due to the fast fluctuation of phase drift (a drift of 2π usually takes a few seconds), this compensation should be done in real-time. A Faraday–Michelson decoy state QKD implementation over 123.6 km has been reported in [59].

Plug & Play Scheme is another improved version of the double Mach–Zehnder interferometer scheme. It has only one Mach–Zehnder interferometer and the light propagates through the same channel and interferometer twice due to the faraday mirror on Alice's side. A conceptual set-up is shown in Fig. 6. We can see that both the polarization drift and the phase drift are automatically compensated. A "Plug & Play" scheme based decoy state QKD implementation over 60 km has been reported in [54].

Nonetheless, the bi-directional design brings complications to security as Eve can make sophisticated operations on the bright pulses sent from Bob to Alice. This is often called the "Trojan horse" attack [72]. Recently, the security of "Plug and Play" QKD system has been proven in [73].

Sagnac Loop Scheme Another bi-directional optical layer design is to use a Sagnac loop where the quantum signal is encoded in the relative phase between the clockwise and counter-clockwise pulses that go through the loop. The typical schematic is shown in Fig. 7.



Quantum Cryptography, Figure 7
Conceptual schematic for Sagnac loop phase-coding BB84 QKD system. PM: Phase Modulator; BS: Beam Splitter; C: Circulator; SPD: Single Photon Detector

Sagnac loop QKD is simple to set up and can be easily used in a network setting with a loop topology. However, its security analysis is highly non-trivial.

Other Quantum Key Distribution Protocols

Given the popularity of the BB84 protocol, why should people be interested in other protocols? There are at least three answers to this question. First, to better understand the foundations of QKD and its generality, it is useful to have more than one protocol. Second, different QKD protocols may have different advantages and disadvantages. They may require different technologies to implement. Having different protocols allows us to compare and contrast them. Third, while it is possible to implement standard BB84 protocol with attenuated laser pulses, its performance in terms of key generation rate and maximum transmission distance is somewhat limited. Therefore, we have to study other protocols. Since, from a practical stand point, the third reason is the most important one, we will elaborate on it in the following paragraph.

The original BB84 [1] proposal requires a single photon source. However, most QKD implementations are based on faint lasers due to the great challenge to build perfect single photon sources. Faint laser pulses are weak coherent states that follow Poisson distribution for the photon number. The existence of multi-photon signals opens up new attacks such as photon-number-splitting attack. The basic idea of a photon-number-splitting attack is that Eve can introduce a photon-number-dependent transmittance. In other words, she can selectively suppress single-photon signals and transmit multi-photon signals to Bob. Notice that, for each multi-photon signal, Eve can beamsplit it and keep one copy for herself, thus allowing her to gain a lot of information about the raw key.

The security of coherent laser based QKD systems was analyzed first against individual attacks in 2000 [45], then eventually for a general attack in 2001 [32] and 2002 [33]. Unfortunately, unconditionally secure QKD

based on conventional BB84 protocol [32,33] will severely limit the performance of QKD systems. Basically, Alice has to attenuate her source so that the expected number μ of photon per pulse is of the same order as the transmittance, η. As a result, the key generation rate will scale only quadratically with the transmittance of the channel.

Some of the protocols discussed in the following subsections may dramatically improve the performance of QKD over standard BB84 protocol. For instance, the decoy state protocol has been proven to provide a key generation rate that scales linearly with the transmittance of the channel and has been successfully implemented in experiments.

We conclude with some simple alternative QKD protocols. In 1992, Bennett proposed a protocol (B92) that makes use of only two non-orthogonal states [2]. A six-state QKD protocol was first noted by Bennett and coworkers [74] and some years later by Bruss [75]. It has an advantage of being symmetric. Even QKD protocols with orthogonal states have been proposed [76]. Efficient BB84 and six-state QKD protocols have been proposed and proven to be secure by Lo, Chau and Ardehali [77]. A Singaporean protocol has also been proposed. Recently, Gisin and co-workers proposed a one-way coherent QKD scheme [78].

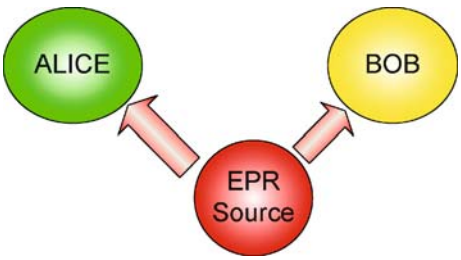
Entanglement-Based Protocols

Proposals In 1991, Ekert proposed the first entanglement based QKD protocol, commonly called E91 [3]. The basic idea is to test the security of QKD by using the violation of Bell's inequality. Note that one can also implement the BB84 protocol by using an entanglement source. Imagine Eve prepares an entangled state of a pair of qubits and sends one qubit to Alice and the second qubit to Bob. Each of Alice and Bob randomly chooses one of the two conjugate bases to perform a measurement.

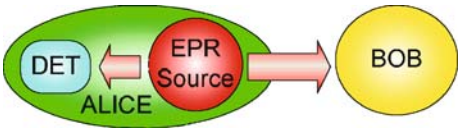
Implementations The key part of entanglement-based quantum cryptography is to distribute an entangled pair (usually EPR pair) to two distant parties, Alice and Bob.

Polarization entanglement is preferred in QKD as it is easy to measure the polarization (typically via polarizing beam splitter). The air has negligible birefringence and thus is the perfect channel for polarization-entanglement QKD.

In free-space QKD, atomspherical turbulence may shift the light beam. Therefore the collection of incident photon is challenging. Usually large diameter optical telescope is needed to increase the collection efficiency.



Quantum Cryptography, Figure 8
Conceptual schematic entanglement-based QKD system with the source in the middle of Alice and Bob



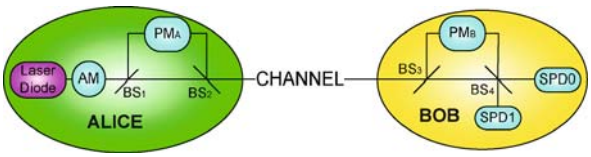
Quantum Cryptography, Figure 9
Conceptual schematic entanglement-based QKD system with the source at Alice's side. DET: Alice's detection system

A standard approach is to put the entanglement source right in the middle of Alice and Bob, see Fig. 8. Once an entangled pair is generated, the two particles are directed to different destinations. Alice and Bob measure the particles locally, and keep the result as the bit value. This approach has potential in the ground-satellite intercontinental entanglement distribution, in which the entanglement source is carried by the satellite and the entangled photons are sent to two distant ground stations. A recent source-in-the-middle entanglement-based quantum communication work over 13 km and is reported in [79].

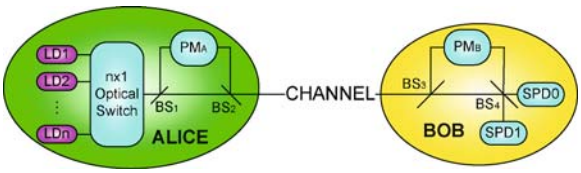
A simpler version is to include the entanglement source in Alice's side locally, see Fig. 9. Once Alice generates an entangled pair, she keeps one particle and send the other to Bob. Both Alice and Bob measure the particle locally and keep the result as the bit value. This approach is significant simpler than the above design because only Bob needs the telescope and compensating parts. A recent experiment of source-in-Alice entanglement-based quantum communication over 144 km open air is reported in [80].

Decoy State Protocols

Proposals Recall that BB84 implemented with weak coherent state has a key generation rate that scales only quadratically with the transmittance. The decoy state protocol can dramatically increase the key generation rate so that it scales linearly with the transmittance. In a decoy state protocol, Alice prepares some decoy states in addition to signal states. The decoy states are the same as the



Quantum Cryptography, Figure 10
Conceptual schematic for decoy state BB84 QKD system (double Mach–Zehnder interferometer phase-coding) with amplitude modulator. PM: Phase Modulator; AM: Amplitude Modulator; BS: Beam Splitter; SPD: Single Photon Detector



Quantum Cryptography, Figure 11
Conceptual schematic for decoy state BB84 QKD system (double Mach–Zehnder interferometer phase-coding) with multiple laser diodes. LDx: Laser Diodes at different intensities; PM: Phase Modulator; BS: Beam Splitter; SPD: Single Photon Detector

signal state, except for the expected photon number. For instance, if the signal state has an average photon number μ of order 1 (e. g. 0.5), the decoy states have an average photon number ν_1, ν_2, etc. The decoy state idea was first proposed by Hwang [47], who suggested using a large ν (e. g. 2) as a decoy state. Our group provided a rigorous proof of security to decoy state QKD[48,49]. Our numerical simulations showed clearly that decoy states provide a dramatic improvement over non-decoy protocols. In the limit of infinitely many decoy states, Alice and Bob can effectively limit Eve's attack to a simple beam-splitting attack. Moreover, we proposed practical protocols. Instead of using a large ν as a decoy state, we proposed using small ν's as decoy states [48]. For instance, we proposed using a vacuum state as the decoy state to test the background and a weak ν to test the single-photon contribution. We and Wang analyzed the performance of practical protocols in detail [50,51,52].

Notice that the decoy state is a rather general idea that can be applied to other QKD sources. For instance, decoy state protocols have recently been proposed in [81,82,83] for parametric down conversion sources. For a comparison of those protocols, see [84].

Implementations The first experimental demonstration of decoy state QKD was reported by our group in 2006 first over 15 km telecom [53] fiber and later over 60 km telecom fiber [54]. Subsequently, the decoy state QKD was further demonstrated experimentally by several groups worldwide [55,56,57,58].

The implementation of decoy state QKD is straightforward. The key part is to prepare signals with different intensities. A simple solution is to use an amplitude modulator to modulate the intensities of each signal to the desired level, see Fig. 10. Decoy state QKD implementations using amplitude modulator to prepare different states are reported in [53,54,55,56,57].

The amplitude modulator has the disadvantage that the preparation of vacuum state is quite challenging. An alternative solution is to use laser diodes of different intensities to generate different states, see Fig. 11. This solution requires multiple laser diodes and high-speed optical switch, and is thus more complicated than the amplitude modulator solution. It is also challenging to guarantee that all the laser diodes are identical so that Eve cannot tell which source generates some specific pulse. Nonetheless, perfect vacuum states can be easily prepared in this way. Decoy state QKD implementations using multiple laser diodes are reported in [58,59]. Decoy state with a parametric down conversion source has been experimentally implemented in [85].

Strong Reference Pulse Protocols

Proposals The proposal of strong reference pulse QKD dated back to Bennett's 1992 paper [2]. The idea is to add a strong reference pulse, in addition to the signal pulse. The quantum state is encoded in the relative pulse between the reference pulse and the signal pulse. Bob decodes by splitting a part of the strong reference pulse and interfering it with the signal pulse. The strong reference pulse implementation can counter the photon number splitting attack by Eve because it removes the *neutral* signal in the QKD system. Recall that in the photon number splitting attack, Eve suppresses single photon signals by sending a vacuum. This works because the vacuum is a neutral signal that leads to no detection. In contrast, in a strong reference pulse implementation of QKD, a vacuum signal is not a neutral signal. Indeed, if Eve replaces the signal pulse by a vacuum and keeps the strong reference pulse unchanged, then the interference experiment by Bob will give *non-zero* detection probability and a random outcome of "0" or "1". On the other hand, if Eve removes both the signal and the reference pulses, then Bob may detect Eve's attack by monitering the intensity of the reference pulse, which is supposed to be strong.

In some recent papers, the unconditional security of B92 QKD with strong reference pulse has been rigorously proven. However, those proofs require Bob's system

to have certain properties and do not apply to standard threshold detectors.

Implementations The B92 [2] protocol is simpler to implement than BB84 [1] protocol. However, its weakness in security limits people's interest on its implementation. A recent implementation of B92 protocol over 200 m fiber is reported in [86].

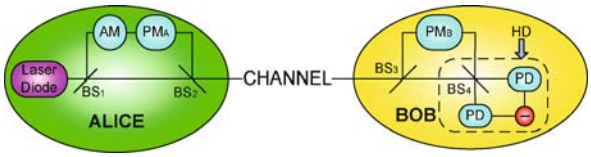
Gaussian-Modulated Coherent State (GMCS) Protocol

Proposals Instead of using discrete qubit states as in the BB84 protocol, one may also use continuous variables for QKD. Early proposals of continuous variables QKD use squeezed states, which are experimentally challenging. More recently, gaussian-modulated coherent states have also been proposed for QKD. Since a laser naturally emits a coherent state, compared to a squeezed state QKD proposal, a GMCS QKD protocol is experimentally more feasible. In GMCS QKD, Alice sends Bob a sequence of coherent state signals. For each signal, Alice draws two random numbers X_A and P_A from a set of Gaussian random number with a mean of zero and a variance of $V_A N_0$ and sends a coherent state $|X_A + iP_A\rangle$ to Bob. Bob randomly chooses to measure either the X quadrature of the P quadrature with a phase modulator and a homodyne detector. After performing his measurement, Bob informs Alice which quadrature he has performed for each pulse, through an authenticated public classical channel. Alice drops the irrelevant data and keeps only the quadrature that Bob has measured. Alice and Bob now share a set of correlated Gaussian variables which they regard as the raw key. Alice and Bob randomly select a subset of their signals and publicly broadcast their data to evaluate the excess noise and the transmission efficiency of the quantum channel. If the excess noise is higher than some prescribed level, they abort. Otherwise, Alice and Bob perform key generation by some prescribed protocol.

An advantage of a GMCS QKD is that every signal can be used to generate a key, whereas in qubit-based QKD such as the BB84 protocol losses can substantially reduce the key generation rate. Therefore, it is commonly believed that for short-distance (say < 15 km) applications, GMCS QKD may give a higher key generation rate.

GMCS QKD has been proven to be secure only against individual attacks. The security of GMCS QKD against the most general type of attack – joint attack – remains an open question.

Implementations GMCS protocol has significant advantage over the BB84 [1] protocol at short distances. It was first implemented by F. Grosshans et al. in 2003 [87]. It was shown to be working with channel loss up to 3.1 dB, which is equivalent to the loss of 15 km telecom fiber. Nonetheless, the strong spectral and polarization dispersion of telecom fiber made it challenging to build up a fiber-based GMCS system. Lodewyck, Debuisschert, Tualle-Brouri, and Grangier built the first fiber-based GMCS system in 2005 [88] but only over a few meters. This distance was largely extended to 14 km by Legré, Zbinden, and Gisin in 2006 [89] with the introduction of the "plug & play" design, which brought questions on its security. The uni-directional GMCS QKD has been later implemented over 5 km optical fiber by Qi, Huang, Qian, and Lo in 2007 [90] and over 25 km optical fiber by J. Lodewyck et al. in 2007 [91].

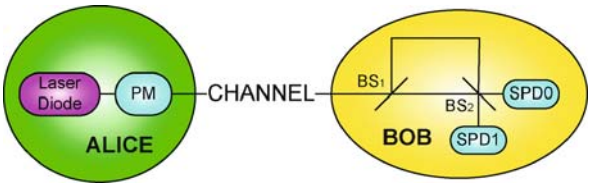


Quantum Cryptography, Figure 12
Conceptual schematic for Gaussian-modulated Coherent State QKD system. PM: Phase Modulator; AM: Amplitude Modulator; BS: Beam Splitter; PD: Photo Diode; HD: Homodyne Detector (inside *dashed box*)

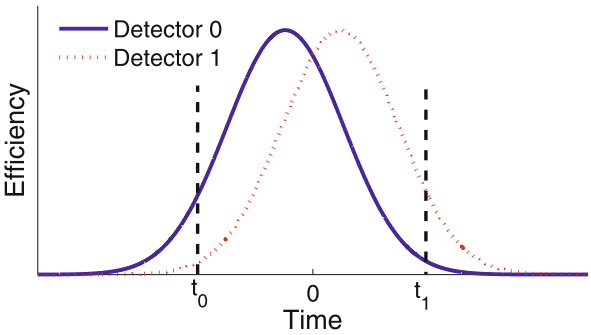
GMCS QKD requires dual-encoding on both amplitude quadrature and phase quadrature, and homodyne detection for decoding, see Fig. 12. Its implementation is in general more challenging than that of BB84 protocol.

Differential-Phase-Shift-Keying (DPSK) Protocols

Proposals In DPSK protocol, a sequence of weak coherent state pulses is sent from Alice to Bob. The key bit is encoded in the relative phase of the adjacent pulses. Therefore, each pulse belongs to two signals. DPSK protocol also defeats the photon number splitting attack by removing the neutral signal. Eve may attack a finite train of signals by measuring its total photon number and then splitting off one photon, whenever the photon number is larger than one. But, since each pulse belongs to two signals, the pulses in the boundary of the train will interfere with the pulses immediately outside the boundary. Therefore, Eve's attack does not allow her to gain full information about all the bit values associated with the train. Moreover, by splitting the signal, Eve has reduced the amplitude of the pulses at the boundary of the train. Therefore, Bob will detect Eve's presence by the higher bit error rates for the bit values between the pulses at the boundary and those just outside the boundary.



Quantum Cryptography, Figure 13
Conceptual schematic for differential phase shift keying QKD system. PM: Phase Modulator; BS: Beam Splitter; SPD: Single Photon Detector



Quantum Cryptography, Figure 14
Conceptual schematic for detection efficiency mismatch in time domain. *X axis*: Time, *Y axis*: Detector Efficiency

While DPSK protocol is simpler to implement than BB84, a proof of its unconditional security is still missing. Therefore, it is hard to quantify its secure key generation rate and perform a fair comparison with, for example, decoy state BB84 protocol. Attacks against DPSK has been studied in, for example, [92,93].

Implementations DPSK protocol is simpler in hardware design than the BB84 [1] protocol as it requires only one Mach–Zehnder interferometer, see Fig. 13. It also has the potential in high-speed applications. Honjo, Inoue, and Takahashi experimentally demonstrated this protocol with a planar light-wave circuit over 20 km fiber in 2004 [94]. This distance was soon extended to 105 km [95] in 2005. In 2007, DPSK scored both the longest and the fastest records in QKD implementations: H. Takesue et al. reported an experimental demonstration of DPSK-QKD over 200 km optical fiber at 10 GHz [61]. However, since a proof of unconditional security is still missing (see last two paragraphs), it is unclear whether the existing experiments generate any secure key. Indeed, the attacks described in [92,93] showed that with only one-way classical post-processing, all existing DPSK experiments are insecure.

Quantum Hacking

Since practical QKD systems exist and commercial QKD systems are on the market, it is important to understand how secure they really are. We remark that there is still a big gap between the theory and practice of QKD. Even though the unconditional security of practical QKD systems with semi-realistic models have been proven [32,33], practical QKD systems may still contain fatal security loopholes. From a historical standpoint, Bennett and Brassard mentioned that the first QKD system [34] was unconditionally secure to any eavesdropper who happened to be deaf! This was because the system made different sounds depending on whether the source was sending a "0" or a "1". Just by listening the sounds, an eavesdropper can learn the value of the final key. This example highlights the existence of side channels in QKD and how easy an eavesdropper might be able to break the security of a QKD system, despite the existence of security proofs.

In this section, we will sketch a few cleverly proposed quantum hacking strategies that are outside standard security proofs and their experimental implementations. We will conclude counter-measures and future outlook. Notice that we will skip eavesdropping attacks that have already been covered by standard security proofs.

Attacks

1) Large Pulse Attack. In a large pulse attack, Eve sends in a strong pulse of laser signal to, for example, Alice laboratory to try to read off Alice's phase modulator setting from a reflected pulse, see [96]. As a result, Eve may learn which BB84 state Alice is sending to Bob.

A simple counter-measure to the large pulse attack is to install an isolator in Alice's system.

2) Faked State Attack. Standard InGaAs detectors suffer from detection efficiency mismatch. More concretely, as noted in Subsect. "Detection Systems", InGaAs detectors are often operated in a gated mode. Therefore, the detection efficiency of each detector is time-dependent. Refer to Fig. 14 for a schematic diagram of the detection efficiencies of two detectors (one for "0" and one for "1") as functions of time. At the expected arrival time, the detection efficiency of the two detectors are similar. However, if the signal is chosen to arrive at some unexpected times, it is possible that the detector efficiencies of the two detectors differ greatly.

The faked state attack proposed by Makarov and coworkers [97] is an intercept-resend attack. In a faked state attack, for each signal, Eve randomly chooses one of the

two BB84 bases to perform a measurement. Eve then sends to Bob a *wrong* bit in the *wrong* basis at a time when the detector for the wrong bit has a low detection efficiency. For instance, if Eve has chosen the rectilinear basis and has found a "0" in the bit value, she then prepares a state "1" in the diagonal basis and sends it to Bob at the arrival time where detector efficiency of the detector for "0" is much higher than that of detector for "1".

Now, should Eve have chosen the wrong basis in her measurement, notice that the detection probability by Bob is greatly suppressed. For instance, in our example, if the correct basis is the diagonal basis, let us consider when happens when Bob measures the signal in the correct basis. Since the bit value resent by Eve is "1" in the diagonal basis and Bob's detector for "1" has a low detection efficiency, most likely Bob will not detect any signal. On the other hand, should Eve have chosen the correct basis in her measurement, Bob has a significant detection efficiency. For instance, in our example, if the correct basis is in fact the rectilinear basis, let us consider what happens when Bob measures in the correct basis. In this case, a bit "1" in the diagonal basis sent by Eve can be re-written as a superposition of a bit "0" and a bit "1" in the rectilinear basis. Since the detector for "0" has a much higher detection efficiency than the detector for "1", most likely Bob will detect a "0". Since "0" was exactly what was originally sent by Alice, Bob will find a rather low bit error rate, despite Eve's intercept-resend attack.

The faked state attack, while conceptually interesting, is hard to implement in practice. This is because it is an intercept-resend attack and as such involves finite detection efficiency in Eve's detectors and precise synchronization between Eve and Alice-Bob's system. For this reason, the faked state attack has never been implemented in practice.

3) Time-shift Attack. The time-shift attack was proposed by Qi, Fung, Lo, and Ma [64]. It also utilizes the detection efficiency mismatch in the time domain, but is much easier to implement than the faked states attack.

As we mentioned in the above section, typical InGaAs-APD detectors usually operate in a gated mode. That is, if the photon hits the detector at unexpected time, the two detectors may have substantially different efficiency. Therefore, Eve can simply shift the arrival time of each signal, creating large efficiency mismatch between "0"s and "1"s.

Let's take a specific example to illustrate this attack: suppose detector 0 has higher efficiency than detector 1 if the signal arrives *earlier* than the expected time, and lower efficiency than detector 1 if the signal arrives *later* than expected. Eve can simply shift the arrival time of each bit by sending it through a longer path or a shorter one. Consider the case in which Eve sends bit i through a shorter path. In this case bit i will hit the detector *earlier* than expected, thus detector 0 has much higher efficiency. If Bob reports a detection event for the ith bit, Eve can make a guess that this bit is a "0" with high probability of success.

Furthermore, Eve can carefully set how many bits should be shifted forward and how many should be shifted backward so that Bob gets similar counts of "0"s and "1"s. In this way, Bob cannot observe a mismatch between the numbers of "0"s and "1"s.

Note that the time-shift attack does not make any measurement on the qubits. Therefore, quantum information is not destroyed. That is, Eve does not change the polarization, the phase, or the frequency of any bit. This means the time-shift attack will not increase the bit error rate of the system in principle. Moreover, since Eve does not need to make any measurement or state preparation, the time-shift attack is practically feasible even with current technology.

The time-shift attack will introduce some loss as the overall detection efficiency is lower if the photon hits the detector at an unexpected time. Nonetheless, Eve can compensate this loss by making the channel more transparent. Notice that, since the quantum channel between Alice and Bob may contain many lossy components such as splices and couplers, it may not be too hard for Eve to make a channel more transparent.

The time-shift attack has been successfully implemented on a commercial QKD system by Zhao, Fung, Qi, Chen, and Lo [65] in 2007. This is the first and so far the only experimentally successful demonstration of quantum hacking on commercial QKD system. It is shown that the system has no-negligible probability to be vulnerable to the time-shift attack. Quantitative analysis shows that the final key shared by Alice and Bob (after the error correction and the privacy amplification of the most general security analysis) has been compromised by Eve.

The success of the time-shift attack in [65] is rather surprising as QKD has been widely believed to be unconditionally secure. The experimental success in quantum hacking highlighted the limit of the whole research program of device-independent security proofs [23] by showing that device-independent security proofs, even if they are found to exist in future, do not apply to a practical QKD system. The success of time-shift attack is not due to some technical imperfection. It is deeply connected with the detection efficiency loophole in the verification of Bell-inequalities. So far the InGaAs detectors have only $\sim 10\%$ detection efficiencies, and the channel connecting Alice and Bob usually has quite large attenuation for long distance communication. The low overall detection efficiency fails the device-independent security proof.

Notice that even non-gated detectors have dead times and generally suffer from detection efficiency loophole. The detection efficiency mismatch is also discussed in [98].

4) Phase remapping attack. In a bi-directional implementation of QKD such as the "Plug and Play" set-up, Eve may attempt to tamper with Alice's preparation process so that Alice prepares four wrong states, instead of the four standard BB84 state. This is called the phase remapping attack and was proposed in [99].

In a "Plug and Play" QKD system, Alice receives a strong pulse from Bob and she then attenuates it to a single-photon level and encodes one of the BB84 state on it. For instance, Alice may encode her state by using a phase modulator. In a phase modulator, the encoded phase is proportional to the voltage applied. In practice, a phase modulator has a finite rise time. For each BB84 setting, one may thus model the applied voltage (and thus the encoded phase) as a trapezium. Ideally, Bob's strong pulse should arrive at the plateau region of the phase modulation, thus getting maximal phase modulation by Alice's phase modulator. Now, imagine that Eve applies a time-shift to Bob's strong pulse so that it arrives in the rise region of the phase modulation graph instead of the plateau region. In this case, Alice has wrongly encoded her phase only partially.

If we assume that the four settings of Alice's encoding (for the four BB84 states) have the same rise region, then by time-shifting Bob's strong pulse, Eve can force Alice to prepare the four state with phase $0, a, 2a, 3a$, rather than $0, \pi/2, \pi, 3\pi/2$. In general, these four states are more distinguishable than the standard BB84 state. Therefore, Eve may subsequently apply an intercept-resend attack to the signal sent out by Alice.

It was proven in [99] that in principle Eve can break the security of the QKD system, without alerting Alice and Bob.

5) Attack by passive listening to side channels. The attack by listening to the sounds made by the source in the first QKD experiment is an example of an attack by passive listening to side channels. Another example is [100]. A counter-measure is to carefully locate all possible side channels and to eliminate them one by one.

6) Saturation Attack. In a recent preprint [101], Makarov studied experimentally how by sending a moderately bright pulse, Eve can blind Bob's InGaAs detector. A simple counter-measure would be for Bob to measure the intensity of the incoming signal.

7) High Power Damage Attack. In Makarov's thesis, it was proposed that Eve may try to make controlled changes in Alice's and Bob's system by using high power laser damage through sending a very strong laser pulse. Again, a simple counter-measure would be for Alice and Bob to measure the intensity of the incoming signals and monitor the properties of various components from time to time to ensure that they perform properly.

Counter-Measures

Once an attack is known, there are often simple counter-measures. For instance, for the large pulse attack, a simple counter-measure would be to add a circulator in Alice's laboratory. As for the faked state attack and time-shift attack, a simple counter-measure would be for Bob to use a four-state setting in his phase modulator. Other counter-measures include Bob applying a random time-shift to his received signals. However, the most dangerous attacks are the *unanticipated* ones.

Notice that it is not enough to say that a counter-measure to an attack exists. It is necessary to actually implement a counter-measure experimentally in order to see how effective and convenient it really is. This will allow Alice and Bob to select a useful counter-measure. Moreover, notice that the implementation of a counter-measure may itself open up new loopholes. For instance, if Bob implements a four-state setting as a counter-measure to a time-shift attack, Eve may still combine a large pulse attack with the time-shift attack to break a QKD system.

Importance of Quantum Hacking

As noted in Sect. "Security Proofs", there has been a lot of theoretical interest on the connection between the security of QKD and fundamental physical principles such as the violation of Bell's inequality. An ultimate goal of such investigations, which has not been realized yet, is to construct a device-independent security proof [23]. Even if such a goal is achieved in future, would any of these theoretical security proofs applies to a quantum key distribution system in *practice*? Unfortunately, the answer is no. As is well-known, the experimental testing of Bell-inequalities often suffers from the detection efficiency loophole [23]. The low detection efficiency of practical detectors not only nullifies security proofs based on Bell-inequality violation, but also gives an eavesdropper a powerful handle to break the security of a practical QKD system. Therefore, the detection efficiency loophole is of both theoretical and practical interest.

A practical QKD system often consists of two or more detectors. In practice, it is very hard to construct detectors of identical characteristics. As a result, two detectors can generally exhibit different detector efficiencies as functions of either one or a combination of variables in the time, frequency, polarization or spatial domains. Now, if an eavesdropper could manipulate a signal in these variables, then

she could effectively exploit the detection efficiency loophole to break the security of a QKD system. In fact, she could even violate a Bell-inequality with only a classical source. In time-shift attack, one consider an eavesdropper's manipulation of the time variable. However, the generality of detection efficiency loophole and detector efficiency mismatch should not be lost.

We should remark that, for eavesdropping attacks, the sky is the limit. The more imaginative one is, the more new attacks one comes up. Indeed, what people have done so far are just scratching the surface of the subject. Much more work needs to be done in the battle-testing of QKD systems and security proofs with testable assumptions. See Sect. "Future Directions".

Beyond Quantum Key Distribution

Besides QKD, many other applications of quantum cryptography have been proposed. Consider, for instance, the millionaires' problem. Two millionaires, Alice and Bob, would like to determine who is richer without disclosing the actual amount of money each has to each other. More generally, in a secure two-party computation, two distant parties, Alice and Bob, with private inputs, x and y respectively, would like to compute a prescribed function $f(x, y)$ in such a way that at the end, they learn the outcome $f(x, y)$, but nothing about the other party's input, other than what can be logically be deduced from the value of $f(x, y)$ and his/her input. There are many possible functions $f(x, y)$. Instead of implementing them one by one, it is useful to construct some cryptographic primitives, which if available, can be used to implement the secure computation of *any* function $f(x, y)$. In classical cryptography, to implement secure two-party computation of a general function will require making additional assumptions such as a trusted third party or computational assumptions. The question is whether we can do unconditionally secure *quantum* secure two-party computations.

Two important cryptographic primitives are namely quantum bit commitment (QBC) and one-out-of-two quantum oblivious transfer (QOT). In particular, it was shown by Kilian [102] that in classical cryptography, oblivious transfer can be used to implement a general two-party secure computation of any function $f(x, y)$. Moreover, in quantum cryptography, it was proven by Yao [103] that a secure QBC scheme can be used to implement QOT securely. For a long time back in the early 1990s, there was high hope that QBC and QOT could be done with unconditional security. In fact, in a paper [104] it was claimed that QBC can be made unconditionally secure. The sky fell around 1996 when Mayers [105] and subsequently, Lo and Chau [106], proved that, contrary to widespread belief at that time, unconditionally secure QBC is, in fact, impossible. Subsequently, Lo [107] proved explicitly that unconditionally secure one-out-of-two QOT is also impossible. Mayers and Lo-Chau's result was a big step backwards and thus a big disappointment for quantum security.

After the fall of QBC and QOT, people turned their attention to quantum coin tossing (QCT). Suppose Alice and Bob are having a divorce and they would like to determine by a coin toss who is going to keep their kid. They do not trust each other. However, they live far away from each other and have to do a coin toss remotely. How can they do so without trusting each other? Classically, coin tossing will require either a trusted third party or making computational assumptions. As shown by Lo and Chau, ideal quantum coin tossing is impossible [108]. Even for the non-ideal case, Kitaev has proven that a strong version of QCT (called *strong* QCT) cannot be unconditionally secure. However, despite numerous papers on the subject (see, for example, [109] and references therein), whether non-ideal *weak* QCT is possible remains an open question.

Other QKD protocols are also of interest. For instance, the sharing of quantum secrets has been proposed in [110]. It is an important primitive for building other protocols such as secure multi-party quantum computation [111]. There are also protocols for quantum digital signatures [112], quantum fingerprinting and unclonable encryption. Incidentally, quantum mechanics can also be used for the quantum sharing of classical secrets [113], conference key agreement and third-man cryptography.

For QKD, so far we have only discussed a point-to-point configuration. In real-life applications, it will be interesting to study QKD in a network setting [114]. Note that the multiplexing of several QKD channels in the single fiber has been successful performed. So has the multiplexing of a classical channel together with a QKD channel. However, much work remains to be done on the design of both the key management structure and the optical layer of a QKD network.

Future Directions

The subject of quantum cryptography is still in a state of flux. We will conclude with a few examples of future directions.

Quantum Repeaters

Losses in quantum channels greatly limit the distance and key generation rate of QKD. To achieve secure QKD over long distances without trusting the intermediate nodes, it is highly desirable to have quantum repeaters. Briefly

stated, quantum repeaters are primitive quantum computers can be perform some form of quantum error correction, thus preserving the quantum signals used in QKD. In more detail, quantum repeaters often rely on the concept of entanglement distillation, whose goal is, given a large number M of noisy entangled states, two parties, Alice and Bob, perform local operations and classical communications to distill out a smaller number (say N) but less noisy entangled states.

The experimental development of a quantum repeater will probably involve the development of quantum memories together with the interface between flying qubits and qubits in a quantum memory.

Ground to Satellite QKD

Another method to extend the distance of QKD is to perform QKD between a satellite and a ground station. If one trusts a satellite, one can even build a global QKD network via a satellite relay. Basically, a satellite can perform QKD with Alice first, when it has a line of sight with Alice. Afterwards, it moves in orbit until it has a line of sight with Bob. Then, the satellite performs a separate QKD with Bob. By broadcasting the XOR of the two keys, Alice and Bob will share the same key. Satellite to ground QKD appears to be feasible with current or near-future technology, for a discussion, see, for example, [62].

With an untrusted satellite, one can still achieve secure QKD between two ground stations by putting an entangled source at the satellite and sending one half of each entangled pair to each of Alice and Bob.

Calculation of the Quantum Key Capacity

Given a specific theoretical model, so far it is not known how to calculate the actual secure key generation rate in a noisy channel. All is known is how to calculate some upper bounds and lower bounds. This is a highly unsatisfactory situation because we do not really know the actual fundamental limit of the system. Our ignorance can be highlighted by a simple open question: what is the highest tolerable bit error rate of BB84 that will still allow the generation of a secure key?

While lower bounds are known [25,26,27] and 25 percent is an upper bound set by a simple intercept-resend attack, we do not know the answer to this simple question.

Notice that this question is of both fundamental and practical interests. Without knowing the fundamental limit of the key generation rate, we do not know what the most efficient procedure for generating a key in a practical setting is.

Multi-party Quantum Key Distribution and Entanglement

Besides its technological interest, QKD is of fundamental interest because it is deeply related to the theory of entanglement, which is the essence of quantum mechanics. So far there have been limited studies on multi-party QKD. Notice that there are many deep unresolved problems in *multi*-party entanglement. It would be interesting to study more deeply multi-party QKD and understand better its connection to multi-party entanglement. Hopefully, this will shed some light on the mysterious nature of multi-party entanglement.

Security Proofs with Testable Assumptions

The surprising success of quantum hacking highlights the big gap between the theory and practice of QKD. In our opinion, it is important to work on security proofs with *testable* assumptions. Every assumption in a security proof should be written down and experimentally verified. This is a long-term research program.

Battle-Testing QKD Systems

Only through battle-testing can we gain confidence about the security of a real-life QKD system. Traditionally, breaking a cryptographic systems is as important as building one. Therefore, we need to re-double our efforts on the study of eavesdropping attacks and their counter-measures.

As stated before, quantum cryptography enjoys forward security. Thanks to the quantum no-cloning theorem, an eavesdropper Eve does not have a transcript of all quantum signals sent by Alice to Bob. Therefore, once a QKD process has been performed, the information is gone and it will be too late for Eve to go back to eavesdrop. Therefore, for Eve to break a real-life QKD system today, it is imperative for Eve to invest in technologies for eavesdropping now, rather than in future.

Acknowledgments

We thank various funding agencies including NSERC, CRC program, QuantumWorks, CIFAR, MITACS, CIPI, PREA, CFI, and OIT for their financial support.

Bibliography

Primary Literature

1. Bennett CH, Brassard G (1984) In: Proceedings of IEEE International Conference on Computers, Systems, and Signal Processing. IEEE, New York, pp 175–179

2. Bennett CH (1992) Phys Rev Lett 68:3121
3. Ekert AK (1991) Phys Rev Lett 67:661
4. Wiesner S (1983) Sigact News 15:78
5. Vernam G (1926) J Am Inst Electr Eng 45:109
6. Shor PW (1997) SIAM J Sci Statist Comput 26:1484
7. Brassard and Crépeau (1996) ACM SIGACT News 27:13
8. Mayers D (2001) J ACM 48:351
9. Lo H-K, Chau HF (1999) Science 283:2050
10. Bennett CH, DiVincenzo DP, Smolin JA, Wootters WK (1996) Phys Rev A 54:3824
11. Deutsch D, Ekert A, Jozsa R, Macchiavello C, Popescu S, Sanpera A (1996) Phys Rev Lett 77:2818
12. Shor P, Preskill J (2000) Phys Rev Lett 85:441
13. Biham E, Boyer M, Boykin PO, Mor T, Roychowdhury V (2000) In: STOC '00: Proceedings of the thirty-second annual ACM symposium on Theory of computing. ACM Press, New York, pp 715–724
14. Ben-Or M (2002) http://www.msri.org/publications/ln/msri/2002/qip/ben-or/1/index.html. Accessed 03 Nov 2008
15. Given a density matrix, ρ, define the von Neumann entropy of ρ as $S(\rho) = -\sum_i \lambda_i \log_2 \lambda_i = -\mathrm{tr}\rho \log_2 \rho$ where λ_i's are eigenvalues of the density matrix ρ
16. http://en.wikipedia.org/wiki/Holevo's_theorem. Accessed 03 Nov 2008
17. Renner R, Koenig R (2005) In: TCC 2005. LNCS, vol 3378. Springer, Berlin, (eprint) quant-ph/0403133
18. Renner R (2005) (eprint) quant-ph/0512258
19. Renner R (2007) Nat Phys 3:645
20. Horodecki K, Horodecki M, Horodecki P, Oppenheim J (2005) Phys Rev Lett 94:160502
21. Horodecki K, Horodecki M, Horodecki P, Leung D, Oppenheim J (2008) IEEE Trans Inf Theory 54(6):2604–2620
22. Koashi M (2007) arXiv:0704.3661
23. Acin A, Brunner N, Gisin N, Massar S, Pironio S, Scarani V (2007) Phys Rev Lett 98:230501
24. Masanes L, Winter A (2006) (eprint) quant-ph/0606049. Accessed 03 Nov 2008
25. Gottesman D, Lo H-K (2003) IEEE Trans Inf Theory 49:457
26. Renner R, Gisin N, Kraus B (2005) Phys Rev A 72:012332
27. Chau HF (2002) Phys Rev A 66:060302
28. Brassard G, Salvail L (1994) Lecture Notes in Computer Science. Springer, vol 765, pp 410–423
29. Bennett CH, Brassard G, Crepeau C, Maurer UM (1995) IEEE Trans Info Theory 41:1915
30. Ben-Or M, Horodecki M, Leung DW, Mayers D, Oppenheim J (2005) In: Kilian J (ed) Theory of Cryptography: Second Theory of Cryptography Conference. TCC 2005. Lecture Notes in Computer Science, vol 3378. Springer, Berlin, pp 386–406
31. Koenig R, Renner R, Bariska A, Maurer U (2007) Phys Rev Lett 98:140502
32. Inamori H, Lütkenhaus N, Mayers D (2007) Eur Phys J D 41:599
33. Gottesman D, Lo H-K, Lütkenhaus N, Preskill J (2004) Quant Info Compu 4:325
34. Bennett CH, Bessette F, Brassard G, Salvail L, Smolin J (1992) J Cryptogr 5:3
35. Townsend PD, Rarity JG, Tapster PR (1993) Electron Lett 29:634
36. Muller A, Breguet J, Gisin N (1993) Europhys Lett 23:383
37. Jacobs BC, Franson JD (1996) Opt Lett 21:1854
38. Franson JD, Lives H (1994) Appl Opt 33:2949
39. Townsend PD (1994) Electron Lett 30:809
40. Muller A, Zbinden H, Gisin N (1995) Nature 378:449
41. Muller A, Zbinden H, Gisin N (1996) Europhys Lett 33:335
42. Muller A, Herzog T, Hutter B, Tittel W, Zbinden H, Gisin N (1997) Appl Phys Lett 70:793
43. Zbinden H, Gautier J-D, Gisin N, Hutter B, Muller A, Tittel W (1997) Electron Lett 33:586
44. Stucki D, Gisin N, Guinnard O, Robordy G, Zbinden H (2002) New J Phys 4:41
45. Lütkenhaus N (2000) Phys Rev A 61:052304
46. Gobby C, Yuan ZL, Shields AJ (2004) Electron Lett 40:1603
47. Hwang WY (2003) Phys Rev Lett 91:057901
48. Lo H-K (2004) In: Proceedings of IEEE International Symposium on Information Theory IEEE, New York, p 137
49. Lo H-K, Ma X, Chen K (2005) Phys Rev Lett 94:230504
50. Ma X, Qi B, Zhao Y, Lo H-K (2005) Phys Rev A 72:012326
51. Wang X-B (2005) Phys Rev Lett 94:230503
52. Wang X-B (2005) Phys Rev A 72:012322
53. Zhao Y, Qi B, Ma X, Lo H-K, Qian L (2006a) Phys Rev Lett 96:070502
54. Zhao Y, Qi B, Ma X, Lo H-K, Qian L (2006b) In: Proceedings of IEEE International Symposium of Information Theory, IEEE, New York, pp 2094–2098
55. Schmitt-Manderbach T et al (2007) Phys Rev Lett 98:010504
56. Rosenberg D et al (2007) Phys Rev Lett 98:010503
57. Yuan ZL, Sharpe AW, Shields AJ (2007) Appl Phys Lett 90:011118
58. Peng C-Z et al (2007) Phys Rev Lett 98:010505
59. Yin Z-Q, Han Z-F, Chen W, Xu F-X, Wu Q-L, Guo G-C (2008) Chin Phys Lett 25:3547
60. Wang Q, Karlsson A (2007) Phys Rev A 76:014309
61. Takesue H et al (2007) Nat Photonics 1:343
62. Villoresi P et al (2004) arXiv:quant-ph/0408067v1
63. Stucki D, Ribordy G, Stefanov A, Zbinden H, Rarity JG, Wall T (2001) J Mod Opt 48:1967
64. Qi B, Fung C-HF, Lo H-K, Ma X (2007a) Quant Info Compu 7:73
65. Zhao Y, Fung C-HF, Qi B, Chen C, Lo H-K (2008) Phys Rev A 78:042333
66. Namekata N, Fujii G, Inoue S (2007) Appl Phys Lett 91:011112
67. Thew RT et al (2006) New J of Phys 8:32
68. Diamanti E, Takesue H, Langrock C, Fejer MM, Yamamoto Y (2006) Opt Express 14:13073, (eprint) quant-ph/0608110
69. Rosenberg D et al (2006) Appl Phys Lett 88:021108
70. Dynes JF, Yuan ZL, Sharpe AW, Shields AJ (2007) Opt Express 15:8465
71. Mo XF, Zhu B, Han ZF, Gui YZ, Guo GC (2005) Opt Lett 30:2632
72. Gisin N, Fasel S, Kraus B, Zbinden H, Ribordy G (2006) Phys Rev A 73:022320
73. Zhao Y, Qi B, Lo H-K (2008) Phys Rev A 77:052327
74. Bennett CH, Brassard G, Breidbart S, Wiesner S (1984) IBM Technical Disclosure Bulletin 26:4363
75. Bruss D (1998) Phys Rev Lett 81:3018
76. Goldenberg L, Vaidman L (1995) Phys Rev Lett 75:1239
77. Lo H-K, Chau HF, Ardehali M (2005) J Cryptology 18:133
78. Gisin N, Ribordy G, Zbinden H, Stucki D, Brunner N, Scarani V (2004) arXiv:quant-ph/0411022v1
79. Peng C-Z et al (2005) Phys Rev Lett 94:150501
80. Ursin R et al (2007) Nat Phys 3:481
81. Mauerer W, Silberhorn C (2007) Phys Rev A 75:050305
82. Wang Q, Wang X-B, Guo G-C (2007) Phys Rev A 75:012312
83. Adachi Y, Yamamoto T, Koashi M, Imoto N (2007) Phys Rev Lett 99:180503

84. Ma X, Lo H-K (2008) New J Phys 10:073018
85. Wang Q et al (2008) Phys Rev Lett 100:090501
86. Mendonça F, de Brito DB, Silva JBR, Thé GAP, Ramos RV (2008) Microw Opt Tech Lett 50:236
87. Crosshans F, Assche GV, Wenger J, Brourl R, Cerf NJ, Grangier P (2003) Nature 421:238
88. Lodewyck J, Debuisschert T, Tualle-Brouri R, Grangier P (2005) Phys Rev A 72:050303
89. Legré M, Zbinden H, Gisin N (2006) Quant Info Compu 6:326
90. Qi B, Huang L-L, Qian L, Lo H-K (2007) Phys Rev A 76:052323. arXiv:0709.3666
91. Lodewyck J et al (2007) Phys Rev A 76:042305
92. Curty M, Zhang LLX, Lo H-K, Lütkenhaus N (2007) Quant Info Compu 7:665
93. Tsurumaru T (2007) Phys Rev A 75:062319
94. Honjo T, Inoue K, Takahashi H (2004) Opt Lett 29:2797
95. Takesue H et al (2005) New J Phys 7:232
96. Gisin N, Fasel S, Kraus B, Zbinden H, Ribordy G (2006) Phys Rev A 73:022320
97. Makarov V, Anisimov A, Skaar J (2006) Phys Rev A 74:022313
98. Larsson J-Å (2002) Quant Info Compu 2:434
99. Fung C-HF, Qi B, Tamaki K, Lo H-K (2007) Phys Rev A 75:032314
100. Lamas-Linares A, Kurtsiefer C (2007) Opt Express 15:9388
101. Makarov V (2007) arXiv:0707.3987v1
102. Kilian J (1988) In: Proceedings of the 20th Annual ACM Symposium on Theory of Computing. ACM, New York, pp 20–31
103. Yao AC-C (1995) In: Proceedings of the 26th Annual ACM Symposium on the Theory of Computing. ACM, New York, p 67
104. Brassard RJG, Crépeau C, Langlois D (1993) In: Proceedings of the 34th Annual IEEE Symposium on the Foundations of Computer Science. IEEE, New York, p 362
105. Mayers D (1997) Phys Rev Lett 78:3414
106. Lo H-K, Chau HF (1997) Phys Rev Lett 78:3410
107. Lo H-K (1997) Phys Rev A 56:1154
108. Lo H-K, Chau HF (1997) Physica D 120:177
109. Mochon C (2005) Phys Rev A 72:022341
110. Cleve R, Gottesman D, Lo H-K (1999) Phys Rev Lett 83:648
111. Ben-Or M, Crépeau C, Gottesman D, Hassidim A, Smith A (2006) In: Proceedings of 47th Annual IEEE Symposium on the Foundations of Computer Science (FOCS'06). IEEE, New York, pp 249–260
112. Gottesman D, Chuang I (2001) arXiv:quant-ph/0105032v2
113. Hillery M, Bužek V, Berthiaume A (1999) Phys Rev A 59:1829
114. Alleaume R et al (2007) quant-ph/0701168. Accessed 03 Nov 2008

Books and Reviews

Brassard G (1994) A Bibliography of Quantum Cryptography. http://www.cs.mcgill.ca/~crepeau/CRYPTO/Biblio-QC.html. Accessed 03 Nov 2008

Gisin N, Thew R (2007) Quantum Communication. Nat Photon 1(3):165–171. On-line available at http://arxiv.org/abs/quant-ph/0703255. Accessed 03 Nov 2008

Gisin N, Ribordy G, Tittel W, Zbinden H (2002) Quantum Cryptography. Rev Mod Phys 74:145–195. On-line Available at http://arxiv.org/abs/quant-ph/0101098. Accessed 03 Nov 2008

Gottesman D, Lo H-K (2000) From Quantum Cheating to Quantum Security. Physics Today, Nov 2000, p 22

Lo H-K, Lütkenhaus N (2007) Quantum Cryptography: from theory to practice Invited paper for Physics In Canada, Sept-Dec 2007. On-line available at http://arxiv.org/abs/quant-ph/0702202. Accessed 03 Nov 2008

Scarani V, Bechmann H-Pasquinucci, Cerf NJ, Dusek M, Lütkenhaus N, Peev M (2008) The security of practical quantum key distribution. arXiv:0802.4155v2

Wikipedia (2008) Quantum Cryptography, http://en.wikipedia.org/wiki/Quantum_cryptography. Accessed 03 Nov 2008

Quantum Dots: Fano Resonances in an Aharonov–Bohm Ring

Arkady M. Satanin[1], Eric R. Hedin[2], Yong S. Joe[2]
[1] Institute for Physics of Microstructures, Russian Academy of Sciences, Nizhny Novgorod, Russia
[2] Center for Computational Nanoscience, Department of Physics and Astronomy, Ball State University, Muncie, USA

Article Outline

Glossary

Nanostructures Modern nanotechnology achieves the arrangement of atoms into structures that are only a few nanometers in size. The following kinds of artificial structures (semiconductor or metallic) may be formed: two- dimensional or layered heterostructures (superlattices); one-dimensional nanowires, or zero-dimensional quantum dots. The characteristic size of systems is between 0.1 and 100 nm.

Quantum dots (QDs) A quantum dot is a semiconductor nanostructure that confines the motion of electrons in all thee spatial directions. The confinement can be due

to electrostatic potentials (generated by external electrodes, doping, strain, or impurities). A quantum dot has discrete allowed energy levels for which the corresponding wavefunctions are spatially localized within the quantum dot.

Wave interference Wave interference is the phenomenon which occurs when two waves meet while traveling in the same medium. If the waves have the same phases, constructive interference takes place, and destructive occurs for waves with opposite phases.

Aharonov–Bohm (AB) effect The Aharonov–Bohm effect is a quantum mechanical phenomenon by which a charged particle is affected by an electromagnetic field which only exists in regions apart from the particle. If there are two propagation paths by which a particle can move without entering the region of space where the magnetic field exists, then a charged quantum particle can still show an observable phase shift in its interference pattern.

Breit–Wigner (BW) resonances The Breit–Wigner resonances arise due to the constructive interference of two counter-propagating waves in the same scattering channel (similar to resonances of the Fabry–Perot interferometer in optics). For instance, the Breit–Wigner resonances occur when an electron moves in the space between two semiconductor layers (barriers) for some particular electron energies (quasi-bound energies). For this case, the Breit–Wigner resonances appear in the transmission. In addition, the scattering amplitude as a function of electron energy possesses a pole for each quasi-bound state in the complex-energy plane. The real part of a pole can be interpreted as the energy of a quasi-bound state and the imaginary part can be connected with the lifetime of this state.

Fano resonances The Fano resonance is a manifestation of the interference between a localized state and the continuum states in the transmission of multi-channel systems. In AB rings, the Fano resonances arise from quantum mechanical interference between the discrete states of the QD in one arm of the interferometer and the continuum in the other arm, characterized by complete transmission and complete reflection. The profile of the Fano asymmetric line-shape in the transmission depends on the strength of the coupling between discrete and continuum states, and on the phase difference between the paths.

Definition of the Subject

Resonance phenomena are a major subject of theoretical and experimental investigations, and the concept of resonances is ubiquitous in physics. The search for new effects related to wave interference and different kinds of resonances in various physical systems continues to be of interest. Interference of a localized wave with propagating states and the resulting Fano resonances in atomic and solid states structures have attracted much attention recently. Now, it is clear that Fano interference is a universal phenomenon because the manifestation of the interference does not depend on the configuration of the material. The natural question then arises: Why are Fano-interference phenomena so interesting in different fields of physics? From the practical point of view, for instance, the resonances can be considered as quantum *probes* that provide important information on the geometric configuration and internal potential fields of low-dimensional structures. Fano interference may potentially be used for the design of new types of quantum electronic or spintronic devices such as Fano-transistors, spin transistors, and Fano-filters for polarized electrons. In addition, Fano phenomena can also be used for lasing without population inversion. The investigation of these resonances is of great importance in the search for optimal working parameters of new devices, such as the resonant diode and the resonant transistor.

Introduction

At present, nanotechnology provides various solid state systems such as Aharonov–Bohm (AB) rings, two-dimensional (2D) electronic waveguides, nanotubes etc., where alternative electronic paths may be realized. In modern laboratories, various AB rings with embedded quantum dots in the arms have been fabricated [28,29,31,32,42,46,47]. If a quantum ring supports coherent transmission, the wave amplitudes through the two arms interfere. The scattering of the waves on the embedded dots, and the interference of waves in the arms results in resonances in the conductance (or transmission) of the ring. At the same time, the electron scattering on the dots produces an additional phase shift. Thus, the interference of the waves coming from the two arms contains information about the amplitude and phase of the waves. In the other words, an AB electronic interferometer operates similarly to an optical interferometer providing illumination of the phase-shift.

It is well known that resonant-transmission phenomena are related to the quasi-bound states of the systems. According to the Breit–Wigner (BW) formalism in nuclear systems, the scattering amplitude as a function of energy possesses a pole for each quasi-bound state in the complex-energy plane [8]. The real part of a pole can be interpreted as the energy of a quasi-bound state and the

imaginary part can be connected with the life time of this state [15]. Thus, BW resonances arise due to the interference of two counter-propagating waves in the same scattering channel (similar to resonances of the Fabry–Perot interferometer in optics).

In contrast to multi-barrier resonant-tunneling structures, quantum nanostructures such as AB rings and 2D-electronic waveguides, where alternative electronic paths may be realized, possess both transmission zeros and resonance poles. This characteristic of a zero-pole pair, called a Fano resonance [14], has been specifically predicted and observed in a hybrid system of an AB ring and a QD both theoretically and experimentally [2,3,4,5,6,13,16,17,18,28,29,30,31,32,34,36,42,43,45,46,47]. In order to fit the experimental data near an asymmetric resonance, the authors of experimental works have used in expression for the conductance G in the form

$$G = G_{\mathrm{b}} \frac{|E - E_0|^2}{(E - E_{\mathrm{R}})^2 + \Gamma^2} , \qquad (1)$$

where G_{b} is the non-resonant conductance, E_{R} is the energy of the resonance, Γ is the width, and E_0 is a complex parameter (in general) – the resonance zero. It has been demonstrated that the pole is an immediate corollary of the electron transition from the bound state to the continuum [45]. Although the nature of the zero is connected with the bound state it also depends on the detail of the interference with the continuum. Notice, that for systems with time reversal symmetry, E_0 always is placed on real axis of energy.

This Fano effect arises from quantum mechanical interference between the discrete state of the QD in one arm and the continuum in other arm. The profile of the Fano asymmetric line-shape in the transmission depends on the strength of the coupling between discrete and continuum states, and on the phase difference between the paths. Here, the scattering amplitude near the zero-pole pair behaves like a dipole, where the pole plays the role of a particle and the zero plays the role of a hole (anti-particle) [16,17,30,34,36,43,45]. The collapse of the particle and hole has been studied in a quasi-one-dimensional constriction with an attractive and finite-size impurity, by modulating the parameters of the system [17,30].

When more than one resonant quasi-bound state is present in a one-channel system, for instance, in a three-barrier system, the resonance levels interact each other and result in the overlapping of resonances [17,19,20,22,23,24,25,26,30,38,39,40,41]. In this situation, therefore, the single BW formula is no longer valid due to the overlapping of resonances. Hence, the interference effects of resonances have been studied by examining the formation of double poles in the transmission amplitude and the effect of the collision of the two poles (or merging of two resonances) associated with the quasi-bound states [19,20,22,23,24,25,26,38,39,40,41].

It was demonstrated recently that the Fano resonance structure can be controlled by changing the confinement parameters of the QD. Transmission through a QD embedded in an AB-ring remains phase-coherent, as indicated by the visibility of the AB-oscillations [28,29,31,32,42,46,47]. An expression for the transmission amplitude through the QD is $t = \sqrt{T_{\mathrm{QD}}} \mathrm{e}^{\mathrm{i}\alpha_{\mathrm{QD}}}$, which is a complex quantity. The phase, α_{QD}, is significant in its effect upon the AB oscillations for a QD embedded in an AB-ring, and has experimentally been seen to exhibit interesting phase-jumps of π as the transmission passes through a resonance. Theoretical analysis of these systems has provided some explanation for the phase behavior seen in experiments [2,3,4,5,6,10,12,13,19,20,21,22,23,24,25,26,27,37,38,39,40,41]. In a two-terminal device, the Onsager relations [35] of time-reversal symmetry and current conservation (unitarity) constrain the transmission phase to values of 0 or π. However, if the two-terminal AB-ring is "opened" by allowing current to flow out through additional terminals, the unitarity condition is broken and it becomes possible to extract meaningful phase information about the QD. Experiments with open rings demonstrate a gradual phase change across the transmission resonances. In particular, Schuster et al. produced a 4-terminal interferometer in a GaAs–AlGaAs heterostructure which showed smooth phase transitions [42].

In this review we shall present two typical approaches for investigation of Fano resonances in AB rings: i) The approach developed in quantum waveguides theory when the arms of the ring are considered as waveguides for electrons and dots are treated as a potential walls embedded in the waveguide; ii) Tight binding theory approach in which the dots are represented by sites which are connected by one-dimensional tight binding chains with different topology. Because both approaches give the same physical effects we shall focus mostly on physical meaning of obtained results. In Sect. "Fano Resonance in a Two Dimensional Electron Waveguide with a Quantum Dot" we give an introduction to Fano phenomena using a simple model of a waveguide with a short range potential well (QD). Applying waveguide theory to a quantum dot in AB rings in Sect. "Model for the Quantum Dots in the Ring: Theory of Waveguides Approach" we will demonstrate the manifestation of the Fano resonance in this system. In Sect. "Effects of Interference Interactions of Fano Resonances on an AB Ring" some novel results regarding Fano resonances in strong overlapping regime are presented.

It is shown that when the overlapping of two Fano resonances takes place in the transmission, two Fano dipoles in the complex-energy plane form a quasi-particle, which behaves as a coupled object – a *Fano quadrupole*. We show a periodic motion of the resonance pole and transmission zero in the complex-energy plane as a magnetic field through the AB ring is changed. In Sect. "Fano Resonance Induced by a Quantum Dot in an Open Three-terminal Interferometer", the properties of Fano resonances in an open AB ring are investigated in the framework of the tight binding model. We shall discuss in this section the phase measurements by using the modes of an open AB interferometer. Section "Future Directions" summarizes the properties of Fano phenomena and its manifestation in AB rings along with future directions.

Fano Resonance in a Two Dimensional Electron Waveguide with a Quantum Dot

In order to see a main feature of the Fano phenomenon associated with the propagating and evanescent waves in a quantum system, we study the propagation of the electron waves in an electronic 2D waveguide of width W arranged along the x-axis [24,30]. The waveguide geometry is schematically depicted in Fig. 1, showing a potential region and an attractive quantum dot (gray-colored area) in the waveguide. Here, the confining potential in the transverse direction is characterized by the function $V_c(y)$ and the attractive potential (dot) by the function $V(x, y)$. There is a complete basis of functions describing the transverse motion $\phi_n(y)$ of an electron with energies, $E_n = \frac{\hbar^2\pi^2 n^2}{2mW^2}$ (with the effective mass m). The electron waves in the perfect waveguide stretched to infinity are described by a combination of a plane wave along the longitudinal direction and confined wave functions in the transverse direction such as $e^{\pm ik_n x}\phi_n(y)$, where the wave vector along the x-direction is $k_n = \sqrt{2m(E-E_n)}/\hbar$, and n is the number of the transverse state. These propagating states can be considered as open channels in the waveguide.

In order to find the wave function of an electron in a waveguide with the dot, we solve the 2D Schrödinger equation

$$-\frac{\hbar^2}{2m}\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)\Psi(x,y) + V_c(y)\Psi(x,y) + V(x,y)\Psi(x,y) = E\Psi(x,y) \quad (2)$$

with plane wave boundary conditions in the leads ($x \to \pm\infty$). It is convenient to expand the wave function in the complete basis of functions describing the transverse motion:

$$\Psi(x,y) = \sum_{n=1}^{\infty} \psi_n(x)\phi_n(y)\,. \quad (3)$$

Substituting Eq. (3) into Eq. (2), we obtain the coupled-channel equations for an electron in the form

$$[b] - \frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2}\psi_n(x) + \sum_{n=1}^{\infty} V_{nn'}(x)\psi_{n'}(x) = (E-E_n)\psi_n(x)\,, \quad (4)$$

where the coupling matrix elements of the dot's potential (which still acts on the x-coordinate) are defined to be

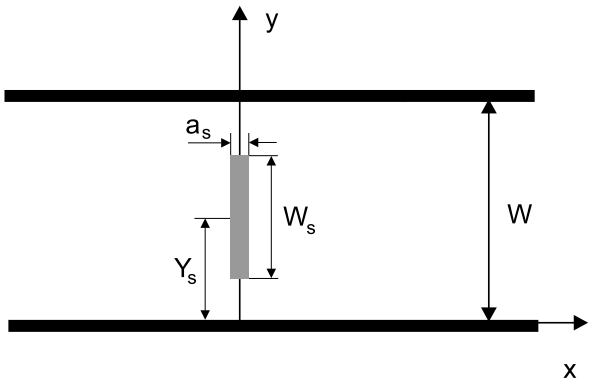
$$V_{nn'}(x) = \int \phi_n(y)V(x,y)\phi_{n'}(y)\,\mathrm{d}\,y\,. \quad (5)$$

Since Eq. (4), which is equivalent to the 2D Schrödinger equation, cannot be solved in general, we use some simplification that allow us to use a resonant perturbation theory [15] in the system under the investigation.

We model the scattering potential as a thin rectangular potential-well by assuming that the longitudinal size of the potential well is much smaller than the characteristic wavelengths of the electron. Then, the matrix elements of the potential can be written as

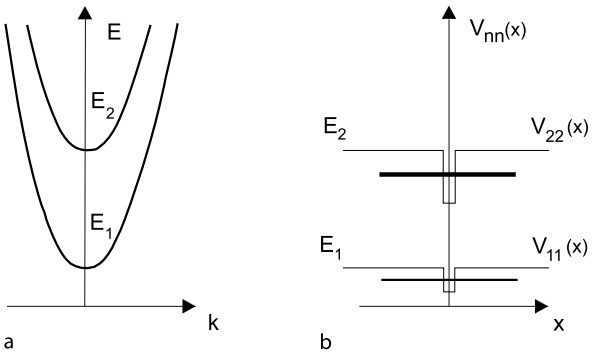
$$V_{nn'}(x) = -\frac{\hbar^2}{m}v_{nn'}\delta(x)\,, \quad (6)$$

where the parameters $v_{nn'}$ of the dot are expressed in an explicit form [30]. It can be shown that the short range



Quantum Dots: Fano Resonances in an Aharonov–Bohm Ring, Figure 1

Schematic illustration of the electron waveguide with embedded quantum dot (*gray-colored area*), where the attractive potential well is centered at $x = 0$ and $y = 0$ and the electron motion is not limited horizontally, $-\infty < x < \infty$, but is confined vertically, $0 < y < W$



Quantum Dots: Fano Resonances in an Aharonov–Bohm Ring, Figure 2

a Energy dispersion relation for an electron in a perfect waveguide and b the diagrams for a bound level (near the first subband) and quasi-bound level (near the second subband) in the dot's effective potential

potential provides the following boundary conditions to be imposed on the multi-component wavefunctions at $x = 0$:

$$\begin{aligned} \psi_n(0+) &= \psi_n(0-)\,, \\ \psi_n'(0+) - \psi_n'(0-) &= -2\sum_{n'=1}^{\infty} \nu_{nn'}\psi_{n'}(0\pm)\,. \end{aligned} \tag{7}$$

Here, we consider the situation when the energy of incoming electron is placed in the interval $E_1\langle E\langle E_2$ (the first energy window), as shown schematically in Fig. 2. If the characteristic value of matrix element V_{12}, describing the coupling between two nearest channels, is small compared to the subband distance $(E_n - E_{n-1})$, then we only need to consider two coupled equations in the first energy window to understand the main physical features of the interference. It is well known that the remaining modes in the waveguide with the attractive impurity only alter the width and position of the resonances and hence play a minor role in Fano phenomenon. Without much difficulties our formulation can be extended for a multi-band approximation.

The wave function in the first channel, obtained from the solutions of the Schrödinger equation, can be written as

$$\psi_1(x) = \begin{cases} a_1 e^{ik_1 x} + b_1 e^{-ik_1 x}\,, & x < 0\,, \\ c_1 e^{ik_1 x}\,, & x > 0\,, \end{cases} \tag{8}$$

where $k_1 = \sqrt{2m(E - E_1)}/\hbar$ is a wave vector in the first channel. Similarly, the wavefunction in the second channel as

$$\psi_2(x) = \begin{cases} b_2 e^{|k_2|x}\,, & x < 0\,, \\ c_2 e^{-|k_2|x}\,, & x > 0\,, \end{cases} \tag{9}$$

where $|k_2| = \sqrt{2m(E_2 - E)}/\hbar$. Notice that the wave function, ψ_2, in the second channel is an evanescent wave. These two waves interfere in the waveguide and the quantum dot plays a role of a mixer of two different types of waves. The undetermined amplitudes appearing in Eqs. (8) and (9) are specified by applying the matching conditions given in Eq. (7). Consequently, we obtain

$$\begin{aligned} (ik_1 + \nu_{11})c_1 + \nu_{12}c_2 &= ik_1 a_1\,, \\ \nu_{12}c_1 + (-|k_2| + \nu_{22})c_2 &= 0\,, \end{aligned} \tag{10}$$

which give

$$c_1 = \frac{ik_1(-|k_2| + \nu_{22})}{(ik_1 + \nu_{11})(-|k_2| + \nu_{22}) - \nu_{12}^2} a_1\,, \tag{11}$$

$$c_2 = -\frac{ik_1\nu_{12}}{(ik_1 + \nu_{11})(-|k_2| + \nu_{22}) - \nu_{12}^2} a_1\,. \tag{12}$$

From Eq. (11) the transmission and reflection amplitudes in the first channel are obtained as

$$t_{11} = \frac{c_1}{a_1} = \frac{ik_1(-|k_2| + \nu_{22})}{(ik_1 + \nu_{11})(-|k_2| + \nu_{22}) - \nu_{12}^2}\,, \tag{13}$$

and

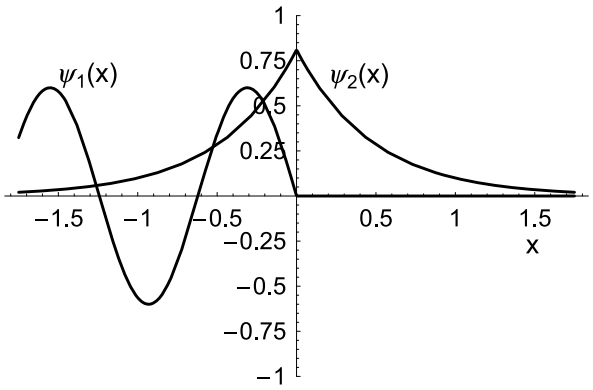
$$r_{11} = \frac{b_1}{a_1} = \frac{-\nu_{11}(-|k_2| + \nu_{22}) + \nu_{12}^2}{(ik_1 + \nu_{11})(-|k_2| + \nu_{22}) - \nu_{12}^2}\,, \tag{14}$$

respectively. As it follows from Eq. (13), the transmission amplitude may vanish if $-|k_2| + \nu_{22} = 0$. When this happens, the reflection amplitude r_{11} is -1 and the energy at which the transmission becomes zero is determined to be

$$E_0 = E_2 - \frac{\hbar^2\nu_{22}^2}{2m}\,. \tag{15}$$

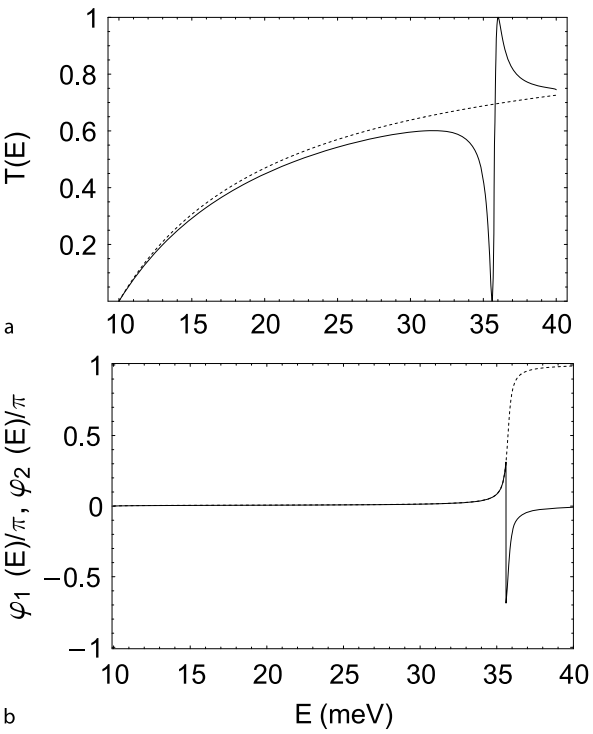
There is a full reflection of the electron wave from a quantum dot when the electron energy is equal to the zero-energy E_0. The wavefunctions in the first and second channels are schematically depicted in Fig. 3 at the zero-energy. Like the classical system, we notice that the position of the amplitude-zero depends on the number of the channels. For instance, if we take into account another closed channel, $n = 3$, by a perturbation, the zero-energy given by Eq. (15) is shifted on the real axis of energy. In the meantime, there is a full transmission of the electron wave through the quantum dot when the reflection amplitude $r_{11} = 0$. If we impose $r_{11} = 0$ in Eq. (14), we get the condition for the reflection-zero, $\nu_{11}(-|k_2| + \nu_{22}) - \nu_{12}^2 = 0$. A real solution to this condition exists at the energy $E_{\max}$

$$E_{\max} = E_2 - \frac{\hbar^2}{2m}\left(\nu_{22} - \frac{\nu_{12}^2}{\nu_{11}}\right)^2\,. \tag{16}$$



Quantum Dots: Fano Resonances in an Aharonov–Bohm Ring, Figure 3

The wave functions of the first and second channels when the electron energy matches with the zero-energy. Note that in this case a full reflection occurs



Quantum Dots: Fano Resonances in an Aharonov–Bohm Ring, Figure 4

a Fano resonance in the transmission for a quantum waveguide with a short-range attractive potential (*solid line*) and a background transmission (*dotted line*). b The phase shift of the amplitude for the electron wave in the first propagating channel (*solid line*) and the second evanescent channel (*dotted line*)

Note that the above expressions (Eqs. (13) and (14)) for the transmission-zero and the reflection-zero energies are exact in the framework of the two-channel approximation.

We have performed a numerical calculation of the transmission using the following parameters of the waveguide and a quantum dot. The width of the waveguide is set to $W = 23.7\,\text{nm}$ and the GaAs effective mass is used as $m = 0.067 m_0$. This gives $E_1 = 10\,\text{meV}$ and $E_2 = 40\,\text{meV}$ for the first two energy levels due to transverse confinement in the waveguide. The parameters of the quantum dot are as follows: $Y_s = 0.55\,\text{W}$ (the position of the dot in the waveguide), $W_s = 0.5\,\text{W}$ (W_s is the transverse width of the dot), and the scattering parameters $a_s V_s = 0.1\,\text{eVnm}$, where $V_s = 100\,\text{meV}$ (V_s is the depth of the attractive potential well) and $a_s = 1\,\text{nm}$ (a_s is the thickness of the potential well). The computed transmission of the system, $T(E) = |t_{11}(E)|^2$, is plotted in Fig. 4a for the chosen energy window, where for numerical purposes the following characteristic energies for the matrix elements of the potential are used: $\hbar^2 \nu_{11}^2/2m \cong 11.33\,\text{meV}$, $\hbar^2 \nu_{22}^2/2m \cong 4.40\,\text{meV}$, and $\hbar^2 \nu_{12}^2/2m \cong 0.34\,\text{meV}$. The pronounced Fano resonant structure (solid line) is clearly shown, i. e. the combined anti-resonance at $E_0 = 35.60\,\text{meV}$ and the nearby resonance peak at $E_\text{R} = 36.01\,\text{meV}$ where the width of resonance line is $\Gamma = 0.19\,\text{meV}$. Notice that if we put our quantum dot at the center of waveguide ($\nu_{12} = 0$), then the interference vanishes and the potential scattering takes place. In this case, only so-called the background profile in the transmission may be seen. This background transmission is also plotted for comparison in Fig. 4a as a dotted line. In Fig. 4b we also show the phase shift of the transmitted electron wave with respect to the incoming wave as a function of the electron energy for the first propagating channel (solid line) and the second evanescent channel (dotted line). One can see that the phase φ_1 in the propagating channel changes by π abruptly at the zero-energy and that it jumps up around the resonance peak, thus gaining essentially no net phase shift after passing through the zero-pole structure. On the other hand, the phase φ_2 of the evanescent channel changes by π rather smoothly over the anti-resonance and resonance structure.

To obtain a simple expression for the transmission amplitude near a zero-pole region, we consider the system in the weak coupling regime (i. e. ν_{12} is assumed to be small in Eq. (13)). Expanding the numerator and denominator of Eq. (13) around the zero and the pole, respectively, one can write the transmission amplitude t_{11} in the desired form

$$t_{11}(E) \sim \frac{E - E_0}{E - E_\text{R} + i\Gamma}\,, \tag{17}$$

where E_R and Γ are the peak position and the width of

the resonance, and E_0 is the zero-energy of the resonance. After performing a perturbation approximation we find that the real part of the resonance pole can be written as $E_R = E_0 + \delta$, where $\delta = \hbar^2 v_{12}^2 v_{11} v_{22}/m(k_1^2 + v_{11}^2)$ and the width $\Gamma = \hbar^2 v_{12}^2 k_1 v_{22}/m(k_1^2 + v_{11}^2)$. (The energy appearing in k_1 is taken at $E = E_R$.) Here, we note that one can neglect the difference between E_{max} and E_R in the weak coupling limit. Furthermore, the expression for the transmission, Eq. (13), can be cast into the canonical Fano form of Eq. (1). The coupling parameter q ($q = v_{11}/k_1$ in our perturbation approximation) measures an asymmetry degree of Fano resonance line shape between the localized sates and the continuum states.

Model for the Quantum Dots in the Ring: Theory of Waveguides Approach

As mentioned in the Introduction, the AB ring plays the role of an interferometer for probing electron states on QD's. For this purpose, the QD's are placed in an arm of the interferometer, whereas the second arm is used as a reference path for the electron waves.

We shall start with a simple model where a QD is embedded in one of its arms, as schematically shown in Fig. 5. Here, the QDs can be formed by two electrodes which play the role of barriers for electrons in the lower arm. We consider both arms of the ring and the leads as perfect waveguides and adopt a single-propagating channel in the quasi-one-dimensional approximation (see the previous section). Propagating waves in the leads and perfect regions of the arms are assumed to be in the form $\psi(x, y) \propto e^{\pm ikx}\varphi(y)$, where x is the local coordinate along the waveguide, y is the transverse coordinate (transverse wave function is $\varphi(y)$), and wave vector $k = \sqrt{2mE}/\hbar$ of an electron with energy E in the open channel (we shift the origin of energy to the band edge of the first subband).

In order to connect incoming and outgoing waves at the junctions of the ring and the leads, we employ a simple junction model [11] where a scattering matrix describes the splitting of the electron wave functions at the junction. Using the amplitudes of electron waves in the ring where the relevant parameters are defined in Fig. 5, the electron transmission from left to right leads through both the upper and lower arms can be expressed by

$$\begin{pmatrix} \bar{A}_u \\ \bar{B}_u \end{pmatrix} = e^{-i\theta/2} M_u \begin{pmatrix} A_u \\ B_u \end{pmatrix}, \qquad \begin{pmatrix} \bar{A}_l \\ \bar{B}_l \end{pmatrix} = e^{-i\theta/2} M_l \begin{pmatrix} A_l \\ B_l \end{pmatrix}. \tag{18}$$

Here, $\theta = 2\pi\Phi/\Phi_0$ describes the phase shift introduced by the magnetic flux Φ threading the AB ring ($\Phi_0 = h/e$ is the elementary flux quantum), and M_u and M_l are the transfer matrices through the upper and lower arms, respectively. In order to find the transfer matrix M_l in the lower arm, we consider the QD that is formed by two short-range potential barriers (V_j, $j = 1, 2$). Then, the transfer matrix of each barrier has form

$$M_j = \begin{pmatrix} 1 - iu_j & -iu_j \\ iu_j & 1 + iu_j \end{pmatrix}, \tag{19}$$

where $u_j = mV_j/\hbar k$ with $j = 1, 2$. The dimensionless matrix element of the potential u_j describes the strength of the repulsive potential barrier. Therefore, the transfer matrix M_l for the lower arm with two QDs can be expressed by

$$M_l = X(L_l - L_2)M_2 X(L_2 - L_1)M_1 X(L_1), \tag{20}$$

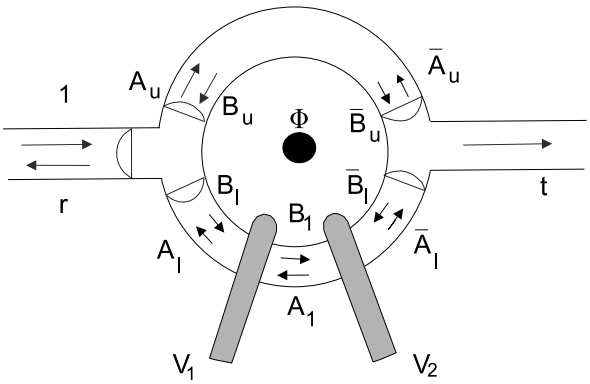
where $X(x) = \mathrm{diag}(e^{ikx}, e^{-ikx})$, L_l is the lower arm length, and the L_1 and L_2 are distances from the left junction to the first and second electrodes, respectively. On the other hand, the transfer matrix M_u for the upper (reference) arm has the simple form, $M_u = X(L_u)$ where L_u is the length of the upper arm.

Using the transfer-matrix and the junction-matrices [11], we can write connections between the amplitudes:

$$t = \sqrt{\varepsilon}(\bar{A}_u + \bar{A}_l), \tag{21}$$

$$A_u = \sqrt{\varepsilon} + aB_u + bB_l, \quad \bar{B}_u = a\bar{A}_u + b\bar{A}_l, \qquad A_l = \sqrt{\varepsilon} + bB_u + aB_l, \quad \bar{B}_l = b\bar{A}_u + a\bar{A}_l. \tag{22}$$

Here, ε plays the role of a coupling parameter between the leads and ring, and the coefficients a and b are expressed



Quantum Dots: Fano Resonances in an Aharonov–Bohm Ring, Figure 5

Geometry of the AB ring with a single dot. Two electrodes in the lower arm produce the depletion regions in the two-dimensional electron gas and play the role of the barriers for electrons with potentials V_1 and V_2

as a function of ε:

$$a = \frac{1}{2}\left(\sqrt{1-2\varepsilon}+1\right)$$

and

$$b = \frac{1}{2}\left(\sqrt{1-2\varepsilon}-1\right).$$

After some matrix manipulations, we obtain the full transmission amplitude $t(E,\Phi)$ analytically as

$$t(E,\Phi) = \frac{\mathrm{i}\varepsilon \mathrm{e}^{\mathrm{i}\theta} N(E,\Phi)}{D(E,\Phi)}. \tag{23}$$

Here, the numerator of Eq. (23) can be written as

$$N(E,\Phi) = \mathrm{e}^{\mathrm{i}(\gamma+2(\delta+\eta))} N_0(E,\Phi), \tag{24}$$

where

$$\begin{aligned} N_0(E,\Phi) = 4\mathrm{i}\{&\sin\gamma + \mathrm{e}^{\mathrm{i}\theta/2}[4u\sin\delta(\cos 2\eta\sin\delta \\ &+ (\cos\delta + u\sin\delta)\sin 2\eta) + \sin 2(\delta+\eta)]\} \end{aligned}$$

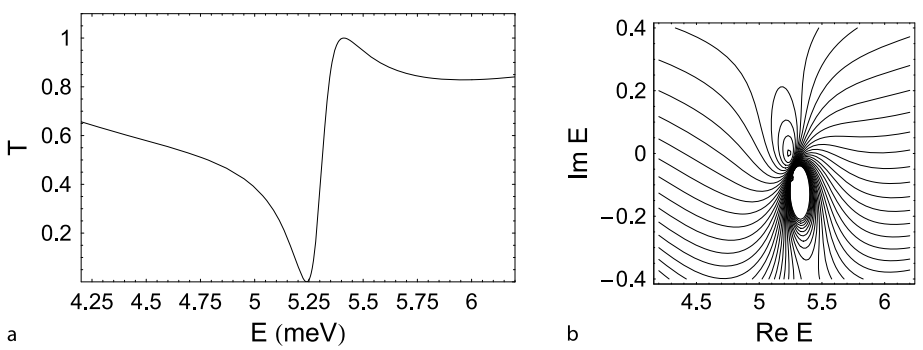
with $u \equiv u_1 = u_2$, $\delta = kL_1$, $\eta = k(L_2 - L_1)$, $\gamma = kL_\mathrm{u}$. On the other hand, the denominator of Eq. (23) can be expressed by

$$\begin{aligned} D(E,\Phi) \equiv D_0(E,\Phi) = {} & \mathrm{e}^{\mathrm{i}(\gamma+2(\delta+\eta))}(1+e^{i\theta}) \\ & + \mathrm{e}^{\mathrm{i}\theta/2}u^2(\mathrm{e}^{4\mathrm{i}\delta} + 4\mathrm{e}^{4\mathrm{i}\eta} + \mathrm{e}^{2\mathrm{i}(\gamma+2\eta)}) \\ & - 4\mathrm{e}^{2\mathrm{i}\delta}u(u-i) + 4(u-\mathrm{i})^2 \\ & - \mathrm{e}^{\mathrm{i}\theta/2}[\mathrm{e}^{2\mathrm{i}\gamma}(u-\mathrm{i})^2 + 4\mathrm{e}^{2\mathrm{i}(\delta+2\eta)}u(u+\mathrm{i}) \\ & - \mathrm{e}^{4\mathrm{i}(\delta+\eta)}(u+\mathrm{i})^2]. \end{aligned}$$

From Eq. (23) we can determine the positions of transmission zeros and resonance poles by setting $N(E,\Phi) = 0$ and $D(E,\Phi) = 0$. We solve these two transcendental equations numerically by using standard routines. In addition, the total transmission probability through the ring as a function of the electron energy and magnetic flux is given by $T = |t(E,\Phi)|^2$ and the conductance is defined by $G = \frac{2e^2}{h}T$ according to the Landauer–Büttiker formalism.

In our calculations, we present results for the following parameters of the ring and dots: the effective electron mass is $m = 0.067m_0$ for GaAs and the geometrical parameters of the ring and positions of gates are chosen to be $L_\mathrm{u} = L_\mathrm{l} = 80\,\mathrm{nm}$, $L_1 = 10\,\mathrm{nm}$ and $L_2 = 70\,\mathrm{nm}$. The scattering parameters describing the QD in the lower arm are used for $a_s V_j = 0.2\,\mathrm{eVnm}$ ($j = 1, 2$), where the width and heights of the barriers are set to $a_s = 4\,\mathrm{nm}$ and $V_1 = V_3 = 50\,\mathrm{meV}$, respectively. Note, that the maximum coupling between the ring and the leads is used with $\varepsilon = \frac{1}{2}$ so that the coefficients of Eq. (23) become $a = -b = \frac{1}{2}$.

The calculated Fano structure is presented in Fig. 6. We show both the Fano resonance in the transmission as a function of electron energy (a) and a contour plot of transmission in the complex-energy plane (b) to illuminate zero-pole structure of the Fano resonance. We have investigated the energy interval near the second quasi-bound state in the two-barrier structure (two electrodes structure). For the chosen parameters of the system, the odd quasi-bound state (second level) has an energy of approximately $E \approx 5.2\,\mathrm{meV}$. In single channel approximation (for instance, when a one-dimensional system with a two-barrier structure is considered) the symmetric Breit–Wigner resonance will be observed in the transmission. Fano resonances in the transmission appear in Fig. 6a due to the quantum interference between the continuum states in the upper arm and the discrete states formed by the QD's states in the lower arm. The contribution of this isolated resonance is described by Eq. (1), with



Quantum Dots: Fano Resonances in an Aharonov–Bohm Ring, Figure 6

The total transmission of a nanoscale AB ring as a function of electron energy (a) and contour plot of the transmission in the complex-energy plane (b). The Fano resonance is shown in a with pole $E_\mathrm{R} = (5.32 - \mathrm{i}0.14)$ meV, and corresponding zero $E_0 = 5.21$ meV

the corresponding parameters of the zero and the pole: $E_R = (5.32 - i0.14)$ meV and zero: $E_0 = 5.21$ meV.

Thus, the single dot produces a Fano resonance in the transmission when the position of the resonance is associated with the level of the dot. The profile depends on the coupling parameters of the dots with the ring.

Effects of Interference Interactions of Fano Resonances on an AB Ring

Next we investigate the effects of the interaction of Fano resonances in the transmission through an AB ring with coupled double QDs. When the overlapping of two Fano resonances takes place in the transmission, two Fano dipoles in the complex-energy plane form a quasi-particle, which behaves as a coupled object – a *Fano quadrupole*. In the regime of strong overlapping resonances, which can be tuned by the interaction parameter between two QDs, the collision of transmission zeros occurs and these zeros leave the real-energy axis and move away in opposite directions in the complex-energy plane. We also obtain an analytical expression of the transmission zeros, and show that these zeros are generally complex when two quasi-bound states lie close together in energy. To explain these effects, an analogous two level system was introduced to demonstrate the feature of interference attraction and repulsion of the zeros. Finally, we show a periodic motion of the resonance pole and transmission zero in the complex-energy plane as a magnetic field through the AB ring is changed.

The model we study is an AB ring where a coupled double QD is embedded in one of its arms, as schematically shown in Fig. 7. Here, double QDs can be formed by three electrodes which play the role of barriers for electrons in the lower arm, and the coupling between two QDs is controlled by the middle electrode.

As shown in the previous section, in order to find the transfer matrix M_l in the lower arm, we consider the coupled QDs that are formed by three short-range potential barriers (V_j, $j = 1, 2, 3$). Therefore, the transfer matrix M_l for the lower arm with two QDs can be expressed by

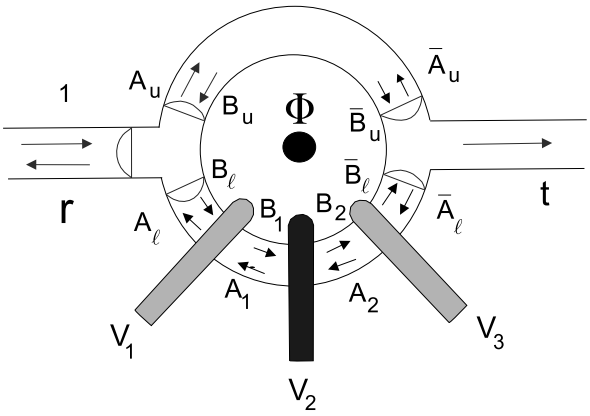
$$M_l = X(L_l - L_3)M_3X(L_3 - L_2)M_2X(L_2 - L_1)M_1X(L_1). \tag{25}$$

After simple calculation we again arrive at Eq. (23), where we now have an additional term, connected with the dots in the form

$$N(E, \Phi) = e^{i(\gamma+2(\delta+\eta))}(N_0(E, \Phi) + \xi N_1(E, \Phi)), \tag{26}$$

and

$$N_1(E, \Phi) = 8iue^{i\theta/2}(\cos\eta \sin\delta + (\cos\delta + 2u\sin\delta)\sin\eta),$$



Quantum Dots: Fano Resonances in an Aharonov–Bohm Ring, Figure 7

Geometry of the AB ring with coupled QDs. Three electrodes in the lower arm produce the depletion regions in the two-dimensional electron gas and play the role of the barriers for electrons with potentials V_1, V_2 and V_3. The interaction between two QDs is controlled by the middle electrode marked by *black color*

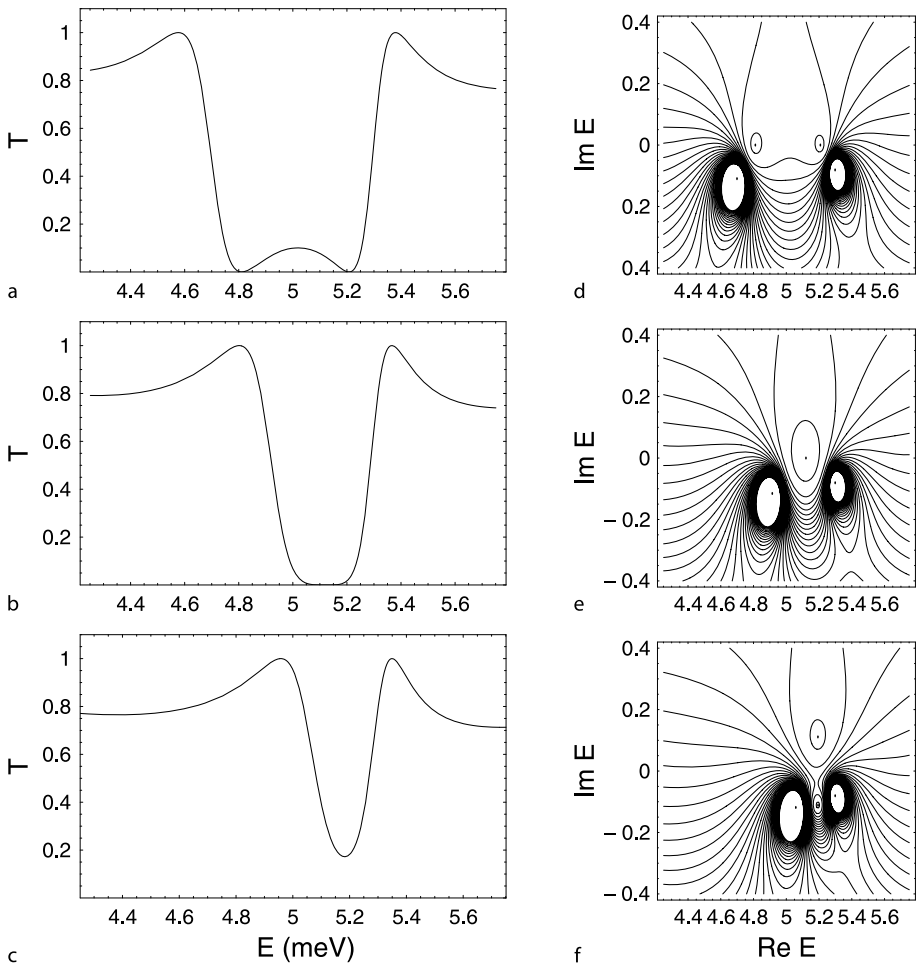
and the interaction parameter between two QDs is $\xi(\equiv u_2/u_1 = V_2/V_1)$. The denominator of Eq. (23) can be expressed for two dots by

$$D(E, \Phi) = D_0(E, \Phi) + \xi D_1(E, \Phi), \tag{27}$$

where

$$\begin{aligned} D_1(E, \Phi) = {} & ie^{i(\theta/2+2\gamma)}u^3(e^{2i\eta} - 1)^2 \\ & - ie^{i(\theta/2-2(\delta+\eta))} \\ & \times \{\sin\delta[3i\cos\eta + \sin\eta(1 + 6iu) \\ & - \cos\delta(\cos\eta + (2u - i3)\sin\eta)^2]\}. \end{aligned}$$

First, we study the interaction of Fano resonances in the transmission through the AB ring in the absence of a magnetic field by investigating the behavior of the transmission amplitude for energy near zero-pole pairs. We consider two quasi-bound states in the double QDs, where even and odd quasi-bound states have the energies E_b and E_a, respectively. In Fig. 8, we show both overlapping of the Fano resonances in the transmission as a function of electron energy and contour plots of the transmission in the complex-energy plane for different values of the interaction parameter ξ. For weak coupling between two QDs ($\xi = 2.5$), two distinct Fano resonances in the transmission appear in Fig. 8a due to the quantum interference between the continuum states in the upper arm and two discrete states from the coupled QDs in the lower arm. The contribution of each separate resonance is described by Eq. (1) with the corresponding parameters of the zero

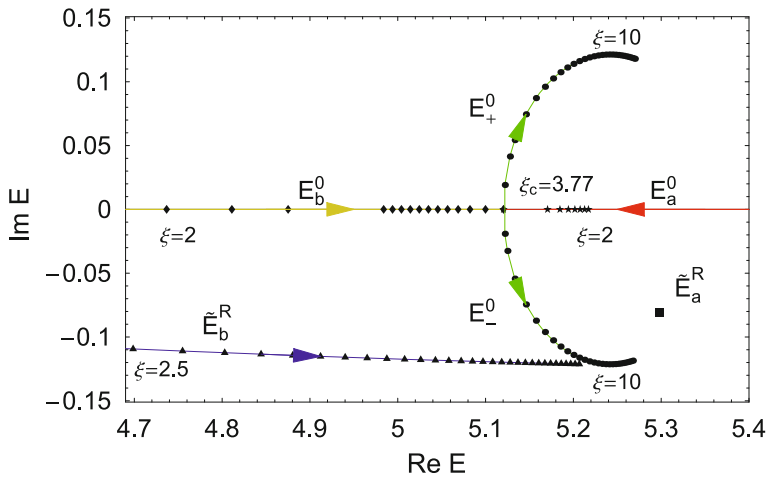


Quantum Dots: Fano Resonances in an Aharonov–Bohm Ring, Figure 8

The total transmission of a nanoscale AB ring as a function of electron energy (the *left column*) and contour plots of the transmission in the complex-energy plane (the *column*) with increasing of the interaction parameter ξ. The two distinct Fano resonances are shown in a and d for the weak overlapping regime ($\xi = 2.5$) with poles $E_b^R = (4.70 - i0.11)$ meV and $E_a^R = (5.30 - i0.08)$ meV (shown by *circles*), and corresponding zeros $E_b^0 = 4.81$ meV and $E_a^0 = 5.21$ meV (both of them are placed on real axis of energy. The collision of Fano resonances and merging of transmission zeros appear in b and e when $\xi = \xi_c = 3.77$, when the zeros have position $E_b^0 = E_a^0 = 5.12$ meV. The pole corresponding with the bounding level approaches to second pole with the increasing of ξ : $E_b^R = (4.91 - i0.12)$ meV and $E_a^R = (5.30 - i0.08)$ meV for $\xi = \xi_c$. In the strong overlapping regime of Fano resonances ($\xi = 5.5$), the transmission zeros move away in opposite directions from the real-energy axis in c and f: $E_+^0 = (5.19 + i0.11)$ meV and $E_-^0 = (5.19 - i0.11)$ meV (the poles are placed $E_b^R = (5.06 - i0.12)$ meV and $E_a^R = (5.30 - i0.08)$ meV)

and the pole. Notice that the width Γ_a of the Fano resonance through the odd quasi-bound state at $E \approx 5.3$ meV is less than the width Γ_b of Fano resonance through the even quasi-bound state at $E \approx 4.7$ meV ($\Gamma_a < \Gamma_b$). The two transmission zeros (E_a^0 and E_b^0) and two resonance poles ($\tilde{E}_a^R = E_a - i\Gamma_a$ and $\tilde{E}_b^R = E_b - i\Gamma_b$) in the complex-energy plane are seen in Fig. 8d, where the two zeros are separately on the real-energy axis. As ξ becomes a critical value $\xi_c = 3.77$, the two Fano resonances in the transmission merge (Fig. 8b) and the transmission zeros move toward each other and collide on the real-energy axis (Fig. 8e). When $\xi = 5.5 > \xi_c$, the minimum of the Fano resonance in the transmission does not reach to zero (Fig. 8c) and the transmission zeros leave the real-energy axis and move away in opposite directions in the complex-energy plane (Fig. 8f).

In order to see the behavior of transmission zeros and resonance poles in detail, we calculate the trajectories of the zeros and poles with the explicit expressions from the equations $N(E, \Phi) = 0$ and $D(E, \Phi) = 0$. In Fig. 9, we



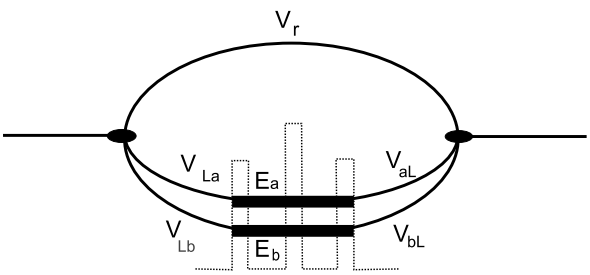
Quantum Dots: Fano Resonances in an Aharonov–Bohm Ring, Figure 9

The trajectories of resonance poles ($\tilde{E}_a^R$, $\tilde{E}_b^R$) and transmission zeros (E_a^0, E_b^0) in the complex-energy plane with increasing of the coupling parameter ξ ($2 < \xi < 10$). When ξ increases, two zeros E_a^0 (*red arrow*) and E_b^0 (*yellow arrow*) move toward each other and the collision and merging take place at $\xi_c = 3.77$. After merging, two zeros move away from the real-energy axis in opposite directions as complex conjugate pairs (denoted by E_+^0 and E_-^0 with *green arrows*). The resonance pole $\tilde{E}_b^R$ associated with even quasi-bound state moves to the higher energy (*blue arrow*), and $\tilde{E}_a^R$ arising from odd quasi-bound state is nearly motionless ($\tilde{E}_a^R \approx (5.30 - i0.08)$ meV)

present the trajectories of the zeros and poles in the complex-energy plane for a variation of the interaction parameter ξ. As ξ increases, one of Fano resonance zeros E_b^0 (shown as diamonds), arising from the even quasi-bound state in the QDs, moves to higher energy and another zero E_a^0 (shown as stars), arising from the odd quasi-bound state in the QDs, moves to lower energy. When $\xi = \xi_c$, the collision and merging of E_b^0 and E_a^0 takes place at $E_a^0 = E_b^0 = 5.12$ meV. In the strong overlapping regime of Fano resonances ($\xi > \xi_c$), the transmission zeros, denoted by E_+^0 and E_-^0 (shown as circles), move away from the real-energy axis in opposite directions as complex conjugate pairs.

It is interesting to note that the behavior of the resonance poles ($\tilde{E}_b^R$ and $\tilde{E}_a^R$) is different from that of the transmission zeros (E_b^0 and E_a^0). The resonance pole $\tilde{E}_b^R$ arising from the even quasi-bound state in the QDs is shifted to higher energy as ξ increases. However, the movement of $\tilde{E}_b^R$ to higher energy is hindered by one of the transmission zeros E_-^0 at $E \approx 5.24$ meV which prevents the collision of the resonance poles $\tilde{E}_b^R$ and $\tilde{E}_a^R$. On the other hand, the resonance pole $\tilde{E}_a^R$ arising from the odd quasi-bound state in the QDs is nearly motionless because the odd quasi-bound state of the double QDs is not strongly affected by ξ, due to short-range potentials in the lower arm of the ring.

In order to confirm the movement of the zeros of the transmission amplitude away from the real axis in the case



Quantum Dots: Fano Resonances in an Aharonov–Bohm Ring, Figure 10

A schematic diagram for even and odd quasi-bound states with energies E_b and E_a, connecting with the junction states of the ring by the appropriate matrix elements V_{Lb}, V_{bL}, V_{aL} and V_{La}. The matrix element of the junctions through the upper arm is denoted by V_r.

of overlapping Fano resonances, we calculate the transmission zeros analytically by employing a simple model shown in Fig. 10. We consider two nearest quasi-bound states in the dots which are even and odd levels with energies E_b and E_a, and bare eigenfunctions ψ_b^0 and ψ_a^0, respectively. These two orthogonal states of the dots in the lower arm can be connected with the junction states of the ring by the appropriate matrix elements V_{Lb}, V_{bL}, V_{aL} and V_{La}. On the other hand, V_r is defined as the matrix element of the junctions through the reference arm. Notice that these matrix elements are generally complex num-

bers and they are dependent on the phase difference between localized wave functions in the dots and propagating waves in the leads. If the energy of an incoming electron from the lead is near the resonant energies of the dots, then the matrix elements that couple an even (odd) quasi-bound state to the left and of the dots are real and have same (opposite) signs: $V_{\mathrm{Lb}} > 0$ and $V_{\mathrm{bL}} > 0$ ($V_{\mathrm{La}} > 0$ and $V_{\mathrm{aL}} < 0$) (this property has been noted in [7] where the electron tunneling through a semiconductor barrier with complex dispersion was investigated). We shall consider what happens with the two level system (see Fig. 10) when we couple it with the leads by effective matrix elements

$$U_{\mathrm{aa}} = \frac{V_{\mathrm{aL}} V_{\mathrm{La}}}{V_{\mathrm{r}}}\,, \quad U_{\mathrm{bb}} = \frac{V_{\mathrm{bL}} V_{\mathrm{Lb}}}{V_{\mathrm{r}}} \quad \text{and} \\ U_{\mathrm{ab}} = U_{\mathrm{ba}} = \frac{V_{\mathrm{aL}} V_{\mathrm{Lb}}}{V_{\mathrm{r}}}\,. \tag{28}$$

Using the Schrödinger equation, it may be shown that the zeros are solutions of the eigenvalue problem

$$\begin{pmatrix} E_{\mathrm{a}} + U_{\mathrm{aa}} & U_{\mathrm{ab}} \\ U_{\mathrm{ba}} & E_{\mathrm{b}} + U_{\mathrm{bb}} \end{pmatrix} \begin{pmatrix} \psi_{\mathrm{a}} \\ \psi_{\mathrm{b}} \end{pmatrix} = E \begin{pmatrix} \psi_{\mathrm{a}} \\ \psi_{\mathrm{b}} \end{pmatrix}\,. \tag{29}$$

In the other words, the problem of finding the zeros from an exact expression for the scattering amplitude is equivalent to the problem of finding zeros from (29). This correspondence takes place because the interaction matrix obeys the property $U_{\mathrm{aa}}U_{\mathrm{bb}} - U_{\mathrm{ab}}U_{\mathrm{ba}} = 0$. The determinant of the system (29) is defined by

$$N(E) = (E-E_{\mathrm{b}})(E-E_{\mathrm{a}}) - (E-E_{\mathrm{b}})U_{\mathrm{aa}} - (E-E_{\mathrm{a}})U_{\mathrm{bb}}\,. \tag{30}$$

Then, the transmission zeros (E^0_{b} and E^0_{a}) can be exactly obtained from the equation $N(E) = 0$, which gives

$$E^0_{a,b} = \frac{1}{2}(\bar{E}_{\mathrm{a}} + \bar{E}_{\mathrm{b}}) \pm \sqrt{(\bar{E}_{\mathrm{a}} - \bar{E}_{\mathrm{b}})^2 + 4U_{\mathrm{aa}}U_{\mathrm{bb}}})\,, \tag{31}$$

where $\bar{E}_{\mathrm{a}} = E_{\mathrm{a}} + U_{\mathrm{aa}}$, $\bar{E}_{\mathrm{b}} = E_{\mathrm{b}} + U_{\mathrm{bb}}$. Because $U_{\mathrm{bb}} > 0$ and $U_{\mathrm{aa}} < 0$ there is an effective interference attraction between levels ($\bar{E}_{\mathrm{a}} = E_{\mathrm{a}} - |U_{\mathrm{aa}}|$, $\bar{E}_{\mathrm{b}} = E_{\mathrm{b}} + U_{\mathrm{bb}}$). An examination of Eq. (31) indicates that when the distance between levels $\bar{E}_{\mathrm{a}} - \bar{E}_{\mathrm{b}}$ is less than the effective interaction, $2\sqrt{|U_{\mathrm{aa}}|U_{\mathrm{bb}}}$, in the other words when

$$E_{\mathrm{a}} - E_{\mathrm{b}} < U_{\mathrm{bb}} + |U_{\mathrm{aa}}| + 2\sqrt{|U_{\mathrm{aa}}|U_{\mathrm{bb}}}\,. \tag{32}$$

E^0_{b} and E^0_{a} are off from the real axis of energy. This implies that when even and odd quasi-bound states lie close together in energy, the transmission zeros move off from the real axis of energy. Using the location of these zeros and poles that are generally complex numbers, the transmission amplitude of Eq. (23) near the resonances in Fano overlapping regime can be expressed as

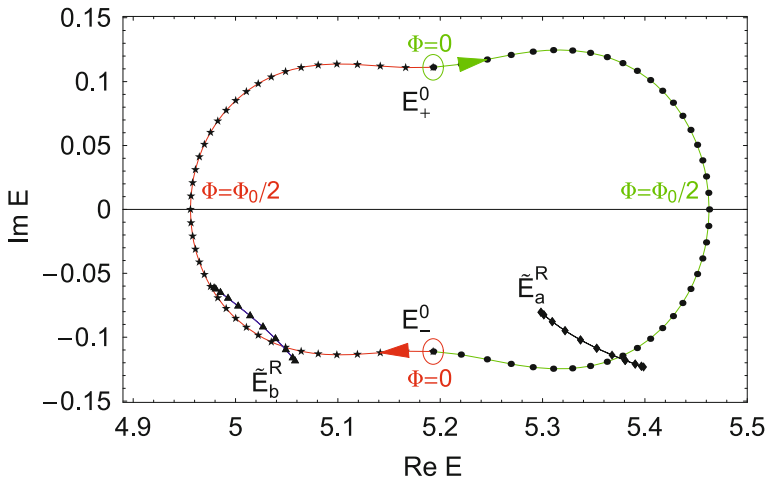
$$t_{\mathrm{F}}(E) \propto \frac{(E - E^0_{\mathrm{a}})(E - E^0_{\mathrm{b}})}{(E - E^+_{\mathrm{b}} \mathrm{i}\Gamma_{\mathrm{b}})(E - E^+_{\mathrm{a}} \mathrm{i}\Gamma_{\mathrm{a}})}\,, \tag{33}$$

which characterizes the transmission line shape in the vicinity of the transmission zeros and resonance poles.

Since the coupled Fano resonances in the transmission can be tuned by the magnetic flux Φ threading the AB ring, we study the magnetic field dependence of the transmission zeros and resonance poles in the strong Fano overlapping regime. The trajectories of the zeros and poles in the complex-energy plane as a function of magnetic flux are shown in Fig. 11 for a given interaction parameter $\xi = 3.9$. As Φ increases, the two zeros at $E^0_+ = (5.19 + \mathrm{i}0.11)\,\mathrm{meV}$ and $E^0_- = (5.19 - \mathrm{i}0.11)\,\mathrm{meV}$ for $\Phi = 0$ start to move to the higher (green arrow with circles) and lower (red arrow with stars) energies, respectively until they reach to the real axis. When $\Phi = \frac{1}{2}\Phi_0$, these complex conjugate pairs become real numbers ($E^0_+ = 5.45\,\mathrm{meV}$ and $E^0_- = 4.95\,\mathrm{meV}$) which indicates full reflection of the transmission. When $\frac{1}{2}\Phi_0 < \Phi < \Phi_0$, these zeros continue to move on the other half of the trajectories to complete a stadium-like orbit. Then, the zero E^0_+ comes to initial position of E^0_- and vice versa for $\Phi = \Phi_0$. In other words, the upper zero E^0_+ circumscribes half of the orbit and then the lower zero E^0_- replaces the same part of the orbit after the flux is changed by one period. On the other hand, the behavior of the resonance poles is quite different. As Φ increases, each pole at $E^{\mathrm{R}}_{\mathrm{b}} = (5.06 - \mathrm{i}0.12)\,\mathrm{meV}$ (shown as triangles) and $E^{\mathrm{R}}_{\mathrm{a}} = (5.30 - \mathrm{i}0.08)\,\mathrm{meV}$ (shown as diamonds) for $\Phi = 0$ moves a short nearly-straight line periodically. (If $\Phi = \frac{1}{2}\Phi_0$, these two poles are located at the opposite end of the starting points along these lines.)

It is worthwhile noting here that the transmission through the ring is a periodical function of magnetic flux, which changes the interference between the parts of the wave function in the arms. Hence, the transmission zeros are more sensitive to the magnetic flux because they are defined by wave interference. However, the resonance poles are defined by quasi-bound energy and move slowly in a magnetic field.

Finally, we present an analytical expression for zero trajectories in the presence of magnetic field using the same model shown in Fig. 10. If a magnetic flux is applied to the loop in the perpendicular direction, each wave function acquires a phase shift on the links. The phase shift on the links connecting the dots with the leads is $\phi/2$ so that the matrix elements between the dots and the leads are



Quantum Dots: Fano Resonances in an Aharonov–Bohm Ring, Figure 11
The trajectories of zeros and poles as a function of a magnetic flux for a fixed interaction parameter $\xi = 3.9$. The two zeros for $\phi = 0$ at $E^0_+ = (5.19 + i0.11)$ meV and $E^0_- = (5.19 - i0.11)$ meV start to move to the (*green arrow*) and left (*red arrow*), respectively, and they reach to the real axis for that is a full reflection of the transmission. These zeros continue to move on the other half of the trajectories and complete a figure-of-eight orbit for $\frac{1}{2}\Phi_0 < \Phi < \Phi_0$. The resonance poles at $E^R_b = (5.06 - i0.12)$ meV and $E^R_a = (5.30 - i0.08)$ meV for $\Phi = 0$ move a short nearly-straight line periodically

replaced by $V_{n,m} \to V_{n,m}e^{\pm i\phi/2}$, where the sign depends on the propagation direction of an electron. On the other hand, the phase shift through the reference arm is ϕ and so we can set $V_r \to V_r e^{-i\phi}$. Then, the effective matrix elements of the two level system in a magnetic field, Eq. (29), can be written as $U_{aa} \to U_{aa}e^{i2\phi}$ and $U_{bb} \to U_{bb}e^{i2\phi}$. The analytical expression for the transmission zeros in the presence of a magnetic field is

$$[b]E^0_{a,b} = \frac{1}{2}\Big(E_a + E_b + (U_{aa} + U_{bb})e^{i2\phi} \pm \sqrt{(E_a - E_b + (U_{aa} - U_{bb})e^{i2\phi})^2 + 4U_{aa}U_{bb}e^{i4\phi}}\Big). \quad (34)$$

This expression indicates the periodic trajectories for the transmission zeros as Φ changes, which is shown in Fig. 11. The analysis of the two level model gives a simple explanation for why the zeros return back to the real axis when the flux equals $\Phi = \frac{1}{2}\Phi_0$. Since for this value of flux the signs of matrix elements are reversed: $U_{bb} < 0$ and $U_{aa} > 0$, the levels are defined now by $\tilde{E}_a = E_a + |U_{aa}|$, and $\tilde{E}_b = E_b - U_{bb}$. This means that the phase shift results in an effective repulsion of the levels and we obtain an inequality $2\sqrt{|U_{aa}|U_{bb}} < \tilde{E}_a - \tilde{E}_b$, which is opposite to (32) (notice that the parameter $2\sqrt{|U_{aa}|U_{bb}}$ has not changed).

We note that for the non-overlapping regime when Fano dipoles exist, the zeros move in separate circular orbits near their associated poles (not shown here). In the strong overlapping regime, however, the zeros act as a coupled object and move in a common *figure-of-eight* orbit around the two coupled poles. This means that two overlapped Fano resonances can be considered as a combined object called a *Fano quadrupole*.

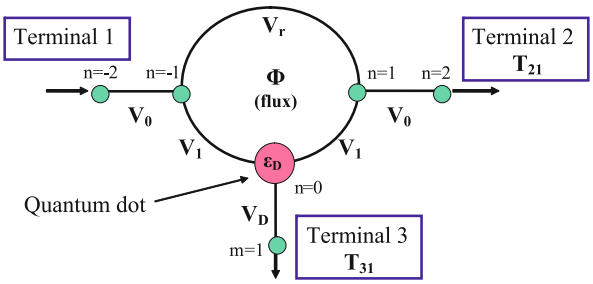
Fano Resonance Induced by a Quantum Dot in an Open Three-Terminal Interferometer

The properties of Fano resonances may be easily demonstrated by the use of simple lattice models in which the dots are represented by sites which are connected by one-dimensional tight-binding chains with different topology. In fact, most of properties of the resonances in the AB ring are the same as in waveguide model. We shall discuss here the phase measurements by using the tight binding model [2,3,4,5,6,13,19,22,41,46].

In this section, we analyze a three-terminal interferometer with an embedded QD in one arm of the AB ring. The general structure studied in our calculations is a three-terminal AB ring with magnetic flux through the ring and an embedded QD, sketched schematically in Fig. 12 where relevant parameters are defined. By discretizing the system spatially with lattice constant a, and denoting the wave function on site n by ψ_n, the Schrödinger equation in the tight-binding approximation can be written as

$$-\sum V_{n,m}\psi_m + \varepsilon_n\psi_n = E\psi_n\,. \quad (35)$$

Here, the sum runs over the nearest neighbors of n, E is the electron energy, and ε_n is the site energy. (In our cal-



Quantum Dots: Fano Resonances in an Aharonov–Bohm Ring, Figure 12

Schematic of the three-terminal interferometer with a QD embedded in one of the arms. In addition to the magnetic flux Φ threading the AB ring, the relevant coupling parameters between sites are defined: the confinement V_1 of the QD, coupling V_D to the third terminal, and coupling V_r through the reference arm of the ring.

culations, the site energies ε_n are set to zero for all sites except for the QD at $n = 0$ which has site energy ε_D.) The parameters $V_{n,m}$ are overlap integrals (or coupling parameters) involving the overlap of the single site, atomic-like wave functions from sites m and n with the single-site potential of site n. In the homogeneous leads, the coupling parameters are all set to $V_0 = 1.0$, which we use throughout the discussion as a unit of energy. In the presence of the magnetic flux Φ, a phase difference between the path through the QD and the path through the reference arm is produced [1]. Therefore, we choose a gauge in which the coupling parameter for each segment of the lower arm is modified as $V_1 \to V_1 e^{\pm i\varphi}$, and the reference arm coupling parameter becomes $V_r e^{\pm 2i\varphi}$ ("+" for counter-clockwise transits around the ring and "−" for clockwise transits). The phase φ is related to the magnetic flux Φ by $2\varphi = \pi\Phi/\Phi_0$.

Let us consider an incoming wave function only from terminal 1, with transmitted waves through the ring into the other terminals:

$$\begin{aligned} \psi_n &= e^{in\theta} + r_{11}e^{-in\theta}\,, \quad && n \le -1\,, \\ \psi_n &= t_{21}e^{in\theta}\,, && n \ge 1\,, \\ \psi_m &= t_{31}e^{im\theta}\,, && m \ge 1\,, \end{aligned} \tag{36}$$

with $\theta = ka$. Here, k is the wave vector that is connected with the energy by the dispersion relation: $E = -2V_0 \cos ka$, t_{21} and t_{31} are the transmission amplitudes from terminal 1 into terminal 2 and 3 respectively; and r_{11} is the reflection amplitude back to terminal 1. Applying the Schrödinger equation to the three sites around the AB ring and also to site $m = 1$ of the third terminal, we obtain the following matrix equation for the complex transmission amplitudes:

$$\begin{pmatrix} V_0 & -V_r e^{-i(2\varphi-\theta)} & -V_1 V_0 e^{i\varphi}/V_D \\ -V_r e^{i(2\varphi+\theta)} & V_0 & -V_1 V_0 e^{-i\varphi}/V_D \\ -V_1 e^{-i(\varphi-\theta)} & -V_1 e^{i(\varphi+\theta)} & -(V_D e^{i\theta} + (E-\varepsilon_D)V_0/V_D) \end{pmatrix} \times \begin{pmatrix} r_{11} \\ t_{21} \\ t_{31} \end{pmatrix} = \begin{pmatrix} -V_0 \\ V_r e^{i(2\varphi-\theta)} \\ V_1 e^{-i(\varphi+\theta)} \end{pmatrix}. \tag{37}$$

Inverting the matrix on the left side of Eq. (37), we can find the unknown reflection and transmission amplitudes: r_{11}, t_{21} and t_{31}. In the same way, we can find the other elements of the scattering matrix: r_{22}, t_{12}, t_{32} and r_{33}, t_{13}, t_{23} when incoming waves are chosen from terminals 2 and 3, respectively. The transmission amplitudes for an electron from terminal (channel) j into terminal (channel) i may be written in the form

$$t_{ij}(\Phi) = \frac{N_{ij}(\Phi)}{D(\Phi)}\,, \tag{38}$$

where we have

$$N_{21}(\Phi) = 2iV_0 \sin\theta e^{2i\varphi}\left[V_r\left(V_0(E-\varepsilon_D) + e^{i\theta}V_D^2\right) - e^{-4i\varphi}V_0V_1^2\right], \tag{39}$$

$$N_{31}(\Phi) = -2iV_0V_1V_D \sin\theta e^{-i\varphi}\left[V_0 + e^{i(4\varphi+\theta)}V_r\right], \tag{40}$$

$$\begin{aligned} D(\Phi) =& e^{i\theta}V_0^2\left(2V_1^2 + V_D^2\right) - e^{3i\theta}V_r^2V_D^2 \\ &+ V_0\left(V_0^2 - e^{2i\theta}V_r^2\right)(E-\varepsilon_D) \\ &+ 2e^{2i\theta}V_0V_rV\cos(4\varphi)\,, \end{aligned} \tag{41}$$

with symmetry conditions

$$D(\Phi) = D(-\Phi)\,, \quad N_{ij}(\Phi) = N_{ji}(-\Phi)\,, \quad \text{and} \quad N_{13}(\Phi) = N_{23}(\Phi)\,. \tag{42}$$

Thus, we have all the transmission coefficients $T_{ij}(\Phi) = |t_{ji}(\Phi)|^2$, which obey the property

$$T_{ij}(\Phi) = T_{ji}(-\Phi)\,. \tag{43}$$

In order to find the non-local conductance of the open ring, we use the Büttiker equations [9]

$$I_i = \frac{2e}{h}\left[(1-R_{ii})\mu_i - \sum_{j\neq i} T_{ij}\mu_j\right], \quad (i,j = 1,2,3)\,, \tag{44}$$

where $R_{ij}(\Phi) = |r_{ji}(\Phi)|^2$ are the reflection coefficients (which may be eliminated from the set of equations by using current conservation: $1 - R_{ii} = \sum_{j\neq i} T_{ij}$), and μ_i are

the chemical potentials of the reservoirs (terminals). The factor of two in Eq. (44) stems from the identical contribution of both electron spin states. Here, it is noteworthy that we consider a typical situation in which two terminals (1 and 2) are used for injection of current and measurements of the conductance $G_{12,12}$ (see Büttiker's notations in Ref. [9]), whereas the potential drop (which is characterized by the resistance $R_{12,13}$) is measured only between terminals 1 and 3. For our purpose, we set the current between terminals 1 and 2 as: $I \equiv I_1 = -I_2$. Thus, terminal 3 represents an ideal probe that draws no net current ($I_3 = 0$). Solving the set of Eq. (44), we find the coefficient

$$G_{12,12} = \frac{2e^2}{h}\left(T_{21} + \frac{T_{23}T_{31}}{T_{31}+T_{32}}\right) \tag{45}$$

between the current I and the bias $U_{12} = (\mu_1 - \mu_2)/e$. The potential drop U_{13} between the terminals 1 and 3 is defined by the resistance $R_{12,13}$: $U_{13} = R_{12,13}I$, where

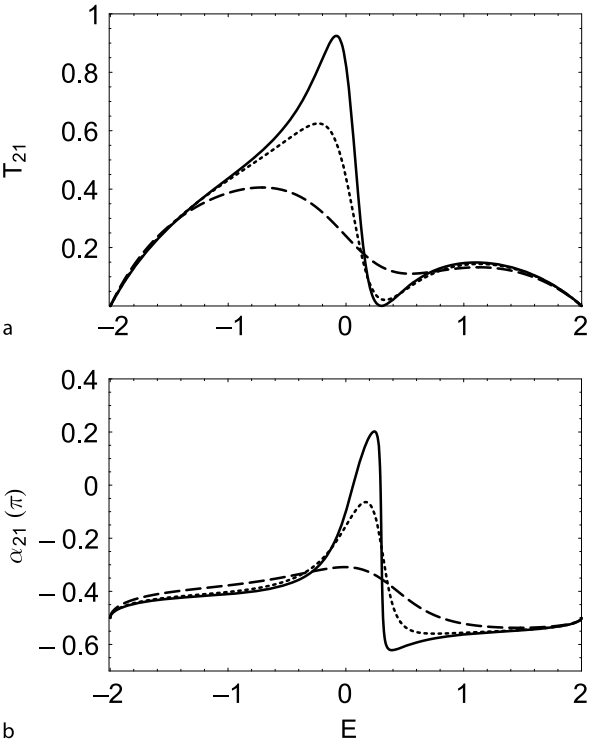
$$R_{12,13} = \frac{h}{2e^2}\left(\frac{T_{32}}{T_{21}T_{31} + T_{23}T_{31} + T_{32}T_{21}}\right) . \tag{46}$$

Here, we present results for the following parameters of the system: $V_1 = 0.3$ (QD confinement), $V_r = 0.3$ (the coupling through the reference arm of the AB ring), and $\varepsilon_D = 0$ (the site energy of the QD which positions the resonance in the center of the allowed energy band).

Now, we study the effect of coupling to the third output terminal on the transmission $T_{21} = |t_{21}|^2$ through the AB ring in the absence of the magnetic flux. Here, the transmission amplitute t_{21} can be calculated from Eq. (38) as

$$t_{21} = \frac{2\mathrm{i}\sin\theta\left[V_1^2 - \mathrm{e}^{4\mathrm{i}\varphi}V_r\left(E - \varepsilon_D + \mathrm{e}^{\mathrm{i}\theta}V_D^2/V_0\right)\right]}{\mathrm{e}^{2\mathrm{i}\theta}V_r V_1^2\left(\mathrm{e}^{6\mathrm{i}\varphi} + \mathrm{e}^{-2\mathrm{i}\varphi}\right)/V_0 + \mathrm{e}^{2\mathrm{i}\varphi}\left[\left(2V_1^2 + V_D^2\right)\mathrm{e}^{\mathrm{i}\theta} - \mathrm{e}^{3\mathrm{i}\theta}V_r^2V_D^2/V_0^2 + (E-\varepsilon_D)\left(V_0 - \mathrm{e}^{2\mathrm{i}\theta}V_r^2/V_0\right)\right]} \tag{47}$$

The behavior of the transmission zero and phase in a three-terminal interferometer as a function of energy for various values of coupling to the third terminal ($V_D = 0.1$ (solid), $V_D = 0.3$ (dotted), and $V_D = 0.6$ (dashed)) is shown in Fig. 13. Unlike a two-terminal closed AB ring with QD's ($V_D = 0$) [22,41], it is seen from Fig. 13a that in an open ring ($V_D \neq 0$) the Fano resonance peak does not reach unity because of energy loss due to the outgoing electrons into the third terminal. In addition, as V_D increases, the Fano zero (obtained from $N_{21} = 0$ in Eq. (39)) shifts progressively further off the real-energy axis into the complex-energy half-plane. The Fano zero can be returned to the real energy axis at discrete values of magnetic flux (see below) as long as V_D is less than a critical value. This flexible control over the transmission resonance features is unavailable in a closed, two-terminal interferometer.



Quantum Dots: Fano Resonances in an Aharonov–Bohm Ring, Figure 13

Transmission T_{21} and transmission phase α_{21} as a function of energy for different values of $V_D = 0.1$ (*solid curve*), $V_D = 0.3$ (*dotted curve*), and $V_D = 0.6$ (*dashed curve*). a As V_D increases, the Fano resonance no longer reaches unity and the Fano zero lifts off the real-energy axis. b The phase jump of π at the transmission resonance diminishes and softens as the ring is opened with coupling to the third terminal

In Fig. 13b, we show the transmission phase as a function of energy for different values of V_D. The transmission phase α_{21}, which can be calculated from Eq. (47) as $\alpha_{21} = \tan^{-1}[\mathrm{Im}(t_{21})/\mathrm{Re}(t_{21})]$, no longer changes abruptly by π at the resonance, as for non-zero V_D, but is shown to progressively soften and to smoothly change by less than π as V_D increases. We attribute this smearing of the abrupt phase jump of π to the fact that current, which flows to the third terminal, breaks unitarity and disrupts the interference effects due to repeated reflections of the electrons from the junctions and back through the ring.

In order to illustrate the transmission phase changing from an abrupt jump to a smooth transition of less than π

in a three-terminal AB ring, we calculate the phase using a simple analytical model which is based on the properties of the Fano resonance. In the vicinity of the Fano resonance in the transmission versus electron energy for an AB ring with an embedded QD, the transmission amplitude has the form [17]

$$t_{21} \cong t_{\mathrm{bg}} \left(\frac{E - \tilde{E}_0}{E - E_{\mathrm{R}} + i\Gamma} \right) . \tag{48}$$

Here, $\tilde{E}_0$ is the position of the transmission zero and E_{R} gives the energy of the pole. The width of the resonance Γ indicates how far the pole is off the real-energy axis, and t_{bg} represents any background contribution to the amplitude. Because the Fano transmission zero lies off the real-energy axis for an open ring, we can write $\tilde{E}_0$ in a complex form: $\tilde{E}_0 = E_0 - \mathrm{i}\gamma$. The transmission amplitude can now be written as the product of two complex terms:

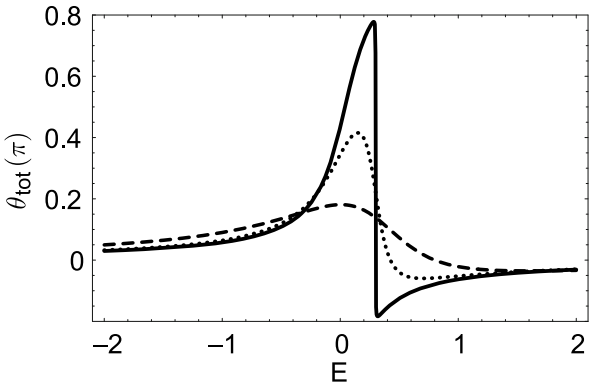
$$t_{21} \cong t_{\mathrm{bg}} \frac{(E - E_0 + \mathrm{i}\gamma)(E - E_{\mathrm{R}} - \mathrm{i}\Gamma)}{(E - E_{\mathrm{R}})^2 + \Gamma^2} = |t_{21}| \mathrm{e}^{\mathrm{i}\theta_{\mathrm{tot}}}, \tag{49}$$

where θ_{tot} is the combined transmission phase from the two complex terms. By separating Eq. (49) into its real and imaginary parts, we obtain an expression for θ_{tot} as

$$\theta_{\mathrm{tot}} = \arctan \left[\frac{(E - E_{\mathrm{R}})\gamma - (E - E_0)\Gamma}{(E - E_{\mathrm{R}})(E - E_0) + \gamma\Gamma} \right] . \tag{50}$$

In Eq. (50) for θ_{tot}, there is no sharp phase jump at a particular value of energy, as exists at $E = E_{\mathrm{R}}$ for a two-terminal closed ring in which the Fano zero is on the real-energy axis ($\gamma = 0$). In Fig. 14, we show plots of θ_{tot} versus E for various coupling parameters V_{D} to the third terminal. It is clearly seen that as V_{D} is increased from zero, the abrupt phase jump of π at the resonance softens and diminishes in magnitude. This indicates that the Onsager relations (unitarity and time reversal-symmetry) are not valid for an open AB interferometer with an embedded QD [2,13].

Since the Fano zero and resonance pole in the transmission can be tuned by the magnetic flux Φ threading the AB ring, we investigate the magnetic flux dependence of transmission T_{21} for a fixed V_{D}. In Fig. 15, the total transmission as a function of electron energy (the left column) and contour plots of the transmission amplitude in the complex-energy plane (the column) with fixed $V_{\mathrm{D}} = 0.3$ are shown for different magnetic flux values: $\Phi/\Phi_0 = 0.0, 0.25, 0.548, 0.75$, and 0.952 (top to bottom). As the magnetic flux is increased from $\Phi = 0$, the Fano zero begins to move on a clockwise orbit around the Fano pole. At $\Phi/\Phi_0 = 0.25$, the zero is positioned directly below the Fano pole in the complex-energy plane.



Quantum Dots: Fano Resonances in an Aharonov–Bohm Ring, Figure 14

Modeled transmission phase θ_{tot} versus energy E for a standard Fano resonance (*solid curve*: $E_0 = 0.3$, $\gamma = 0.0005$, $E_{\mathrm{R}} = 0.04$, $\Gamma = 0.192$), and for modified Fano resonances (*dotted curve*: $E_0 = 0.3$, $\gamma = 0.1$, $E_{\mathrm{R}} = 0.05$, $\Gamma = 0.3$; *dashed curve*: $E_0 = 0.38$, $\gamma = 0.45$, $E_{\mathrm{R}} = 0.1$, $\Gamma = 0.75$) based on the approximate positions of the Fano zeros and poles for $V_{\mathrm{D}} = 0.01, 0.3, 0.6$, respectively. The phase jump of π at the transmission resonance diminishes and softens as V_{D} increases

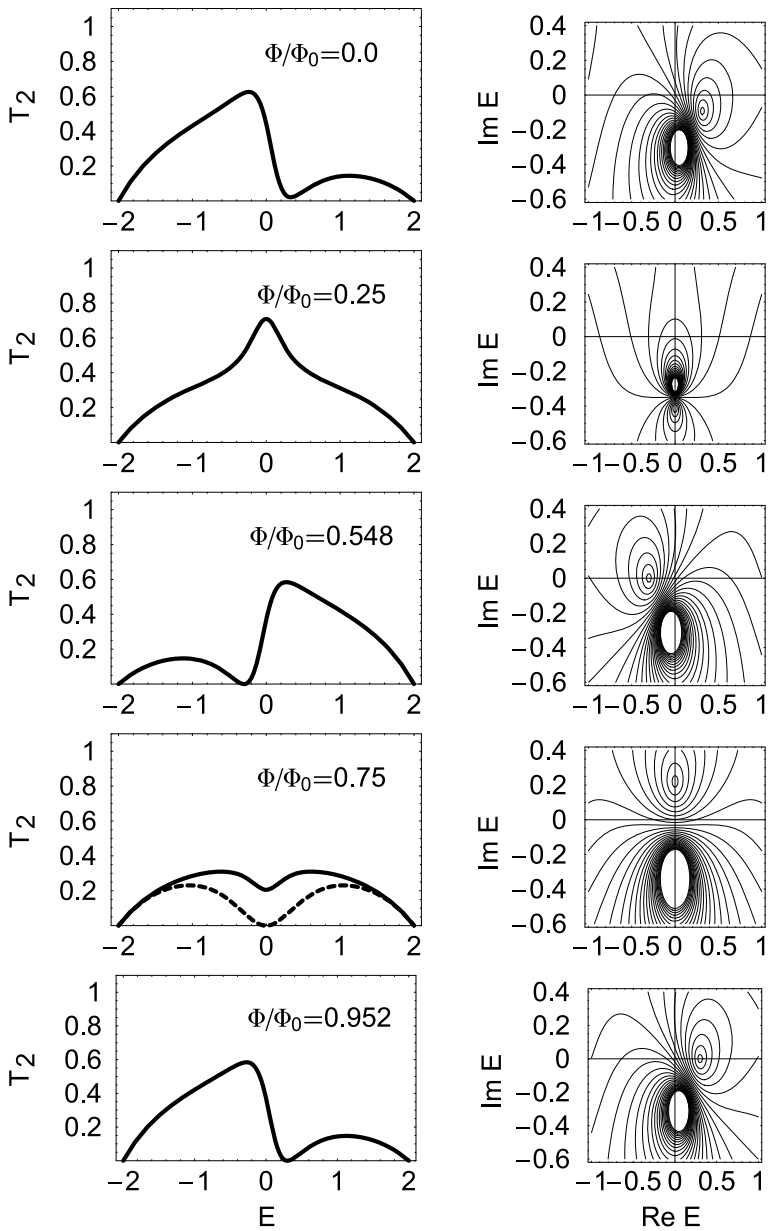
As Φ continues to increase, the Fano zero moves back up towards the real-energy axis and *crosses* the axis at $\Phi/\Phi_0 = 0.548$. When $\Phi/\Phi_0 = 0.75$, the Fano zero arrives directly above the Fano pole, attaining its most positive imaginary value. As the flux is further increased towards $\Phi/\Phi_0 = 1.0$, the Fano zero again *crosses* the real-energy axis at $\Phi/\Phi_0 = 0.952$ on the way back to its position from which it started at $\Phi/\Phi_0 = 0$.

It is interesting to note from Fig. 15 that for a fixed value of V_{D}, there exist two values of magnetic flux for which the Fano zero *crosses* the real-energy axis. By setting $V_{\mathrm{r}}(V_0(E - \varepsilon_{\mathrm{D}}) + \mathrm{e}^{\mathrm{i}\theta} V_{\mathrm{D}}^2) - \mathrm{e}^{-4\mathrm{i}\varphi} V_0 V_1^2 = 0$ from Eq. (39), the analytical expression for the energy values of the Fano zeros (E_0) and the corresponding normalized magnetic flux values (Φ/Φ_0) in terms of the coupling parameter V_{D} can be obtained as

$$E_0 = \pm \sqrt{\frac{\left(V_1^2 V_0/V_{\mathrm{r}}\right)^2 - V_{\mathrm{D}}^4}{V_0^2 - V_{\mathrm{D}}^2}} \qquad \text{and}$$

$$\cos\left[2\pi\left(1 - \frac{\Phi}{\Phi_0}\right)\right] = \pm \sqrt{\frac{1 - \left(V_{\mathrm{r}} V_{\mathrm{D}}^2/V_0 V_1^2\right)^2}{1 - V_{\mathrm{D}}^4/\left(2V_0^2 - V_{\mathrm{D}}^2\right)^2}} . \tag{51}$$

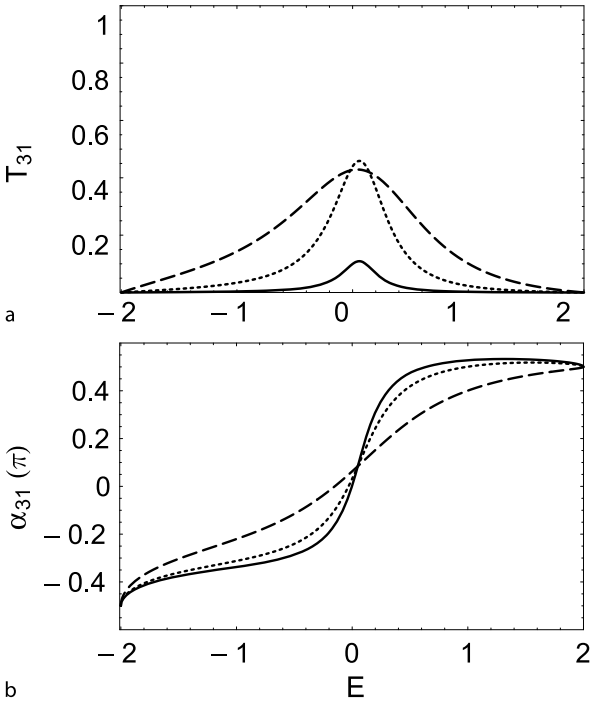
Here, it is required that both the real and imaginary parts of N_{21} from Eq. (39) be zero. For the parameters used in Fig. 15 ($V_0 = 1.0$, $V_1 = V_{\mathrm{r}} = 0.3$, and $V_{\mathrm{D}} = 0.3$), the two Fano zeros are at $E_0 = -0.30$ and 0.30 when



Quantum Dots: Fano Resonances in an Aharonov–Bohm Ring, Figure 15

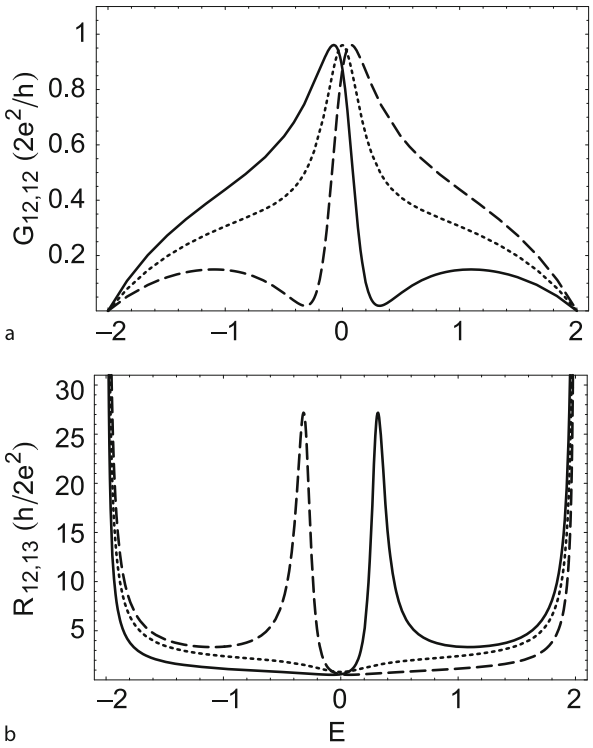
The total transmission as a function of electron energy (the *left column*) and contour plots of the transmission amplitude in the complex-energy plane (the *column*) with fixed $V_D = 0.3$, for different magnetic flux $\Phi/\Phi_0 = 0.0, 0.25, 0.548, 0.75$, and 0.952 (*top to bottom*). The Fano zero moves directly downward and *crosses* the real-energy axis at $V_D = V_D^{crit}$, shown in the *dashed curve* for $\Phi/\Phi_0 = 0.75$

$\Phi/\Phi_0 = 0.548$ and 0.952, respectively. Notice, however, that there is a critical value of V_D^{crit} [$V_D^{crit} = V_1\sqrt{V_0/V_r}$, obtained from requiring E_0 to be real in Eq. (51)], which is the *maximum* value of V_D for which there is the possibility of placing the Fano zero on the real-energy axis at any value of flux. We show in Fig. 15 that the Fano zero moves directly downward and *crosses* the real-energy axis at V_D^{crit} (see the dashed transmission curve for $\Phi/\Phi_0 = 0.75$). When $V_D > V_D^{crit}$, the Fano zero passes into the negative complex-energy half-plane and there is no value of flux which can bring the Fano zero back to the real-energy axis.



Quantum Dots: Fano Resonances in an Aharonov–Bohm Ring, Figure 16

Transmission T_{31} and transmission phase α_{31} as a function of energy for different values of $V_D = 0.1$ (*solid curve*), $V_D = 0.3$ (*dotted curve*), and $V_D = 0.6$ (*dashed curve*). a The BW resonances, which arise from the fact that the amplitude t_{31} does not have zeros in the energy plane, are seen for a variation of V_D. b The phase change near the BW resonance softens as V_D increases



Quantum Dots: Fano Resonances in an Aharonov–Bohm Ring, Figure 17

a The conductance $G_{12,12}$ and b the resistance $R_{12,13}$ are depicted as a function of electron energy with a fixed $V_D = 0.1$ for different magnetic flux $\Phi/\Phi_0 = 0.0$ (*solid curve*), 0.25 (*dotted curve*), and 0.5 (*dashed curve*). As Φ increases, the swing from Fano to BW resonance (or vice versa) appears in the conductance $G_{12,12}$ and the resistance $R_{12,13}$ increases dramatically near the zero of the Fano resonance

In contrast to amplitude t_{21}, the cross amplitudes t_{31} and t_{32} do not have zeros in the actual region of energy plane (see Eq. (40)) and hence, the behavior of the amplitudes near the pole is expected to be similar to that of the amplitudes near a BW resonance. The transmission T_{31} and the transmission phase α_{31} in the absence of a magnetic flux for different coupling parameters, $V_D = 0.1, 0.3$, and 0.6, are shown in Fig. 16 as solid, dotted, and dashed curves, respectively. A simple BW resonance peak in T_{31}, which is less than unity, can clearly be seen in Fig. 16a, and a corresponding phase change α_{31} near the BW resonance is depicted in Fig. 16b. Notice here that α_{31} at the resonance monotonically softens as V_D increases, but the BW peak near $E \approx 0$ has a maximum amplitude at $V_D = 0.5$.

Finally, we investigate the magnetic flux dependence of the conductance $G_{12,12}$ and the resistance $R_{12,13}$ for an open ring with a fixed V_D. In Fig. 17a, the conductance $G_{12,12}$ as a function of electron energy E with a fixed $V_D = 0.1$ is shown for different values of magnetic flux $\Phi/\Phi_0 = 0.0$ (solid), 0.25 (dotted), and 0.5 (dashed). As Φ increases, a transition from Fano resonance (asymmetry parameter $q < 0$, peak $\rightarrow$ dip) to BW resonance and then back to Fano resonance ($q > 0$, dip $\rightarrow$ peak) in $G_{12,12}$ can be observed as a sequence. The Fano resonance produces a very strong influence on the resistance $R_{12,13}$. As shown in Fig. 17b, the resistance increases dramatically when the electron energy approaches the zero energy E_0 of the Fano resonance. The appearance of the peak in the resistance near the zero of the Fano resonance is connected with almost full reflection of the electron waves traveling from terminal 1 to terminal 2. This indicates that at this Fermi energy there is an additional interference resistance in the circuit region between terminals 1 and 3.

Future Directions

Electron transmission through quantum dots and Aharonov–Bohm rings has shown a rich resonance structure,

which includes Fano resonances when the QD is embedded in one arm of the AB ring. The Fano resonance is a manifestation of interference between the localized quasi-bound states of the QD in one arm and the continuum states in the other arm, characterized by both complete transmission and complete reflection. This characteristic of resonance structure (a zero-pole pair) can be controlled by changing the confinement parameters of the QD. Transmission through a QD embedded in an AB ring remains phase-coherent, as indicated by the visibility of the AB oscillations. The intrinsic phase of the QD is significant in its relation to the AB oscillations when the QD is embedded in the AB-ring, and it has experimentally been seen to exhibit interesting phase-jumps of π in a *two-terminal* system when the conductance of the AB ring reaches a peak. In a two-terminal device, the Onsager relations of time-reversal symmetry and current conservation (unitarity) constrain the transmission phase to values of 0 or π. However, if the two-terminal AB-ring is "opened" by allowing current to flow out through additional terminals, the unitarity condition is broken and it becomes possible to extract meaningful phase information about the QD.

We examine new effects resulting from the interaction of Fano resonances in the transmission of an Aharonov–Bohm ring with two embedded quantum dots. As the interaction parameter between two quantum dots is modulated, two Fano dipoles (a resonance zero-pole pair) in the complex-energy plane form a new quasi-particle, which behaves as a coupled object called a *Fano quadrupole*. In the strong overlapping regime of the Fano resonances, the collision and merging of resonance zeros takes place and these zeros move off from the real axis of energy in complex conjugate pairs. A simple two-level model demonstrating the interference attraction and repulsion of the zeros is introduced. The periodic motion of both transmission zeros and resonance poles as a function of a magnetic field is discussed. Although we have investigated the collision of two resonances, our approach may be applied to *multi-resonant* cases in which Fano-complexes in the transmission may be demonstrated.

These predicted effects may be observed in nanoscale devices with geometrical dimensions smaller than the elastic mean free paths. In these nanoscopic systems, electron transport is ballistic and phase coherence can be preserved. Because of this merit of the nanosystem, there have been many studies on electron transmission characteristics in semiconductor structures. Especially, interference of electron waves and Fano resonances in nanostructures have been extensively studied both experimentally and theoretically [2,3,4,5,6,12,13,19,20,21,22,23,24,25,26,27,28,29,31,32,37,38,39,40,41,42,46,47]. For instance, a quantum interference experiment for a quantum dot embedded in an AB ring fabricated in a two-dimensional AlGaAs/GaAs heterostructure was recently performed by Kobayashi et al. [31]. They have studied unique properties of the Fano effect on the phase and coherence of electrons traversing the AB ring. They have also reported in this tunable Fano experiment that the Fano line-shape in the electron transmission through the AB ring is tunable by an external control such as a magnetic field. In other words, the relative phase between a discrete level in the QD and the continuum state changes the Fano-type line shapes which is characterized by a complex number for the asymmetric parameter. Notice that as a result of their small size, the investigated dots are operated in Coulomb blockade regime. Meanwhile, because only a few electrons in the channels were effective in the transport through the dot, the Fano resonances have been observed in the transmission [28,29,31,32,42,46,47]. In these references, it was pointed out that single electron interference effects may be responsible for the formation of resonances.

AB rings with embedded quantum dots may be used for conductance control of quantum interference devices. Fano interference may potentially be used for the design of new types of quantum electronic or spintronic devices such as Fano-transistors [18], spin transistors, and Fano-filters for polarized electrons [44]. In addition, Fano phenomena can also be used for lasing without population inversion [33]. The developed Fano interference theory for electrons in AB rings will allow us to gain an essential understanding of the operation of novel quantum devices, and will open new opportunities for applications.

Acknowledgments

We thank G. Klimeck and C.S. Kim for helpful conversations and collaboration. The work of AMS was supported in part by the Russian Basic Research Foundation Grant No. 05-02-16762.

Bibliography

Primary Literature

1. Aharonov Y, Bohm D (1959) Significance of electromagnetic potentials in the quantum theory. Phys Rev 115:485–491
2. Aharony A, Entin-Wohlman O, Imry Y (2003) Measuring the transmission phase of a quantum dot in a closed interferometer. Phys Rev Lett 90:156802–156805
3. Aharony A, Entin-Wohlman O, Imry Y (2003) Phase measurements in Aharonov–Bohm interferometers. Turk J Phys 27:299–312
4. Aharony A, Entin-Wohlman O, Imry Y (2005) Phase measurements in open and closed Aharonov–Bohm interferometers. Physica E 29:283–288

5. Aharony A, Entin-Wohlman O, Halperin BI, Imry Y (2002) Phase measurement in the mesoscopic Aharonov–Bohm interferometer. Phys Rev B 66:115311–115318
6. Aharony A, Entin-Wohlman O, Otsuka T, Katsumoto S, Aikawa H, Kobayashi K (2006) Breakdown of phase rigidity and variations of the Fano effect in closed Aharonov–Bohm interferometers. Phys Rev B 73:195329, 1–8
7. Bowen RC, Frensley WR, Klimeck G, Lake RK (1995) Transmission resonances and zeros in multiband models. Phys Rev B 52:2754–2765
8. Breit G, Wigner E (1936) Capture of slow neutrons. Phys Rev 49:519–531
9. Büttiker M (1986) Four-terminal phase–coherent conductance, Phys Rev Lett 57:1761–1764
10. Clerk AA, Waintal X, Brouwer PW (2001) Fano resonances as a probe of phase coherence in quantum dots. Phys Rev Lett 86:4636–4639
11. Datta S (1995) Electronic transport in mesoscopic systems. Cambridge University Press, Cambridge
12. Deo PS, Jayannavar AM (1996) Phase of Aharonov–Bohm oscillations in conductance of mesoscopic systems. Mod Phys Lett B 10:787–792
13. Entin-Wohlman O, Aharony A, Imry Y, Levinson Y, Schiller A (2002) Broken unitarity and phase measurements in Aharonov–Bohm interferometers. Phys Rev Lett 88:166801–166805
14. Fano U (1961) Effects of configuration interaction on intensities and phase shifts. Phys Rev 124:1866–1878
15. Goldberg M, Watson K (1964) Collision theory. Wiley, New York
16. Gurvitz SA, Levinson YB (1993) Resonant reflection and transmission in a conducting channel with a single impurity. Phys Rev B 47:10578–10587
17. Kim CS, Satanin AM, Joe YS, Cosby RM (1999) Resonant tunneling in a quantum waveguide: effect of a finite-size attractive impurity. Phys Rev B 60:10962–10970
18. Göres J, Goldhaber-Gordon D, Heemeyer S, Kastner MA, Shtrikman H, Mahalu D, Meirav U (2000) Fano resonances in electronic transport through a single-electron transistor. Phys Rev B 62:2188–2194
19. Hedin E, Joe YS, Satanin AM (2007) Control of Fano resonances and phase of a multi-terminal Aharanov–Bohm ring with three embedded quantum dots, J Comput Electron 6:323–327
20. Hedin E, Cosby RM, Satanin AM, Joe YS (2005) Electron wave interferometry through an asymmetric Aharonov–Bohm ring. J Appl Phys 97:063712–063716
21. Hofstetter W, König J, Schoeller H (2001) Kondo correlations and the Fano effect in closed Aharonov–Bohm interferometers. Phys Rev Lett 87:156803–156807
22. Joe YS, Hedin E, Satanin AM (2007) Manipulation of resonances in an open three-terminal interferometer with an embedded quantum dot. Phys Rev B 76:085419, 1–6
23. Joe YS, Kim J, Satanin AM (2006) Resonance characteristics through double quantum dots embedded in series in an Aharonov–Bohm ring. J Phys D: Appl Phys 39:1766–17667
24. Joe YS, Satanin AM, Kim CS (2006) Classical analogy of Fano resonance. Phys Scr 74:259–266
25. Joe YS, Satanin AM, Klimeck G (2005) Interactions of Fano resonances in the transmission for an Aharonov–Bohm ring with two embedded quantum dots in the presence of a magnetic field. Phys Rev B 72:115310–115316
26. Joe YS, Kim JS, Hedin E, Cosby RM, Satanin AM (2005) Fano resonance through quantum dots in tunable Aharonov–Bohm Rings. J Comput Electron 4:129–133
27. Kang K (1999) Phase evolution of the transmission coefficient in an Aharonov–Bohm ring with Fano resonance. Phys Rev B 59:4608–4611
28. Katsumoto S, Kobayashi K, Aikawa H, Sano A, Iye Y (2003) Quantum coherence in quantum dot-Aharonov–Bohm ring hybrid systems. Superlattices Microstruct 24:151–157
29. Keyser UF, Borck S, Haug RJ, Bichler M, Abstreiter G, Wegscheider W (2002) Aharonov–Bohm oscillations of a tunable quantum ring. Semiconduct Sci Technol 17:L22–L24
30. Kim CS, Roznova ON, Satanin AM, Shtenberg VB (2002) Interference of quantum states in electronic waveguides with impurities. JETP 94:992–1007
31. Kobayashi K, Aikawa H, Katsumoto S, Iye Y (2002) Tuning of the Fano effect through a quantum dot in an Aharonov–Bohm interferometer. Phys Rev Lett 88:256806–256809
32. Kobayashi K, Aikawa H, Sano A, Katsumoto S, Iye Y (2004) Fano resonance in a quantum wire with a side-coupled quantum dot. Phys Rev B 70:035319–035325
33. Nikonov DE, Imamoglu A, Scully MO (1999) Fano interference of collective excitations in semiconductor quantum wells and lasing without inversion. Phys Rev B 59:12212–12215
34. Nöckel JU, Stone AD (1994) Resonance line shapes in quasi-one-dimensional scattering. Phys Rev B 50:17415–17432
35. Onsager L (1931) Reciprocal relations in irreversible processes. Phys Rev 38:2265–2279
36. Porod W, Shao ZA, Lent CS (1993) Resonance-antiresonance line shape for transmission in quantum waveguides with resonantly coupled cavities. Phys Rev B 48:8495–8498
37. Ryu CM, Cho SY (1998) Phase evolution of the transmission coefficient in an Aharonov–Bohm ring with Fano resonance. Phys Rev B 58:3572–3575
38. Satanin AM, Joe YS (2005) Fano-interference and resonances in open systems. Phys Rev B 71:205417, 1–12
39. Satanin AM, Joe YS (2005) Manipulating of resonances in conductance of an electron waveguide with antidotes. J Comput Electron 4:149–153
40. Satanin AM, Joe YS (2007) Fano-interference of decaying states with continuum in nanostructures. J Superconduct Novel Magn 20:179–182
41. Satanin AM, Hedin E, Joe YS (2006) Collision of Fano resonances: an exact solvable model. Phys Lett A 349:45–52
42. Schuster R, Buks E, Heiblum M, Mahalu D, Umansky V, Shtrikman H (1997) Phase measurement in a quantum dot via a double-slit interference experiment. Nature 385:417–421
43. Shao Z, Porod W, Lent CS (1994) Transmission resonances and zeros in quantum waveguide systems with attached resonators. Phys Rev B 49:7453–7465
44. Song JF, Ochiai Y, Bird JP (2003) Fano resonances in open quantum dots and their application as spin filters. Appl Phys Lett 82:4561–4563
45. Tekman E, Bagwell PF (1993) Fano resonances in quasi-one-dimensional electron waveguides. Phys Rev B 48:2553–2559
46. Yacoby A, Schuster R, Heiblum M (1996) Phase rigidity and $h/2e$ oscillations in a single-ring Aharonov–Bohm experiment. Phys Rev B 53:9583–9586
47. Yacoby A, Heiblum M, Mahalu D, Shtrikman H (1995) Coherence and phase sensitive measurements in a quantum dot. Phys Rev Lett 74:4047–4050

Books and Reviews

Adair RK, Bockelman CK, Peterson RE (1949) Experimental corroboration of the theory of neutron resonance scattering. Phys Rev 76:308

Cerdeira F, Fjeldly TA, Cardona M (1973) Effect of free carriers on zone-center vibrational modes in heavily doped p-type Si. II Optical modest. Phys Rev B 8:4734–4745

Dittes F-M (2000) The decay of quantum systems with a small number of open channels. Phys Rept 339:215–311

Fano U, Rau ARP (1986) Atomic collisions and spectra. Academic, Orlando

Feshbach H (1958) Unified theory of nuclear reactions. Ann Phys (NY) 5:357–390

Quantum Dot Spin Transistors, Self-consistent Simulation of

Dmitriy V. Melnikov, Jean-Pierre Leburton
Department of Electrical and Computer Engineering & Beckman Institute for Advanced Science and Technology, University of Illinois, Urbana, USA

Article Outline

Glossary

Quantum dot (QD) A man-made nanostructure, generally made with semiconductor materials, in which the motion of particles (such as conduction band electrons) is bound in all three spatial directions. As a result of this 3D spatial confinement, quantum dots exhibit a discreet energy spectrum with particle wave functions localized within the quantum dot.

Coulomb blockade The increased resistance experienced at small bias voltages by an electronic device comprising of at least two tunnel junctions (terminals), which is manifested in the electrostatic blockade of the current by the charge accumulation on a small conducting island between the terminals.

Single electron transistor An electronic device characterized by two tunnel junctions ("source" and "drain") between which a conducting "island" (QD) is located. The electrostatic potential of the QD island is controlled by a third electrode (the "base"). By changing the controlling voltage on the base electrode, transport channels for electron current through the QD can be opened or closed. When the channel is open, an electron can tunnel into the island on an available energy level from the source contact and then subsequently tunnels out to the drain electrode (single-electron transport).

Spin blockade The suppression (partial or total) of the current through a QD device populated by electrons and caused by the Pauli exclusion principle.

Definition of the Subject

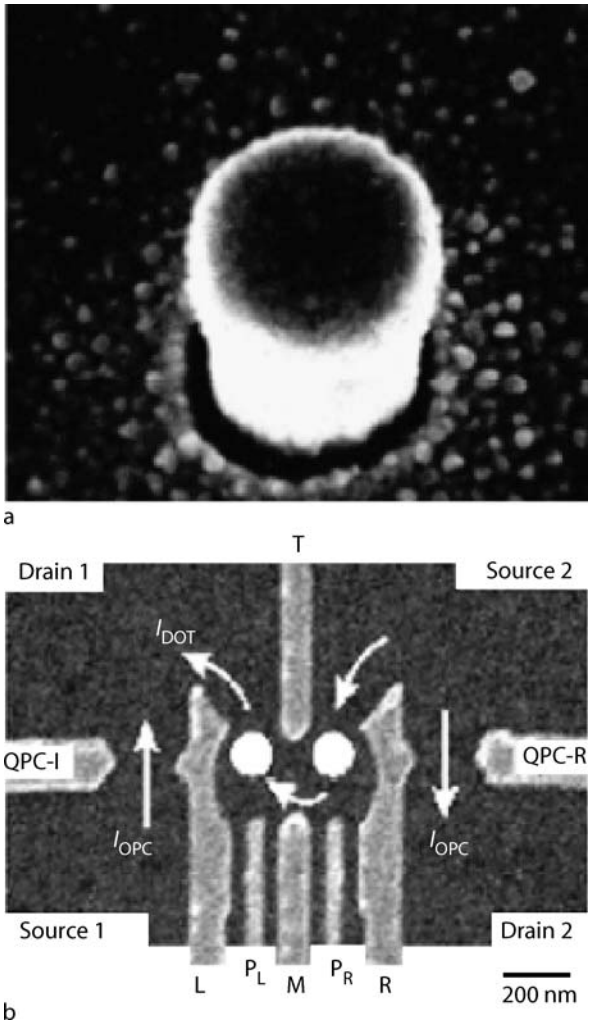
The electronic and transport properties of quantum dot spin transistors are presented with emphasis on single-electron tunneling and shell structure. A comprehensive modeling approach based on two methods is developed: (1) quantum dot electronic structure and confinement potential are determined from the self-consistent solution of the Poisson and Schrödinger equations within the spin-density-functional theory in magnetic fields and (2) a quantum transport model based on numerically exact diagonalization of the many-body Schrödinger equation is used to describe transport properties of quantum dots. In the linear transport regime characterized by a small applied source-drain voltage, single-electron tunneling through the quantum dot reveals the existence of a shell structure in the ground state electron addition spectra which magnetic field dependence is determined by competition among the many-body interaction effects, confinement potential strength and magnetic field induced localization of electrons. In the non-linear transport regime, the information about excitation states of quantum dots can be extracted from the transport spectra. The current in this case is dominated by the asymmetry in the tunneling barriers, spin selection rules and the overlap between various many-body states. The very good agreement between calculated and measured currents allows one to put in correspondence numerous features in transport and energy spectra of quantum dots.

Introduction

In recent years few-electron semiconductor quantum dots (QDs) have attracted considerable attention because these nanostructures exhibit many features characteristic of real atomic systems. QDs are also promising candidates for future device applications [1]. In these man-made systems conduction electrons are held together in a finite region of space by the confinement potential that is usually created by the hetero-structure barriers and/or the electrostatic potential of remote dopant charge distributions modulated by external gate voltages.

Among several types of QDs, we consider the so-called "gated" quantum dots in which controlled single-electron charging and shell structure was first observed [2]. In vertical QDs, a small, quasi-two-dimensional (quasi-2D) electron island (dot) is formed in a mesa-structure pillar (Fig. 1a) between two heterostructure barriers (see Fig. 4). The size of the island (or the electron occupation of the QD) can be changed electrostatically by applying a voltage to the gate wrapped around the vertical structure. The current flows vertically (hence the term "vertical QD") through the heterostructure in response to a bias applied between the source and drain contacts on top and bottom of the mesa-structure. The fact that this device has three terminals is reminiscent of a field-effect transistor where the base terminal corresponds to the side gate in the QD device and controls the single-electron current through the heterostructure.

Another method frequently utilized to fabricate QDs is the lithographic patterning of gates, i. e., the deposition of metal electrodes on a heterostructure surface (Fig. 1b). By biasing the top gate electrodes, the two-dimensional (2D) electron gas formed at the hetero-interface between different materials such as AlGaAs and GaAs is depleted underneath them, thereby creating an island of non-zero electron density that can be further fine-tuned by changing voltages on nearby gates [3]. This device is called lateral or planar QD because the current flows in the plane of the 2D electron gas.



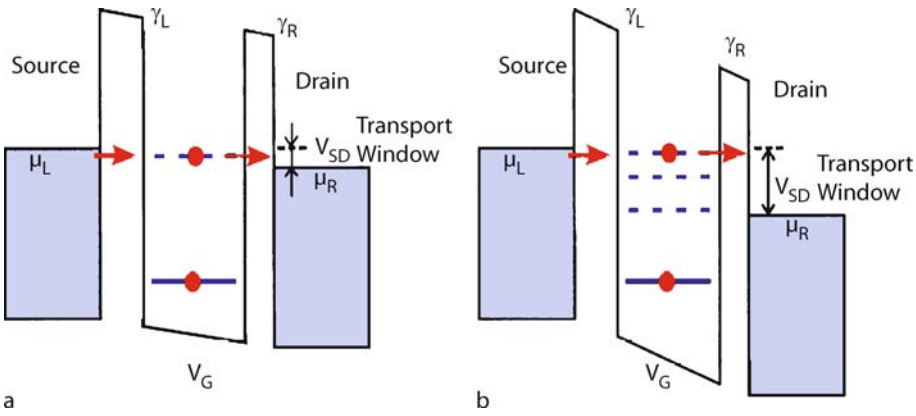
Quantum Dot Spin Transistors, Self-consistent Simulation of, Figure 1

Scanning-electron micrograph of a the vertical QD structure (analogous to the one used in [2]) and b planar system of two laterally coupled QDs [3]

Single-Electron Tunneling

Single-electron effects have been known for a long time. In his famous 1911 experiments, Millikan [4] measured the value of the electron charge observing the falling rate of charged oil drops. Systematically single-electron charging effects in electron tunneling were first studied by transport measurements on thin films of small metallic grains [5,6,7]. In 1975, Kulik and Shekter [8] showed that in a two-terminal system, the current through a small grain at low bias voltages is blocked by the accumulated charge, whereas the differential conductance varies periodically when a larger source-drain bias is applied. This so-called Coulomb blockade and the Coulomb "staircase" were later observed for the first time in granular systems [9] and thin-film tunnel junctions [10].

In principle, QDs are ideal systems for the investigation of single-electron tunneling which can be accomplished by connecting the dot island to surrounding reservoirs [11,12,13]. Figure 2a schematically shows a QD electron island connected to the leads through hetero-barriers as in the vertical QD structure (it can also be electrostatic barriers as in planar QD structures), and a side gate biased



Quantum Dot Spin Transistors, Self-consistent Simulation of, Figure 2

Single-electron transport through the QD ($N = 2$) in a the linear regime ($V_{SD} \to 0$) and b non-linear regime (V_{SD} is large). In both cases the QD contains one electron while one channel (three channels in b) open for the tunneling of the second electron. The left and right barrier permeabilities are γ_L and γ_R respectively

with a controlling voltage V_G that changes positions of the energy levels in the QD with respect to the chemical potentials of the left (L) and right (R) leads μ_L and μ_R. The electro-chemical potential of the dot filled with N electrons (the solid line in Fig. 2), is given by

$$\mu_{QD}(N) = E(N) - E(N-1)\,, \tag{1}$$

where $E(N)$ is the ground-state energy (at zero temperature) of the N-electron QD. As the leads (the source and the drain contacts) are weakly coupled to the QD, the electrochemical potentials μ_L and μ_R become different when a constant source-drain bias V_{SD} is applied between the leads, and a transport window of width $\mu_L - \mu_R = eV_{SD}$ opens up.

In the linear transport regime, the transport window eV_{SD} is much smaller than the average spacing among the quantum states, so that only the ground state of the QD contribute to the conductance. By changing the voltage V_G on the side gate, one can achieve alignment of $\mu_{QD}(N)$ with the transport window so that electrons can subsequently tunnel in and out of the QD; this situation corresponds to a conductance maximum when the number of electrons in the QD cycles between N and $N + 1$. Otherwise the current is blocked; this scenario corresponds to zero conductance (and current) when the number of electrons N in the dot is fixed, and it increases to $N + 1$ each time a conductance maximum is crossed. This mechanism of discrete charging and discharging of the QD leads to Coulomb blockade oscillations in the conductance as a function of the gate voltage. The distance between neighboring Coulomb peaks [14] equals the difference in the electrochemical potentials of a QD containing $N + 1$ and N electrons:

$$\Delta_2(N) = \mu_{QD}(N+1) - \mu_{QD}(N)\,. \tag{2}$$

In the non-linear transport regime, i. e. for a larger transport window eV_{SD} and at a fixed V_G, additional peaks in differential conductance traces occur. They are due to the excitations in the N-electron QD system as electrons can tunnel in and out of the QD via the ground as well as low lying excited states (Fig. 2b). By plotting the positions of these peaks as a function of V_{SD} and V_G, a characteristic diamond-shaped structure is observed, which provides information about the ground and excited states of the QD [15,16,17,18,19]. Note that unlike optical absorption (radiation) processes in real atoms where electron excitations are created by photons, the excitation mechanism in artificial QD atoms is electron-only, e. g., excitations can be created when one electron tunnels out from the ground state, while another one tunnels in the excited state. However, in general, the analysis of the various features observed in the tunneling current through the QD in terms of the excitation energies in the non-linear transport regime is a challenging task that we address in Sect. "Tunneling Spectroscopy of a Few-Electron QD in the Non-Linear Transport Regime".

Shell Structure

In the three-dimensionally (3D) confined systems such as QDs energy quantization and interaction among the confined particles leads to the existence of a "shell structure". It can be illustrated on the following example of a simple

two-dimensional harmonic-oscillator confinement potential:

$$V_{\text{conf}}(x, y) = \tfrac{1}{2} m^* \omega^2 \left(x^2 + y^2\right) , \tag{3}$$

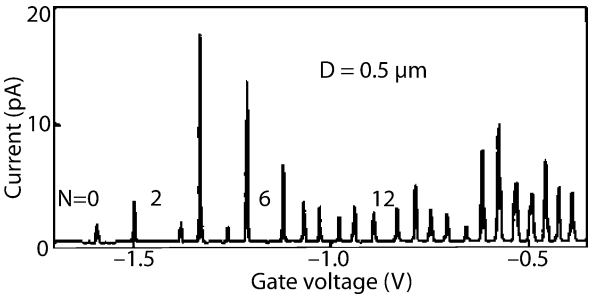
which can be considered as an empirical mean-field potential in which N particles with effective mass m^* move independently and where ω determines the confinement strength. The corresponding single-particle energy levels are obviously

$$E_{n_x,n_y} = \hbar\omega \left(n_x + n_y + 1\right) , \tag{4}$$

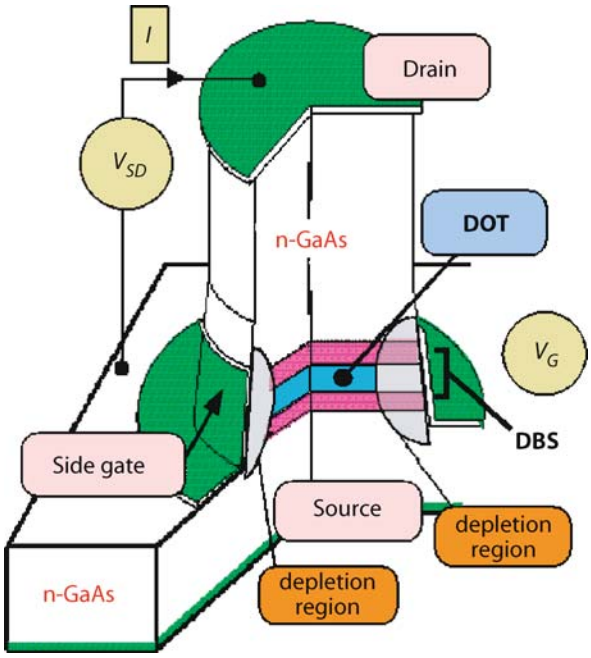
where one easily recognizes the $(N_0 + 1)$-fold state degeneracy with respect to the quantum number $N_0 = n_x + n_y = 0, 1, 2, \ldots$ By filling the states with non-interacting electrons and by accounting for the Pauli principle, one can obtained closed shells for a sequence of $N = 2, 6, 12, 20, \ldots$ particles. For these configurations, a particular stability is achieved as the degeneracy of the shell is resolved. Adding one more electron to a closed shell results in the single occupation of an orbital belonging to the next higher shell which makes this configuration less stable since as a larger amount of energy than for the closed shell configurations needs to be supplied to the system. Despite its simplicity, this example illustrates some of the basic features of few-electron QD systems such as the existence of electron shells and the stability of closed-shell configurations.

The periodic table of elements (Mendeleev table), with its groups of elements characterized by similar chemical properties, is the most well-known example of shell structure in nature. Atomic shells are clearly seen in the pronounced maxima of the ionization energies of neutral atoms for atomic numbers $Z(= N) = 2, 10, 18, \ldots$ corresponding to the noble gases He, Ne, Ar, etc. (the 3D spherical symmetry of the rigid confinement caused by the strong $1/r$ Coulomb potential of the nucleus results in closed shells forming at different values N than in the above 2D example). The shells are populated according to first Hund's rule stating that the spin is maximized for half-filled orbitals due to the Pauli principle and the repulsive Coulomb interaction [20].

In a clear analogy with atoms, nuclei [21] or atomic clusters [22], shell structure was observed in the transport spectra of semiconductor QDs for the first time by Tarucha et al. [2] in vertical structures (see Figs. 1a and 4). The electron number N in the QD was varied one-by-one starting from $N = 0$ by increasing the negative voltage V_G applied to the side gate (at $V_G = 0$ the QD was already populated by a large number of electrons). A current I flowed through the QD sandwiched between two heterostructure barriers in response to a very small bias V_{SD} (linear transport regime) between the source and drain contacts. By measuring the current as a function of the side gate voltage, current peaks corresponding to single-electron tunneling events were observed. The separation between consecutive peaks is proportional to the difference in energy needed to add an electron to the dot already confining N particles (Eq. (2)). The current-voltage characteristics plotted in Fig. 3 has large inter-peak separations at electron numbers $N^* = 2$, 6, and 12 which correspond to



Quantum Dot Spin Transistors, Self-consistent Simulation of, Figure 3
Current oscillations due to the Coulomb blockade in the linear transport regime measured in the vertical QD system [2]



Quantum Dot Spin Transistors, Self-consistent Simulation of, Figure 4
Schematic representation of the vertical QD mesa-structure [2]

closed shells for the 2D harmonic oscillator potential (3). The smaller inter-peak separations at the midshell regions for $N = 4, 9, \ldots$ is a consequence of electron spin alignment due to first Hund's rule (for details, see Sect. "QD Ground State Charging in Magnetic Fields").

Experimental Techniques

In semiconductor QDs, numerous experimental techniques for probing single-electron charging effects on single dots or arrays of dots have been implemented which allow a detailed spectroscopic study of the ground and excited states of individual artificial atoms. Originally, far-infrared (FIR) and optical spectroscopy were applied to arrays [23,24] as well as to individual quantum dots [25] because of their success in atomic and molecular physics. However, the parabolic dependence of the confinement potential and the long wave-length of the radiation suppresses most transitions in the excited states by dipole selection rules [26]. In many cases, it was found that the FIR absorption spectra are indicative of a non-interacting electron system [23,24], in which case Kohn's theorem [27] indicates that the effects of electron-electron interactions in a QD can be observed only if the anharmonicity of the confinement is sufficiently strong [28].

On the other hand, transport-based methods are expected to be largely free of the optical spectroscopy limitations which makes them more attractive for studying QDs. The first QD capacitance measurements were reported by Smith et al. [29] and single-electron capacitance spectroscopy has been used for both arrays [30,31,32] and individual quantum dots [33]. In these experiments, electron tunneling into the QD was observed upon increasing the positive bias on the top plate. The spacing between conductance peaks was approximately constant reflecting conventional Coulomb blockade effect similar to earlier experiments [34] in larger, lateral quantum dots. For smaller biases, however, the distance between consecutive peaks increased, and the spacing between them became nonuniform [33,35]. These deviations from the equidistant Coulomb blockade spectra were attributed to the QD energy spectrum quantization which will be explained in details below (see Sect. "QD Ground State Charging in Magnetic Fields").

Many-Body Calculations of the QD Electronic Structure

In general, electrons in GaAs-based QD mesa-structures originate from remote ionized shallow impurities in the (doped) leads and are confined around the lowest energy minimum of the semiconductor conduction band. The electron density in the leads is low with the mean electron-electron distance ~ 10 nm validating the effective-mass approximation: the conduction electrons have an effective mass m^* and their Coulomb interaction is screened with the static dielectric constant ε of the semiconductor in question.

From a theoretical point of view, interacting electrons confined in a quasi-2D QD form a seemingly simple many-body problem . The electron-electron correlations give rise to numerous intriguing QD properties [36]. Kumar et al. [37] calculated the effective single-particle confinement for a QD created by square-shaped metallic gates using a self-consistent approach, where the electrostatic confinement potential was obtained from a self-consistent solution of the combined Hartree and Poisson equations for the whole 3D QD mesa-structure. In their approach, electrons in the QD region were treated fully quantum-mechanically via the solution of Schrödinger equation (neglecting image charge effects as well as electron correlations) while the charge density in the leads was described semi-classically. They found that in the limit of small particle numbers confined in the QD, the effective confinement has a symmetry close to circular, even if the dot region was defined by a square-shaped gate pattern. On the basis of their work, the simple isotropic harmonic oscillator (see Eq. (3) above) was adopted as the "standard" 2D QD *model* potential for future numerous QD electronic structure calculations. Macucci, Hess, and Iafrate [38] extended the work [37] by incorporating the exchange and correlation contributions in electron-electron interaction within the density-functional theory (DFT) and found a shell-like structure for the electro-chemical potentials $\mu_{\mathrm{QD}}(N)$.

However, it turns out that the addition energy calculations based on a model 2D confinement potential and the experimental data start to deviate when the electron number N increases [39]. In this respect, analysis of a series of experimental addition energy spectra for 14 different structures with diameters between 0.44 and 0.6 μm, which are similar to the QD device used by Tarucha et al. [2], finds [40] strong device-to-device variations: While all structures show the first shell at $N = 2$, only 71% of them show shells at both $N = 2$ and 6, 64% at $N = 2$, 6, and 12, and 21% at $N = 2$, 6, 12, and 20. These observations seem to indicate that each QD has "a mind of its own," and one should be cautious with a quantitative comparison between experiment and theoretical calculations based on a fixed model confinement potential. Possible reasons for the disagreement are either non-parabolicity of the confining potential or the unavoidable inaccuracies in device fabrication that disturb the perfect circular symmetry of the QD.

These deviations, at least in part, can be accounted for by performing a full scale 3D analysis of the QD mesa-structure without any a priori assumptions about the shape or strength of the confinement potential (as it was done in the above [40]). In this multi-scale approach spin-density-function theory (SDFT) is utilized to treat the many body interactions among the electrons in the QD fully quantum mechanically including correlation and spin effects by solving Kohn–Sham (KS) equations. On the other hand, a semi-classical Thomas–Fermi approach is used to describe the influence of the charge distribution outside the QD region. As a result, the QD confinement potential and the quantized energy spectrum are obtained directly from the self-consistent solution of the non-linear Poisson and Kohn–Sham (KS) equations with device boundary conditions.

DFT calculations were also performed [41] for studying the electron structure and statistical properties of the level spacings in dots containing ~ 100 electrons. However, in this approach the 3D KS equations were separated into 2D and 1D equations by taking into account quasi-2D nature of the QD confinement. Self-consistent procedure [42] was also used for solving full 3D Poisson and Schrödinger equations for a few-electron cylindrical QDs. Here the N-electron Schrödinger equation was solved by the unrestricted Hartree–Fock method and the ground state charging energies (chemical potentials) of the QD in magnetic fields were computed [43].

In the following chapters a theoretical analysis of the QD electronic and transport properties is given. The discussion is based on the calculations performed for the vertical QD structure described in Sect. "Vertical QD Device Structure". An approach suitable to the simulation of the QD ground state properties which is based on the self-consistent solution of the Poisson and Schrödinger equations within the SDFT is presented in Sect. "Self-consistent Approach". The ground state addition energy spectra and shell structure in magnetic fields are computed and compared with experimental data in Sect. "QD Ground State Charging in Magnetic Fields". Next a description of a numerically exact diagonalization of the many-particle Schrödinger equation (Sect. "Exact Diagonalization of the Many-Electron Schrödinger Equation") and the quantum transport model (Sect. "Quantum Transport Model") is provided followed by Sect. "Tunneling Spectroscopy of a Few-Electron QD in the Non-Linear Transport Regime," where the relationship between the computed current and energy spectra in the non-linear transport regime (excited state spectroscopy) is established and comparison with measured data is given. Finally, Sect. "Conclusion" contains concluding remarks.

Vertical QD Device Structure

Single-electron tunneling controlled by external gates is best illustrated on the example of a gated vertical QD structure schematically shown in Fig. 4 which is similar to the QD used in the pioneering experiments of Tarucha et al. [2]. The mesa-structure was fabricated from a double-barrier heterostructure (DBS) by etching techniques [44], and the electron "puddle" (QD) was located in the quantum well between two hetero-barriers that separated it from the outside environment. A metal Schottky gate (side gate) was wrapped around the base of the circular pillar with a diameter of 0.5 μm. The energy gap between conductance and valence bands was also reduced by including 5% In in the 12 nm GaAs quantum well sandwiched between the two $Al_{0.22}Ga_{0.78}As$ barriers which were nominally 7.5 and 9.0 nm thick. The presence of In lowered the bottom of the conduction band below that of the n-doped GaAs leads (source and drain contacts) so that the lowest energy level in the QD was below the Fermi level of the contacts, i. e., electrons could accumulate in the dot even if no source-drain bias voltage V_{SD} was applied. This made it possible to study electron transport at very small source-drain bias voltages. The side gate voltage V_G changed the effective diameter of the island, i. e., it controlled the strength of the effective confinement potential, thereby allowing one-by-one change in the QD electron population (single-electron QD charging). As the effective diameter of the QD is much larger than its thickness in the vertical direction perpendicular to the hetero-barriers, the motion of the electrons in the vertical z-direction is "frozen" so that only the ground state in that direction is occupied, and it is the energy quantization in lateral x–y-plane which is affected by the side gate voltage.

Self-consistent Approach

We use the SDFT [45] to describe the ground state properties of electrons confined in the QD, in which the charge density $\rho(\mathbf{r})$ is calculated after solving KS equations for electrons with spin up ($\uparrow$) and down ($\downarrow$):

$$\hat{H}^{\uparrow(\downarrow)}\psi^{\uparrow(\downarrow)}(\mathbf{r}) = \varepsilon^{\uparrow(\downarrow)}\psi^{\uparrow(\downarrow)}(\mathbf{r}), \tag{5}$$

where the single-particle Hamiltonian $\hat{H}^{\uparrow(\downarrow)}$ is given as

$$\hat{H}^{\uparrow(\downarrow)} = \hat{T} + \phi(\mathbf{r}) + \Delta E_c(\mathbf{r}) + v_{xc}^{\uparrow(\downarrow)}(\mathbf{r}) \tag{6}$$

with $\hat{T}$ being the kinetic energy operator which in the presence of an external magnetic field $\mathbf{B}$ reads as:

$$\hat{T} = \frac{1}{2}\left(-i\hbar\nabla + \frac{e}{c}\mathbf{A}\right)\frac{1}{m^*(\mathbf{r})}\left(-i\hbar\nabla + \frac{e}{c}\mathbf{A}\right). \tag{7}$$

Here $m^*(\mathbf{r})$ is the position dependent electron effective mass, and $\mathbf{A} = (B/2)(y, -x)$ is the vector potential in symmetric gage for the the magnetic field oriented along the z-direction perpendicular to the QD plane (Zeeman splitting is neglected for clarity). In Eq. (6), $\Delta E_c(\mathbf{r})$ stands for the conduction band offset between the different materials. For the vertical QD structure described in Sect. "Vertical QD Device Structure", its values are fixed at 180 and −40 meV for $Al_{0.22}Ga_{0.78}As/GaAs$ and $In_{0.05}Ga_{0.95}As/GaAs$ interfaces, respectively (note that for the real structures with non-zero doping in the GaAs contact regions, the value of $\Delta E_c(\mathbf{r})$ for $In_{0.05}Ga_{0.95}As/GaAs$ is difficult to determine precisely). In general, this method is not limited to single vertical QD structures but can be quite straightforwardly applied to modeling any layered semiconductor structures such as those used for the planar (lateral) QDs [46] as well as systems of double [47] and triple vertical QDs [48].

The exchange-correlation potential $v_{xc}^{\uparrow(\downarrow)}(\mathbf{r})$ in Eq. (5) is computed within the local spin density approximation (LSDA) [49] and does not explicitly depend on magnetic field. Comparison of DFT results with calculations using current-spin density functional theory [50] for 2D systems showed that this approximation is reliable over a wide range of magnetic field, although at higher fields, effects of paramagnetic currents in $v_{xc}^{\uparrow(\downarrow)}(\mathbf{r})$ should become more important [51].

The potential $\phi(\mathbf{r}) = \phi_{ext}(\mathbf{r}) + \phi_{ion}(\mathbf{r}) + \phi_H(\mathbf{r})$ in Eq. (6) is the sum of the external potential $\phi_{ext}(\mathbf{r})$ due to the applied voltage, screening potential $\phi_{ion}(\mathbf{r})$ arising from the ionized impurities in the structure, and Hartree potential $\phi_H(\mathbf{r})$ accounting for the repulsive electron-electron interactions. It is obtained from the solution of the Poisson equation:

$$\nabla\epsilon(\mathbf{r})\nabla\phi(\mathbf{r}) = 4\pi\rho(\mathbf{r})\,, \tag{8}$$

where $\epsilon(\mathbf{r})$ is the dielectric constant of the medium, and $\rho(\mathbf{r})$ is the charge density which inside the QD region is equal to

$$\rho(\mathbf{r}) = -e\left(\sum_{\text{occup}}\left|\psi^{\uparrow}(\mathbf{r})\right|^2 + \sum_{\text{occup}}\left|\psi^{\downarrow}(\mathbf{r})\right|^2\right), \tag{9}$$

with the summations spanning occupied states for electrons with spin up and down (the number of those states is, in general, different). Outside the QD region, charge distribution is determined from electron $n(\mathbf{r})$ and hole $p(\mathbf{r})$ densities calculated within the semi-classical Thomas–Fermi approximation [52]:

$$n(\mathbf{r}) = \frac{4}{\sqrt{\pi}}\left(\frac{2\pi m_e(\mathbf{r})T}{h^2}\right)^{3/2} F_{1/2}\left[\frac{-e\phi(\mathbf{r}) + \Delta E_c(\mathbf{r})}{T}\right], \tag{10}$$

$$p(\mathbf{r}) = \frac{4}{\sqrt{\pi}}\left(\frac{2\pi m_h(\mathbf{r})T}{h^2}\right)^{3/2} F_{1/2}\left[\frac{e\phi(\mathbf{r}) - E_G(\mathbf{r}) - \Delta E_c(\mathbf{r})}{T}\right], \tag{11}$$

$$F_{1/2}[\eta] = \int_0^\infty \frac{dx\, x^{1/2}}{1 + \exp(x - \eta)}\,, \tag{12}$$

where $E_G(\mathbf{r})$ and $m_h(\mathbf{r})$ are the band gap and the hole effective mass of the constituent materials, and $T = 0.1$ K is the temperature. These densities are screened by the (completely) ionized donors and acceptors $N_D^+(\mathbf{r})$ and $N_A^-(\mathbf{r})$ distributed in the QD leads:

$$\rho(\mathbf{r}) = -e\left[N_D^+(\mathbf{r}) - N_A^-(\mathbf{r}) + p(\mathbf{r}) - n(\mathbf{r})\right]. \tag{13}$$

Since QD regions are usually much smaller than the physical dimensions of the device, the KS wavefunctions actually vanish long before reaching the device boundaries. This allows us to embed a local region in the global mesh for solving the KS equations. This local region is chosen to be large enough to ensure vanishing wavefunctions on its boundaries. For the Poisson equation (8), zero electric field on the lower part of the structure buried in the substrate and on the top contact plane are used as a boundary condition. On other surfaces, not covered by the side gate, the potential ϕ is set equal to the Schottky barrier value $V_S = 0.9$ eV. Boundary values of the potential on the side gate are equal to the Schottky barrier value modified by the applied gate bias, $V_S - V_G$. The system of KS and Poisson equations (5), (8) is solved iteratively until a self-consistent solution for the KS orbitals $\psi^{\uparrow(\downarrow)}(\mathbf{r})$ and eigenvalues $\varepsilon^{\uparrow(\downarrow)}$ is obtained.

The calculations are performed on a parallel platform by means of the finite element method (FEM) with trilinear polynomials on a variable size grid [53]. The advantages of FEM utilization are the ability to systematically improve the accuracy by expanding the basis set and its inherent variational nature. Due to the last reason, calculated energy differences between two similar configurations are usually more accurate than total energy computed for each system separately since each energy is already an upper bound to the exact value.

In our approach Poisson's equation (8) is solved by means of the damped Newton–Raphson method while the

generalized eigenvalue problem obtained after discretization of the KS equation (5) is tackled by means of a subspace iteration method based on a Rayleigh–Ritz analysis [53]. The small number of required eigenpairs (~ 10) made this approach sufficient. A parallel conjugate-gradient method preconditioned with block Jacobi with an incomplete LU factorization on the blocks is utilized for solving the resulting matrix equations. In the presence of a magnetic field, the matrix obtained from the KS equation is Hermitian and the hermitian-conjugate method with the same preconditioner as above had to be used in solving the eigenvalue problem. Compared to the ordinary conjugate gradient method with a Jacobi preconditioner, this approach gives rise to at least an order of magnitude increase in performance especially when working with Hermitian matrices [54].

After the eigenvalues are determined, the ground state electron charging diagram can be calculated by directly comparing the total energy $E(N)$ of the N-electron system with the system containing $N-1$ electrons, the difference of which gives the required value of the chemical potential $\mu_{\mathrm{QD}}(N)$ (see Eq. (1)). One can also use Slater's rule (transition state technique) in order to calculate $\mu_{\mathrm{QD}}(N)$ [55]:

$$\mu_{\mathrm{QD}}(N) = E(N) - E(N-1) \approx \varepsilon(1/2)\,, \tag{14}$$

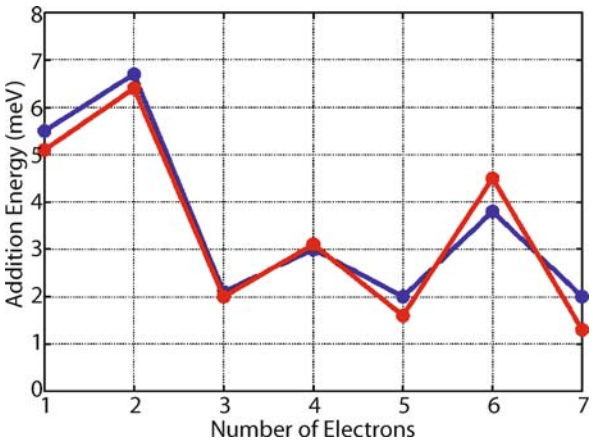
where $\varepsilon(1/2)$ is the eigenvalue of the state with half occupation in the system with $N-1/2$ electrons (the transition state). By varying the side gate voltage V_{G}, both $E(N)$ and $\mu_{\mathrm{QD}}(N)$ can be changed. If $\mu_{\mathrm{QD}}(N) < 0$, then the N-electron configuration is stable, otherwise the number of electrons in the QD is $N-1$. The side gate bias $V_{\mathrm{G}}(N)$ at which $\mu_{\mathrm{QD}}(N) = 0$ gives the charging voltage for the Nth electron (or equivalently the boundary between the QD stable configurations with N and $N-1$ electrons).

An external magnetic field induces crossings among the states with different total spins, i. e., the spin configuration of the system changes while the number of electrons in the QD remains constant. In this regard, one should note that Slater's rule (14) can only be used when there is a change in the occupation of a single eigenlevel. If the change in the electron number is accompanied by spin rotations of individual electrons, i. e., the occupation of several KS levels is simultaneously changed (this happens, e. g., in $N = 2$ system with $S = 0$ at $B \sim 5$ T in Fig. 6 when the addition of a third electron results in $S = 3/2$ state), Slater's rule should be invoked several times. However, such situations involving several transition states are infrequent so that overall, the utilization of Slater's rule significantly reduces the amount of computational time to calculate the full charging diagram of a QD in magnetic fields [56].

From the $\mu_{\mathrm{QD}}(N)$ value, the ground state electron addition energies can be determined by computing the QD capacitive energy [57] given by Eq. (2). This quantity is evaluated for side gate voltage $V_{\mathrm{G}}(N+1)$ corresponding to the addition of the $(N+1)$th electron to the system so that $\Delta_2(N) = -\mu_{\mathrm{QD}}(N)$. In the following chapter, we discuss in detail is the variation of the charging voltage $V_{\mathrm{G}}(N)$ and addition energy as functions of the electron number N in a QD and magnetic fields.

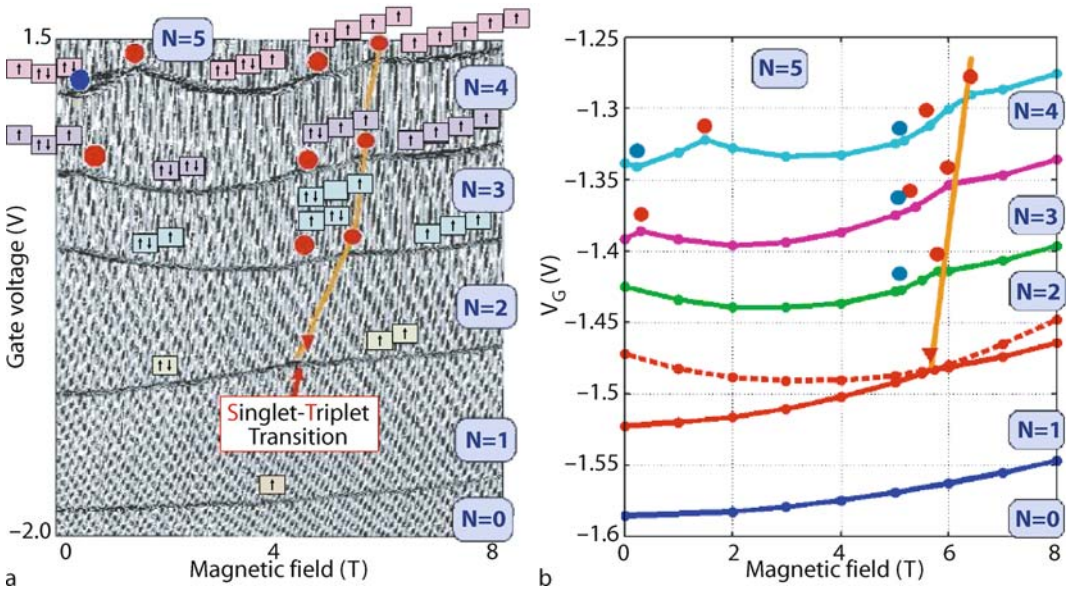
QD Ground State Charging in Magnetic Fields

The computed electron addition energy spectra $\Delta_2(N)$ of the circular QD in the absence of a magnetic field are shown in Fig. 5 together with the corresponding experimental data from [58]. The spectrum exhibits pronounced maxima for two and six electrons due to the first and second shell closures characteristic of QDs with parabolic 2D circular confinement. The peaks and valleys are a consequence of the interplay between confinement and many-body effects. For $N = 2$, the lowest single-particle state ε_1 is fully occupied. The third electron populates the next available eigenstate with $\varepsilon_2 > \varepsilon_1$, thus making the addition of the electron energetically more costly than for the case of the second electron. The same situation is repeated for $N = 6$ when the second shell is closed. A smaller peak at $N = 4$ is due to the fulfillment of Hund's first rule: The total spin of the N-electron system is equal to zero (singlet) for $N = 2$ and 6 and to one (triplet) for $N = 4$. The



Quantum Dot Spin Transistors, Self-consistent Simulation of, Figure 5

Addition energy spectra for the circular QD at zero magnetic field. *Red line* represents the results of calculations while *blue line* stands for the experimental data



Quantum Dot Spin Transistors, Self-consistent Simulation of, Figure 6

a Experimental [15] and b calculated ground state charging diagrams in magnetic fields. Dashed curves show the lowest excited state for $N = 2$ (see also Fig. 8). *Red circles* define boundaries of regions with different total spin and angular momentum in the ground state of the *N*-electron system. *Blue circles* depict features in the the *N*-electron ground state due to reconstruction of the $(N - 1)$-electron system. *Orange curves* show the magnetic field above which the *N*-electron system is fully spin-polarized

agreement between the calculated and experimental spectra is very good, for both peaks and valleys. From the energy separation between the two lowest adjacent eigenvalues, the confinement strength $\hbar\omega \sim 6$ meV in the empty QD can also be deduced.

In Fig. 6 the charging voltage $V_G(N)$ for $N = 1$–5 electrons as a function of magnetic field (charging diagram) is shown. Comparison between experimental (Fig. 6a) and calculated (Fig. 6b) data shows that overall agreement is very good, albeit the confinement potential at zero magnetic field in our model structure is somewhat weaker, i. e., the curves are more closely spaced. The charging voltages generally (in the considered range) increase with magnetic field since the effective confinement becomes stronger. The curves corresponding to the charging of $N \geq 2$ electrons exhibit "cusps" due to various magnetic field induced spin and angular momentum transitions. For $N = 2$, ∇ marks the magnetic field of the singlet-triplet transition (see also Fig. 8), below which the singlet is the ground state while above the triplet is the ground state. For the $N = 3$ charging curve arising from the addition of the third electron to the two-electron system, the cusp near 5 T is due to a change in the ground state configuration of the two-electron system at $B \approx 5.6$ T (the singlet-triplet transition) which affects the addition energy of the third electron. The shift from 5.6 T to 5.0 T is due to a screening effect and is consistent with the experiment (Fig. 6a). The cusp near 6 T reflects an increase in the total spin of the three-electron system from $S = 1/2$ to $S = 3/2$, namely below this point two electrons are spin-up, and one is spin-down while above this point all three electrons in the QD are spin-up and form a spin-polarized system.

In general, the rightmost cusp in charging diagram (Fig. 6) always corresponds to complete spin polarization of the electron system. The magnetic field at which the formation of this state occurs increases with the number of electrons since a stronger field is required to overcome the large kinetic energy accompanying single occupation of consecutive orbitals. Cusps in the N-electron curves in the vicinity of $B = 5$ T are either due to changes in the $(N - 1)$-electron configuration or due to single spin rotations in the N-electron system. The two cusps around $B \approx 0.25$ T in the $N = 4$ and $N = 5$ curves mark the breakdown of Hund's first rule filling in the four-electron system and respective change in the addition energy of the fifth electron due to the decrease of exchange energy in the four electron system, similar to the $N = 3$ curve. The cusp at $B \approx 1.5$ T in the $N = 5$ curve can be understood in terms of the Fock–Darwin spectrum [59]. Around that point, an electron undergoes a transition moving to a state with higher angular momentum while keeping the total spin value constant.

Exact Diagonalization of the Many-Electron Schrödinger Equation

In order to gain insight in the single-electron tunneling characteristics in the non-linear transport regime at large source-drain bias values, it is necessary first to calculate the energy spectrum of a N-electron system. Since the above described self-consistent SDFT-based approach is insufficient for this purpose as the DFT is predominantly the ground state theory [45], we utilize the numerically exact diagonalization of the corresponding many-electron Hamiltonian:

$$\hat{H} = \sum_{i=1}^{N} \left[-\frac{\hbar^2}{2m^*} \left(\nabla_i - \frac{ie}{\hbar c} \mathbf{A}_i \right)^2 + V_{\text{conf}}(\mathbf{r}_i) \pm \frac{1}{2} g \mu_B B \right] + \sum_{i<j} \frac{e^2}{\epsilon |\mathbf{r_i} - \mathbf{r_j}|}, \quad (15)$$

where m^* and ϵ stand for the electron effective mass and dielectric constant in $In_{0.05}Ga_{0.95}As$ respectively. $\mathbf{A} = (1/2)(Bx, -By, 0)$ is the vector potential in the symmetric gage for the magnetic field B oriented along the z-direction. The term $\pm\frac{1}{2} g \mu_B B$ accounts for Zeeman splitting with $g = -0.44$ being the electron g-factor.

In this equation the confinement potential $V_{\text{conf}}(\mathbf{r})$ is assumed for simplicity to be two-dimensional as the width of the vertical QD is much smaller than the lateral extension of the electron "puddle" (see, e. g., Sect. "Vertical QD Device Structure" and [60]). Due to lithographic and nature imperfections [15,61], in the following calculations it is also assumed to be slightly elliptic, i. e.

$$V_{\text{conf}}(\mathbf{r}) = \tfrac{1}{2} m^* \left[\omega_x^2 x^2 + \omega_y^2 y^2 \right] \quad (16)$$

where the confinement energies in the x- and y-directions are taken to be $\hbar\omega_x = 5.3$ and $\hbar\omega_y = 5.65$ meV, respectively, for comparison with experimental structures [15,40,61]. Note that it is also feasible to perform hybrid calculations [60] where the confinement potential is first obtained for an empty QD ($N = 0$) by the self-consistent method of Sect. "Self-consistent Approach" and then the electronic structure of the QD populated with N electrons is computed by numerically diagonalizing the corresponding many-particle Hamiltonian (15) with thus obtained potential. In a few-electron vertical QD structure, this separation works well because the electron system in the QD is sufficiently well isolated from the environment so that interaction between the electrons in the QD and the charges in the outside regions does not affect electron confinement strongly, and it becomes possible to keep the confinement potential frozen and independent on the electron number [60,62,63].

We diagonalize the above Hamiltonian (15) by expanding the N-electron wave function for the αth state in terms of $N \times N$ Slater determinants [64]:

$$\Psi_\alpha(N) = \sum_{i\ldots n} c^\alpha_{i\ldots n} \begin{vmatrix} \varphi_i(\mathbf{r}_1, s_{z1}) & \ldots & \varphi_n(\mathbf{r}_1, s_{z1}) \\ \vdots & & \vdots \\ \varphi_i(\mathbf{r}_N, s_{zN}) & \ldots & \varphi_n(\mathbf{r}_N, s_{zN}) \end{vmatrix}. \quad (17)$$

Here the basis wave function $\varphi_i(\mathbf{r}_j, s_{zj})$ is the product of a 2D anisotropic harmonic oscillator eigenfunction (with frequencies being adjustable parameters) and a spin wave function. Each quantum number $i = (n_x, n_y, s)$ corresponds to the set of the 2D harmonic oscillator (n_x, n_y) and spin s quantum numbers. The summation is carried over all permutations $\{i \ldots n\}$ available for the particular spin state, and the coefficients $c^\alpha_{i\ldots n}$ are determined after the minimization of $\langle \Psi_\alpha | \hat{H} | \Psi_\alpha \rangle$ for a given state which leads to the generalized eigenvalue problem with dense Hermitian matrices. We found that the harmonic oscillator frequencies smaller than the confinement strengths work best due to the fact that the Coulomb interaction tends to flatten out the effective potential [65]. In case of a 2D circular confinement, the Coulomb matrix elements can be evaluated analytically yielding four-fold series. For the anisotropic 2D potential (or in 3D case), the matrix elements are also expressed through the fourfold (sixfold in 3D) series but the auxiliary one-dimensional exponential integral evaluated numerically by means of the Gauss–Kronrod quadrature has to be used to compute their final numerical values [66].

The above method of numerically exact diagonalization of the many-electron Hamiltonian (15) also known as full configuration-interaction (CI) approach yields very accurate values for the eigenenergies and eigenfunctions provided that the single-particle basis set (17) is chosen to be large enough [67]. The downside of this method resides in the size and complexity of the Hamiltonian matrix that grow very quickly with the number of electrons ($\propto M^N$ with M being the number of the single-particle basis states). As a result, it is usually used for QDs with highly symmetric confinement potential containing a small number of electrons ($\lesssim 10$). The most crucial step in this approach is the construction of the many-particle wave function which can be achieved by using either

(1) The single-particle wave functions of the corresponding Hamiltonian or

(2) Analytic wave functions of an elementary confinement potential such as a 2D (3D) harmonic oscillator (as used in this work) or Fock–Darwin states [16,65,68] or
(3) Multi-center expansion of the single-particle wave functions [69].

Computations using method (1) are usually performed fully numerically on a grid [70] and can in principle be applied to various complex confinement potentials [71]. The disadvantage of this approach is the absence of a systematic way to build up an adequate basis set that ensures fast convergence in the computed energy values and the inherent loss of accuracy and dramatic increase in the computational time required for the numerical evaluation of the Coulomb integrals as the basis set gets larger. A hybrid approach in which the single-particle contributions to the Hamiltonian with the real confinement potential (obtained by the method described in Sect. "Self-consistent Approach" for $N = 0$) were computed numerically with the harmonic oscillator basis set while the Coulomb integrals were evaluated analytically has been recently used to simulate double QD systems [72].

Quantum Transport Model

Since the tunneling hetero-barriers in the vertical QD mesa-structure are large (with tunneling times ~ 1 ns [73]), only sequential tunneling of electrons is of interest [11,12,13]. In case of small tunneling barriers such as in lateral (planar) QD systems [74], the probability of electron co-tunneling [75] through the QD increases and should also be taken into account in the current calculations [76,77]. This indicates that in order to describe transport properties of the vertical QD, once the eigenspectrum of the N-electron system is obtained, we need to compute only the sequential current I [13]:

$$I = -e \sum_{\alpha\beta} \Gamma_{\alpha\beta} \left[P_\alpha(N) + P_\beta(N-1) \right] \left[f_{\mathrm{L}}(\mu_{\alpha\beta} - \mu_{\mathrm{L}}) - f_{\mathrm{R}}(\mu_{\alpha\beta} - \mu_{\mathrm{R}}) \right] . \quad (18)$$

In this equation $\mu_{\alpha\beta} = E_\alpha(N) - E_\beta(N-1)$ is the energy difference between the αth N-electron and βth $(N-1)$-electron energies (the electro-chemical potential, cf. Eq. (1)) while the Fermi distributions $f_{\mathrm{L(R)}}$ determine the energy level occupation for electrons tunneling in the QD from the left L (right R) lead (or from the source and drain contact, respectively) with chemical potential μ_{L} (μ_{R}). $\Gamma_{\alpha\beta}$ is the effective tunneling rate, which in the random phase approximation (neglecting phase correlations between electrons in the leads and the QD) can be written as a product of the electron "bare" single-particle tunneling rate $\gamma_{\mathrm{L(R)}}$ due to the electron permeating the barriers separating the QD from the leads (see Fig. 2) and the overlap matrix element [78,79]:

$$\sum_i \left| \left\langle \Psi_\alpha(N) \left| a_i^\dagger \right| \Psi_\beta(N-1) \right\rangle 25 \right|^2 , \quad (19)$$

where the operator $a_i^\dagger$ is responsible for the creation of an electron in the QD in the ith single-particle state. The overlap matrix element is equal to unity in the case of the non-interacting particles but, in general, correlations among electrons (due to the Coulomb interaction and/or spin effects) can reduce its value down to zero.

The state occupation factors $P_\alpha(N)$ in Eq. (18) can be determined from the steady state solution of the coupled master (or rate) equations [11,12,80,81]:

$$\frac{\mathrm{d}P_\alpha(N)}{\mathrm{d}t} = -\sum_\beta \left[R_{(\alpha,N)\to(\beta,N\pm1)} P_\alpha(N) - R_{(\beta,N\pm1)\to(\alpha,N)} P_\beta(N \pm 1) \right] , \quad (20)$$

where the transition rates are ($\kappa = L, R$)

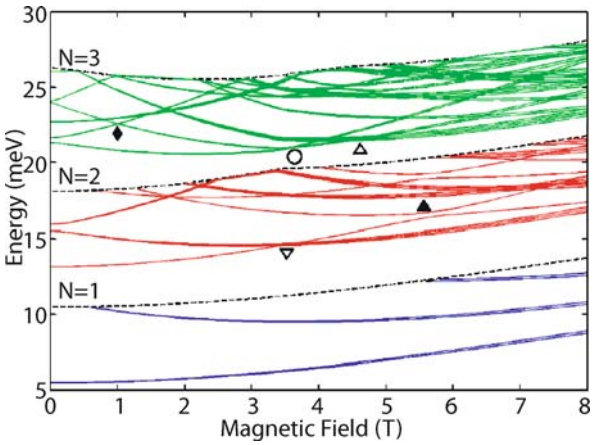
$$R_{(\alpha,N)\to(\beta,N\pm1)} = \sum_{\kappa\in(L,R)} \Gamma_{\alpha\beta} f(\mu_{\alpha\beta} - \mu_\kappa) , \quad (21)$$

$$R_{(\beta,N\pm1)\to(\alpha,N)} = \sum_{\kappa\in(L,R)} \Gamma_{\alpha\beta} \left[1 - f(\mu_{\alpha\beta} - \mu_\kappa) \right] . \quad (22)$$

In the following chapter, calculations of the transport spectra for the QD with the number of electrons $N \leq 3$ and for $V_{\mathrm{SD}} = |\Delta\mu| = |\mu_{\mathrm{L}} - \mu_{\mathrm{R}}| = 5$ meV are presented. In accordance with parameters of the experimental QD structure (Sect. "Vertical QD Device Structure"), the emitter and collector barriers are assumed to be highly asymmetrical with a ratio of permeabealities $\gamma_{\mathrm{L}}/\gamma_{\mathrm{R}} \approx 30$ [11]. A large number of eigenstates (up to 24) is included in solution of the master equation (20) to describe properly degeneracy effects in the addition energies arising when $\mu_{\alpha\beta}$ become equal for specific combinations of states α and β.

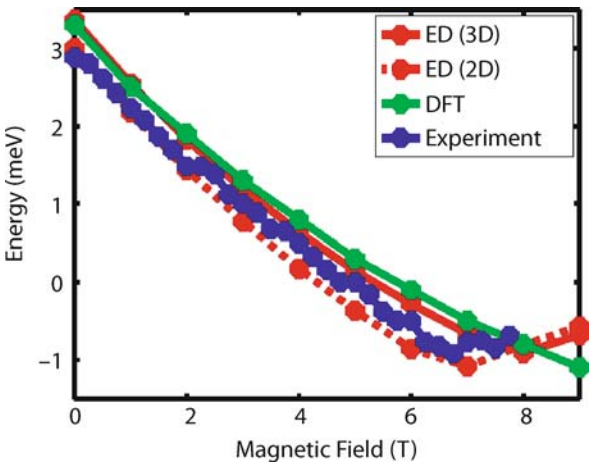
Tunneling Spectroscopy of a Few-Electron QD in the Non-Linear Transport Regime

In Fig. 7 we show the energy diagram for $N = 1, 2, 3$ electrons. Only the levels with energy within 5 meV of the ground state are plotted. The diagram becomes more complex with increasing N as the interplay between confinement, electron-electron interaction, and magnetic field smears out the electronic spectra. Of special interest is the behavior of the lowest singlet $S = 0$ and triplet $S = 1$ energies for $N = 2$ as their energy difference, the exchange



Quantum Dot Spin Transistors, Self-consistent Simulation of, Figure 7

Energy spectrum $E(N) - E_{GS}(N-1)$ for $N = 1, 2, 3$ electrons in magnetic fields. Here $E_{GS}(N)$ is the ground state energy of the N-electron system, $E_{GS}(N=0) = 0$. *Dashed lines* are the upper boundaries of the transport windows $E_{GS}(N) + |V_{SD}|$, $|V_{SD}| = 5$ meV. Zeeman splitting of the energy levels is visible at higher magnetic fields as multiple closely separated lines. ▲ shows the magnetic field at which the exchange energy magnitude starts to decrease. All other symbols are discussed in the text



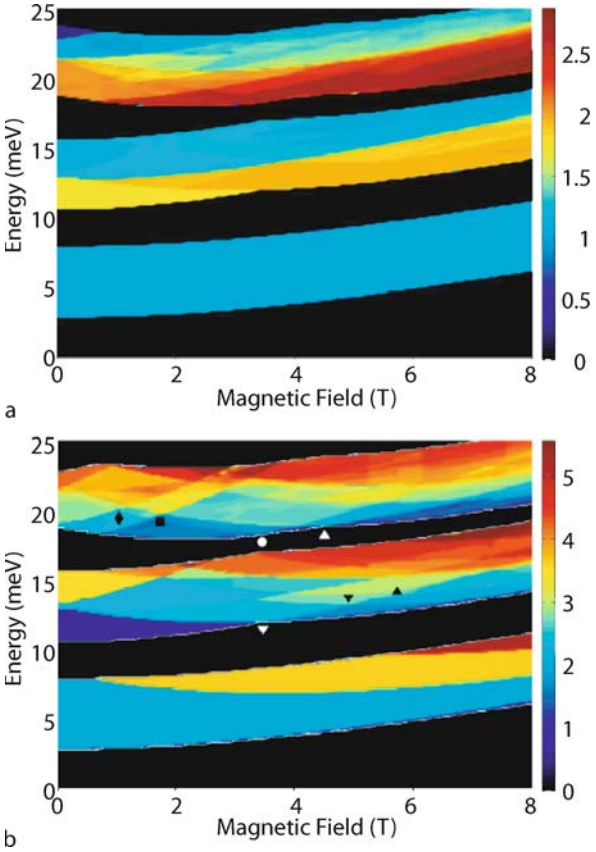
Quantum Dot Spin Transistors, Self-consistent Simulation of, Figure 8

Exchange energy J in magnetic fields. *Dashed red curve* is obtained from the ED calculations including 3D effects, *solid red curve* is the results of the 2D ED while *green curve* is for the results of the DFT calculations (Sect. "Self-consistent Approach"). *Blue curve* is the experimental data [84]

energy J, is of interest for quantum computation schemes involving double QDs [71,72,82]. The exchange energy J as a function of the applied magnetic field computed by both the numerically exact digonalization of the two-electron Hamiltonian (Sect. "Exact Diagonalization of the Many-Electron Schrödinger Equation") and the self-consistent approach based on the spin-density-functional theory (Sect. "Self-consistent Approach") is shown in Fig. 8. One can see that all three computed J-curves agree very well with the experimental data. However, the ED calculations that account for the QD finite width in the vertical dimension (3D ED) have a slight lead over the pure 2D ED results and the self-consistent SDFT-based calculations. The latter method fares progressively worse with increasing magnetic fields, probably because of the lack of the proper accounting of the magnetic field effects in the exchange-correlation potential. In particular, it fails to reproduce a "kink" in the exchange energy at ~ 7 T visible in both experimental data and ED calculations and which finds its origin in the crossing between the two singlet states with different angular momenta [83].

When magnetic fields get larger, density of the electronic states also increases due to the level compression in the lowest Landau band [39]. In the three-electron system, the transition between the two lowest doublets ($S = 1/2$) in the three-electron system (○) around 3.5 T is apparent in the energy spectrum (Fig. 7) while above ~ 4.5 T the system becomes fully spin-polarized (△) and forms a spin quartet ($S = 3/2$). Note also the transition between two doublet levels with different values of the angular momentum (♦) at about 1 T; its relationship with the current traces shown in Fig. 9 will be discussed below. In general, all states shown in Fig. 7 may be involved in the transport process; however, a host of physical reasons make most of them very difficult to distinguish in the measured transport spectra ("dark transport states").

The computed tunneling current in magnetic fields is shown in Fig. 9 for forward ($V_{SD} > 0$, Fig. 9a) and reverse ($V_{SD} < 0$, Fig. 9b) source-drain biases V_{SD}. In the former case, electrons injected in the QD through the thin (forward bias) rather than the thick (reverse bias) emitter barrier. The regions with finite current (the so-called current stripes) of width $|V_{SD}| = 5$ meV, in which the number of electrons in the QD cycles between N and $N-1$ (see Sect. "Introduction" for the general description of the sequential transport through the QD), are separated by black regions with zero current. Inside each stripe, the current variations are reflected by different colors depending on the current increase (color shifts towards the red) or decrease (color shift towards the blue). One can immediately see the striking differences between the forward and reverse bias spectra: the stripes in the latter one have a much more complex structure with a larger number of visible current features. Indeed, since in the single-elec-



Quantum Dot Spin Transistors, Self-consistent Simulation of, Figure 9

Tunneling current (in arb. units) as a function of the gate voltage potential eV_G (vertical axis) and magnetic field B (horizontal axis) for a $V_{SD} > 0$, and b for $V_{SD} < 0$. ▲ and ▼ mark the magnetic fields at which the lowest region with the constant current and the magnitude of the exchange energy, respectively, start to decrease. The symbols ♦ and are discussed in the text while all other symbols are the same as in Fig. 7

tron tunneling process, the QD electronic configuration fluctuates between N- and $(N-1)$-electron systems, the carrier transmission depends on the detailed occupation of the many-body energy states [11,85]. Hence, in a structure with asymmetric barriers, electrons injected in the QD through the thick rather than the thin emitter barrier give rise to different electron configurations, as in the former case the dot will be in the $(N-1)$-electron state most of the time (because of the thin collector barrier allowing easy escape of electrons from the QD), while in the latter case it will be predominantly occupied by N electrons [11].

This behavior is clearly visible in the first stripe for $V_{SD} > 0$ where the current remains unchanged even though there are several single-particle levels present within the bias window (cf. Fig. 7). However, these levels becomes clearly discernable for $V_{SD} < 0$. In the limiting case of the barrier permeabealities $\gamma_L \gg \gamma_R$ with $f_{L(R)} = 0(1)$, one can find approximate solutions of the master equation (20) as $P_\alpha(1) \approx 0$, $P(0) \approx 1$ for $V_{SD} < 0$ and as $P_\alpha(1) \approx 1/M_{N=1}$, $P(0) \approx 0$ for $V_{SD} > 0$ ($M_{N=1}$ is the number of "active" states in the transport window at the given V_G and B, $\alpha = 1, \ldots, M_{N=1}$), i. e., all one-electron levels are occupied with the same probability. Substituting these occupation factors in the equation for the current (18), one can indeed see that in the former case ($V_{SD} < 0$) the current exhibits step-like increase (broadened by the temperature) when a new energy level dips below the Fermi energy as each new level represents an additional transport channel. On the other hand, when $V_{SD} > 0$, the current remains independent on the number of the levels in the transport window since the sum $\sum_{\alpha=1}^{M_{N=1}} P_\alpha$ (see Eq. (18)) remains largely unchanged.

Similar interpretation can be given for other N, although in these cases the various values of the overlap matrix elements (19) complicate the matter. In the second stripe for $V_{SD} > 0$ (Fig. 9a), the dominant feature is the sharp decrease of the current around the middle of the bias window (the sharp transition from yellow to blue). The large drop in the current is due to the non-zero occupation of the first excited $N = 1$ state which induces a redistribution of the electrons among available $N = 2$ energy levels, i. e., the occupation factors $P_\alpha(N)$, $P_\beta(N-1)$ change, so that the tunneling process associated with this state leaves a visible footprint in the transport spectrum. Physically this corresponds to the situation in which an electron tunnels out of the $N = 2$ ground state within the stripe and leaves the system with the other electron in the excited state [86]. Such process has no corresponding counterpart in the N-electron energy spectrum computed with respect to the ground $(N-1)$-electron state (Fig. 7). In the background of this large current drop, smaller changes in the current due to other processes are also visible (Fig. 9a). In particular, spin effects manifest themselves as a total spin blockade [87] of the current through specific levels when $S(N) - S(N-1) > 1/2$ or as a partial spin blocking of the transport channels for $|S_z(N) - S_z(N-1)| > 1/2$ which lead to current suppression when the corresponding state become available, i. e., the current decreases when the number of channels in transport window increases. These effects become especially clear for smaller values of the source drain bias ($V_{SD} \sim 1$ mV) when the contributions from individual transport channels become discernible in both experimental and calculated current stripes [72].

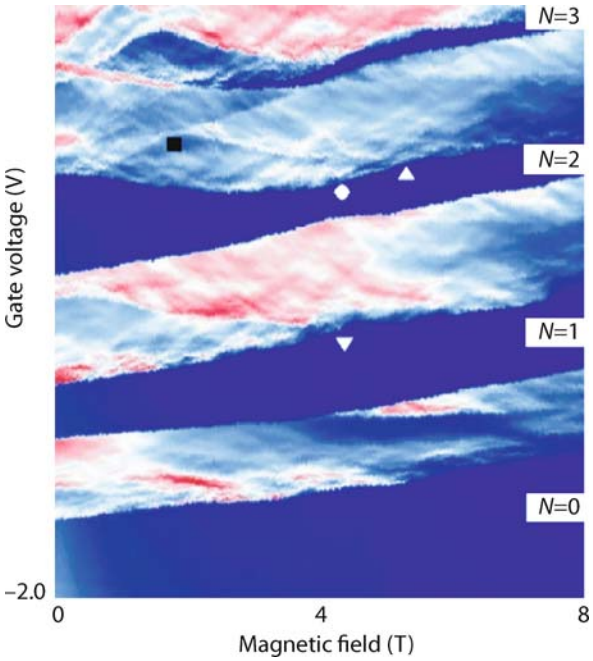
For reverse source-drain bias ($V_{SD} < 0$, Fig. 9b), the current in the second stripe tends to increase with V_G, though this increase is neither gradual nor monotonic. In

this case, the current variation due to the excited $N = 1$ state is barely noticeable, and in its absence transport processes involving other $N = 2$ energy levels become more apparent. Note that this drastic difference between the current in the two stripes (for $V_{SD} > 0$ and $V_{SD} < 0$) occurs mostly because of the large difference in the emitter and collector barrier permeabealities $\gamma_{L(R)}$. When $\gamma_L \sim \gamma_R$ the current spectrum becomes even more complex due to the mixture of these two distinct tunneling processes.

In principle, important information about the QD electronic structure – such as the behavior of the exchange energy in magnetic fields – can be extracted from the transport spectra, but caution should be exerted. For instance, if one naively assumes that the current in the Nth stripe may only change when a N-electron level enters into the transport window (Fig. 7), then the exchange energy could be simply determined by tracing the boundaries of the lowest region in the $N = 2$ stripe characterized by a constant current. This is because the exchange energy is given by the difference between the two lowest $N = 2$ electron states with different spins which are bound to be occupied first upon increasing gate voltage. By inspecting the current spectrum of Fig. 9b, one can easily see that this region starts to shrink above 5 T (▼), which, according to the above argument, should indicate the decrease in the exchange energy (in magnitude) above this magnetic field. In reality, however, this energy difference continues to increase up to 7 T (Fig. 8) and only then decreases (▲): It is due to contributions from excited $N = 1$ states at ~ 5 T for $V_{SD} = 5$.

Another example of inconsistency between the current and energy spectra can be observed in the third stripe, where a prominent kink at about 2 T (marked by ■ in Fig. 9b) has no counterpart in the energy spectrum shown in Fig. 7. Analysis of the electrochemical potentials shows that the crossing between two $\mu_{\alpha\beta}$ curves originating from the lowest $N = 2$ triplet/the $N = 3$ quartet and the lowest $N = 2$ triplet/$N = 3$ doublet at $B \sim 2$ T is responsible for this kink. This feature was also observed experimentally (Fig. 10) but was interpreted simply in terms of the crossing between the two $N = 3$ doublet levels [15]; however, the corresponding kink in the energy diagram at ~ 1 T (marked as ♦ in Figs. 7 and 9) is barely noticeable in the current spectrum.

The above examples indicate that simplistic interpretation of the transport spectra (assignment of features in the measured current) in terms of the many-particle energies can be inadequate and, in general, for the accurate analysis of the QD transport processes in the non-linear regime at finite V_{SD} [84] the quantum-mechanical overlap between various many-body states [13,78], their non-equilibrium occupancies [11,12,13] and the inherent asymmetry of the double barrier hetero-structure [11] should all be considered simultaneously.



Quantum Dot Spin Transistors, Self-consistent Simulation of, Figure 10

Experimental current stripes [15]. All symbols are the same as in Figs. 7 and 9b. Overall, these early measurements do not show any noticeable features above ∼ 5 T and suffer from the extreme variations in the current due to the random dopant distribution in the leads [15]

Conclusion

While QD artificial atoms offer high potential for quantum electronics and continue to be a fast-growing area of research, they provide exciting opportunities for investigating the basic physical properties of interacting many-body system in vertical structures. Single-electron transport spectroscopy within the Coulomb blockade regime bears signatures of individual tunneling events through the QD which provide invaluable tools to investigate the interaction between incident electrons and resident particles. In the few-electron regime, a comparison between experimental electron addition energies and theoretical calculations confirms the existence of electronic shells occupied according to Hund's rules, as in real atoms. Closed shells are particularly stable as the electron addition energy to access the next shell is large, implying the existence of a "noble-gas" structure for specific numbers of electrons.

In contrast to real atoms, the confinement potential in vertical QDs is relatively weak and long-range so that the electronic structure of QDs can be easily probed by relatively small magnetic fields that induce transitions between the various electronic states as the number of particles and the strength of field are varied. In particular, transitions from a low-spin state to a completely spin-polarized system have been observed experimentally in single QDs and computed theoretically within a self-consistent approach based on SDFT.

Non-linear transport spectroscopy in QDs has been used to probe the excited states of a few-electron system. In this case excitations in the QD electronic structure are electron-only (Auger-like), unlike in real atoms where the excitations are the result of electron-photon interaction. The theoretical description of non-linear transport processes in QDs requires the simultaneous knowledge of the energy spectrum and the overlaps between various many-body states. Here, the significant increase in computational effort pays off by a good agreement between experiments and theory as well as by new insights in the electronic and transport properties of QDs.

Future Directions

Recently a novel type of artificial atoms based on electron confinement in semiconductor nanowires (quantum wires) has emerged. The principal difference from the quasi-2D QDs is the 3D nature of the electron confinement potential that bounds the particles wave functions in all three spatial dimensions. Similarly to 2D QDs, electric gates are used to confine the electron motion along the nanowire direction leading to the formation of a quantum dot in the quantum wire (QDQW). Accurate control over the electron population and the Coulomb diamonds characterizing single electron charging have been recently demonstrated in these systems [88]. From the simulation point of view, modeling QDQWs is more challenging as the geometry do not easily permit omission of the third dimension and, consequently, a simplified 2D model is inadequate. In this respect, the 3D self-consistent approach discussed in this work is more suitable as it does not rely on 2D approximation.

Within traditional 2D QD devices, more complicated systems of multiple dots, such as triple QDs, are now being investigated. They represent natural steps in creating QD networks with potential applications in single-electron logic circuits. In these devices a very large parameter space (sizes of individual QDs, number of electric gates and their positions, applied biases, etc) makes both simulation and design of the triple dot system very difficult, thereby leading to a wealth of possible electronic structure configurations such as spin density-waves and double charging of electrons [48], usually not present in single QD devices. Our preliminary results [48] also showed that in order to ensure a sizable coupling among the constituent QDs, a very careful optimization of the mesa structure physical dimensions should be performed; this can be easily achieved with the methods discussed in this article.

Acknowledgments

We are grateful to R. M. Martin and D. G. Austing for helpful discussions. This work was supported by DARPA-QuIST program, MCC through the NSF, and NCSA.

Bibliography

1. Likharev KK (1999) Proc IEEE 87:606
2. Tarucha S, Austing DG, Honda T, van der Hage RJ, Kouwenhoven LP (1996) Phys Rev Lett 77:3613
3. Elzerman JM, Hanson R, Greidanus JS, van Beveren LHW, De Franceschi S, Vandersypen LMK, Tarucha S, Kouwenhoven LP (2003) Phys Rev B 67:161308(R)
4. Millikan RA (1911) Phys Rev 32:349
5. Gorter CJ (1951) Physica (Amsterdam) 17:777
6. Giaever I, Zeller HR (1968) Phys Rev Lett 20:1504; Zeller HR, Giaever I (1969) Phys Rev 181:789
7. Lambe J, Jaklevic RC (1969) Phys Rev Lett 22:1371
8. Kulik IO, Shekhter RI (1975) Zh Eksp Teor Fiz 68:623; Sov Phys JETP 41:308
9. Kuzmin LS, Likharev KK (1987) Pis'ma Zh Eksp Teor Fiz 45:250; Lett JETP 45:495
10. Fulton TA, Dolan GJ (1987) Phys Rev Lett 59:109
11. Averin DV, Korotkov AN, Likharev KK (1991) Phys Rev B 44:6199; Su B, Goldman VJ, Cunnigham JE (1992) ibid 46:7644
12. Beenakker CW (1991) Phys Rev B 44:1646
13. Kinaret JM, Meir Y, Wingreen NS, Lee PA, Wen X-G (1992) Phys Rev B 46:4681
14. Kastner MA (1993) Phys Today 46(1):24
15. Kouwenhoven LP, Oosterkamp TH, Danoesastro MWS, Eto M, Austing DG, Honda T, Tarucha S (1997) Science 278:1788
16. Kyriakidis J, Ladriere PM, Ciorga M, Sachrajda AS, Hawrylak P (2002) Phys Rev B 66:035320
17. Hanson R, Vandersypen LMK, Willems LH, van Beveren Elzerman JM, Vink IT, Kouwenhoven LP (2004) Phys Rev B 70:241304(R)
18. Zumbühl DM, Marcus CM, Hanson MP, Gossard AC (2004) Phys Rev Lett 93:256801
19. Ellenberger C, Ihn T, Yannouleas C, Landmann U, Ensslin K, Driscoll D, Gossard AC (2006) Phys Rev Lett 96:126806
20. Landau LD, Lifshitz E (1977) Quantum mechanics: non-relativistic theory. Pergamon Press, Oxford
21. Goeppert-Mayer M (1949) Phys Rev 75:1969
22. Ekardt W (1999) Metal clusters. Wiley, New York
23. Sikorski C, Merkt U (1989) Phys Rev Lett 62:2164
24. Meurer B, Heitmann D, Ploog K (1992) Phys Rev Lett 68:1371
25. Brunner K, Bockelmann U, Abstreiter G, Walther M, Böhm G, Gänkle T, Weimann G (1992) Phys Rev Lett 69:3216
26. Maksym PA, Chakraborty T (1990) Phys Rev Lett 65:108

27. Kohn W (1959) Phys Rev 115:1160; (1961) ibid 123:1242
28. Gudmundsson V, Gerhardts RR (1991) Phys Rev B 43:12098
29. Smith TP III, Lee KY, Knoedler CM, Hong JM, Kern DP (1988) Phys Rev B 38:2172
30. Hansen W, Smith TP III, Lee KY, Brum JA, Knoedler CM, Kern DP (1989) Phys Rev Lett 62:2168
31. Silsbee RH, Ashoori RC (1990) Phys Rev Lett 64:1991
32. Ashoori RC, Silsbee RH, Pfeiffer LN, West KW (1992) Nanostructures and mesoscopic systems. In: Reed M, Kirk W (eds) Proceedings of the int. symp., Santa Fe, May, 1991. Academic, San Diego
33. Ashoori RC, Störmer HL, Weiner JS, Pfeiffer LN, Pearton SJ, Baldwin KW, West KW (1992) Phys Rev Lett 68:3088
34. Meirav U, Kastner MA, Wind SJ (1990) Phys Rev Lett 65:771
35. Ashoori RC, Störmer HL, Weiner JS, Pfeiffer LN, Baldwin KW, West KW (1993) Phys Rev Lett 71:613
36. Bryant GW (1987) Phys Rev Lett 59:1140
37. Kumar A, Laux SE, Stern F (1990) Phys Rev B 42:5166
38. Macucci M, Hess K, Iafrate GJ (1993) Phys Rev B 48:17354; Macucci M, Hess K, Iafrate GJ (1995) Appl Phys J 77:3267
39. Reimann SM, Koskinen M, Kolehmainen J, Manninen M, Austing DG, Tarucha S (1999) Eur Phys J D 9:105
40. Matagne P, Leburton J-P, Austing DG, Tarucha S (2001) Phys Rev B 65:085325
41. Stopa M (1996) Phys Rev B 54:13767
42. Bednarek S, Szafran B, Adamowski J (2001) Phys Rev B 64:195303
43. Szafran B, Bednarek S, Adamowski J (2001) Phys Rev B 65:035316
44. Austing DG, Honda T, Tarucha S (1996) Semicond Sci Technol 11:388
45. Jones RO, Gunnarsson O (1989) Rev Mod Phys 61:689
46. Zhang L-X, Matagne P, Leburton J-P, Hanson R, Kouwenhoven LP (2004) Phys Rev B 69:245301
47. Ravishankar R, Matagne P, Leburton J-P, Martin RM, Tarucha S (2004) Phys Rev B 69:035326
48. Kim J, Melnikov DV, Leburton J-P, Austing GD, Tarucha S (2006) Phys Rev B 74:035307
49. Perdew JP, Wang Y (1992) Phys Rev B 45:13244
50. Ancilotto F, Austing DG, Barranco M, Mayol R, Muraki K, Pi M, Sasaki S, Tarucha S (2003) Phys Rev B 67:205311
51. Vignale G, Rasolt M (1987) Phys Rev Lett 59:2360
52. Sze SM (1981) Physics of semiconductor devices, 2nd edn. Wiley Interscience, New York
53. Bathe K-J (1982) Finite element procedures in engineering analysis. Prentice Hall, New Jersey
54. Fettig J, Melnikov DV, Sobh N, Leburton J-P (unpublished)
55. Slater JC (1972) Adv Quantum Chem 6:1
56. Fonseca LRC, Jimenez JL, Leburton J-P, Martin RM (1998) Phys Rev B57:4017
57. Lee I-H, Ahn K-H, Kim Y-H, Martin RM, Leburton J-P (1999) Phys Rev B 60:13720
58. Austing DG, Sasaki S, Tarucha S, Reimann SM, Koskinen M, Manninen M (1999) Phys Rev B 60:11514
59. Reimann S, Manninen MM (2002) Rev Mod Phys 74:1283
60. Melnikov DV, Leburton J-P (2006) Phys Rev B 73:155301
61. Kouwenhoven LP, Austing DG, Tarucha S (2001) Rep Progr Phys 64:701
62. Bruce NA, Maksym PA (2000) Phys Rev B 61:4718
63. Szafran B, Bednarek S, Adamowski J (2003) Phys Rev B 67:115323
64. Mikhailov SA (2002) Phys Rev B 65:115312
65. Reimann SM, Koskinen M, Manninen M (2000) Phys Rev B 62:8108
66. Drouvelis PS, Schmelcher P, Diakonos FK (2004) Phys Rev B 69:155312
67. Chakraborti T (1999) Quantum dots: Survey A of the properties of artificial atoms. North Holland, Amsterdam
68. Saarikoski H, Harju A (2005) Phys Rev Lett 94:246803
69. Szafran B, Peeters FM, Bednarek S, Adamowski J (2004) Phys Rev B 69:125344
70. Bellucci D, Rontani M, Troiani F, Goldoni G, Molinari E (2004) Phys Rev B 69:201308(R)
71. Stopa M, Marcus CM (2008) Nano Lett 8:1778
72. Melnikov DV, Leburton J-P, Taha A, Sobh N (2006) Phys Rev B 74:041309(R)
73. Fujisawa T, Austing DG, Tokura Y, Hirayama Y, Tarucha S (2002) Nature 419:278
74. De Franceschi S, Sasaki S, Elzerman JM, van der Wiel WG, Tarucha S, Kouwenhoven LP (2001) Phys Rev Lett 86:878
75. Averin DV, Odintsov AA (1989) Phys Lett A 140:251
76. Sukhorukov EV, Burkard G, Loss D (2001) Phys Rev B 63:125315
77. Thielmann A, Hettler MH, König J, Schön G (2005) Phys Rev Lett 95:146805
78. Pfannkuche D, Ulloa SE (1995) Phys Rev Lett 74:1194
79. Palacios JJ, Martín-Moreno L, Chiappe G, Louis E, Tejedor C (1994) Phys Rev B50:R5760
80. Bonet E, Deshmukh MM, Ralph DC (2002) Phys Rev B 65:045317
81. Muralidharan B, Ghosh AW, Datta S (2006) Phys Rev B 73:155410
82. Petta JR, Johnson AC, Taylor JM, Laird EA, Yacoby A, Lukin MD, Marcus CM, Hanson MP, Gossard AC (2005) Science 309:2180
83. Wagner M, Merkt U, Chaplik AV (1992) Phys Rev B 45:R1951
84. Melnikov DV, Fujisawa T, Austing DG, Tarucha S, Leburton J-P (2008) Phys Rev B 77:165340
85. Grabert H, Devoret MH (eds) (1992) Single electron tunneling, Series ASINATO, Series B, vol 294. Plenum Press, New York
86. Agam O, Wingreen NS, Altshuller BA, Ralph DC, Tinkham M (1997) Phys Rev Lett 78:1956
87. Weinmann D, Häusler W, Kramer B (1995) Phys Rev Lett 74:984
88. Björk MT, Thelander C, Hansen AE, Jensen LE, Larsson MW, Wallenberg LR, Samuelson L (2004) Nano Lett 4:1621

Quantum Error Correction and Fault Tolerant Quantum Computing

Markus Grassl[1], Martin Rötteler[2]
[1] Institute for Quantum Optics and Quantum Information, Austrian Academy of Sciences, Innsbruck, Austria
[2] NEC Laboratories America, Inc., Princeton, USA

Article Outline

Glossary

Ancilla qubits Refers to qubits that are used to facilitate quantum computations, but serve neither as input nor as output of the computation. The two main uses of ancilla qubits are a) for the purpose of a scratch pad during a quantum computation, i. e., states that are initialized in a known state and can be used during a computation such that they are returned to the initial state at the end, and b) for the purpose of quantum error correction where they serve as space for holding the error syndrome.

Bit-flip error An error that affects a single qubit and interchanges the basis states of that qubit.

CSS code A special class of quantum codes named after their inventors Calderbank, Steane, and Shor. CSS codes allow to construct quantum codes from certain classical error-correcting codes.

CNOT gate The controlled-NOT gate is an example of a quantum gate. It is a two qubit gate and – together with single qubit gates – can be shown to be a universal gate for quantum computation.

Error syndrome A classical bit-vector that describes a quantum error that has affected a quantum code. Contrary to the classical case where there is a one-to-one correspondence between syndromes and errors, in the quantum case a syndrome does in general not uniquely identify a quantum error.

Fault-tolerant quantum computing The discipline that studies how to perform reliable quantum computations with imperfect hardware.

Phase-flip error An error that affects a single qubit and gives a relative phase of -1 to the basis states of that qubit. Contrary to the bit-flip error, this error has no classical analog.

Quantum channel A term describing a general physical operation that can affect the state of a quantum-mechanical system. Another name for quantum channels are completely-positive trace preserving maps.

Quantum error-correcting code (QECC) A method to introduce redundancy to a quantum mechanical system in such a way that certain errors that are affecting the system can be detected or corrected.

Quantum circuits Operations that a quantum computer can carry out. Quantum circuits are composed of quantum gates and, when acting on an n qubit system, correspond to unitary matrices of size $2^n \times 2^n$.

Quantum gate A basic operation that a quantum computer can carry out. Common choices for quantum gates are certain unitary matrices of size 2×2 (single qubit gates) or 4×4 (two qubit gates).

Qubit Shortened form of *quantum bit*. A physical object that can support states in a two-dimensional complex Hilbert space. Contrary to the classical case where a bit refers to both, the physical object and the states it can hold, a qubit refers to the physical object only. Qubits are the basic units of a quantum computer's memory.

Stabilizer code A class of quantum codes that is defined as the joint eigenspace of a group of commuting operators. Stabilizer codes give rise to many examples of quantum codes, comprise the class of CSS codes as a special case, and are equivalent to additive codes over $GF(4)$ that are self-orthogonal with respect to the Hermitian inner product.

Threshold theorem An important result of the theory of fault-tolerant quantum computing, stating that there is a threshold value for this noise level such that arbitrarily long quantum computations become possible if the gates have a noise level that is under the threshold.

Transversal gates Special class of operations acting on encoded quantum data. Transversal gates have the important property to exhibit benign behavior in case errors happen during the application of the gate.

Definition of the Subject

Quantum error correction offers a solution to the problem of protecting quantum systems against noise induced by interactions with the environment or caused by imperfect control of the system. The need for error correction arises not only in communication, when quantum information is sent over some distance, but also in locally, when storing and processing quantum information. Fault-tolerant quantum computing builds on quantum error correction and denotes techniques that allow computations to be performed on a quantum system with faulty gates as well as storage errors. Without mechanisms for quantum error correction and fault-tolerance, quantum computing would be impossible even for moderate error rates.

The idea of quantum error correction was first conceived in a paper [64] by Shor in 1995 in which a particular quantum code was given that encodes one quan-

tum bit (qubit) into nine quantum bits, while being able to correct against one arbitrary error on one of these nine qubits. Before long, Bennett et al. [11] developed a theory of error correction describing a quantum code that encodes one quantum bit into five while still being able to correct against arbitrary single qubit errors. Around the same time, Calderbank, Shor [15] and Steane [66] constructed quantum codes from suitable pairs of classical codes. Named after their inventors, this important class of error correcting codes is called CSS codes.

Other noteworthy work in the early days of quantum error correction were the contributions by Calderbank et al. [14], which related quantum codes to classical codes over the finite field $GF(4)$ that are additively closed and self-orthogonal with respect to the Hermitian inner product, and by Gottesman [25,26], who developed the theory of stabilizer codes. These two constructions are equivalent.

As far as fault-tolerant quantum computing is concerned, again the first method was given by Shor [65] who introduced the idea of performing a universal set of quantum gates fault-tolerantly, including syndrome measurements required for quantum error correction. Around the same time, ideas based on concatenation of quantum codes were used by Aharonov and Ben-Or [2], Kitaev [42], Gottesman [26], Knill, Laflamme and Zurek [45], and Preskill [58], to derive fault-tolerant schemes that work even if only faulty gates are available. The characteristic feature of all these constructions is that there is a certain value for the error rate such that if the actual error rate is below this value, then arbitrarily long quantum computations can be performed with any desired accuracy, whereas if the actual error rate is above this value, then quantum computing is impossible with the same scheme. This watershed type behavior is characterized in the threshold theorem which is a central result in fault-tolerant quantum computing. As the task to determine the exact value of the threshold is daunting, much of work has been done to get upper and lower bounds as well as to perform numerical simulations that give indications of the threshold value.

Introduction

In the early days of quantum computing, Haroche and Raimond asked the poignant question whether the dream of quantum computing could ever be realized in a real physical system or if "the large-scale quantum machine … is the experimenter's nightmare" [39]. At the time the article was written, the first quantum error-correcting code had just been proposed [64]. However, Haroche and Raimond argued that "the implementation of error-correcting codes will become exceedingly difficult" given any detection efficiency less than 100%. It was only later that it was shown that even with imperfect quantum memory and imperfect quantum operations it is possible to implement an arbitrarily long quantum computation, provided that the failure probability of each element is below a certain threshold [2,45].

Here we provide an overview of the ingredients leading to fault-tolerant quantum computation (FTQC). In the first part, we present the theory of quantum error-correcting codes (QECCs) and, in particular, two important classes of QECCs: CSS codes and stabilizer codes. Both are related to classical error-correcting codes, so we start with some basics from this area. In the second part of the article, we present a high-level view of the main ideas of FTQC and the threshold theorem.

For the background of quantum computing in general, we refer the reader to the related articles in this volume as well as the book by Nielsen and Chuang [56].

Basics of Classical Error Correction

Block Codes

Before discussing the principles of quantum error correction, we briefly summarize some results from classical error correction. For more information see MacWilliams and Sloane [53]. Error correction is part of information theory whose foundations were laid by Claude Shannon in his landmark paper "A Mathematical Theory of Communication" [63]. In that paper, Shannon introduced the basic mathematical concepts for communication systems that are used to send messages from one point in space or time to another point:

> "*The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have meaning; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one selected from a set of possible messages. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design.*"

Hence we can model the set of all possible messages, for example, by a set of binary strings of fixed length.

Definition 1 (block code) A *block code* B of length n is a subset of all possible sequences of length n over an al-

phabet $\mathcal{A}$, i. e., $B \subseteq \mathcal{A}^n$. The *rate*

$$R = \frac{\log|B|}{\log|\mathcal{A}^n|} = \frac{\log|B|}{n\log|\mathcal{A}|}$$

of the code is a measure of the amount of information that is transmitted per symbol in the code.

While the rate of the code should be as high as possible, using only a proper subset of all possible messages allows us to correct some errors. On the highest level of abstraction, one distinguishes only whether a symbol is transmitted correctly or not. This yields to the following measure for the distance between two sequences of length n.

Definition 2 (Hamming distance) The *Hamming distance* between two sequences $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ equals the number of positions where $\mathbf{x}$ and $\mathbf{y}$ differ, that is,

$$d_{\mathrm{H}}(\mathbf{x}, \mathbf{y}) := |\{i \colon 1 \le i \le n \mid x_i \ne y_i\}| \,.$$

If the alphabet contains a symbol zero, the *Hamming weight* of a sequence is defined as the number of non-zero elements of the sequence.

If we send two copies of every symbol to the receiver, we get a *repetition code* of length $n = 2$ and rate $R = 1/2$. The Hamming distance between any two codewords is two. In this situation, the receiver can detect an error if at most one of the symbols has been corrupted during transmission. If we repeat every symbol three times ($n = 3$, $R = 1/3$), the receiver can even deduce the correct message from the majority of the symbols, assuming that no more than one symbol is erroneous. The general situation is as follows:

Theorem 1 *Let B be a block with minimum distance d, that is, any two codewords differ in at least d positions. Then one can either detect errors that affect less than d positions or correct errors that affect strictly less than $d/2$ positions.*

Proof As the minimum distance of the code is d, changing up to $d-1$ positions of a codeword does not yield another codeword. So in order to detect up to $d-1$ errors, it is sufficient to check whether the received word is a codeword or not. In order to correct errors, note that the spheres of radius $\lfloor (d-1)/2 \rfloor$ around any codeword are disjoint. In the error correction process, every word in that sphere will be mapped to the corresponding codeword. □

In this setting, finding a good error-correcting code, by which we mean a code with both high rate and high minimum distance, amounts to packing spheres of words of length n with radius $\lfloor (d-1)/2 \rfloor$ with respect to Hamming distance. In general, the resulting code will be a set of codewords without further structure so that to succeed we would have to store the list of all codewords.

Linear Block Codes

Developing more efficient ways of describing all codewords and testing whether a given word lies in the code requires that the code has some additional structure. First, we assume that the alphabet of the code is a finite field $GF(q)$ with q elements. Second, we require that any linear combination of two codewords is again a codeword. The resulting code is a linear block code, denoted by $C = [n, k, d]$ (for more details see [53]). Here k is the dimension of the code as a subspace of the vector space $GF(q)^n$, and d denotes the minimum distance of the code. Instead of listing all q^k codewords, it is sufficient to specify a basis with k linearly independent vectors of the linear space. Alternatively, a subspace of dimension k is uniquely described as the space of solutions of $n-k$ linearly independent homogeneous linear equations.

Definition 3 (generator matrix/parity check matrix) A *generator matrix* of a linear block code $C = [n, k, d]$ is a matrix G with k rows and n columns of full rank such that the row span of G equals the code C. A *parity check matrix* is a matrix H with $n-k$ rows and n columns of full rank such that the row-nullspace of H equals the code C.

The generator matrix G provides both a compact description of a linear block code and an efficient way of encoding a message which can be represented by a vector $\mathbf{i}$ of length k. The corresponding codeword is obtained by the linear mapping $\mathbf{i} \mapsto \mathbf{c} := \mathbf{i}G$. The parity check matrix provides an efficient way to detect errors.

Theorem 2 (error syndrome) *Let H be a parity check matrix of the linear block code $C = [n, k, d]$. Then a vector $\boldsymbol{v}$ is a codeword if and only if the error syndrome $\mathbf{s} := \mathbf{v}H^t$ is zero. Furthermore, for an erroneous codeword $\mathbf{v} = \mathbf{c} + \mathbf{e}$ the error syndrome depends only on the error $\boldsymbol{e}$, but not on the codeword $\boldsymbol{c}$.*

Proof By definition, the code C equals the row-nullspace of the parity check matrix H. Furthermore, any erroneous codeword can be written as $\mathbf{v} = \mathbf{c} + \mathbf{e}$. Then we have

$$\mathbf{s} = \mathbf{v}H^t = (\mathbf{c} + \mathbf{e})H^t = \mathbf{c}H^t + \mathbf{e}H^t = \mathbf{e}H^t \,.$$

□

Later we will need the concept of the *dual code* which also motivates the definition of the parity check matrix with n columns and $n-k$ rows.

Proposition 1 (dual code) *Let $C = [n, k, d]$ be a linear block code. Then the (Euclidean) dual code $C^{\perp}$ is the space of all vectors that are orthogonal to all codewords with respect to the Euclidean inner product $\mathbf{v} \cdot \mathbf{w} := \sum_{i=1}^{n} v_i w_i$, i. e.*

$$C^{\perp} = \{\mathbf{v} \colon \mathbf{v} \in GF(q)^n \mid \mathbf{v} \cdot \mathbf{c} = 0 \quad \textit{for all} \ \ \mathbf{c} \in C\} \, .$$

If G is a generator matrix and H is a parity check matrix for C, then G is a parity check matrix and H is a generator matrix for $C^{\perp}$, i. e., the role of G and H is interchanged.

The parity check matrix can also be used to compute the minimum distance of a linear block code.

Theorem 3 *If any $d-1$ columns of the parity check matrix H of a linear block code C are linearly independent, then the minimum distance of the code is at least d.*

Proof First note that for a linear code, the minimum distance equals the minimum Hamming weight of a nonzero codeword as $d_H(\mathbf{x}, \mathbf{y}) = d_H(\mathbf{x} - \mathbf{y}, \mathbf{0})$. Assume that $\mathbf{c}$ is a codeword of Hamming weight $d-1$, that is, there are $d-1$ indices $i_1, \ldots, i_{d-1}$ such that $c_{i_j} \neq 0$. The syndrome is computed as $0 = \mathbf{c}H^t = c_{i_1} h^{(i_1)} + \cdots + c_{i_{d-1}} h^{(i_{d-1})}$, where $h^{(i)}$ denotes the ith column of H. Hence we have a non-trivial linear combination of $d-1$ columns of H that is zero, contradicting the fact that any $d-1$ columns of H are linearly independent. □

Hamming Codes

The previous theorem can be used to construct codes with a prescribed minimum distance d. For a code that can correct a single error, the minimum distance d must be at least three, that is, any two columns of the parity check matrix must be linearly independent. For the simplest case in which the field has two elements $GF(2) = \{0, 1\}$, that is, all operations are modulo two, it is sufficient that all columns of H are distinct and nonzero. This yields the following family of single error-correcting codes (see [38,53]).

Proposition 2 (binary Hamming code) *The rth binary Hamming code is a linear binary block code of length $n = 2^r - 1$, dimension $k = 2^r - r - 1$, and minimum distance $d = 3$. A parity check matrix H of the Hamming code is a matrix whose $2^r - 1$ columns are all nonzero binary vectors of length r.*

The columns of H can be arranged such that the ith column equals the binary expansion of i. Then for an error $\mathbf{e}$ of weight one, the syndrome $\mathbf{e}H^t$ equals the binary expansion of the position of the error. For $r = 3$, we get the following parity check matrix

$$H = \begin{pmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix} .$$

For the error $\mathbf{e} = (0, 0, 1, 0, 0, 0, 0)$, the syndrome is $\mathbf{e}H^t = (0, 1, 1)$, that is, the binary expansion of three.

In general it is difficult to deduce the error of smallest Hamming weight—which is often the most likely error—from the error syndrome. More precisely, given a binary parity check matrix H, an error syndrome $\mathbf{s}$, and a positive integer w it is NP complete to decide whether there is an error vector $\mathbf{e}$ whose weight does not exceed w such that $\mathbf{e}H^t = \mathbf{s}$ [12]. Nonetheless, for some classes of codes, such as BCH codes or Reed–Solomon codes, there exist efficient algorithms for the correction of all errors up to a certain weight (see [53]).

Basic Ideas of Quantum Error Correction

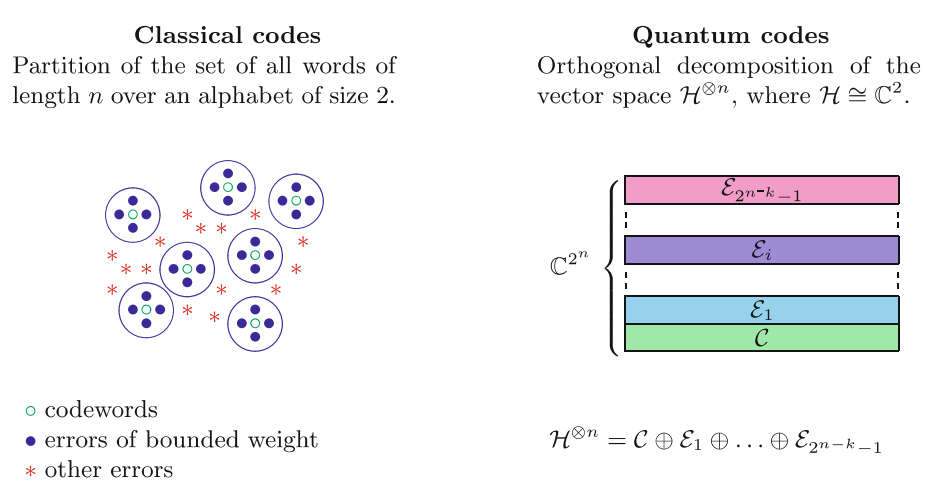
In the classical setting of error correction, information is represented, for example, by a binary string of length n. The extremal case is that we have only one bit that is either zero or one. In the context of quantum information, the simplest quantum system is modeled by a two-dimensional complex vector space. A *quantum bit* (or *qubit*, for short) corresponds to a normalized vector in this space (for more details see the book by Nielsen and Chuang [56]). The qubit is given by

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle, \quad \text{where} \ \ |\alpha|^2 + |\beta|^2 = 1, \ \ \alpha, \beta \in \mathbb{C} \, .$$

Here $|0\rangle$ and $|1\rangle$ denote two orthogonal vectors in the two-dimensional vector space. While these basis states correspond to the two classical values 0 and 1 of a bit, a qubit can be in a *superposition* of both $|0\rangle$ and $|1\rangle$. When the qubit is measured with respect to the basis $\{|0\rangle, |1\rangle\}$, the result is either "0" or "1" with probability $|\alpha|^2$ and $|\beta|^2$, respectively.

As we have seen in Subsect. "Block Codes", a simple classical one-error correcting code can be obtained by sending the information three times and taking a majority decision at the receiver's end. However, this does not work in the context of quantum information. First, it is not possible for the sender to compute copies of an unknown quantum state (see [69]). The main idea of this so-called *no-cloning theorem* is as follows.

Theorem 4 *There is no quantum operation that maps an arbitrary quantum state $|\psi\rangle$ and a fixed state $|\phi_0\rangle$ to two independent copies $|\psi\rangle|\psi\rangle$.*



Quantum Error Correction and Fault Tolerant Quantum Computing, Figure 1
Similarities between classical and quantum error-correcting codes

Proof Let $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$. Two independent copies of $|\psi\rangle$ are given by

$$|\psi\rangle|\psi\rangle = \big(\alpha|0\rangle + \beta|1\rangle\big) \otimes \big(\alpha|0\rangle + \beta|1\rangle\big) = \alpha^2|00\rangle + \alpha\beta|01\rangle + \alpha\beta|10\rangle + \beta^2|11\rangle \,.$$

If $|\psi\rangle$ is one of the basis states, we have $|0\rangle|\phi_0\rangle \mapsto |00\rangle$ and $|1\rangle|\phi_0\rangle \mapsto |11\rangle$. By the linearity of quantum mechanics, starting with a superposition we get $\big(\alpha|0\rangle + \beta|1\rangle\big)|\phi_0\rangle \mapsto \alpha|00\rangle + \beta|11\rangle$. This equals $|\psi\rangle|\psi\rangle$ only when $\alpha = 0$ or $\beta = 0$. □

Second, even if independent copies of a quantum state $|\psi\rangle$ are sent (for example, if the sender knows how to prepare the state $|\psi\rangle$), the direct quantum mechanical analogue of a majority decision is not possible for the receiver. Instead, the receiver may, for example, check whether the joint quantum state of all received copies is invariant under permutation of the copies. This allows us to detect and correct some errors [9].

Although the classical repetition code does not have a direct quantum mechanical analogue, we will see that quantum error correction can be related to classical error correction codes. For this we recall that, in general, a classical code is given by a proper subset of the finite set of all possible messages of fixed length. In contrast, quantum information is represented by an arbitrary normalized vector in a complex vector space. In order to achieve the possibility of correcting quantum errors, a quantum error-correcting code must be a proper subspace of a larger vector space. Similar to the problem of packing spheres for classical codes, the subspace has to be chosen in such a way that the spaces corresponding to the quantum errors do not overlap (see Fig. 1). Before addressing this question in more detail in Sect. "Conditions for Quantum Error Correction", we present a simple example.

A Simple Example and Shor's Nine-Qubit Code

The Three-Qubit Code

The shortest classical code that encodes one bit and can correct one error is the repetition code of length three which coincides with the second binary Hamming code. Hence a parity check matrix H and a generator matrix G are given by

$$H = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix} \quad \text{and} \quad G = \begin{pmatrix} 1 & 1 & 1 \end{pmatrix} .$$

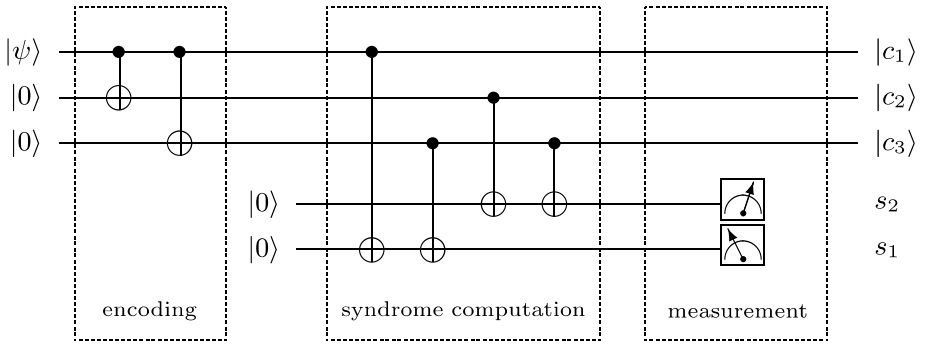
We use the codewords (000) and (111) of the binary code to define the basis states of a three-qubit code, that is, the encoding operation is given by

$$\begin{aligned} \mathsf{C}\colon \ \mathbb{C}^2 &\to (\mathbb{C}^2)^{\otimes 3} \\ |0\rangle &\mapsto |000\rangle \\ |1\rangle &\mapsto |111\rangle \,. \end{aligned}$$

Hence a superposition $\alpha|0\rangle + \beta|1\rangle$ is encoded as $\mathsf{C}(\alpha|0\rangle + \beta|1\rangle) = \alpha|000\rangle + \beta|111\rangle$. The encoding transformation C can be implemented using two controlled-NOT (CNOT) gates (see [56]). The CNOT gate is given by

$$\text{CNOT}\colon |x\rangle|y\rangle \mapsto |x\rangle|x \oplus y\rangle \,,$$

where $x \oplus y$ denotes addition modulo two (XOR). Hence the second (target) qubit is flipped if and only if the first (control) qubit is one. The states $|000\rangle$ and $|111\rangle$ span the



Quantum Error Correction and Fault Tolerant Quantum Computing, Figure 2
Quantum circuit for the three-qubit code. Two ancilla qubits are used for the encoding, and another two for the error syndrome. The measurement yields two classical syndrome bits

quantum code C which is a two-dimensional subspace of $\mathbb{C}^2 \otimes \mathbb{C}^2 \otimes \mathbb{C}^2$. A classical error flips a bit, interchanging 0 and 1. The quantum mechanical analogue is given by the matrix

$$\sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

interchanging the basis states $|0\rangle$ and $|1\rangle$. Applying σ_x to at most one of the three subsystems we obtain the following states:

error	state	subspace
no error	$\alpha\|000\rangle + \beta\|111\rangle$	$(I \otimes I \otimes I)\, C =: C_0$
1st position	$\alpha\|100\rangle + \beta\|011\rangle$	$(\sigma_x \otimes I \otimes I)\, C =: C_1$
2nd position	$\alpha\|010\rangle + \beta\|101\rangle$	$(I \otimes \sigma_x \otimes I)\, C =: C_2$
3rd position	$\alpha\|001\rangle + \beta\|110\rangle$	$(I \otimes I \otimes \sigma_x)\, C =: C_3$

(1)

The four different cases yield four mutually orthogonal subspaces, that is, the Hilbert space of three qubits can be decomposed as follows:

$$\mathbb{C}^2 \otimes \mathbb{C}^2 \otimes \mathbb{C}^2 = (I \otimes I \otimes I)\, C \oplus (\sigma_x \otimes I \otimes I)\, C \oplus (I \otimes \sigma_x \otimes I)\, C \oplus (I \otimes I \otimes \sigma_x)\, C\,.$$

In principle it is possible to construct a quantum mechanical observable whose eigenspaces are the four two-dimensional spaces C_i in (1). The corresponding projective measurement projects onto one of these spaces and provides information about the error, but preserves the superposition within the spaces. Alternatively, we can compute information about the error using two auxiliary qubits (ancillae). Recall that for the binary Hamming code, the error syndrome $\mathbf{s} = \mathbf{e}H^t$ equals the binary expansion of the position of the error, provided that there is at most one error. The computation of the error syndrome can also be implemented using CNOT gates (see Fig. 2). The gates in the box labeled "syndrome computation" implement the transformation

$$|\mathbf{c}\rangle|00\rangle = |c_1c_2c_3\rangle|0\rangle|0\rangle \mapsto |c_1c_2c_3\rangle|c_2 \oplus c_3\rangle|c_1 \oplus c_3\rangle = |\mathbf{c}\rangle|\mathbf{c}H^t\rangle\,. \quad (2)$$

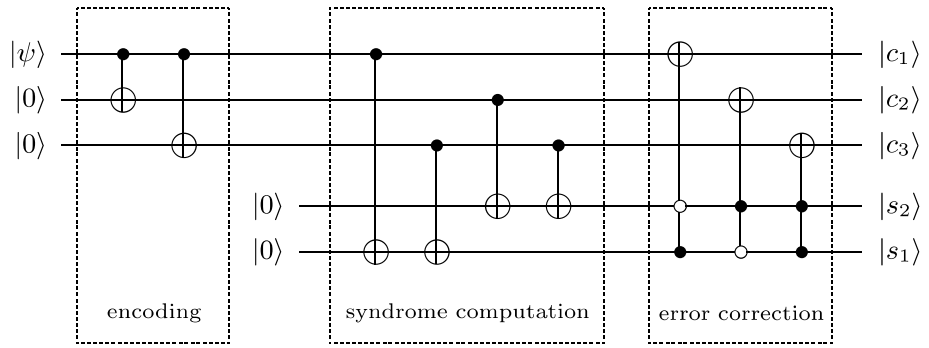
Measuring the two ancilla qubits yields a two-bit syndrome $\mathbf{s} = (s_2s_1)$ encoding the position of the error. This classical information can be used to correct the error. Instead of measuring the syndrome qubits, one can use controlled quantum operations to correct the errors as illustrated in Fig. 3.

While this three-qubit code allows us to correct the quantum mechanical analogue of a single bit-flip error, it cannot correct an arbitrary single qubit error. For instance, measuring any of the three qubits with respect to the standard basis $\{|0\rangle, |1\rangle\}$ for the encoded state $|\Phi\rangle = \alpha|000\rangle + \beta|111\rangle$ has the same effect as measuring the unencoded qubit $\alpha|0\rangle + \beta|1\rangle$. In order to turn the three-qubit code into a code that can correct the effect of measuring one qubit, we first use the following identities for projection onto the basis states:

$$|0\rangle\langle 0| = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} = \frac{1}{2}(I + \sigma_z) \quad \text{and}$$

$$|1\rangle\langle 1| = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} = \frac{1}{2}(I - \sigma_z)\,, \quad (3)$$

where $\sigma_z = \left(\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}\right)$. The relation between σ_x and σ_z is given by $\sigma_z = H\sigma_xH$, where $H = 1/\sqrt{2}\left(\begin{smallmatrix} 1 & 1 \\ 1 & -1 \end{smallmatrix}\right)$. Hence the Hadamard matrix H turns bit-flip errors σ_x into phase-flip errors σ_z. Since the three-qubit code C is able to correct a single bit-flip error, the code $C^{\text{phase}} = (H \otimes H \otimes H)\, C$ can correct a single phase-flip error. By linearity, this code



Quantum Error Correction and Fault Tolerant Quantum Computing, Figure 3
Quantum circuit for the three-qubit code. In contrast to Fig. 2, the error is corrected without measuring the syndrome qubits

can also correct the effect of measuring one subsystem. Assume that we measured the first qubit of the encoded state $|\Phi\rangle$ and the result was 0. Using (3), for the state after the measurement we compute

$$\big(|0\rangle\langle 0| \otimes I \otimes I\big)|\Phi\rangle = \frac{1}{2}|\Phi\rangle + \frac{1}{2}\big(\sigma_z \otimes I \otimes I\big)|\Phi\rangle\,,$$

that is, the state is a superposition of the states corresponding to "no error" and "phase error at the first position". Measuring the error syndrome projects either onto the error free state $|\Phi\rangle$ or onto the state with a single phase-flip error that can be corrected. The measurement result 1, as well as measuring one of the other qubits, can be treated similarly.

Shor's Nine-Qubit Code

The three-qubit code C and its Hadamard transformed version C^{phase} of the previous section can correct a single bit-flip error or a single phase-flip error, respectively. However, none of the codes can correct both types of errors. For an encoded state $|\Psi\rangle = \alpha|000\rangle + \beta|111\rangle$ of the three-qubit code, a single phase-flip error σ_z results in the state $(\sigma_z \otimes I \otimes I)|\Psi\rangle = \alpha|000\rangle - \beta|111\rangle$. In terms of the *encoded* or *logical* basis states $|0\rangle_L = |000\rangle$ and $|1\rangle_L = |111\rangle$ corresponding to the encoding of $|0\rangle$ and $|1\rangle$, respectively, a single phase-flip has the effect of an *encoded* σ_z operation. Note that the operations $\sigma_z \otimes I \otimes I$, $I \otimes \sigma_z \otimes I$, and $I \otimes I \otimes \sigma_z$ all have the same effect on the code. Hence, in order to correct also for phase-flip errors, we can add another layer of encoding using the code C^{phase} with the encoded basis states

$$\begin{aligned}|0\rangle &\mapsto \frac{1}{2}(|000\rangle + |011\rangle + |101\rangle + |110\rangle)\\ |1\rangle &\mapsto \frac{1}{2}(|001\rangle + |010\rangle + |100\rangle + |111\rangle)\,.\end{aligned} \tag{4}$$

Each qubit in (4) is replaced by its encoded version with respect to the three-qubit code C, that is, we get the encoding

$$\begin{aligned}|0\rangle &\mapsto \frac{1}{2}(|000000000\rangle + |000111111\rangle\\ &\qquad + |111000111\rangle + |111111000\rangle)\\ |1\rangle &\mapsto \frac{1}{2}(|000000111\rangle + |000111000\rangle\\ &\qquad + |111000000\rangle + |111111111\rangle)\,.\end{aligned} \tag{5}$$

This nine-qubit code has been constructed by Shor [64]. It turns out it can correct an arbitrary single-qubit error. For this, first note that the errors σ_x and σ_z can be corrected independently on the two levels of encoding, respectively. Therefore, the code can correct not only these errors, but also their combination $\sigma_x\sigma_z$. Together with identity we have the matrices

$$I = \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}, \quad \sigma_x = \begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix},$$
$$\sigma_z = \begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}, \quad \text{and} \quad \sigma_x\sigma_z = \begin{pmatrix}0 & -1\\ 1 & 0\end{pmatrix}. \tag{6}$$

Any 2×2 matrix can be written as linear combination of the four matrices in (6). Similar to the arguments following (3), it can be shown that in order to correct an arbitrary single-qubit error it is indeed sufficient to correct the errors corresponding to the matrices in (6). The quantum error-correcting code given by (5) is denoted by $C = [\![9, 1, 3]\!]$, indicating that one qubit is encoded into nine qubits and that the minimum distance of the code is three.

Before presenting further constructions for quantum codes, we will discuss necessary and sufficient conditions for quantum error correction.

Conditions for Quantum Error Correction

Quantum Channels

Errors in quantum systems are due to interaction of the system with its environment or imperfection in the control of quantum operations. The latter can be modeled by a perfect operation that additionally depends on the state of the environment. The Hilbert space of the combined system is given by $\mathcal{H}_{\text{sys/env}} = \mathcal{H}_{\text{system}} \otimes \mathcal{H}_{\text{environment}}$. If the dimension of both spaces is sufficiently large, we can assume that the initial state is a pure state. Additionally, we assume that there are no initial correlations between the system and its environment, that is, the initial state is a product state $|\phi\rangle_{\text{sys}}|\varepsilon\rangle_{\text{env}}$. Again, for sufficiently large dimensions of the Hilbert spaces we can assume that the dynamics of the joint system are given by a unitary transformation $U_{\text{sys/env}}$. The resulting state of the system is obtained by tracing out the environment:

$$\rho_{\text{out}} = \text{Tr}_{\text{env}}\left(U_{\text{sys/env}}(|\phi\rangle\langle\phi| \otimes |\varepsilon\rangle\langle\varepsilon|)U^{\dagger}_{\text{sys/env}}\right) . \quad (7)$$

The output state in (7) can equivalently be expressed as a function of the input state $\rho_{\text{in}} = |\phi\rangle\langle\phi|$ in the form

$$\rho_{\text{out}} = \sum_i E_i \, \rho_{\text{in}} \, E_i^{\dagger} ,$$

where the operators E_i are the so-called *error operators* or *Kraus operators* [47]. They depend on both the initial state $|\varepsilon\rangle$ of the environment and the unitary interaction $U_{\text{sys/env}}$. In the following, some important special cases of quantum channels are presented.

Example (depolarizing channel) The *depolarizing channel* on the Hilbert space $\mathcal{H}$ with error parameter p, $0 \le p \le 1$, is given by

$$\rho \mapsto (1-p)\cdot\rho + p \cdot I/\dim\mathcal{H} .$$

The input ρ is transmitted faithfully with probability $1-p$. With probability p, the state is replaced by the completely random state $I/\dim\mathcal{H}$. Note that even in this case, the probability of measuring a particular pure state $|\psi\rangle$ is $1/\dim\mathcal{H} \neq 0$.

While the depolarizing channel treats all input states uniformly, the next quantum channel is basis-dependent.

Example (dephasing channel) The *dephasing channel* on the Hilbert space $\mathcal{H}$ with orthonormal basis $\mathcal{B} = \{|\mathbf{b}_i\rangle : i \in \mathcal{I}\}$ and error parameter p, $0 \le p \le 1$, is given by

$$\rho \mapsto (1-p)\cdot\rho + p\sum_{i\in\mathcal{I}} |\mathbf{b}_i\rangle\langle\mathbf{b}_i|\rho|\mathbf{b}_i\rangle\langle\mathbf{b}_i| .$$

With probability p, the channel performs a projective measurement with respect to the basis $\mathcal{B}$. is derived from the fact that this is equivalent to randomizing the phases of the basis states. The dephasing channel allows us to perfectly transmit classical information by encoding the information as basis states. Coherent superpositions of basis states, however, are changed into classical mixtures.

The final example is a channel that provides the side-information that an error has occurred.

Example (quantum erasure channel [35]*)* The *quantum erasure channel* on the Hilbert space $\mathcal{H}$ with error parameter p, $0 \le p \le 1$, is given by

$$\rho \mapsto (1-p)\cdot\rho + p\cdot|\epsilon\rangle\langle\epsilon| ,$$

where $|\epsilon\rangle$ is a quantum state in the Hilbert space $\mathcal{H}' \supset \mathcal{H}$ that is orthogonal to all states in $\mathcal{H}$. The input ρ is transmitted faithfully with probability $1-p$. With probability p, the state is replaced by the state $|\epsilon\rangle\langle\epsilon|$. As $|\epsilon\rangle$ is orthogonal to all states in $\mathcal{H}$, the receiver can perform a measurement which detects that an error has occurred.

The quantum mechanical analogue of a memoryless channel is a *product channel* which is defined for a quantum system with n subsystems of, say, equal dimension, that is, on $\mathcal{H} = \mathcal{H}_0^{\otimes n}$. The product channel is given by n uses of a channel on $\mathcal{H}_0$ which acts independently on each of the n subsystems. If the channel on $\mathcal{H}_0$ is given by the error operators $\mathcal{E}_0 = \{E_i : i \in \mathcal{I}\}$, the error operators of the product channel on $\mathcal{H}$ are

$$\mathcal{E} = \mathcal{E}_0^{\otimes n} := \{E_{i_1}\otimes E_{i_2}\otimes\ldots\otimes E_{i_n} : (i_1, i_2, \ldots, i_n) \in \mathcal{I}^n\}.$$

Characterization of Quantum Codes

As we have seen in Fig. 1, a quantum error-correcting code is a subspace C of the Hilbert space $\mathcal{H}$. What is more, the full Hilbert space $\mathcal{H}$ can be decomposed into mutually orthogonal unitary images of C, corresponding to different error events (see Eq. (1)). In general, the error operators E_i describing the quantum channel need not be unitary. This leads to the question whether a subspace C of $\mathcal{H}$ is a quantum error-correcting code (QECC) for a given quantum channel.

Theorem 5 (QECC characterization [44]) *Let Q be a quantum channel on $\mathcal{H}$ with error operators $\{E_i : i \in \mathcal{I}_Q\}$. A subspace $C \le \mathcal{H}$ with orthonormal basis $\{|\mathbf{c}_i\rangle : i \in \mathcal{I}_C\}$ is a quantum error-correcting code for Q if and only if the following conditions hold:*

$$\forall k, \ell \in \mathcal{I}_Q \forall i \neq j \in \mathcal{I}_C : \langle\mathbf{c}_i|E_k^{\dagger}E_\ell|\mathbf{c}_j\rangle = 0 \quad (8a)$$

$$\forall k,\ell \in \mathcal{I}_Q \forall i,j \in \mathcal{I}_C\colon \langle \mathbf{c}_i|E_k^\dagger E_\ell|\mathbf{c}_i\rangle = \langle \mathbf{c}_j|E_k^\dagger E_\ell|\mathbf{c}_j\rangle =: \alpha_{k\ell} \in \mathbb{C} \quad (8b)$$

Denoting by $P_C := \sum_{i\in\mathcal{I}_C} |\mathbf{c}_i\rangle\langle\mathbf{c}_i|$ *the projection onto the code C, we obtain the following equivalent condition which is independent of the basis of the code:*

$$\forall k,\ell \in \mathcal{I}_Q\colon P_C E_k^\dagger E_\ell P_C = \alpha_{k\ell} P_C\,.$$

In principle, the proof of Theorem 5 implies an algorithm that allows the correction of errors (see [31]). However, as error correction is NP hard in the classical case we cannot expect to have an efficient algorithm for the more general situation of the quantum case.

If the conditions (8) are fulfilled, then the errors corresponding to the operators E_i can be corrected. It must be stressed that this implies that one can correct any operator that is a linear combination of the error operators E_i. For this, we show that the conditions (8) are linear in the error operators. Consider the new error operators

$$A := \sum_k \lambda_k E_k \quad \text{and} \quad B := \sum_l \mu_l E_l$$

which are arbitrary linear combinations of the E_i. Using (8) we compute

$$\begin{aligned}\langle c_i|A^\dagger B|c_j\rangle &= \sum_{k,l} \overline{\lambda_k}\mu_l \langle c_i|E_k^\dagger E_l|c_j\rangle \\ &= \sum_{k,l} \overline{\lambda_k}\mu_l \delta_{i,j}\alpha_{k,l} \\ &= \delta_{i,j}\cdot\alpha'(A,B)\,,\end{aligned}$$

where $\alpha'(A,B) \in \mathbb{C}$ is some constant depending on the operators A and B only. From this it also follows that it is sufficient to check the conditions (8) for a vector space basis of $\mathcal{E}$ and hence for a finite set of errors. For qubit systems, the Pauli matrices

$$X = \sigma_x = \begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}, \quad Y = \sigma_y = \begin{pmatrix}0 & -\mathbf{i}\\ \mathbf{i} & 0\end{pmatrix}, \quad \text{and} \quad Z = \sigma_z = \begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}, \quad (9)$$

together with identity form a vector space basis of all operators in $\mathbb{C}^{2\times 2}$. For a quantum code using n qubits, we consider the tensor product of Pauli matrices and identity as the so-called *error basis*. The number of tensor factors different from identity is referred to as the *number of errors* or the *weight of an error*.

Quantum Codes from Classical Codes

CSS Codes

The three-qubit code of Sect. "The Three-Qubit Code" is based on the classical triple repetition code. Both codes can correct a single bit-flip error. The Hadamard transformation yields the code C^{phase} which can correct a single phase-flip error. In the following we present a similar construction of quantum codes based on linear binary block codes, but the resulting codes will allow both the correction of bit-flip errors and phase-flip errors. For the construction we need the following lemma.

Lemma 1 *Let* $C \le GF(2)^n$ *denote a k-dimensional linear subspace of* $GF(2)^n$ *and let* $\mathbf{a},\mathbf{b} \in GF(2)^n$ *be two arbitrary binary vectors. Furthermore, by* $H_{2^n} := H^{\otimes n}$, *where* $H = 1/\sqrt{2}\left(\begin{smallmatrix}1 & 1\\ 1 & -1\end{smallmatrix}\right)$, *we denote the Hadamard transformation on n qubits. Then the state*

$$|\psi\rangle := \frac{1}{\sqrt{|C|}}\sum_{\mathbf{c}\in C}(-1)^{\mathbf{a}\cdot\mathbf{c}}|\mathbf{c}+\mathbf{b}\rangle$$

is mapped by the Hadamard transformation to

$$H_{2^n}|\psi\rangle = \frac{(-1)^{\mathbf{a}\cdot\mathbf{b}}}{\sqrt{|C^\perp|}}\sum_{\mathbf{d}\in C^\perp}(-1)^{\mathbf{b}\cdot\mathbf{d}}|\mathbf{d}+\mathbf{a}\rangle\,.$$

Proof The Hadamard transformation on n qubits can be written as

$$H_{2^n} = \frac{1}{\sqrt{2^n}}\sum_{\mathbf{x},\mathbf{y}\in GF(2)^n}(-1)^{\mathbf{x}\cdot\mathbf{y}}|\mathbf{x}\rangle\langle\mathbf{y}|\,,$$

where $\mathbf{x}\cdot\mathbf{y}$ denotes the inner product of the binary vectors $\mathbf{x}$ and $\mathbf{y}$. Then

$$\begin{aligned}H_{2^n}|\psi\rangle &= \frac{1}{\sqrt{2^n|C|}}\sum_{\mathbf{x},\mathbf{y}\in GF(2)^n}(-1)^{\mathbf{x}\cdot\mathbf{y}}|\mathbf{x}\rangle\langle\mathbf{y}|\sum_{\mathbf{c}\in C}(-1)^{\mathbf{a}\cdot\mathbf{c}}|\mathbf{c}+\mathbf{b}\rangle \\ &= \frac{1}{\sqrt{2^n|C|}}\sum_{\mathbf{x},\mathbf{y}\in GF(2)^n}\sum_{\mathbf{c}\in C}(-1)^{\mathbf{x}\cdot\mathbf{y}+\mathbf{a}\cdot\mathbf{c}}|\mathbf{x}\rangle\langle\mathbf{y}|\mathbf{c}+\mathbf{b}\rangle \\ &= \frac{1}{\sqrt{2^n|C|}}\sum_{\mathbf{x}\in GF(2)^n}\sum_{\mathbf{c}\in C}(-1)^{\mathbf{x}\cdot(\mathbf{c}+\mathbf{b})+\mathbf{a}\cdot\mathbf{c}}|\mathbf{x}\rangle \\ &= \frac{1}{\sqrt{2^n|C|}}\sum_{\mathbf{x}\in GF(2)^n}(-1)^{\mathbf{b}\cdot\mathbf{x}}|\mathbf{x}\rangle\sum_{\mathbf{c}\in C}(-1)^{(\mathbf{x}+\mathbf{a})\cdot\mathbf{c}} \\ &\overset{(*)}{=} \frac{|C|}{\sqrt{2^n|C|}}\sum_{\mathbf{x}\in C^\perp+\mathbf{a}}(-1)^{\mathbf{b}\cdot\mathbf{x}}|\mathbf{x}\rangle \\ &= \frac{(-1)^{\mathbf{a}\cdot\mathbf{b}}}{\sqrt{|C^\perp|}}\sum_{\mathbf{d}\in C^\perp}(-1)^{\mathbf{b}\cdot\mathbf{d}}|\mathbf{d}+\mathbf{a}\rangle\,.\end{aligned}$$

In $(*)$ we have used that the sum $\sum_{\mathbf{c}\in C}(-1)^{\mathbf{x}\cdot\mathbf{c}}$ vanishes if and only if $\mathbf{x} \neq C^\perp$. □

Lemma 1 shows that the Hadamard transformation not only changes phase-flip errors into bit-flip errors, it also maps superpositions of all codewords of the linear binary code C to superpositions of all codewords of the dual code $C^\perp$ (cf. Proposition 1).

Example (seven-qubit code) The 3rd binary Hamming code $C = [7, 4, 3]$ (cf. Proposition 2) contains its dual code $C^\perp = [7, 3, 4]$, that is, $C^\perp \subset C$. Hence we can partition the codewords of C into two cosets of $C^\perp$ as follows:

$$C = \left(C^\perp + \mathbf{x}_0\right) \dot\cup \left(C^\perp + \mathbf{x}_1\right) ,$$
$$\text{where } \mathbf{x}_0 = (0000000) \quad \text{and } \mathbf{x}_1 = (1111111) .$$

The Hamming weight of all codewords of $C^\perp$ is even, while the weight of all vectors in the coset $C^\perp + \mathbf{x}_1$ is odd. Based on this decomposition, we define the following encoding:

$$|0\rangle \mapsto |0\rangle_L = \frac{1}{\sqrt{|C^\perp|}} \sum_{\mathbf{c}\in C^\perp} |\mathbf{c}+\mathbf{x}_0\rangle = \frac{1}{\sqrt{|C^\perp|}} \sum_{\mathbf{c}\in C^\perp} |\mathbf{c}\rangle \tag{10a}$$

$$|1\rangle \mapsto |1\rangle_L = \frac{1}{\sqrt{|C^\perp|}} \sum_{\mathbf{c}\in C^\perp} |\mathbf{c}+\mathbf{x}_1\rangle . \tag{10b}$$

Hadamard transformation of these states yields

$$H_{2^7}|0\rangle_L = \frac{1}{\sqrt{|C|}} \sum_{\mathbf{c}\in C} (-1)^{\mathbf{c}\cdot\mathbf{x}_0}|\mathbf{c}\rangle = \frac{1}{\sqrt{|C|}} \sum_{\mathbf{c}\in C} |\mathbf{c}\rangle$$
$$= \frac{1}{\sqrt{2}} \left(|0\rangle_L + |1\rangle_L\right) \tag{11a}$$

$$H_{2^7}|1\rangle_L = \frac{1}{\sqrt{|C|}} \sum_{\mathbf{c}\in C} (-1)^{\mathbf{c}\cdot\mathbf{x}_1}|\mathbf{c}\rangle = \frac{1}{\sqrt{2}} \left(|0\rangle_L - |1\rangle_L\right) . \tag{11b}$$

A superposition $|\psi\rangle = \alpha|0\rangle_L + \beta|1\rangle_L$ of the logical qubits is a superposition of words of the Hamming code $C = [7, 4, 3]$. Similar to the transformation (2) it is possible to compute an error syndrome for the bit-flip errors using a parity check matrix of the Hamming code. Measuring the error syndrome provides information about the position of a single bit-flip, allowing us to correct this error. From (11) it can be seen that the Hadamard transformation of the state $|\psi\rangle$ is again a superposition of words of the Hamming code, so a single phase-flip error can be corrected as well. Similar to Shor's nine-qubit code, for this seven-qubit code $C = [\![7, 1, 3]\!]$ given by (10), bit-flips and phase-flips can be corrected independently (for more details see [31]).

Equation (11) additionally shows that applying the Hadamard transformation to all seven qubits corresponds to a Hadamard transformation of the logical qubits $|0\rangle_L$ and $|1\rangle_L$. Similarly, applying σ_x or σ_z to all qubits transforms the encoded qubits like encoded versions of σ_x and σ_z, respectively (see also Sect. "Transversal Gates").

The generalization of this construction principle yields so-called CSS codes. This class of codes was independently derived by Calderbank and Shor [15] and Steane [66,67].

Theorem 6 (CSS code) *Let $C_1 = [n, k_1, d_1]$ and $C_2 = [n, k_2, d_2]$ be linear binary codes of length n, dimension k_1 and k_2, respectively, and minimum distance d_1 and d_2, respectively, with $C_2^\perp \subseteq C_1$. Furthermore, let $\mathcal{W} = \{\mathbf{w}_1, \ldots, \mathbf{w}_K\} \subset GF(2)^n$ be a system of representatives of the cosets of $C_2^\perp$ in C_1.*

The $K = 2^{k_1-(n-k_2)}$ mutually orthogonal states

$$|\psi_i\rangle = \frac{1}{\sqrt{|C_2^\perp|}} \sum_{\mathbf{c}\in C_2^\perp} |\mathbf{c}+\mathbf{w}_i\rangle \tag{12}$$

span a quantum error-correcting code $C = [\![n, k, d]\!]$ with $k := k_1 + k_2 - n$. The code corrects at least $\lfloor (d_1 - 1)/2 \rfloor$ bit-flip errors and simultaneously at least $\lfloor (d_2 - 1)/2 \rfloor$ phase-flip errors. Its minimum distance is $d \geq \min\{d_1, d_2\}$.

Proof An arbitrary combination of bit-flip and phase-flip errors can be written as

$$\mathbf{e} := \left(\sigma_x^{e_{x,1}}\sigma_z^{e_{z,1}}\right) \otimes \ldots \otimes \left(\sigma_x^{e_{x,n}}\sigma_z^{e_{z,n}}\right) , \tag{13}$$

where the binary vectors $\mathbf{e}_x$ and $\mathbf{e}_z$ indicate the positions with bit-flip and sign-flip errors, respectively. An arbitrary state of the CSS code is a superposition of the encoded basis states $|\psi_i\rangle$ in (12). Rewriting the superposition shows that the state is a superposition of codewords of the binary code C_1:

$$|\psi\rangle = \sum_{i=1}^{K} \alpha_i|\psi_i\rangle = \sum_{i=1}^{K} \alpha_i' \sum_{\mathbf{c}\in C_2^\perp} |\mathbf{c}+\mathbf{w}_i\rangle = \sum_{\mathbf{c}\in C_1} \beta_{\mathbf{c}}|\mathbf{c}\rangle . \tag{14}$$

Combining (13) and (14) the erroneous state reads

$$\mathbf{e}|\psi\rangle = \sum_{\mathbf{c}\in C_1} \beta_{\mathbf{c}}(-1)^{\mathbf{c}\cdot\mathbf{e}_z}|\mathbf{c}+\mathbf{e}_x\rangle . \tag{15}$$

Using a parity check matrix H_1 for the linear binary code C_1 we can implement the mapping

$$S_1 : |\mathbf{x}\rangle|\mathbf{y}\rangle \mapsto |\mathbf{x}\rangle|\mathbf{x}H_1^t + \mathbf{y}\rangle .$$

Applying S_1 to the state (15) and $n-k_1$ ancilla qubits in the state $|0\rangle$, we get

$$\sum_{\mathbf{c}\in C_1} \beta_{\mathbf{c}}(-1)^{\mathbf{c}\cdot\mathbf{e}_z}|\mathbf{c}+\mathbf{e}_x\rangle|(\mathbf{c}+\mathbf{e}_x)H_1^t\rangle$$
$$= \Big(\sum_{\mathbf{c}\in C_1} \beta_{\mathbf{c}}(-1)^{\mathbf{c}\cdot\mathbf{e}_z}|\mathbf{c}+\mathbf{e}_x\rangle\Big)\otimes|\mathbf{e}_xH_1^t\rangle$$
$$= \big(\mathbf{e}|\psi\rangle\big)\otimes|\mathbf{e}_xH_1^t\rangle\,. \quad (16)$$

As the syndrome $\mathbf{s}_x = \mathbf{e}_xH_1^t$ depend only on the error $\mathbf{e}_x$, not on the codeword $\mathbf{c}$, the state in (16) is a tensor product. Hence we can measure the syndrome of the bit-flip errors without disturbing the state $\mathbf{e}|\psi\rangle$. A classical algorithm for the decoding of the code C_1 can then be used to deduce the error vector $\mathbf{e}_x$ from the syndrome $\mathbf{s}_x$. In order to correct phase-flip errors, we recall that the Hadamard transformation interchanges σ_x and σ_z. Applying the Hadamard transformation to the erroneous state (15) we get

$$H_{2^n}\mathbf{e}|\psi\rangle = (H_{2^n}\mathbf{e}H_{2^n})H_{2^n}\left(\sum_{i=1}^{K}\alpha_i'\sum_{\mathbf{c}\in C_2^\perp}|\mathbf{c}+\mathbf{w}_i\rangle\right)$$
$$= \sum_{i=1}^{K}\alpha_i''\sum_{\mathbf{c}\in C_2}(-1)^{\mathbf{c}\cdot\mathbf{e}_x}(-1)^{\mathbf{c}\cdot\mathbf{w}_i}|\mathbf{c}+\mathbf{e}_z\rangle\,.$$

The last equation follows using Lemma 1. Similar to (16) we can use a parity check matrix H_2 of the binary code C_2 to define the mapping $S_2\colon |\mathbf{x}\rangle|\mathbf{y}\rangle \mapsto |\mathbf{x}\rangle|\mathbf{x}H_2^t+\mathbf{y}\rangle$. Using $n-k_2$ additional ancilla qubits, we obtain the state

$$\big(H_{2^n}\mathbf{e}|\psi\rangle\big)\otimes|\mathbf{e}_xH_1^t\rangle\otimes|\mathbf{e}_zH_2^t\rangle\,.$$

Undoing the Hadamard transformation on the first n qubits we finally get

$$\big(\mathbf{e}|\psi\rangle\big)\otimes|\mathbf{e}_xH_1^t\rangle\otimes|\mathbf{e}_zH_2^t\rangle$$
$$= \Big(\sum_{\mathbf{c}\in C_1}\beta_{\mathbf{c}}(-1)^{\mathbf{c}\cdot\mathbf{e}_z}|\mathbf{c}+\mathbf{e}_x\rangle\Big)\otimes|\mathbf{e}_xH_1^t\rangle\otimes|\mathbf{e}_zH_2^t\rangle\,.$$

Again we can measure the syndrome $\mathbf{s}_z = \mathbf{e}_zH_2^t$ and use a decoding algorithm for the linear binary code C_2 to find the error vector $\mathbf{e}_z$. □

Stabilizer Codes

The construction of CSS codes uses two binary codes which define a decomposition of the set of all binary strings of length n. Using the binary strings as labels of quantum states, one obtains a decomposition of the complex vector space $(\mathbb{C}^2)^{\otimes n}$ (cf. Fig. 1). Such a decomposition can also be defined via eigenspaces of operators.

Let $\mathcal{P}_n$ denote the group which is generated by tensor products of n Pauli matrices (cf. (9)) and identity. Every element $g\in\mathcal{P}_n$ has a unique representation of the form

$$g = i^c\cdot\sigma_x^{g_{x,1}}\sigma_z^{g_{z,1}}\otimes\cdots\otimes\sigma_x^{g_{x,n}}\sigma_z^{g_{z,n}}\,,$$

where $c\in\{0,1,2,3\}$ and $\mathbf{g}_x$ and $\mathbf{g}_z$ are binary vectors of length n. Two elements g and h of $\mathcal{P}_n$ either commute or anti-commute, that is, $gh=\pm hg$.

Let S be an Abelian subgroup of $\mathcal{P}_n$, that is, any two elements of S commute. Furthermore, we assume that S does not contain $-I$. Then S has 2^{n-k} elements and there are $n-k$ generators. The spectrum of every generator g_i is $\{+1,-1\}$. Furthermore, there are 2^{n-k} common eigenspaces, each of dimension 2^k, which can be labeled by the eigenvalues of the generators. We choose one of these eigenspaces as our quantum code C, e. g., the eigenspace with all eigenvalues $+1$. For the analysis of the error-correcting properties of this code, recall that any two elements in $\mathcal{P}_n$ either commute or anti-commute. Assume that $E\in\mathcal{P}_n$ anti-commutes with some generator g_i of S. Then for every state $|\psi\rangle\in C$ we have

$$g_i(E|\psi\rangle) = -Eg_i|\psi\rangle = -E|\psi\rangle\,, \quad (17)$$

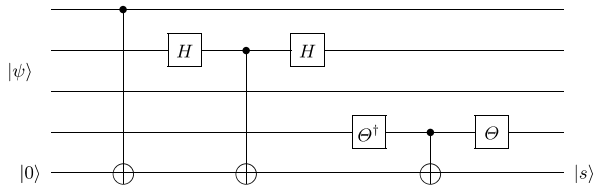
that is, $E|\psi\rangle$ lies in the eigenspace of g_i with eigenvalue -1 which is orthogonal to the code. Hence every error E that anti-commutes with at least one of the generators g_i of S can be detected. From the definition of S it follows that $E|\psi\rangle = |\psi\rangle$ for all $E\in S$ and every state $|\psi\rangle\in C$, that is, these "errors" E have no effect on the code. However, if $E\notin S$ commutes with all elements in S, then E preserves the code space, but acts non-trivially on it. When designing a code, we want the weight of these undetectable errors to be large, that is, we want the number of tensor factors of E that are different from identity to be high.

This construction of so-called *stabilizer codes* has been found independently by Gottesman [25] and Calderbank et al. [14]. In the latter paper, a connection to additive codes over the finite field $GF(4)$ has been established which allows us to use results from classical coding theory for the construction of good quantum codes.

Theorem 7 (stabilizer codes) *Let S be an Abelian subgroup of the n-qubit Pauli group $\mathcal{P}_n$ that does not contain $-I$. The stabilizer code C with stabilizer S is the common eigenspace with eigenvalue + 1 of all operators in S. The dimension of C is 2^k if $|S| = 2^{n-k}$. The normalizer $\mathcal{N}(S)$ is defined as*

$$\mathcal{N}(S) = \{x\colon x\in\mathcal{P}_n \mid x^{-1}Sx = S\}$$
$$= \{x\colon x\in\mathcal{P}_n \mid xg = gx \ \textit{for all}\ g\in S\}\,.$$

The normalizer has 2^{n+k} elements and contains the stabilizer S. If the weight of all elements of the set $\mathcal{N}(S)\setminus S$ is at



Quantum Error Correction and Fault Tolerant Quantum Computing, Figure 4
Quantum circuit to measure the eigenvalue of $g = \sigma_z \otimes \sigma_x \otimes I \otimes \sigma_y$

least d, then all errors of weight strictly less than d can either be detected or have no effect on the code. Equivalently, one can correct all errors of weight up to $\lfloor (d-1)/2 \rfloor$. The code is denoted by $C = [\![n, k, d]\!]$.

In a manner similar to error correction for CSS codes, error correction for stabilizer codes is based on measuring a syndrome of the error. Recall that an error can be detected if it anti-commutes with one of the generators g_i of the stabilizer S. Then the erroneous state lies in the eigenspace of g_i with eigenvalue -1 (see (17)). Hence measuring the eigenvalues of the generators g_i provides information about the error E. Since the Pauli matrices are both unitary and Hermitian, the generators define a quantum mechanical observable. However, if the weight of the generator g_i is m, one would have to measure an m-qubit observable. Alternatively, one can use one ancilla qubit per generator g_i to compute an error syndrome. Assume that we want to measure the eigenvalue of $g = \sigma_z \otimes \sigma_x \otimes I \otimes \sigma_y$. For this, we can use the quantum circuit shown in Fig. 4.

The eigenstates of σ_z are $|0\rangle$ and $|1\rangle$ with eigenvalues $+1$ and -1, respectively. Hence the first CNOT gate flips the state of the ancilla qubit if and only if the first qubit is in the eigenspace of σ_z with eigenvalue -1. The control qubit of the next CNOT gate is conjugated with the Hadamard matrix which interchanges σ_x and σ_z. Hence the ancilla qubit is flipped if and only if the second qubit is in the eigenspace of σ_x with eigenvalue -1. In order to measure the eigenvalue of σ_y, the control of the final CNOT gate is conjugated with the matrix $\Theta = 1/\sqrt{2}\begin{pmatrix} 1 & -\mathbf{i} \\ -\mathbf{i} & 1 \end{pmatrix}$ that maps σ_y to σ_z. If measuring the ancilla qubit $|s\rangle$ yields the result 0 or 1, the state of the first four qubits is projected onto the eigenspace of g with eigenvalue $+1$ or -1, respectively.

On the one hand, the elements of the normalizer $\mathcal{N}(S)$ which do not lie in the stabilizer S correspond to errors that cannot be detected. On the other hand, these elements can be used to perform operations on the code, as we will see in the following example.

Example (five-qubit code) The shortest quantum code that can correct an arbitrary single-qubit error is the code $C = [\![5, 1, 3]\!]$ which is a stabilizer code. The stabilizer S of this code is generated by

$$g_1 = \sigma_z \otimes \sigma_x \otimes \sigma_x \otimes \sigma_z \otimes I = ZXXZI\,,$$
$$g_2 = I \otimes \sigma_z \otimes \sigma_x \otimes \sigma_x \otimes \sigma_z = IZXXZ\,,$$
$$g_3 = \sigma_z \otimes I \otimes \sigma_z \otimes \sigma_x \otimes \sigma_x = ZIZXX\,,$$
$$g_4 = \sigma_x \otimes \sigma_z \otimes I \otimes \sigma_z \otimes \sigma_x = XZIZX\,.$$

The normalizer of S is generated by $g_1, \ldots, g_4$ together with

$$h_1 = \sigma_x \otimes \sigma_z \otimes \sigma_x \otimes I \otimes I = XZXII\,,$$
$$h_2 = \sigma_z \otimes \sigma_z \otimes I \otimes \sigma_x \otimes I = ZZIXI\,.$$

Both h_1 and h_2 commute with all g_i, and h_1 and h_2 anti-commute with each other. Hence h_1 and h_2 act on the code as encoded versions of σ_x and σ_z.

We close this section by noting that CSS codes are also stabilizer codes. The stabilizer of the seven-qubit code of Example 4 is generated by

$$g_1 = XIXIXIX\,, \qquad g_4 = ZIZIZIZ\,,$$
$$g_2 = IXXIIXX\,, \qquad g_5 = IZZIIZZ\,,$$
$$g_3 = IIIXXXX\,, \qquad g_6 = IIIZZZZ\,.$$

The additional generators of the normalizer are given by

$$h_1 = XXXXXXX\,, \qquad h_2 = ZZZZZZZ\,.$$

As already mentioned at the end of Example 4, h_1 and h_2 correspond to the encoded version of σ_x and σ_z, respectively.

Techniques for Fault-Tolerant Quantum Computing

The theory of quantum codes as described in the previous sections is a prerequisite to be able to achieve a larger goal, namely, reliable quantum computations performed with imperfect hardware. Imperfections arise from the fact that every physical device is subject to noise. This includes memory elements as well as gate elements. Not only is memory affected by errors, but the operations by which this memory is modified can also be faulty, so the task of devising reliable quantum computations is very delicate. The idea behind fault-tolerant quantum computing (FTQC) is to achieve this very goal, provided that the noise level introduced by the memory and the gates is not too high.

It should be noted that FTQC is quite different from the techniques used in classical fault-tolerance and dependability of systems (see [7] for an overview). In general, the supply of techniques that can be drawn from in the

quantum case is less rich than in the classical case, where fault-tolerance techniques are known on basically all levels of abstraction, for hardware as well as for software. The reason for this difference is that in the quantum case there are strong restrictions on how quantum information can be maintained. One example might be the "no cloning theorem" that was already mention as Theorem 4. Not having the ability to copy quantum information rules out many classical fault-tolerance techniques such as error-detection for rollback recovery [23].

The results shown in FTQC have the flavor of some of the results shown in the early days of gate level fault-tolerance in the classical case, such as the celebrated threshold result for faulty NAND gates due to J. von Neumann [55]. An important result of the theory of FTQC is that also in the quantum case there is a threshold value for this noise level such that arbitrarily long quantum computations become possible if the gates have a noise level that is under the threshold. We will briefly discuss the current state of the art of estimates of this threshold value from below and above in Sect. "The Threshold Theorem".

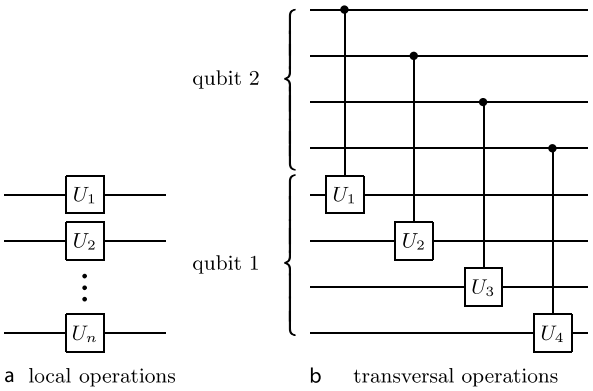
FTQC comprises methods to perform the following tasks on encoded quantum states in such a way that errors do not accumulate. This means that during the application of a single gate, errors do not spread out in an uncontrolled fashion over the whole circuit. Instead, errors stay locally confined so that they can be taken care of by error correction routines that are applied periodically to the system:

- Preparation of initial states and ancilla states
- Algorithms for quantum error-correction
- Measurement of quantum states
- Universal set of quantum gates.

Throughout, in FTQC it is assumed that quantum data is never given in a form where it is unencoded. Instead, it is assumed that whenever a qubit consisting of two (logical) basis states $|0\rangle_L$, $|1\rangle_L$ is given, that these states are encoded into some higher-dimensional Hilbert space and that the need never arises to decode this quantum information from its encoded physical representation to its logical representation.

Transversal Gates

FTQC requires careful design of how quantum information is encoded. A first indication of the inherent difficulties is given by the observation that errors have the tendency to be propagated through quantum gates. The CNOT gate might serve as a simple example. A simple calculation reveals that it propagates (even when working perfectly) a single X-error on the control qubit to a double X-error on both control and target qubit. Conversely, the same CNOT gate propagates a single Z-error on the target qubit to a double Z-error on control and target qubit. A basic requirement for all quantum gates that are used to operate on encoded quantum data is that they should not propagate errors, that is, errors should stay locally confined to a small set of qubits (see Fig. 5).



Quantum Error Correction and Fault Tolerant Quantum Computing, Figure 5

Fault-tolerant quantum gates. Shown are two examples, namely *a* local operations that operate on encoded qubits without entangling them, thereby avoiding any error propagation between them, and *b* transversal gates which in this case are given by a sequence of controlled gates. An error in one of these gates will have only a local effect to at most two of the encoded qubits and therefore can be corrected by one round of quantum error-correction provided the qubits are encoded using a quantum code that can correct at least one error. In both *a* and *b* the gates $U_1, U_2, \ldots$, are arbitrary unitaries

It was shown in [27] how to perform encoded Clifford operations transversally on any stabilizer code. Clifford operations are those quantum operations that leave the group of tensor products of Pauli matrices unchanged under conjugation. Clifford operations are not universal for quantum operations. For certain inputs, they can even be efficiently simulated classically [1,5]. However, it was also shown in [27] that, using transversal gates together with local measurements, a universal set of quantum gates can be implemented on any stabilizer code.

Teleporting Gates

While in the beginnings of FTQC universal gate sets were realized using ad hoc constructions involving measurements [65], later it was realized that teleportation [10] also can be used to implement universal sets of quantum gates. This has the following unique advantage: suppose that a particular gate U has to be applied during the com-

putation. Then as shown in [30], under some conditions (which characterize a set of gates more general than Clifford gates, hence a universal set of gates), such a quantum gate can be "precomputed" and stored into a quantum state $|\psi_U\rangle$. Once the gate U has to be applied, this state $|\psi_U\rangle$ is used in a generalized teleportation scheme which can be shown to have the same effect as applying U. The advantage of this method of *teleporting the gate U* is that $|\psi_U\rangle$ can be prepared and verified offline. Then only very high fidelity specimens of $|\psi_U\rangle$ are actually used in the computation.

Fault-Tolerant Error Correction

Several fault-tolerant methods are known to measure the syndrome of a quantum code. We first describe Shor's method which relies on generalized cat states and which can be applied to any stabilizer code. Next, we describe Steane's method which is applicable whenever the code is a CSS code. Finally, we describe Knill's method which is based on teleportation and post-selection.

Shor's Method In Fig. 4 we have seen an example of how the eigenvalue of a generator of the stabilizer can be measured. The principle underlying this example is to compute a controlled operation for each of the non-trivial Pauli operators contained in the generator and to compute the output of this controlled gate into a fresh qubit which is then subsequently measured. Unfortunately, this method is not fault-tolerant as the CNOT has the mentioned property of propagating X-errors from control qubits to target qubits and Z-errors from target qubits to control qubits. Therefore, a different method is needed to implement a fault-tolerant measurement of stabilizer generators. As shown by Shor [65], such a method of fault-tolerant syndrome measurements exists, provided that a supply of cat states, that is, quantum states of the form $1/\sqrt{2}(|0\ldots0\rangle + |1\ldots1\rangle)$, is available. Such a cat state is first transformed using the Hadamard transform applied to all qubits, yielding the equal superposition of all even weight binary words. Next, the same controlled operations as in Fig. 4 are applied but instead of having the same target qubit, they are controlled to individual qubits of the cat state. Afterwards, all ancilla qubits are measured in the standard basis, and the parity of the measurement results is computed classically.

The remaining problem is how to ensure the supply of cat states. This is done in a separate, offline procedure which first performs an encoding of a cat state followed by suitable tests to verify that, indeed, the correct state has been produced.

Steane's Method A different method can be applied if the given quantum code is a CSS code. In this so-called Steane method [68], first an X-error correction is performed, followed by a Z-error correction. Here, these two error correction routines are performed in a very special way. To correct X-errors, an ancilla is prepared in an encoded state $|+\rangle_L = \frac{1}{\sqrt{2}}(|0\rangle_L + |1\rangle_L)$ that corresponds to the superposition of all logical codewords. Then transversal CNOTs from the encoded codeword (control) into the ancillas (target) are applied. Afterwards, the ancilla is measured in the standard basis and, if necessary, a correction is applied. For the Z-errors, the corresponding operations in the Hadamard basis are applied.

A fault-tolerant measurement for a CSS code derived from $C_2^\perp \subseteq C_1$ is defined with respect to the orthogonal projectors onto the states $|\psi_i\rangle$ as in Eq. (6):

$$|i\rangle_L = |\psi_i\rangle = \frac{1}{\sqrt{|C_2^\perp|}} \sum_{\mathbf{c}\in C_2^\perp} |\mathbf{c} + \mathbf{w}_i\rangle \, .$$

First, we describe a way of measuring $|i\rangle_L$ that is fault-tolerant but does not implement the orthogonal projectors given by the CSS basis states (6). Measuring all qubits of $|\psi_i\rangle$ in the computational basis yields a random codeword of the classical code C_1, possibly with some error added. Using a classical error-correction strategy for C_1, this error can be corrected, resulting in a random element of a coset $C_2^\perp + \mathbf{w}_i$. Then, computing the error-syndrome with respect to the classical code $C_2^\perp$, it is possible to infer what i was. In order to obtain a measurement that also implements the orthogonal projection operators, we use ancilla qubits which can be prepared in the state $|0\rangle_L$ and tested offline [68]. Next, the quantum information is transferred from the encoded qubits to the ancilla using transversal CNOTs. Finally, the ancillas are measured in the computational basis and the information about i is computed as described above.

Knill's Method Knill [43] suggested a scheme for fault-tolerant quantum computation based on error detection and post-selection. Knill's scheme involves a "sieving" step in which ancillas are prepared in a suitable state that is later used to teleport a gate into a computation. The gates involved in this sieving phase are protected using a concatenated code that is error-detecting. If an error occurs in this phase, the preparation of the ancilla state is aborted and a new attempt is started. A rigorous threshold for Knill's scheme was given in [4] where it was shown that $\delta \geq 1.04 \times 10^{-3}$ is a lower bound on the threshold. It should be noted that Knill's numerical simulations [43] indicate a much higher value of the threshold of about

$\delta \approx 3 \times 10^{-2}$, albeit at the price of a significant (but additive) hardware overhead.

Both Knill's and Steane's schemes have in common that they are based on post-selection, a feature that also underlies the proposed implementation of a quantum computer based on linear optical elements [46].

The Threshold Theorem

In its simplest form, FTQC is achieved by recursively encoding qubits using a fixed quantum code C. This is the idea underlying several papers in which accuracy thresholds for quantum computing are proven. We briefly describe the reasoning as to why a threshold for FTQC exists and conclude with some historical remarks and pointers to further reading.

Intuition Behind the Threshold Theorem

Let C be a quantum code that encodes only one qubit into several qubits, that is, it is an $[\![n, 1, d]\!]$ which can correct up to $t = \lfloor \frac{d-1}{2} \rfloor$ errors. Encoding one qubit with respect to this code will lead to an improved error rate of p^t over the given error rate of p. Repeating this process h times results in a quantum code C^h with parameters $[\![n^h, 1, d^h]\!]$. Although the code C^h has a very poor rate, it leads to a dramatic reduction of the error probability with a relatively small price in terms of overhead: for h levels of concatenation the error probability becomes $p_*(p/p_*)^{2^h}$, where p_* denotes the threshold probability. This threshold is given by $p_* = 1/c_*$, where c_* roughly denotes the number of events that can cause a logical error under the constraining assumptions made about the physical error model.

It should be noted that, in FTQC, usually one parameter is given, namely the probability p of any gate failing during the computation. This probability p is actually a property of a given, fixed set of universal quantum gates. The quantity $1 - p$ is given as the minimum of the probability of projecting the state obtained from the faulty gate onto the correct quantum state, where the minimum is taken over all gates in the universal set and all states of the Hilbert space.

History of the Threshold Theorem

The first method for FTQC was given by Shor [65], who introduced the idea of alternating error-correction and actual quantum gates in order to do long quantum computations that can tolerate more noise than unencoded operations. He introduced and used universal fault-tolerant gates and showed how to measure the syndrome of a stabilizer code fault-tolerantly by using cat states.

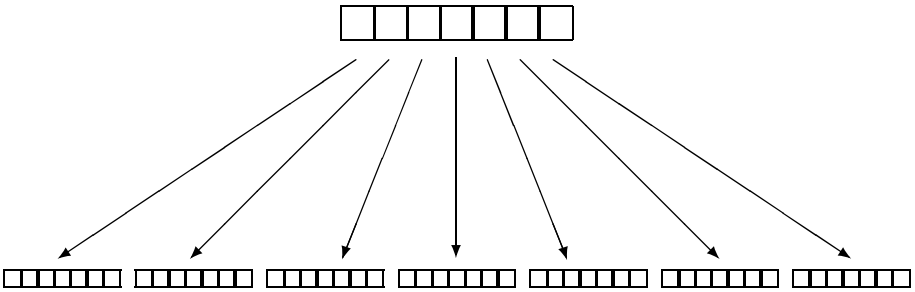
Around the same time, ideas based on concatenation of quantum codes were used by Aharonov and Ben-Or [2], Kitaev [42], Gottesman [26], Knill, Laflamme and Zurek [45], and Preskill [58], to derive FTQC schemes that can perform arbitrarily long computations provided the given error rate lies below a certain value, namely the threshold for FTQC. In these early works, the QECCs used had to be at least two-error correcting. This requirement was later relaxed and it was shown by Reichardt [62] and Aliferis, Gottesman, and Preskill [3] that also one-error correcting quantum codes yield a positive threshold. This allowed us to obtain much higher estimates of the threshold.

In the following years, several attempts have been made to obtain good bounds on the true value of the threshold. Starting with [70] there have been attempts to compute numerical approximations to the true value of the threshold via Monte-Carlo simulations. For this, a particular quantum error-correcting code and a particular error-correction strategy are fixed. Then, in a classical simulation, errors are introduced randomly and it is checked to see how many of them cause uncorrectable errors. This result can be then used to obtain an estimate of the threshold. This technique was later criticized [3] as being prone to overestimating the true threshold value. A further disadvantage is that more advanced error models such as local non-Markovian noise cannot be captured by this method.

On the other hand, in FTQC, there has been a tradition of giving rigorous, analytical proofs for bounds on the threshold by analyzing the recursion. Examples for this method include [2,45]. Typically, this gives a fairly low bound for the true value of the threshold because all error effects are assumed to be worst-case. A recently-studied intermediate model of proving bounds on the threshold makes educated guesses that certain effects are negligible, then calculates the threshold based on that. An advantage of this method is that it never underestimates the true value of the threshold while giving significantly higher values.

Similar to the DiVincenzo criteria for building a quantum computer [21], there is a list of requirements and desiderata for universal FTQC attributable to Gottesman [29] which we repeat here. The point of this list is that any quantum computer that performs its computations in a fault-tolerant fashion must have the following features:

- The gate error rates must be low.
- The architecture must support the ability to perform operations in parallel.



Quantum Error Correction and Fault Tolerant Quantum Computing, Figure 6
Concatenated quantum codes. Shown is one level of concatenation for the seven qubit $[\![7, 1, 3]\!]$ code. Each of the seven physical qubits in the first layer is replaced by seven qubits encoded again into a $[\![7, 1, 3]\!]$ quantum code. If the initial error rate is given by ε, then iterating this construction h times yields a quantum code with parameters $[\![7^h, 1, 3^h]\!]$ and a resulting error rate of roughly $c_*^{2^h-1}\varepsilon^{2^h}$

- There must be a way of remaining in, or of returning to, the computational Hilbert space, thereby preventing leakage errors.
- There must be a source of fresh initialized qubits during the computation.
- The error scaling must be benign, that is, the error rates must not increase as the computer gets larger. Also, there must not be any large-scale correlated errors during the computation.

Further Reading

For an overview of results on FTQC including noise thresholds for models based on post-selection and a detailed treatment of many FTQC constructions, we recommend Reichardt's Ph D thesis [61]. That reference also gives a discussion of alternative techniques to achieve FTQC such as topological quantum computing [20]. For an overview on developments on non-Markovian models, extended rectangles and gadgets used for fault-tolerant error correction, the reader is referred to [3]. Complementing the lower bound results on the threshold are results which give upper bounds, that is, results which determine noise levels above which no quantum computation is possible. Results in this direction were first obtained by Razborov [60] and have been further improved [40]. Finally, for a comparison between different methods for FTQC based on short block codes combined with concatenation, see the recent paper [19].

Further Aspects

In this article we have considered only qubit systems. Both quantum error-correction and fault-tolerant quantum computation can be generalized to quantum systems that are composed of higher dimensional systems (see [6,28,36]). The stabilizer formalism has also been extended from block codes to quantum convolutional codes [17,24,57]. This class of codes allows us to encode and decode a stream of quantum information without partitioning it into blocks of fixed size. For both quantum block codes and quantum convolutional codes there are efficient ways for encoding quantum information [18,33,37].

Building on the connection between stabilizer codes and classical additive codes, various constructions of QECCs have been proposed (for an overview see [41]). Among those are cyclic codes, for which efficient encoding and sub-optimal decoding algorithms are known [32], as well as families of good QECCs based on classical codes from algebraic geometry [54]. More recently, the use of classical LDPC codes for quantum error correction has been proposed [16,52]. Since classical LDPC codes achieve very good performance, it is hoped that their quantum versions result in a high threshold for FTQC. Allowing non-additive classical codes in the construction of QECCs can lead to QECCs whose dimension is larger than those of stabilizer codes [34,59].

Instead of using active methods to obtain information about an error and subsequently correct the error, some physical systems allow for passive error-protection. If the interaction of the quantum system with its environment possesses some symmetry, the quantum information can be stored in a so-called *decoherence free subspace* on which the environment has no effect [22,51,71]. A generalization of this concept is given by *decoherence free subsystems* (more details can be found in the review article [50]). Combining the ideas of active quantum error correction and encoding information into subsystems yields to *operator quantum error correction* [48,49]. Within this framework it is possible to derive simplified active error-cor-

recting schemes where the error is only partially corrected while the residual error affects only the state of a gauge subsystem [8]. These codes yield to a higher threshold for FTQC.

In a communication scenario, quantum information can be sent via teleportation. For this one needs entangled quantum states shared by the sender and receiver. It is possible to distill these entangled states from noisy entangled states [11]. More recently, the construction of QECCs which combine active error correction with pre-shared entanglement has been proposed [13].

Bibliography

1. Aaronson S, Gottesman D (2004) Improved simulation of stabilizer circuits. Phys Rev A 70(5):052328
2. Aharonov D, Ben-Or M (1997) Fault-tolerant quantum computation with constant error. In: Proceedings of 29th ACM symposium on theory of computing (STOC'97). ACM, El Paso, pp 176–188
3. Aliferis P, Gottesman D, Preskill J (2006) Quantum accuracy threshold for concatenated distance-3 codes. Quantum Inf Comput 6(2):97–165
4. Aliferis P, Gottesman D, Preskill J (2008) Accuracy threshold for postselected quantum computation. Quantum Inf Comput 8:181–244
5. Anders S, Briegel HJ (2006) Fast simulation of stabilizer circuits using a graph-state representation. Phys Rev A 73:022334
6. Ashikhmin A, Knill E (2001) Nonbinary quantum stabilizer codes. IEEE Trans Inf Theory 47(7):3065–3072
7. Avizienis A, Laprie JC, Randell B, Landwehr C (2004) Basic concepts and taxonomy of dependable and secure computing. IEEE Trans Dependable Secur Comput 1(1):11–33
8. Bacon D (2006) Operator quantum error-correcting subsystems for self-correcting quantum memories. Phys Rev A 73:012340
9. Barenco A, Berthiaume A, Deutsch D, Ekert A, Macciavello C (1997) Stabilization of quantum computations by symmetrization. SIAM J Comput 26(5):1541–1557
10. Bennett CH, Brassard G, Crepeau C, Jozsa R, Peres A, Wootters W (1993) Teleporting an unknown quantum state via dual classical and Einstein–Podolsky–Rosen channels. Phys Rev Lett 70(13):1895–1899
11. Bennett CH, DiVincenzo DP, Smolin JA, Wootters WK (1996) Mixed-state entanglement and quantum error correction. Phys Rev A 54(5):3824–3851
12. Berlekamp ER, McEliece RJ, van Tilborg HCA (1978) On the inherent intractability of certain coding problems. IEEE Trans Inf Theory 24(3):384–386
13. Brun T, Devetak I, Hsieh MH (2006) Correcting quantum errors entanglement. Science 314(5798):436–439
14. Calderbank AR, Rains EM, Shor PW, Sloane NJA (1998) Quantum error correction via codes over GF(4). IEEE Trans Inf Theory 44(4):1369–1387
15. Calderbank AR, Shor PW (1996) Good quantum error-correcting codes exist. Phys Rev A 54(2):1098–1105
16. Camara T, Ollivier H, Tillich JP (2005) Constructions and performance of classes of quantum LDPC codes. Preprint quant-ph/0502086
17. Chau HF (1998) Quantum convolutional codes. Phys Rev A 58(2):905–909
18. Cleve R, Gottesman D (1997) Efficient computations of encodings for quantum error correction. Phys Rev A 56(1):76–82
19. Cross AW, DiVincenzo DP, Terhal BM (2007) A comparative code study for quantum fault tolerance. Preprint arXiv:0711.1556 [quant-ph]
20. Dennis E, Kitaev A, Landahl A, Preskill J (2002) Topological quantum memory. J Math Phys 43:4452
21. DiVincenzo DP (2000) The physical implementation of quantum computation. Fortschr Phys 48(9–11):771–783
22. Duan LM, Guo CC (1997) Preserving coherence in quantum computation by pairing quantum bits. Phys Rev Lett 79(10):1953–1956
23. Elnozahy EN, Alvisi L, Wang YM, Johnson DB (2002) A survey of rollback-recovery protocols in message-passing systems. ACM Comput Surv 34(3):375–408
24. Forney GD Jr, Grassl M, Guha S (2007) Convolutional and tail-biting quantum error-correcting codes. IEEE Trans Inf Theory 53(3):865–880
25. Gottesman D (1996) A class of quantum error-correcting codes saturating the quantum hamming bound. Phys Rev A 54(3):1862–1868
26. Gottesman D (1997) Stabilizer codes and quantum error correction. Ph D thesis, California Institute of Technology, Pasadena
27. Gottesman D (1998) A theory of fault-tolerant quantum computation. Phys Rev A 57(1):127–137
28. Gottesman D (1999) Fault-tolerant quantum computation with higher-dimensional systems. Chaos, Solitons & Fractals 10(10):1749–1758
29. Gottesman D (2005) Requirements and desiderata for fault-tolerant quantum computing: Beyond the DiVincenzo criteria. http://www.perimeterinstitute.ca/personal/dgottesman/FTreqs.ppt
30. Gottesman D, Chuang IL (1999) Demonstrating the viability of universal quantum computation using teleportation and single-qubit operations. Nature 402:390–393
31. Grassl M (2002) Algorithmic aspects of quantum error-correcting codes. In: Brylinski RK, Chen G (eds) Mathematics of quantum computation. CRC, Boca Raton, pp 223–252
32. Grassl M, Beth T (2000) Cyclic quantum error-correcting codes and quantum shift registers. Proc Royal Soc London A 456(2003):2689–2706
33. Grassl M, Rötteler M (2006) Non-catastrophic encoders and encoder inverses for quantum convolutional codes. In: Proceedings 2006 IEEE International Symposium on Information Theory (ISIT 2006), Seattle, pp 1109–1113
34. Grassl M, Rötteler M (2008) Quantum Goethals-Preparata codes. In: Proceedings 2008 IEEE International Symposium on Information Theory (ISIT 2008), Toronto, pp 300–304
35. Grassl M, Beth T, Pellizzari T (1997) Codes for the quantum erasure channel. Phys Rev A 56(1):33–38
36. Grassl M, Beth T, Rötteler M (2004) On optimal quantum codes. Int J Quantum Inf 2(1):55–64
37. Grassl M, Rötteler M, Beth T (2003) Efficient quantum circuits for non-qubit quantum error-correcting codes. Int J Found Comput Sci 14(5):757–775
38. Hamming RW (1986) Coding and information theory. Prentice-Hall, Englewood Cliffs

39. Haroche S, Raimond JM (1996) Quantum computing: Dream or nightmare? Phys Today 49(8):51–52
40. Kempe J, Regev O, Unger F, de Wolf R (2008) Upper bounds on the noise threshold for fault-tolerant quantum computing. Preprint arXiv:0802.1464 [quant-ph]
41. Ketkar A, Klappenecker A, Kumar S, Sarvepalli PK (2006) Nonbinary stabilizer codes over finite fields. IEEE Trans Inf Theory 52(11):4892–4914
42. Kitaev A (1997) Quantum computations: Algorithms and error correction. Russ Math Surv 52(6):1191–1249
43. Knill E (2005) Quantum computing with realistically noisy devices. Nature 434:39–44
44. Knill E, Laflamme R (1997) Theory of quantum error-correcting codes. Phys Rev A 55(2):900–911
45. Knill E, Laflamme R, Zurek WH (1998) Resilient quantum computation: Error models and thresholds. Proc Royal Soc London Series A, 454:365–384. Preprint quant-ph/9702058
46. Knill E, Laflamme R, Milburn G (2001) A scheme for efficient quantum computation with linear optics. Nature 409:46–52
47. Kraus K (1983) States, effects, and operations. Lecture notes in physics, vol 190. Springer, Berlin
48. Kribs DW, Laflamme R, Poulin D (2005) Unified and generalized approach to quantum error correction. Phys Rev Lett 94(18):180501
49. Kribs DW, Laflamme R, Poulin D, Lesosky M (2006) Operator quantum error correction. Quantum Inf Comput 6(3–4): 382–399
50. Lidar DA, Whaley KB (2003) Decoherence-free subspaces and subsystems. In: Benatti F, Floreanini R (eds) Irreversible quantum dynamics, Lecture Notes in Physics, vol 622. Springer, Berlin, pp 83–120
51. Lidar DA, Chuang IL, Whaley KB (1998) Decoherence-free subspaces for quantum computation. Phys Rev Lett 81(12): 2594–2597
52. MacKay DJC, Mitchison G, McFadden PL (2004) Quantum computations: Algorithms and error correction. IEEE Trans Inf Theory 50(10):2315–2330
53. MacWilliams FJ, Sloane NJA (1977) The theory of error-correcting codes. North-Holland, Amsterdam
54. Matsumoto R (2002) Improvement of Ashikhmin–Litsyn–Tsfasman bound for quantum codes. IEEE Trans Inf Theory 48(7):2122–2124
55. von Neumann J (1956) Probabilistic logic and the synthesis of reliable organisms from unreliable components. In: Shannon CE, McCarthy J (eds) Automata studies. Princeton University Press, Princeton, pp 43–98
56. Nielsen MA, Chuang IL (2000) Quantum computation and quantum information. Cambridge University Press, Cambridge
57. Ollivier H, Tillich JP (2003) Description of a quantum convolutional code. Phys Rev Lett 91(17):177902
58. Preskill J (1998) Reliable quantum computers. Proc Royal Soc London A, 454:385–410. Preprint quant-ph/9705031
59. Rains EM, Hardin RH, Shor PW, Sloane NJA (1997) Nonadditive quantum code. Phys Rev Lett 79(5):953–954
60. Razborov A (2004) An upper bound on the threshold quantum decoherence rate. Quantum Inf Comput 4(3):222–228
61. Reichardt BW (2006) Error-detection-based quantum fault tolerance against discrete Pauli noise. Ph D thesis, University of California, Berkeley
62. Reichardt BW (2006) Fault-tolerance threshold for a distance-three quantum code. In: Proceedings of the 2006 International Colloquium on Automata, Languages and Programming (ICALP'06). Lecture Notes in Computer Science, vol 4051. Springer, Berlin, pp 50–61
63. Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27:379–423, 623–656, http://cm.bell-labs.com/cm/ms/what/shannonday/paper.html
64. Shor PW (1995) Scheme for reducing decoherence in quantum computer memory. Phys Rev A 52(4):R2493–R2496
65. Shor PW (1996) Fault-tolerant quantum computation. In: Proceedings of 35th Annual Symposium on Fundamentals of Computer Science (FOCS'96). IEEE Press, Burlington, pp 56–65. Preprint quant-ph/9605011
66. Steane AM (1996) Error correcting codes in quantum theory. Phys Rev Lett 77(5):793–797
67. Steane AM (1996) Simple quantum error correcting codes. Phys Rev A 54(6):4741–4751
68. Steane AM (1997) Active stabilization, quantum computation and quantum state synthesis. Phys Rev Lett 78(11):2252–2255
69. Wootters WK, Zurek WH (1982) A single quantum cannot be cloned. Nature 299(5886):802–803
70. Zalka C (1996) Threshold estimate for fault tolerant quantum computation. Preprint quant-ph/9612028
71. Zanardi P, Rasetti M (1997) Noiseless quantum codes. Phys Rev Lett 79(17):3306–3309

Quantum Impurity Physics in Coupled Quantum Dots

Rok Žitko[1], Janez Bonča[1,2]

[1] Jožef Stefan Institute, Ljubljana, Slovenia

[2] Faculty of Mathematics and Physics, University of Ljubljana, Ljubljana, Slovenia

Article Outline

Glossary

Quantum dot device nanoscopic electronic device resembling a transistor which incorporates a quantum dot as the central active element; sometimes also called single electron transistor. A quantum dot is an extremely small puddle of electrons which can be con-

sidered as an artificial atom since the confinement of electrons leads to quantized energy levels: the electrons form orbitals much like the electrons in orbit around an atomic nucleus. Gate-defined semiconductor quantum dots provide precisely tunable physical realizations of quantum impurity models.

Quantum impurity system system of a localized magnetic impurity in interaction with itinerant free electrons from a conduction band of an otherwise clean metal. It can be described using an idealized quantum impurity model such as Kondo or Anderson model.

Tunneling transmission of electrons from one electrode to another through classically-forbidden potential barriers such as thin insulators or empty space. Tunneling is a characteristic quantum phenomenon that is commonly at play on the nanoscopic scale.

Kondo effect Kondo effect is a many-particle effect which occurs in quantum impurity systems due to increased spin-flip scattering of the conduction band electrons on the magnetic impurity at low temperatures. It leads to various anomalies in thermodynamic and dynamic properties. In the context of the electronic transport through a quantum dot, the Kondo effect is reflected in enhanced conductance (zero-bias anomaly) at reduced temperatures.

Channel In the context of impurity physics, a channel is a set of energy levels in the conduction band which are coupled to the impurity. Several independent channels may be coupled to a single impurity. In the context of quantum dots, the relevant channels may be identified with the conduction channels in the leads attached to the nanostructure, however the number of channels in the effective impurity problem may be lower than the number of physical conduction channels.

Quantum phase transition A quantum phase transition is a zero-temperature phase transition triggered by tuning system parameters. While thermal phase transitions occur due to thermal fluctuations, quantum phase transitions emerge from zero-point quantum fluctuations in the ground state.

Particle-hole symmetry Idealized impurity models exhibit particle-hole symmetry if the model remains unchanged when all occupied levels are mapped into unoccupied levels and vice versa. This occurs for half-filled systems, when precisely one electron occupies each impurity on the average.

Definition of the Subject

Advances in the field of nanoscience and nanotechnology empower us with new tools for probing electronic systems of increasingly small sizes. Nowadays one can, for example, measure electrical conduction of semiconductor quantum dots with lateral extent of a few 10 nm, single molecules, and even individual atoms trapped between two electrodes. Nanodevices of practical interest typically consist of an active element (such as a quantum dot [1,2,3]) weakly coupled to two conducting leads by tunneling junctions so that electric current can flow through the device. The active element confines a small number of electrons. Particularly interesting is the case where this number is an odd integer; the excess single electron is then unpaired and carries magnetic moment. It has been recently demonstrated that a quantum dot of this type behaves as an artificial magnetic atom which can be experimentally tuned using electrodes [4,5,6]. The advantage of performing experiments on such artificial atoms is that various effects that depend on the number of electrons can be studied simply by changing voltages applied on gate electrodes, rather than performing experiments on different chemical elements.

It is now possible to produce nanodevices consisting of a small number of quantum dots which are coupled by tunneling junction between each other and to external electrodes [7,8,9,10,11,12]. Multiple quantum dot systems can be used to study various magnetic effects, such as antiferromagnetic and ferromagnetic ordering, Kondo screening, and other phenomena in which the role of electron-electron interactions is essential. This provides insight into the behavior of similar macroscopic magnetic systems and reveals how magnetic behavior scales from the atomic size. Furthermore, these devices are interesting in their own right as candidates for quantum information storage and processing. They represent the ultimate degree of miniaturization of electronic devices and they are likely to evolve into the building blocks of the circuitry of tomorrow.

Introduction

Real metal is never an ideally clean and homogeneous material. Instead, any metal sample invariably contains a finite concentration of various impurities. Impurities affect resistivity of the metal particularly at low temperatures when electron scattering of thermal origin is suppressed and the residual scattering on static impurities determines at which value the resistivity ultimately saturates. It was remarked very early that in some samples the resistance after the initial decrease unexpectedly increases at the lowest temperatures: this behavior constitutes the "problem of the resistance minimum" [13]. Further experimental work indicated that such anomalies are due to the pres-

ence of magnetic impurities, such as iron, cobalt or manganese, in a non-magnetic metal host. Theoretical understanding was lacking until the work of J. Kondo in 1964 who had shown that the rate of scattering events in which the magnetic moment of the impurity is changed (magnetic or spin-flip scattering) surprisingly increases as the temperature is lowered [14]; this behavior of impurity systems became known as the Kondo effect. More detailed understanding became possible with the development of advanced theoretical tools based on the idea of the renormalization group by P. W. Anderson, K. G. Wilson and others [15].

Impurity models and the Kondo effect are widely studied for several reasons. The Kondo effect is one of the very few non-trivial many-particle effects where an intensive theoretical effort eventually resulted in a very good and detailed understanding. In fact, the Kondo problem was historically the primary motivation for the development of many widely applicable theoretical approaches and has driven the progress in the field of the many-particle physics for many decades. More generally, the impurity models have attracted the condensed matter community due to their unexpectedly complex and rich behavior. On a more practical level, Kondo physics plays an important role in many complex materials which may have practical applications. The Kondo screening of local moments namely competes with magnetic ordering; the result of this competition determines the magnetic properties of materials at low temperatures. Fermi and non-Fermi liquid behaviors, ferromagnetic and antiferromagnetic correlations, and diverse behavior of heavy fermion systems [16] are the outcome of the competition between the Kondo effect and magnetic exchange interaction [8].

More recently, the Kondo problem became popular due to the advances in the field of nanoscience and nanotechnology. It is now possible to perform electron transport measurements on very small systems, such as quantum dots [17], segments of carbon nanotubes [18], single molecules with an embedded magnetic ion [19] and in the extreme case even single magnetic atoms deposited on the surface of a normal metal [20,21,22]. The Kondo effect was predicted to occur in quantum dots in late 1980s [23,24] and experimentally observed a decade later [4,5]. Again, it was found that the Kondo effect leads to transport anomalies at low temperatures. By studying the Kondo physics in systems where parameters can be continuously tuned, we better understand systems where such control is not possible, as in the case of bulk materials. Quantum dot systems are thus a laboratory for studying various effects driven by strong electron correlations.

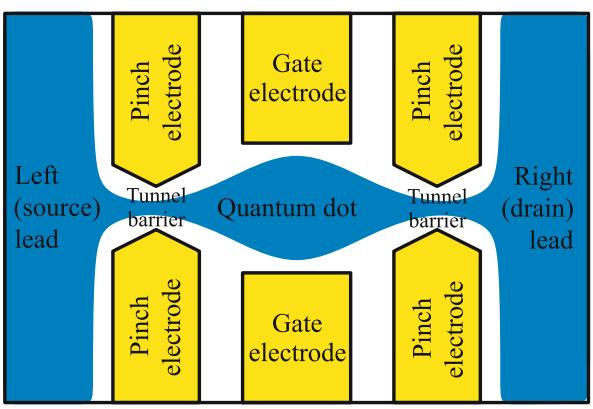
Quantum Dots as Impurity Systems

Semiconductor Quantum Dots

Particularly interesting devices are quantum dots patterned in high-quality semiconductor heterostructures. In heterostructures a subsurface layer of high-electron-mobility two-dimensional electron gas is confined near the interface between gallium arsenide (GaAs) and aluminum gallium arsenide (AlGaAs) [25]. Such crystal structures may be grown very accurately one atomic layer at a time by molecular beam epitaxy [26]. Lateral quantum dots [27] are then defined by patterning metallic gates on the sample surface (Fig. 1). Using a sufficiently negative gate voltage, the two-dimensional electron gas is depleted in the region below the electrode and a barrier is formed: quantum dot is said to be electrostatically defined. This "split-gate"' technique is also used to build quantum point contacts, quantum wires and similar devices [28]. By changing the voltage on the pinch electrodes, the strength of the coupling of the dot with the electron gas in the leads is controlled. By applying voltage on the gate electrode near the quantum dot region, the number of electrons confined in the dot can be accurately tuned. Nowadays it is possible to fabricate few-electron quantum dots and exactly control the number of electrons starting from zero.

Quantum Impurity Models

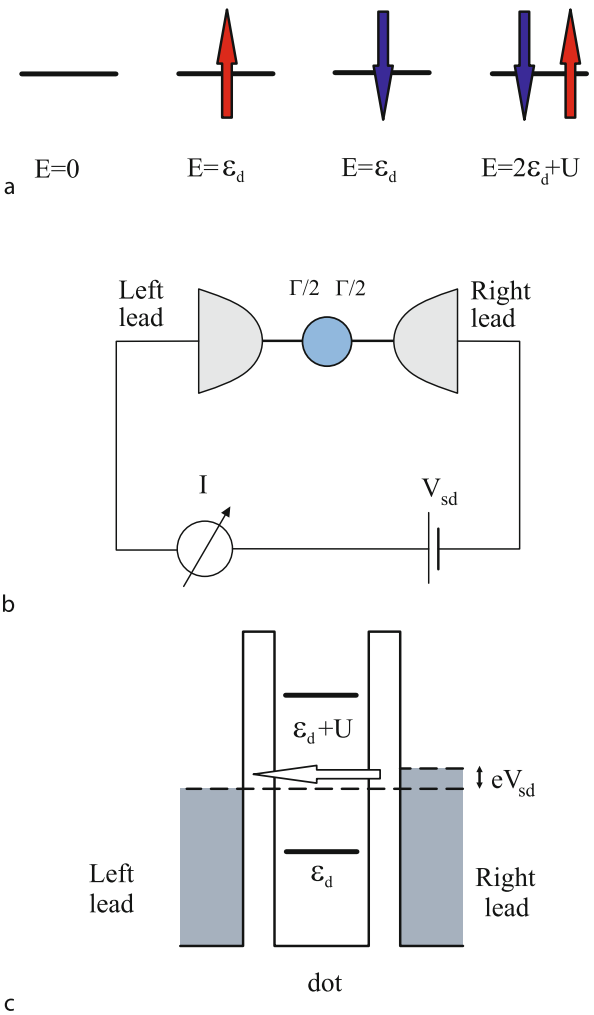
An idealized quantum impurity model describes a single point-like impurity (a zero-dimensional defect) in an otherwise homogeneous host environment composed of a gas of particles that form a continuum of extended states. The impurity is assumed to have internal degrees of free-



Quantum Impurity Physics in Coupled Quantum Dots, Figure 1
Schematic representation of a semiconductor quantum dot electrostatically defined by the voltages applied on surface metal gate electrodes

dom (such as intrinsic angular momentum, or "spin") and interacts with the continuum particles. A paradigmatic quantum impurity model is the Kondo model for a magnetic impurity atom, such as cobalt, embedded in a host metal which is non-magnetic, such as copper; the magnetic impurity interacts with the conduction band electrons via anti-ferromagnetic exchange interaction. Generalized quantum impurity models may involve several impurities or more complex, non-homogeneous environment. The theoretical significance of the quantum impurity models stems from their ubiquitous applicability to a vast array of physical systems such as bulk Kondo systems, heavy-fermion compounds and other strongly correlated systems, dissipative two-level systems, single magnetic impurities and quantum dots.

In nanoscopic electronic devices the electron-electron interactions are particularly strong and they induce interesting many-particle effects, among them the Kondo effect which appears to be a relatively generic feature of nanodevices [4,19,29,30]. As in bulk systems, the Kondo effect gives rise to various anomalies in the thermodynamic and transport properties, in particular to increased conductance through nanostructures. The conductance through a quantum dot in the Kondo regime is in agreement with theoretical predictions that such dots behave rather universally as single magnetic impurities [17] and can be modelled using single impurity Anderson and Kondo models [17,31]. Quantum dots thus serve as tunable realizations of the quantum impurity models.



Quantum Impurity Physics in Coupled Quantum Dots, Figure 2
Representations of the single impurity Anderson model for a quantum dot

Anderson Impurity Model

Due to electron confinement, the quantum-mechanical energy levels in a quantum dot form a series of discrete quantized levels. We focus on the electrons in the levels closest to the Fermi level in the leads, i. e. in the last occupied and first unoccupied orbital states of the dot. In the simplest case, a single electron level is relevant and a quantum dot with an odd number of confined electrons is expected to behave as a spin-1/2 magnetic impurity, similar to magnetic ions [32].

In the formalism of the second quantization, the Hamiltonian for interacting electrons in the quantum dot is

$$H_{\mathrm{dot}} = \epsilon_{\mathrm{d}}(n_\uparrow + n_\downarrow) + U n_\uparrow n_\downarrow \,. \tag{1}$$

ϵ_{d} is the energy of the electron orbital in the quantum dot (also named "on-site energy"), U is the strength of the effective electron-electron repulsion between two electrons in the same orbital, and the number operator n_μ is defined as $n_\mu = d_\mu^\dagger d_\mu$, where $d_\mu^\dagger$ and d_μ are the creation and annihilation operators; the spin index μ takes values $\mu = \pm 1/2$ or, equivalently, $\mu = \uparrow, \downarrow$. The on-site energy ϵ_{d} can be regulated using gate voltages which allows the charge state (occupancy) on the dots to be tuned. We may rewrite the Hamiltonian in an equivalent but more symmetric manner as

$$H_{\mathrm{dot}} = \delta n + \frac{U}{2}(n-1)^2 \,, \tag{2}$$

where $n = n_\uparrow + n_\downarrow$ and $\delta = \epsilon_{\mathrm{d}} + U/2$. For $\delta = 0$, the model is particle-hole symmetric and the level is occupied by a single electron on the average. The four possible configurations of the Anderson model and their energies are represented in Fig. 2a.

The conduction bands in the leads are described as

$$H_{\text{band}} = \sum_{k\mu\alpha} \epsilon_k c^\dagger_{k\mu\alpha} c_{k\mu\alpha} \,. \tag{3}$$

ϵ_k is the energy of an electron with wave-vector k in left ($\alpha = L$) or right ($\alpha = R$) lead described by the creation/annihilation operator pair $c^\dagger_{k\mu\alpha}$ and $c_{k\mu\alpha}$. A conduction band behaves as a sea of electrons: all states below some energy (Fermi level) are occupied, while all other high-energy states are empty. When a source-drain bias voltage V_{sd} is applied on the leads (Fig. 2b), the Fermi levels are displaced and the electrons in an energy interval of width eV_{sd} will attempt to flow from the lead with higher Fermi level through the quantum dot to the other lead, Fig. 2c. Tunneling of electrons through the junctions is described by the Hamiltonian

$$H_{\text{coupling}} = \sum_{k\mu\alpha} V_{k\alpha} (c^\dagger_{k\mu\alpha} d_\mu + d^\dagger_\mu c_{k\mu\alpha}) \,, \tag{4}$$

where $V_{k\alpha}$ are the amplitudes for electron tunneling from lead α to the dot. The Anderson impurity model is then given by the sum $H = H_{\text{dot}} + H_{\text{band}} + H_{\text{coupling}}$.

Assuming that the coupling to the left and right electrode is equal, only symmetric combinations $c^\dagger_{k\mu\text{L}} + c^\dagger_{k\mu\text{R}}$ of conduction band electrons play a role at small bias voltage V_{sd}, while antisymmetric combinations $c^\dagger_{k\mu\text{L}} - c^\dagger_{k\mu\text{R}}$ are decoupled [23]. The use of a single channel Anderson model is then justified and the index α is unnecessary. This simplification occurs only for simple systems; in general, systems of coupled quantum dots are true multichannel quantum impurity problems.

Often the approximation of taking a constant hopping $V_k \equiv V$ is taken. Further simplification consists of considering the conduction band to have a constant density of states $\rho = 1/(2D)$, where $2D$ is the bandwidth. The hybridization strength Γ which characterizes how strongly the impurity is coupled to the conduction band is then also a constant, $\Gamma = \pi\rho V^2$.

Validity of the approximation of describing the electron in the highest occupied electron level in the quantum dot using the Anderson model has been experimentally well tested: the temperature, magnetic field and gate and bias voltage dependence of the conductance through quantum dots may be described by the simple Anderson model, however the agreement is qualitative, not quantitative [17].

In some parameter regimes, Anderson model reduces to a simpler Kondo model. Kondo model consists of a single spin in interaction with the conduction band:

$$H = \sum_{k\mu} \epsilon_k c^\dagger_{k\mu} c_{k\mu} + J\boldsymbol{S}\cdot\boldsymbol{s}(0) \,, \tag{5}$$

where $J \approx 8V^2/U$ is the effective Kondo antiferromagnetic exchange constant, $\boldsymbol{S}$ is the impurity spin-1/2 operator and $\boldsymbol{s}(0)$ is the spin density of the conduction band electrons at the position of the impurity. Despite their seeming simplicity, Anderson and Kondo models are both difficult many-particle problems.

Multi-Impurity Models

Several quantum dots (artificial atoms) can be interconnected to form an "artificial molecule" [33,34]. Systems of multiple coupled impurities are realizations of generalized Kondo models where more exotic types of the Kondo effect may occur. The research in this field has recently intensified due to a multitude of new experimental results; the multi-impurity magnetic nanostructures under study are not only systems of multiple quantum dots [8,9,10,35,36], but also clusters of magnetic adsorbates on surfaces of noble metals (such as Ni dimers [37], Ce trimers [38] and molecular complexes [39]). Systems of two impurities are the simplest systems where one can study a number of very interesting effects, such as the effects of inter-dot electron hopping, formation of ionic or covalent inter-dot bonds, and the competition between magnetic ordering and Kondo screening, leading to quantum phase transitions [40,41]. Recently, few-electron triple quantum dot structures have also been fabricated [11,12] and even more complex multi-dot nanostructures can in principle also be assembled.

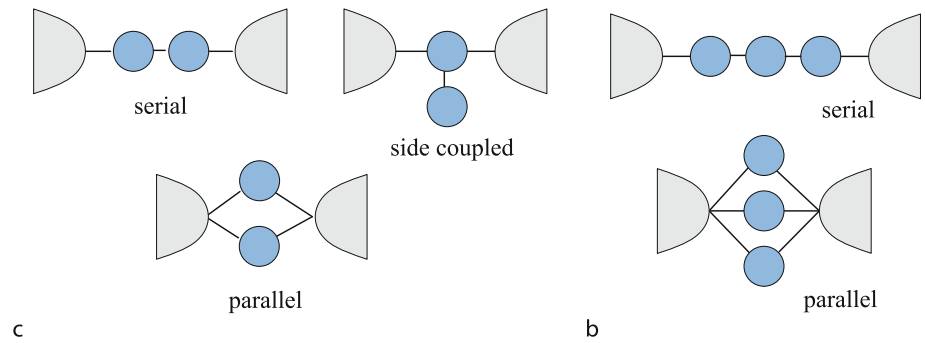
Systems of multiple quantum dots can be modeled by suitably generalizing the Anderson model. The properties depend in an essential way on the coupling topology, i. e. on how the various impurities are inter-connected, as represented in the examples in Fig. 3. The system can be modeled using discrete lattice models as an impurity cluster in contact with two conduction leads. Each dot (indexed by the subscript i) is described using a Hamiltonian

$$H_{\text{dot},i} = \delta_i n_i + \frac{U_i}{2}(n_i - 1)^2 \,. \tag{6}$$

Junctions between the dots are described by "hopping Hamiltonian"

$$H_{\text{hopping}} = \sum_{\langle i,j\rangle,\mu} t_{i,j} \left(d^\dagger_{i,\mu} d_{j,\mu} + d^\dagger_{j,\mu} d_{i,\mu} \right) \,, \tag{7}$$

and junctions between the dots and the conduction leads by suitable generalizations of Eq. (4). The impurity Hamil-



Quantum Impurity Physics in Coupled Quantum Dots, Figure 3
Representations of various multiple-impurity generalizations of the Anderson model

tonian is thus similar in form to the Hubbard model for correlated systems.

Theoretical Tools

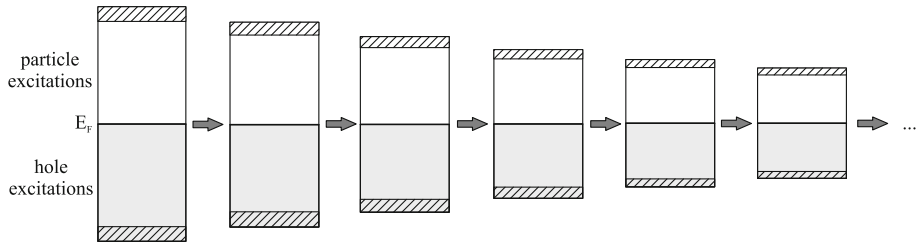
Most quantum impurity models are non-perturbative: the commonly used technique of expanding a problem in terms of a small perturbation around an exactly solvable non-interacting model cannot be applied in all parameter regimes due to divergences [13]. The difficulties occur in particular at low temperatures where the systems have anomalous properties. New techniques have been developed to tackle this problem: large-N expansion [42], Bethe-Ansatz [43,44], bosonization-refemionization [45], conformal field theory [46,47], variational methods [48,49] and various renormalization group techniques [15,50,51,52]. Large-N techniques (such as slave-boson mean-field-theory and various improvements) allow in many cases to obtain results in closed form and often provide a qualitatively correct description; in the case of multi-impurity models, however, these methods may fail or they become impractical. Bethe-Ansatz approach provides an exact solution to the thermodynamics of the Kondo model; furthermore, very recently a method to calculate non-equilibrium dynamics was developed [53]. It seems, however, difficult to expand this approach to general multi-impurity models. Bosonization-refermionization technique has been instrumental in providing additional exact results at some special points in the parameter space, however they are less useful for exploring generic problems. Conformal field theory approach based on the non-Abelian bosonization has provided important conceptual insights into the nature of the Kondo effect: the impurity degrees of freedom are engulfed by the continuum and the only residual effect are the modified boundary conditions for continuum electron scattering at the impurity site. The actual implementation of this approach depends from case to case and has not yet been performed for complex multi-impurity problems. Variational methods were the first approach that allowed to study dynamics of quantum impurity models and has been recently generalized to multi-impurity models [54,55,56]. The difficulty in this approach is to correctly describe physics at very low energy scales. Since quantum impurity models become strongly renormalized at low temperatures, the development of renormalization group methods was essential in building correct understanding of the nature of the low-temperature behavior. These methods range from simple scaling of model parameters [52], to mapping to a particle gas [50,51], and finally to Wilson's numerical renormalization group [15,57].

Renormalization

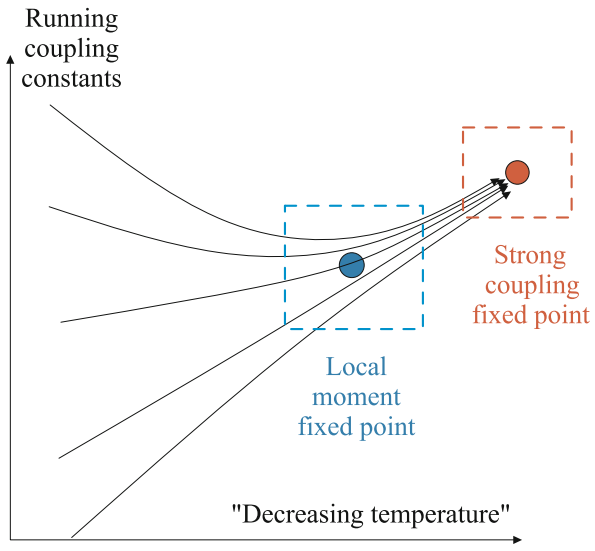
The renormalization is a way of describing and understanding the relation between the different ways a physical system behaves at different energy scales [58]. To study a system at low energies, the irrelevant high-energy degrees of freedom are eliminated from the problem ("integrated out") to obtain an effective description in terms of modified, "renormalized" coupling constants g_i which specify the strengths of various interaction terms in the Hamiltonian, Figs. 4 and 5. The renormalization process can be described using scaling equations, which typically take the form of a system of partial differential equations

$$\partial g_j/\partial l = \beta_j(\{g_i\}) \,, \tag{8}$$

where l is a "running parameter" which runs towards $-\infty$ as the energy scale is decreased and β_j are "beta" functions. Negative beta function corresponds to a relevant coupling constant which grows at low energies, i. e. to an interaction which becomes important at low temperatures. Positive beta function corresponds to an irrelevant coupling



Quantum Impurity Physics in Coupled Quantum Dots, Figure 4
Cutoff renormalization: the particle and hole excitations from the hatched regions at the top and bottom of the conduction band are integrated out to obtain an effective Hamiltonian at lower energy scale



Quantum Impurity Physics in Coupled Quantum Dots, Figure 5
Schematic representation of the renormalization flow in the Anderson model. The *horizontal direction* represents the direction of decreasing energy scale (temperature), while the *vertical direction* represents the multi-dimensional space of the effective Hamiltonians (which can be considered to be parametrized by some large set of coupling constants). When the system is near a fixed point *(dashed boxes)*, its properties can be described by a perturbative expansion around the fixed-point Hamiltonian. The diagram also illustrates the idea of universality: even for widely different original microscopic Hamiltonians, the low-temperature behavior of the systems in the same universality class is essentially the same

constant, which diminishes at low energies. If the coupling constants change only little as the renormalization procedure is performed, the system is said to be near a "fixed point". As the temperature is reduced, the system typically crosses over several times between different fixed points which correspond to particular kinds of system's behavior at different temperature scales, until it ultimately ends up in a stable fixed point which describes the essence of the low energy physics [58].

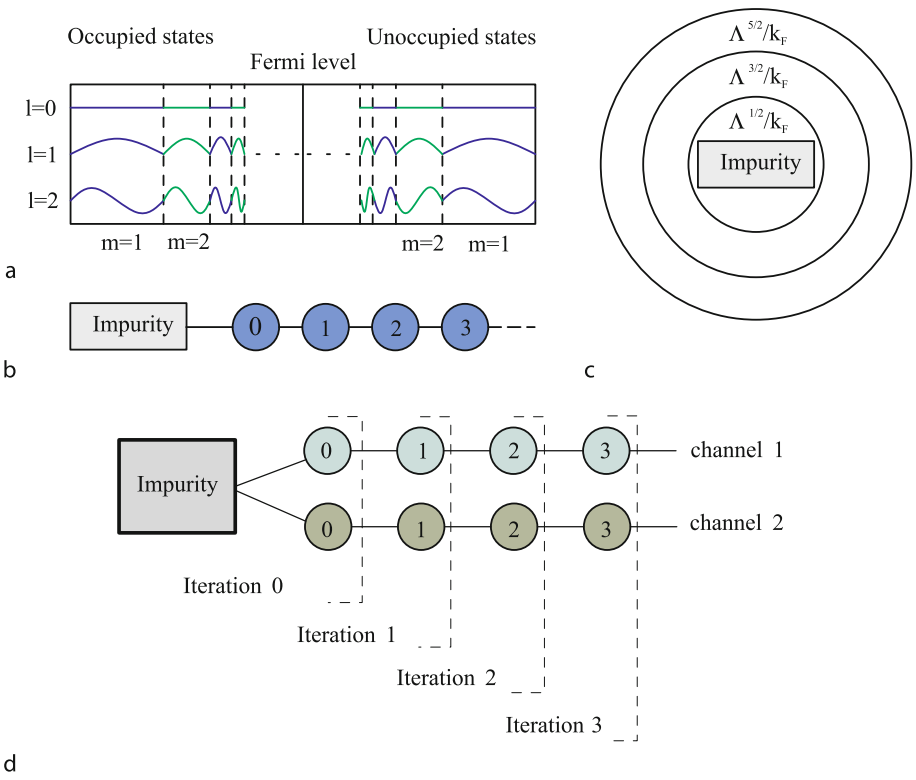
Generally a simple effective Hamiltonian arises from more complicated ones. A set of model Hamiltonians with the same low energy behavior constitute a universality class, see also Fig. 5. Renormalization is thus an essential ingredient in model building in many-particle theory.

Numerical Renormalization Group

The numerical renormalization group (NRG) was developed in 1970s by K. G. Wilson as a way of numerically exactly solving the Kondo problem [15]. It was later successfully extended to Anderson model and other quantum impurity models [59]. The NRG makes possible to compute the spectrum of excitations of the system, thermodynamic quantities such as magnetic and charge susceptibilities, entropy, and specific heat, dynamic quantities such as spectral functions, dynamical charge and spin susceptibilities, and expectation values of operators such as impurity occupancy, charge fluctuations and spin-spin correlations. The NRG is a non-perturbative method and as such does not suffer from various divergencies as other techniques do. It provides information about the behavior on all temperature scales, from the high-temperature perturbative regime to the low-temperature strong-coupling regime. It can be applied to multi-impurity and multi-channel problems; the complexity of the problem that is still manageable depends on the skillful use of the symmetries present in the problem and ultimately on the available computational resources.

The NRG consists of several steps (Fig. 6):

- Reduction of the quantum impurity problem to an effective one-dimensional problem. Since the impurity is by assumption a zero-dimensional point-like object, it always effectively couples to a continuum of states which can be parametrized by a single variable (or a finite number of such continua that we denote as "channels").
- The one-dimensional continuum of states is discretized into bins (intervals) of geometrically decreasing widths



Quantum Impurity Physics in Coupled Quantum Dots, Figure 6

Numerical renormalization group. a Logarithmic discretization. b Chain Hamiltonian (one-channel case). c Onion-shell representation of Wannier orbitals around the impurity. d Chain Hamiltonians and the successive iterations in the NRG procedure: one site from each channel is added during each renormalization group transformation

proportional to Λ^{-m}, where the parameter Λ controls the fineness of the discretization and m is the bin index, Fig. 6a. The continuum limit is recovered for $\Lambda = 1$, while in practical calculations $\Lambda \gtrsim 2$ is used. In each interval, a spectral Fourier decomposition is performed (index l in Fig. 6a). In practical calculations, only the lowest $l = 0$ Fourier mode is retained in each interval, i. e. an interval of states is represented by the energy-averaged state. This procedure is named logarithmic discretization since the continuum degrees of freedom near the Fermi level are described with a logarithmic accuracy. A further transformation allows the problem to be formulated as an impurity attached to a one-dimensional chain of sites with exponentially decreasing hopping parameters, Fig. 6b. The sites in this chain can be interpreted as forming "onion shell"-like orbitals encircling the impurity, Fig. 6c.

- Iterative diagonalization of the chain Hamiltonian is performed, Fig. 6d. The first step consists of an exact diagonalization of the initial cluster, typically composed of the impurity sites and one chain site for each continuum channel in the problem. Additional sites are then added consecutively, one from each channel in every iteration: a new Hamiltonian is constructed and diagonalized exactly. In NRG, this procedure represents the renormalization group transformation and the iteration corresponds to the renormalization flow.
- The problem of the exponential growth of the size of the Hilbert space with the number of sites in the chain is alleviated by truncating the number of states retained at each iteration to a predefined small number of the order thousand. This turns out to be a good approximation for quantum impurity problems since there is little mixing between low-energy and high-energy excitations as the chain sites are added at each step (this property is known as the energy-scale separation).

Each iteration corresponds to the behavior of the system on a temperature scale $T_N \propto \Lambda^{-N/2}$, where N is the iteration number. The full description of the system at step N consists of the eigenstates and irreducible matrix elements for creation operators $f^\dagger_{N\mu\alpha}$ in the chain. This description

is clearly much more complex compared to that in the simple renormalization approach where a small set of running coupling constants is used; the advantage is that the NRG is unbiased and, in some sense, essentially exact.

Quantum Transport and Kondo Physics

Experiments

Experiments probing fundamental properties of semiconductor quantum dots are typically performed in helium-3 dilution refrigerators at extremely low temperatures in the range of 100 mK or even less. At low temperatures, electrons occupy distinct energy levels and the Coulomb energy plays a crucial role [26]. Performing experiments at the lowest attainable temperatures is important since the energy resolution of spectroscopic techniques used is limited solely by the sample temperature [26]. Systems are characterized by performing gate-voltage and bias-voltage sweeps (gated transport spectroscopy), or by magnetic spectroscopy. This allows to obtain information about the energy levels, number of confined electrons, and electron–electron repulsion. Furthermore, finite-bias current can be approximately related to the impurity spectral function at finite frequencies.

Three elements affect the transport properties of coupled quantum dots in a characteristic manner: quantum coherence, discrete nature of the electric charge and strong electron–electron interactions.

In nanodevices made of very clean semiconductors the coherence length of electrons at low temperatures exceeds the size of the device through which the electric current flows; electrons then travel coherently through the system and behave in a wave-like manner so that quantum mechanical interference effects can occur. As the electrons scatter only off the boundaries (walls) of the device, rather than on the defects or phonons, the transport is said to be ballistic.

The conductance through nanoscopic constrictions is often found to be quantized in terms of the conductance quantum,

$$G_0 = 2e^2/h = e^2/\pi\hbar \approx 12.9\,\mathrm{k}^{\cdot -1}\,. \tag{9}$$

This is the conductance of a fully transmitting single-mode conduction channel taking into account both spin orientations and is experimentally measured in quantum point contacts and quantum wires. In lateral quantum dots the tunnel barriers from the 2DEG to the quantum dot are obtained by successively pinching off the propagating channels using the gate electrodes. When the last channel is pinched off, the Coulomb blockade regime develops. In this regime, only one channel from each lead is coupled to the dot.

Coulomb Blockade and Cotunneling

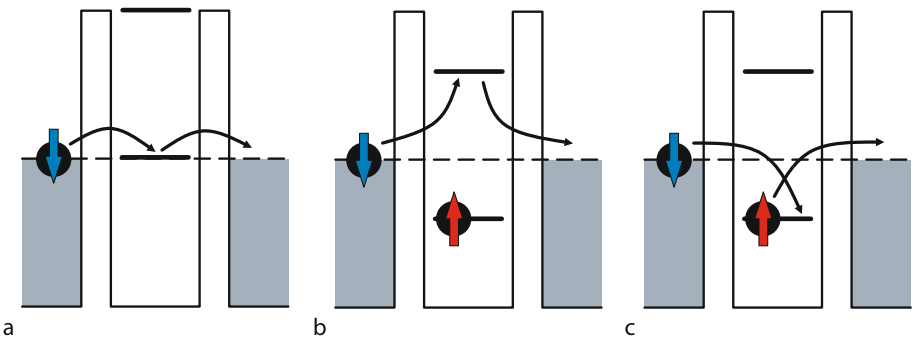
According to the analogy between a quantum dot and an atom, we expect that removing an electron from the dot (or adding it) takes energy as this is similar to the ionization of an atom. The transfer of an electron electrically charges the dot and increases the electrostatic energy by $E_C = e^2/2C$ where C is the effective capacitance between the dot and the surrounding electrodes. If the available energy is lower than the charging energy E_C (i. e. for small voltage drop across the system and for low temperature), the conductance is suppressed. This is the Coulomb blockade effect. Unless the energies of quantum dot configurations with N and $N+1$ confined electrons happen to be aligned by suitably tuning the gate voltages (Fig. 7a), the current can flow only by cotunneling (high-order processes in hybridization strength Γ) through the virtual state with excess energy $\sim E_C$ [26,60], Fig. 7b.

Cotunneling is an electric conduction process whereby an electron makes a virtual transition to a high-energy excited intermediate state in the quantum dot to travel from source to destination electrode in a single quantum step. It is to be opposed to a sequential tunneling process, where the electron makes a real transition to an energetically accessible state inside the device and the tunneling proceeds in two steps. Cotunneling is a characteristically quantum phenomenon related to the Heisenberg's uncertainty principle and becomes relevant at low temperatures. Occupation of the virtual state is allowed for a short time, $\sim h/E$, where h is the Planck constant and E the energy cost involved.

In spin-flip co-tunneling process the impurity spin is effectively flipped from spin up to spin down, or vice versa: electron with a given spin orientation tunnels in, while another electron with the opposite spin orientation tunnels out, Fig. 7c. Processes of this type are responsible for the emergence of the Kondo effect.

Conductance Formulas

Particularly important transport quantity is the conductance in the limit of zero source-drain bias voltage $G = \lim_{V_{\mathrm{sd}}\to 0} I/V_{\mathrm{sd}}$, i. e. the linear response of the system to an imposed bias. Linear conductance is an equilibrium property of the system which can be reliably calculated for impurity models. Finite-bias problems require non-equilibrium techniques which are not yet developed to a comparable degree.



Quantum Impurity Physics in Coupled Quantum Dots, Figure 7
a First-order tunneling, b cotunneling, c cotunneling with a spin-flip

For the purpose of theoretical modeling, an electronic nanodevice may be idealized and considered as a very small scatterer embedded between two metallic contacts. According to R. Landauer, the conductance of a coherent mesoscopic device is related to the transmission probability for incident electrons [61], i. e. to the scattering properties of the impurity region. At zero temperature, the conductance is simply proportional to the transmission probability [62,63]

$$G(T=0) = G_0 |S_{\mathrm{RL}}|^2 \,, \tag{10}$$

where S_{RL} is the right-left component of the scattering matrix, i. e. the amplitude for the scattering of an electron from right to left lead.

In the vast majority of the quantum impurity problems the system behaves at low temperatures as a local Fermi liquid even if it is strongly renormalized. Fermi liquid systems are described in terms of weakly interacting quasiparticles and are fully characterized by the quasiparticle scattering phase shifts which quantify how the quasiparticles scatter in the quantum dot structure [17,64,65]. In the absence of the magnetic field, a single phase shift δ_{qp} per channel is required. Matrix element S_{RL} can be expressed in terms of the phase shifts and an additional angle parameter θ which depends on the symmetry of the problem, yielding the following conductance formula [66]:

$$G = G_0 \sin^2(2\theta) \sin^2(\delta^a_{\mathrm{qp}} - \delta^b_{\mathrm{qp}}) \,. \tag{11}$$

For left-right symmetric problems, it is found that $\theta = \pi/4$ and that the relevant channels are formed by the symmetric (even) and antisymmetric (odd parity) linear combinations of the conduction electron states from left and right lead, so that

$$G = G_0 \sin^2(\delta^{\mathrm{even}}_{\mathrm{qp}} - \delta^{\mathrm{odd}}_{\mathrm{qp}}) \,. \tag{12}$$

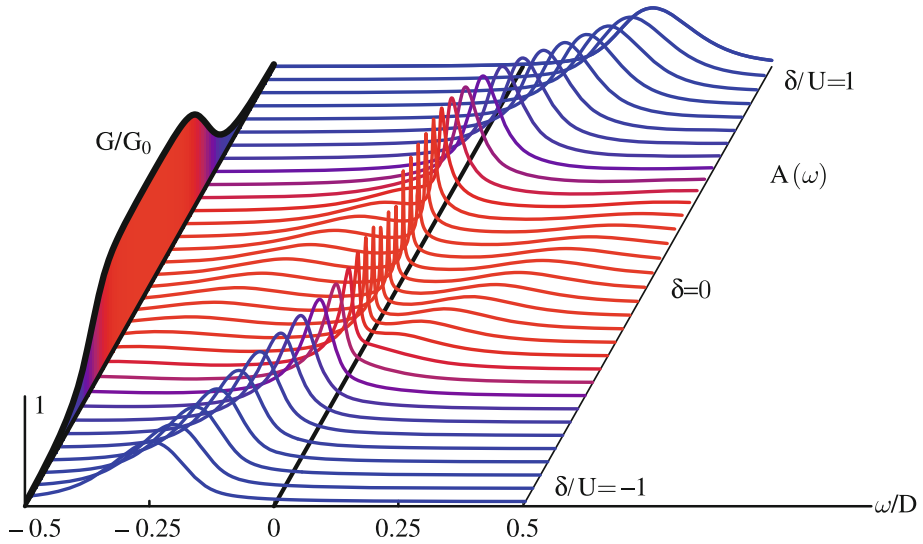
Conductance can also be computed from the impurity spectral functions using the Meir–Wingreen formula [67]:

$$G = G_0 \left(-\mathrm{Im}\mathrm{Tr}\left[\boldsymbol{\Gamma}\mathbf{G}^r\right]\right) \,, \tag{13}$$

where $\boldsymbol{\Gamma}$ is a coupling matrix and $\mathbf{G}^r$ is the matrix of retarded Green's functions of the impurity region. Green's function is essentially the Fourier transform of the probability amplitude for adding an electron to the impurity and extracting it at a later time. The imaginary part of the Green's function is the impurity spectral function. Peaks in the spectral function correspond to electronic excitations of the quantum dot and the value of the spectral function at the Fermi level is directly related to the conductance at zero temperature and can be related to the quasiparticle phase shifts in Fermi liquid systems. Meir–Wingreen formula is actually more general and it is in particular valid also for systems which are not Fermi liquids.

The Kondo Effect

The Kondo effect arises due to strongly enhanced spin-flip scattering of the conduction band electrons on the impurity at low temperatures. In conventional bulk Kondo systems this leads to increased resistivity since electrons scatter isotropically in all directions which impedes the flow of the current. As a consequence, the temperature dependence of the resistivity is non-monotonic and exhibits a minimum at small temperatures, which was the first experimentally observed manifestation of the Kondo effect [14]. Curiously, in quantum dot systems the increased scattering leads to the opposite behavior: at very low temperatures the conductance increases up to the theoretical limit of one conductance quantum, G_0 [68],. The origin of this seeming discrepancy lies in the reduced dimensionality of the problem. In quantum dot problems, scattering is effectively one-dimensional: backwards (reflection back to



Quantum Impurity Physics in Coupled Quantum Dots, Figure 8
Spectral functions $A(\omega)$ and conductance through a quantum dot described by the single-impurity Anderson model for a range of parameters δ with $U/D = 0.5$, $\Gamma/U = 0.08$. Color of each spectral function corresponds to the value of the conductance

the lead) or forwards (transmission to the other lead). The scattering increases in forward direction, which in this case corresponds to increased electric current.

The temperature scale where the scattering increases is called the Kondo temperature, T_K. The Kondo effect is not a phase transition, but rather a cross-over, therefore the change in conductance is a very smooth function of the temperature. The Kondo temperature is a non-analytic function of model parameters, $T_K \propto \exp(-1/\rho J)$, where ρ is the density of states in the conduction band at the Fermi level and J is the effective Kondo exchange constant. It may be noted that the exponential dependence of T_K reflects the non-perturbative nature of this problem.

At temperatures much below T_K, the impurity spin is screened by the conduction band electrons and the system as a whole is non-magnetic. Properties below T_K are universal and can be described using functions of argument T/T_K; a single parameter T_K fully characterizes the system instead of the microscopic ϵ_d, U, Γ, etc.

The Kondo exchange scattering processes generate an additional resonance in the impurity spectral function at the Fermi level. This "Kondo resonance" is of many-particle origin: the correlated behavior of a large number of electrons is required to produce it. Since properties at low temperatures are predominantly determined by the electron states near the Fermi level, the Kondo resonance significantly modifies the behavior of the system. In particular it leads to the predicted increase of the conductance through the quantum dot, Fig. 8.

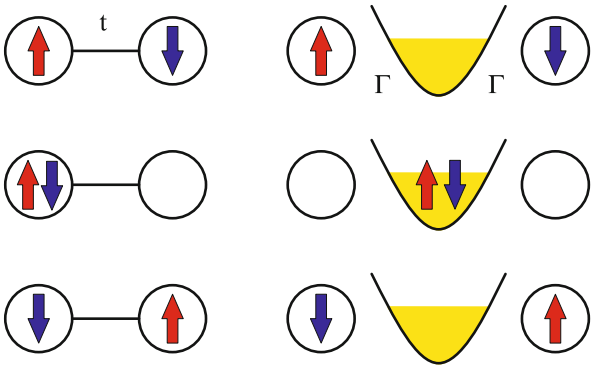
The Kondo effect is clearly a magnetic effect related to electron spin. As such, it is strongly perturbed when an external magnetic field is applied. The Kondo resonance splits, the zero-bias voltage is reduced and there are two conductance peaks at finite bias. The characteristic magnetic-field dependence of the transport properties is an ultimate proof that Kondo physics is at play in a nanostructure.

Competing Physical Effects

Physical systems tend to reduce their entropy as the temperature is lowered. In the context of systems of coupled quantum impurities with spin degrees of freedom, this is most often achieved either by Kondo screening, or by magnetic ordering of some kind [69]. Both mechanism of relieving the entropy can be in competition which leads to interesting behavior [69].

Inter-Impurity Magnetic Interactions

There are several possible origins of the inter-impurity exchange interaction in quantum impurity models. One is the super-exchange mechanism which is mediated by the direct inter-impurity electron hopping (tunneling). Virtual excursions of an electron from one impurity to another modify the energy: it is reduced if the spins are anti-aligned, so that the effective exchange interaction is antiferromagnetic. It may be represented by an interaction term $J_{\text{eff}}\boldsymbol{S}_1 \cdot \boldsymbol{S}_2$, where $\boldsymbol{S}_i$ is the impurity spin operator on



Quantum Impurity Physics in Coupled Quantum Dots, Figure 9
Processes leading to effective inter-impurity exchange interaction. a Superexchange interaction due to electron tunneling between the dots. b RKKY interaction mediated by the conduction band

dot i. For two dots decoupled from the leads, the exchange constant is given by the expression

$$J_{\text{eff}} = \frac{t}{2}\left(\sqrt{\left(\frac{U}{t}\right)^2 + 16} - \frac{U}{t}\right) \approx \frac{4t^2}{U}\,, \tag{14}$$

where t is the inter-impurity hopping parameter (tunneling amplitude).

Another important mechanism is the Ruderman–Kittel–Kasuya–Yosida (RKKY) effective exchange interaction. One impurity polarizes the conduction band electrons in its vicinity; these electrons in turn interact with the other impurity. The intensity and the sign of the resulting exchange interaction depends on the inter-impurity separation. It is ferromagnetic at very short distances, it oscillates with a period proportional to $1/k_F$ where k_F is the Fermi momentum, and decays as $(k_F R)^{-D}$ where D is the effective dimensionality of the conduction band electron gas.

Effects of the Dot-Lead and Inter-Dot Coupling Topology

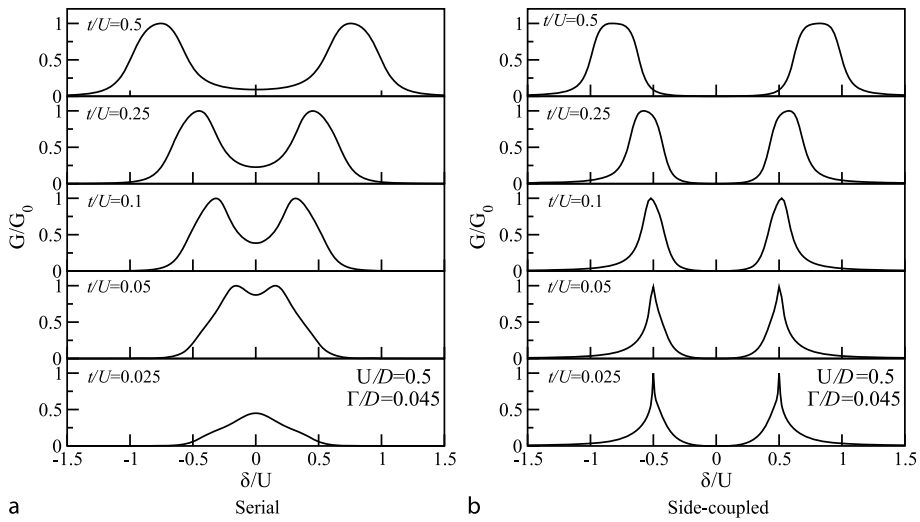
In multiple dot systems, the coupling of the quantum dots between each other and to the conduction leads affects the conductance in a non-trivial way since the entire system behaves in a quantum coherent way and no part of the system may be considered separately from other parts. Simple circuit theory is *not* applicable. The effects of the coupling topology can be conveniently studied in the case of double quantum dot [70]. As an illustration, we will consider the significantly different behavior of serial and side-coupled double quantum dot [70,71].

Serial Double Quantum Dot The zero-temperature conductance of two quantum dots coupled in series between two electrodes is shown in Fig. 10a for a range of the inter-dot hopping parameters t. The conductance is calculated from the scattering phase shifts obtained in NRG calculations. For large t, the coupling between the dots is strong and the system behaves as an artificial molecule composed of two atoms. When the system is occupied by an odd number of electrons, only the unpaired electron plays an important role and the system may be mapped to an effective single-impurity Anderson model where the role of the in purity orbital is played by the bonding (symmetric) or anti-bonding (anti-symmetric) "molecular orbital". Thus for $t/U = 0.5$, the conductance is high for $|\delta/U| = 0.5, \ldots, 1.2$ when the system is occupied by 1 or 3 electrons. For $|\delta/U| < 0.5$ the conductance is low.

As t decreases, the system starts to behave as two localized magnetic moments and may be approximately described by the two-impurity Kondo model. It is found that the conductance at the particle-hole symmetric point attains the theoretical limit of G_0 at the point where the inter-impurity exchange interaction J_{eff} is comparable to the scale of the Kondo temperature for one impurity coupled to a single conduction lead, T_{K}, i. e. for $J_{\text{eff}} \sim T_{\text{K}}$ (which corresponds to $t/U \sim 0.05$, see Fig. 10a). In the true two-impurity Kondo model, this point in the parameter space corresponds to a quantum phase transition between a regime of local antiferromagnetic singlet and a regime of separate Kondo screening of each impurity moment. In the double quantum dot system, however, this quantum phase transition is replaced by a smooth crossover due to charge transfer between the two conduction leads [72].

Finally, for very small t, the conductance tends to zero for all values of δ as the system becomes separated in two parts which no longer communicate.

Side-Coupled Double Quantum Dot In the side-coupled configuration, the first quantum dot is embedded between source and drain electrodes while the second dot is coupled to the first through a tunneling junction; there is no direct coupling of the second dot to the leads. By changing the gate voltage and the inter-dot tunneling rate, this system can be tuned to one of the following low-temperature regimes: i) a non-conducting local spin-singlet state, ii) the conventional Kondo regime with odd number of electrons occupying the dots, iii) the two-stage Kondo regime with two confined electrons, or iv) a valence-fluctuating state [71]. In addition, at finite temperatures a Fano resonance appears in the conductance; its origin lies in the sudden filling of the side-



Quantum Impurity Physics in Coupled Quantum Dots, Figure 10
Zero-temperature conductance G/G_0 of the double quantum dot system in a serial configuration and b side-coupled configuration as a function of the gate voltage for a range of the inter-dot hopping parameters t

coupled dot when its on-site energy crosses the Fermi level [71].

For large inter-dot tunneling coupling t, there are two wide regimes where the conductance is enhanced due to the conventional Kondo effect, for example in the ranges $|\delta/U| = 0.5, \ldots, 1.5$ for $t/U = 0.5$ (Fig. 10b) when the dot is occupied by 1 or 3 electrons. These regimes are separated by a low-conductance regime where the localized spins of two electrons are antiferromagnetically coupled for $|\delta/U| \lesssim 0.5$. For large t, the side-coupled and serial configurations of quantum dots thus have qualitatively similar properties.

For small t, in the two stage Kondo regime the two local moments are screened at different Kondo temperatures [71,73,74,75]. The *two-stage Kondo effect* is a generic name for successive Kondo screening of the impurity local moments at different temperatures [66,71,73,74,76,77,78,79]. This term has been used in two different (but closely related) contexts: 1) two-step screening of a $S = 1$ spin in the presence of two channels [76], 2) two step screening of two local moments in the single-channel case [73,77,78]. In the first case, the first-stage Kondo screening is an underscreened spin-1 Kondo effect which reduces the spin to 1/2, while the second-stage Kondo screening is a perfect-screening spin-1/2 Kondo effect which leads to a spin singlet ground state [77,80]. This first case is relevant when the lowest-energy impurity configuration is a spin triplet. In the second case, at a higher Kondo temperature $T_K^{(1)}$ the Kondo effect occurs on the more strongly coupled impurity; the Fermi liquid quasiparticles associated with the Kondo effect on the first impurity participate in the Kondo screening of the second impurity on an exponentially reduced Kondo temperature scale $T_K^{(2)}$ [71,73,74]. This case occurs when the lowest-energy configuration is a singlet, but there is a nearby excited triplet state [78].

In the double quantum dot system, the two-stage Kondo effect occurs when the effective exchange interaction between the dots is such that $J_{\text{eff}} < T_K$, where $T_K = T_K^{(1)}$ is the Kondo temperature of the single-impurity Anderson model that describes impurity 1 (without impurity 2) [71,74]. The second Kondo crossover then occurs at

$$T_K^{(2)} = c_2 T_K^{(1)} \exp(-c_1 T_K^{(1)}/J_{\text{eff}}) \,. \tag{15}$$

Constants c_1 and c_2 are of the order of 1 and they are problem-dependent. The spectral function $A_1(\omega)$ of impurity 1 increases at $\omega \sim T_K^{(1)}$, but then drops at $\omega \sim T_K^{(2)}$, i. e. there is a gap in the spectral function and the system is non-conducting at zero temperature.

The conductance increases to G_0 on the temperature scale of $T_K^{(1)}$, then drops to zero on the scale of $T_K^{(2)}$. The conductance can be high at finite temperatures even in the vicinity of the particle-hole symmetric point, $\delta = 0$, if $T_K^{(2)} < T < T_K^{(1)}$.

Capacitive Coupling and Charge Ordering

The effect of the inter-impurity electron repulsion (induced by capacitive coupling between two quantum dots)

may be modeled using the following Hamiltonian term:

$$H_{\text{dots}} = \sum_{i=1}^{2} H_{\text{dot},i} + U_{12}(n_1 - 1)(n_2 - 1)\,. \qquad (16)$$

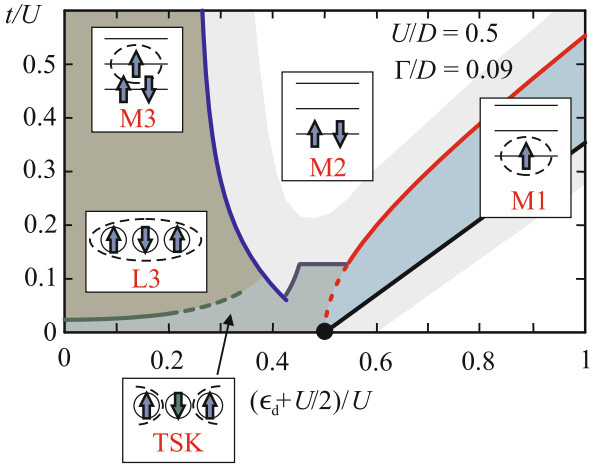
The inter-impurity repulsion is not an important perturbation as long as $U_{12} < U$; finite U_{12} only modifies the Kondo temperature, while the behavior of the system remains qualitatively unchanged [81]. For $U_{12} > U$ the electrons can lower their energy by forming on-site singlets and the system enters the *charge-ordering regime* [82]. The system behaves in a peculiar way at the transition point $U_{12} = U$, when an intermediate temperature fixed point with a six-fold symmetry of states appears. In serial dots, this leads to an exotic SU(4) Kondo effect [56,82]. For parallel dots, however, the coupling of impurities to the leads breaks the orbital symmetry and conventional Kondo screening occurs [81].

Universal Behavior Versus Complex Particularities

Near the particle-hole symmetric point (or, equivalently, at half filling when one electron occupies each quantum dot on the average), systems consisting of even or odd number of quantum dots have radically different behavior due to the distinct properties of integer and half-integer spin states. The half-integer spin states are always degenerate and quantum dot systems with such impurity configuration tend to exhibit some form of the Kondo effect for any coupling strength; the zero-temperature conductance of systems of an odd number of dots will tend to be high. In systems with an even number of quantum dots, however, the range of half filling is generally associated with Mott–Hubbard insulating behavior [65]: the conductance for a half-filled system decreases exponentially with electron–electron repulsion U [83]. Actual behavior also crucially depends on the coupling topology. The cases of serial and parallel dots will be considered in the following.

Linear Chains of Quantum Dots

The simplest non-trivial system with an odd number of quantum dots consists of three quantum dots coupled in series between two conduction leads. This triple quantum dot system is usually modelled as a three-site Hubbard chain. The special feature of this system is the presence of two equivalent screening channels combined with two-stage Kondo screening and/or magnetic ordering. Triple quantum dot structures have been manufactured in recent years and the analysis of their stability diagrams demonstrates that a description in terms of a Hubbard-like model is in deed a good approximation [11,12].

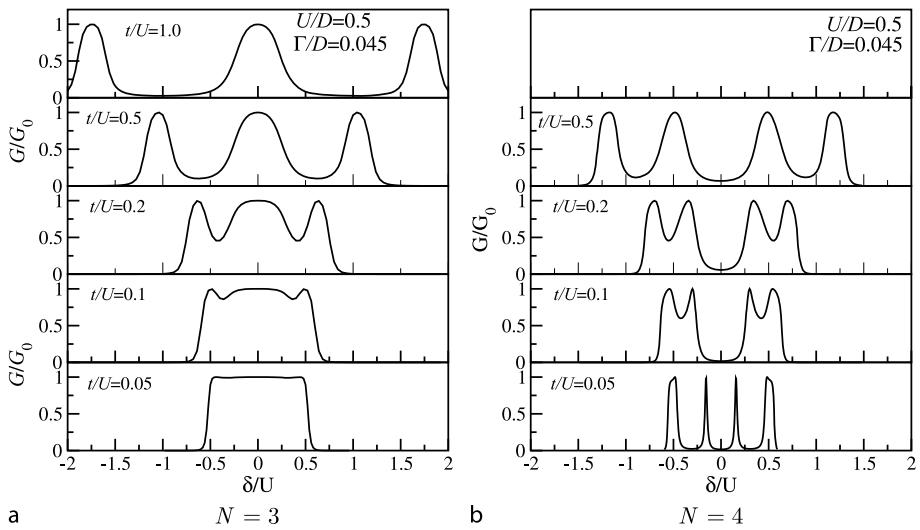


Quantum Impurity Physics in Coupled Quantum Dots, Figure 11
M1, M3: molecular-orbital Kondo regime with $\langle n\rangle \sim 1, 3$. M2: non-conductive even-occupancy state. L3: local Kondo regime with $\langle n\rangle \sim 3$. TSK: two-stage Kondo regime. Due to the particle-hole symmetry of the problem, the diagram is mirror-symmetric with respect to the $\delta = \epsilon_d + U/2 = 0$ axis; for negative $\delta < 0$ we thus find M4 non-conductive regime and M5 molecular-orbital Kondo regime

The behavior of the system depends strongly on the values of the on-site energies and on the inter-impurity hopping. Based on extensive calculations using several complementary methods, a phase diagram has been established, Fig. 11 [55]. It indicates the parameter ranges where the zero-temperature conductance is high.

For strong inter-impurity coupling t, the system may be mapped to an effective single-impurity Anderson model where the role of the impurity orbital is played by the bonding, non-bonding, or anti-bonding "molecular orbital". In this regime, the conductance is high when the occupancy is odd, and it is nearly zero when the occupancy is even, see Fig. 12a for $t/U = 1.0$, 0.5 and 0.2. For smaller inter-impurity coupling ($t/U \lesssim 0.1$), the molecular orbital description becomes inappropriate as the local behavior of the spins becomes important. The system then behaves as a necklace of magnetic atoms, rather than as a strongly-bound molecule.

When there are three electrons in the dots (i. e. for $\delta = 0$) and the coupling t is gradually decreased, the system crosses over from the molecular orbital M3 regime ($t \gtrsim U$) to the antiferromagnetic spin-chain L3 regime ($J_{\text{eff}} \sim t$), and finally to the two-stage Kondo (TSK) regime ($J_{\text{eff}} < T_{\text{K}}^{(1)}$), see Fig. 11. In the spin-chain regime, the three spins lock at $T \sim J_{\text{eff}}$ into a rigid spin-1/2 spin-chain state; at lower temperature, this collective spin is screened by the conventional spin-1/2 Kondo effect. In

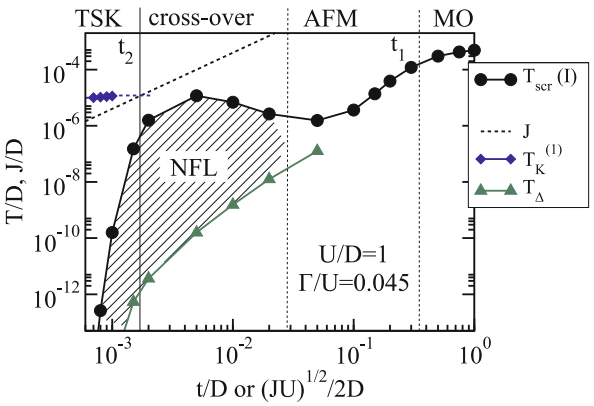


Quantum Impurity Physics in Coupled Quantum Dots, Figure 12
Conductance G/G_0 of the triple and quadruple quantum dot systems as a function of the gate voltage for a range of inter-dot hopping parameters t

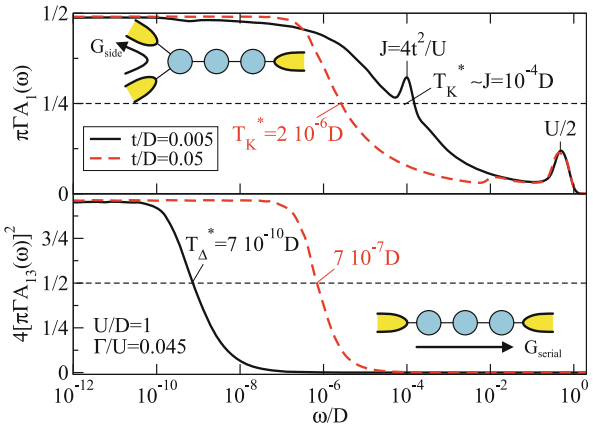
the two-stage Kondo regime, the spins on the first and third sites are screened at higher Kondo temperature $T_K^{(1)}$, then the spin on the central site is screened at an exponentially reduced second Kondo temperature $T_K^{(2)} \propto T_K^{(1)} \exp(-cT_K^{(1)}/J_{\text{eff}})$, where $J_{\text{eff}} \approx 4t^2/U$.

Antiferromagnetic and two-stage Kondo regimes are separated by a cross-over region with unusual properties at finite temperatures where the system approaches the so-called *two-channel Kondo model* non-Fermi liquid fixed point [75]. The non-Fermi liquid behavior emerges as the temperature is decreased below the Kondo temperature (T_{scr} screening temperature in Fig. 13) and disappears below the temperature T_Δ below which the behavior again corresponds to that of a Fermi liquid system.

Non-Fermi liquid behavior can be experimentally detected by measuring the differential conductance in a three-terminal configuration (see the insets in Fig. 14). The qualitative temperature dependence of the zero-bias conductance through the system can be approximately inferred from the frequency dependence of the spectral functions. The conductance through the system (from left to right conduction lead) is given by $G_{\text{serial}}/G_0 \approx 4(\pi\Gamma A_{13})^2$ [84] and the conductance through a side dot in the three-terminal configuration by $G_{\text{side}}/G_0 \approx \pi\Gamma A_1$ [67]. The appropriately normalized spectral densities are shown in Fig. 14 for the cases of cross-over regime with a non-Fermi liquid region and the antiferromagnetic regime with no discernible non-Fermi liquid behavior. When non-Fermi liquid fixed point is approached (for $t/D = 0.005$ and $T_\Delta \lesssim T \lesssim T_{\text{scr}}$), the conductance $G_{\text{side}} \sim G_0/2$, while $G_{\text{serial}} \sim 0$. The increase of the conductance G_{serial} through the system at $T \lesssim T_\Delta$ is concomitant with the cross-over from the non-Fermi liquid to Fermi liquid fixed point [72]. In the antiferromagnetic regime with no non-Fermi liquid region, both conductances increase below the same temperature scale, i. e. at $T \lesssim T_{\text{scr}}$.



Quantum Impurity Physics in Coupled Quantum Dots, Figure 13
Cross-over scales of triple quantum dot as function of the inter-dot coupling. The magnetic screening temperature T_{scr} is defined by $T_{\text{scr}}\chi(T_{\text{scr}})/(g\mu_B)^2 = 0.07$; it is equal to the Kondo temperature when screening is due to the single-channel Kondo effect. T_Δ is defined through $s_{\text{imp}}(T_\Delta)/k_B = \ln 2/4$, where $s_{\text{imp}}(T)$ is the impurity contribution to the total entropy at temperature T. Here $\ln 2/4$ is half the impurity entropy in the non-Fermi liquid fixed point



Quantum Impurity Physics in Coupled Quantum Dots, Figure 14
Dynamic properties of triple quantum dot in the antiferromagnetic *(dashed lines)* and in the cross-over regime *(full lines)*. *Upper panel*: on-site spectral function $A_1(\omega)$ of the left dot. *Lower panel*: out-of-diagonal spectral function $A_{13}(\omega)$ squared. Temperature T_Δ^* is of order T_Δ, T_K^* is of order T_{scr}

The conductance for four quantum dots coupled in series is shown in Fig. 12b. For large t the description in terms of molecular orbitals is again appropriate and we observe four conductance peaks [65]. There is a wide region of low conductance around the particle-hole symmetric point which corresponds to the Hubbard gap, and two pairs of conductance peaks which correspond to the Hubbard sub-bands. As t is reduced, the two inner peaks become rapidly extremely sharp, while the outer peaks centered at $|\delta/U| = 1/2$ narrow down more progressively. For very small t, the system is fully insulating at zero temperature for (almost) all values of δ.

On one hand, at zero temperature short chains of even and odd number of dots in a chain are seen to have widely different properties. On the other hand, for a very large number of dots, i. e. in the limit of a macroscopic system, insulating behavior is expected at half-filling irrespective of the even or odd parity of N. These two contrasting predictions can be reconciled by considering the order of taking the $T \to 0$ and $N \to \infty$ limits [65]. Taking the $T \to 0$ limit first, the even/odd alternation is obtained. If the $N \to \infty$ limit is taken first, which is the physically correct procedure, the conductance will vanish since the Kondo temperature (at which the conductance would increase for odd N) decreases with N [65].

Parallel Quantum Dots

Systems of N parallel quantum dots (see Fig. 3 for $N = 2$ and $N = 3$ cases) can be described by the multi-impurity single-channel Anderson model [81]. This model is defined by $H = H_{\rm band} + \sum_{i=1}^{N} H_i$, where $H_{\rm band}$ describes the conduction band and

$$H_i = \delta n_i + \frac{U}{2}(n_i - 1)^2 + V \sum_{k\mu} (c_{k\mu}^\dagger d_{i\mu} + d_{i\mu}^\dagger c_{k\mu}), \quad (17)$$

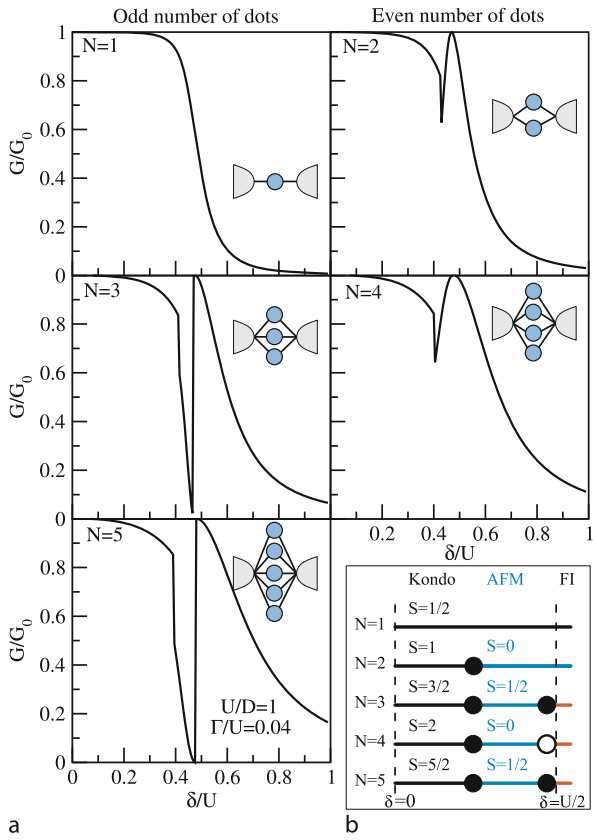
describe the N quantum dots.

It is assumed that all impurities hybridize with the same left-right symmetric combinations of states from both leads with a constant hybridization function $\Gamma = \pi\rho V^2$ [81]. This model may equally be applicable to other system where the RKKY interaction is ferromagnetic, for example to clusters of neighboring magnetic adatoms on metallic surfaces [38,85,86].

It is assumed the inter-dot tunneling coupling t and capacitive coupling (inter-dot charge repulsion) U_{12} are negligible, so that all dots are equivalent. At low temperature, the multi-impurity Anderson model maps to a multi-impurity Kondo model. At the particle-hole symmetric point, $\delta = 0$, the conduction-band-mediated RKKY exchange interaction is ferromagnetic, $J_{\rm RKKY} \sim U(\rho J_K)^2 = (64/\pi^2)(\Gamma^2/U)$, therefore the impurity spins order and the system effectively behaves as a single-impurity spin-$N/2$ Kondo model which undergoes spin-$N/2$ Kondo effect. This behavior is named the strong coupling (SC) regime. The Kondo temperature is approximately the same irrespective of the number of the impurities N. The residual spin at zero-temperature is $N/2 - 1/2$ if there is no coupling to additional screening channels. The ferromagnetically ordered regime and the ensuing spin-$N/2$ Kondo effect are fairly robust against various perturbations. Very strong perturbations lead, however, to quantum phase transitions of different kinds [81].

For very large δ/U, the impurities are unoccupied and the system is in the so-called frozen-impurity (FI) fixed-point with no residual spin. In the single-impurity ($N = 1$) case, the SC and FI fixed points are closely related: they belong to the same class of fixed points which differ in the strength of the potential scattering of the conduction band electrons on the impurity [87]. For multiple impurities ($N \geq 2$), however, the SC and FI lines of fixed points are qualitatively different (each corresponding to a different residual spin) and must be separated by at least one quantum phase transition [88].

At $\delta = 0$, the systems are fully conducting for any N and there is a wide plateau of high conductance associated with the spin-$N/2$ Kondo effect, Fig. 15 [81]. While the $N = 1$ system smoothly crosses over from the highly-conducting Kondo regime to the non-conducting FI regime, in the multi-impurity case we observe sharp



Quantum Impurity Physics in Coupled Quantum Dots, Figure 15
a Zero-temperature conductance through systems of N parallel quantum dots as a function of the gate voltage. Only $\delta > 0$ is shown due to the symmetry of the problem. b Zero-temperature phase diagram delimiting the different regimes as a function of the gate voltage. Filled circles (●) correspond to quantum phase transitions visible in the conductance curves, while the empty circle (○) denotes the phase transition with no associated conductance discontinuity

conductance discontinuities: one discontinuity for N even and two discontinuities for N odd. The conductance culminates in a unitary peak slightly below $\epsilon = 0$ (i. e. below $\delta/U = 1/2$) for all $N \geq 2$. The origin of this peak is simply potential scattering. The magnetic field B has a strong effect on the Kondo plateau: the conductance is significantly reduced as soon as B is of the order of the Kondo temperature T_K. The potential scattering peak, however, is only affected by extremely high fields of the order of U.

Conductance discontinuities find their counterparts in the jumps of the total electron occupancy and spin-spin correlation functions $\langle \boldsymbol{S}_i \cdot \boldsymbol{S}_j \rangle$, $i \neq j$; a new feature, however, is the existence of two points of discontinuity for $N = 4$ even though the conductance only exhibits one. In the Kondo regime for $\delta < \delta_{c1}$, the systems are nearly half-filled and spins are ferromagnetically ordered [81]. As we cross δ_{c1}, the occupancy slightly decreases and the spin correlations turn from ferromagnetic to antiferromagnetic. For $N \geq 3$, a second discontinuity occurs at somewhat higher δ_{c2}; its characteristic property is that the occupancy changes by almost exactly $N-2$, from $N-1$ to 1. According to the Friedel sum rule, a change in the average total impurity occupancy by n is mirrored in a change of the scattering phase shift by $\Delta\delta_{\text{q.p.}} = n\pi/2$. This explains the conductance jump from $G = G_0 \sin^2[(N-1)\pi/2] = 0$ to $G = G_0 \sin^2(\pi/2) = G_0$ in the case of odd $N \geq 3$ and the absence of the second conductance discontinuity for even $N \geq 4$, as 1 and $N-1$ are both odd integers. It is remarkable that the second quantum phase transition occurs precisely at the point where the conductance is extremal.

For $N \geq 3$, the N-impurity Anderson model thus undergoes two phase transitions. The first transition separates the ferromagnetically ordered regime and associated spin-$N/2$ Kondo screening from the antiferromagnetically ordered regime and (for odd N) Kondo screening of the spin-1/2 moment [89]. The second transition reflects the instability of the phases with the occupancy in the interval $1 < \langle n_{\text{tot}} \rangle < N-1$. Furthermore, for odd N the system abruptly switches from being fully conducting to zero conductance; this would facilitate the experimental observation of similar effects in quantum dot systems.

Future Directions

Further research in quantum transport theory will likely be centered at non-equilibrium, time-dependent and finite-temperature properties of interacting impurity systems. This is required to better understand transport at finite applied dc source-drain bias or for ac bias in real nanodevices. While the basic transport formalism is well developed, calculations of non-equilibrium properties of correlated system is still a formidable problem. The recently developed time-dependent NRG technique [90] appears very promising in this respect.

In recent years, the interest in quantum impurity physics has intensified once again due to an observation that extended lattice models of correlated electron systems may be exactly mapped in the limit of infinite lattice connectivity to effective impurity models subject to self-consistency conditions. This forms the foundation of a rapidly developing technique which has become known as the dynamical mean-field theory [91]. Results obtained in this way are believed to be a good approximation to the true behavior of such systems.

Experimental research is making progress toward creating artificial materials consisting of a large number of interconnected quantum dots. Studies of such systems would shed light on the behavior of extended bulk correlated materials. Furthermore, one could study how to tune material parameters to obtain desirable properties.

Bibliography

1. McEuen PL (1996) Artificial atoms: New boxes for electrons. Science 278:1729
2. Kouwenhoven LP, Marcus CM, McEuen PL, Tarucha S, Westervelt RM, Wingreen NS (1997) Electron transport in quantum dots. In: Sohn LL, Kouwenhoven LP, Schön G (eds) Mesoscopic electron transport. E, vol 345. NATO ASI, Kluwer, Dordrecht, pp 105–214
3. Kouwenhoven L, Marcus C (1998) Quantum dots. Phys World June 1998
4. Goldhaber-Gordon D, Shtrikman H, Mahalu D, Abusch-Magder D, Meirav U, Kastner MA (1998) Kondo effect in a single-electron transistor. Nature 391:156
5. Cronenwett SM, Oosterkamp TH, Kouwenhoven LP (1998) A tunable kondo effect in quantum dots. Science 281:540
6. Kouwenhoven L, Glazman L (2001) Revival of the kondo effect. Phys World Jan
7. Waugh FR, Berry MJ, Mar DJ, Westervelt RM, Campman KL, Gossard AC (1995) Single-electron charging in double and triple quantum dots with tunable coupling. Phys Rev Lett 75:705
8. Jeong H, Chang AM, Melloch MR (2001) The kondo effect in an artificial quantum dot molecules. Science 293:2221
9. Chen JC, Chang AM, Melloch MR (2004) Transition between quantum states in a parallel-coupled double quantum dot. Phys Rev Lett 92:176801
10. Craig NJ, Taylor JM, Lester EA, Marcus CM, Hanson MP, Gossard AC (2004) Tunable nonlocal spin control in a coupled-quantum dot system. Science 304:565
11. Gaudreau L, Studenikin SA, Sachrajda AS, Zawadzki P, Kam A, Lapointe J, Korkusinski M, Hawrylak P (2006) Stability diagram of a few-electron triple dot. Phys Rev Lett 97:036807
12. Korkusinski M, Gimenez IP, Hawrylak P, Gaudreau L, Studenikin SA, Sachrajda AS (2007) Topological hund's rules and the electronic properties of a triple lateral quantum dot molecule. Phys Rev B 75:115301 2007
13. Hewson AC (1993) The Kondo Problem to Heavy-Fermions. Cambridge University Press, Cambridge
14. Kondo J (1964) Resistance minimum in dilute magnetic alloys. Prog Theor Phys 32:37
15. Wilson KG (1975) The renormalization group: Critical phenomena and the kondo problem. Rev Mod Phys 47:773
16. Stewart GR (1984) Heavy-fermion systems. Rev Mod Phys 56:755
17. Pustilnik M, Glazman L (2004) Kondo effect in quantum dots. J Phys: Condens Matter 16:R513
18. Nygard J, Cobden DH, Lindelof PE (2000) Kondo physics in carbon nanotubes. Nature 408:342
19. Liang W, Shores MP, Bockrath M, Long JR, Park K (2002) Kondo resonance in a single-molecule transistor. Nature 417:725
20. Madhavan V, Chen W, Jamneala T, Crommie M, Wingreen NS (1998) Tunneling into a single magnetic atom: Spectroscopic evidence of the kondo resonance. Science 280:567
21. Li J, Schneider W-D, Berndt R, Delley B (1998) Kondo scattering observed at a single magnetic impurity. Phys Rev Lett 80:2893
22. Manoharan HC, Lutz CP, Eigler DM (2000) Quantum mirages formed by coherent projection of electronic structure. Nature 403:512
23. Glazman LI, Raikh ME (1988) Resonant kondo transparency of a barrier with quasilocal impurity states. JETP Lett 47:452
24. Ng TK, Lee PA (1988) On-site coulomb repulsion and resonant tunneling. Phys Rev Lett 61:1768
25. Kouwenhoven LP, Oosterkamp TH, Danoesastro MWS, Eto M, Austing DG, Honda T, Tarucha S (1997) Excitation spectra of circular, few-electron quantum dots. Science 278:1788
26. Ashoori RC (1996) Electrons in artificial atoms. Science 379:413
27. Kastner MA (1992) The single-electron transistor. Rev Mod Phys 64:849
28. Thornton TJ, Pepper M, Ahmed H, Andrews D, Davies GJ (1986) One-dimensional conduction in the 2d electron gas of a gaas-algaas heterojunction. Phys Rev Lett 56:1198
29. Park J, Pasupathy AN, Goldsmith JI, Chang C, Yaish Y, Petta JR, Rinkoski M, Sethna JP, Abruna HD, McEuen PL, Ralph DC (2002) Coulomb blockade and the kondo effect in single-atom transistors. Nature 417:722
30. Yu LH, Natelson D (2004) The kondo effect in c_{60} single-molecule transistors. Nanoletters 4:79
31. Schmid J, Weis J, Eberl K, v Klitzing K (2000) Absence of odd-even parity behavior for kondo resonance in quantum dots. Phys Rev Lett 84:5824
32. Anderson PW (1961) Localized magnetic states in metals. Phys Rev 124:41
33. Schedelbeck G, Wegscheider W, Bichler M, Abstreiter G (1997) Coupled quantum dots fabricated by cleaved edge overgrowth: From artificial atoms to molecules. Science 278:1792
34. Oosterkamp TH, Fujisawa T, van der Wiel WG, Ishibashi K, Hijman RV, Tarucha S, Kouwenhoven LP (1998) Microwave spectroscopy of a quantum-dot molecule. Nature 395:873
35. Holleitner AW, Blick RH, Hüttel AK, Eberl K, Kotthaus JP (2002) Probing and controlling the bonds of an artificial molecule. Science 297:70
36. van der Wiel WG, De Franceschi S, Elzerman JM, Fujisawa T, Tarucha S, Kouwenhoven LP (2003) Electron transport through double quantum dots. Rev Mod Phys 75:1
37. Madhavan V, Jamneala T, Nagaoka K, Chen W, Li J-L, Louie SG, Crommie MF (2002) Observation of spectral evolution during the formation of ni_2 kondo molecule. Phys Rev B 66:212411
38. Jamneala T, Madhavan V, Crommie MF (2001) Kondo response of a single antiferromagnetic chromium trimer. Phys Rev Lett 87:256804
39. Wahl P, Diekhoner L, Wittich G, Vitali L, Schneider MA, Kern K (2005) Kondo effect of molecular complexes at surfaces: ligand control of the local spin coupling. Phys Rev Lett 95:166601
40. Jones BA, Varma CM, Wilkins JW (1988) Low-temperature properties of the two-impurity kondo hamiltonian. Phys Rev Lett 61:125
41. Affleck I, Ludwig AWW, Jones BA (1995) Conformal-field-theory approach to the two-impurity kondo problem: Comparison with numerical renormalization-group results. Phys Rev B 52:9528
42. Bickers NE (1987) Review of techniques in the large-*n* expansion for dilute magnetic alloys. Rev Mod Phys 59:845
43. Andrei N, Furuya K, Lowenstein JH (1983) Solution of the kondo problem. Rev Mod Phys 55:331

44. Tsvelick AM, Wiegmann PB (1983) Exact results in the theory of magnetic alloys. Adv Phys 32:453
45. Gogolin AO, Nersesyan AA, Tsvelik AM (1999) Bosonization and strongly correlated systems. Cambridge University Press, Cambridge
46. Affleck I (1990) A current algebra approach to the kondo effect. Nucl Phys B 336:517
47. Affleck I, Ludwig AWW (1991) The kondo effect, conformal field theory and fusion rules. Nucl Phys B 352:849
48. Varma CM, Yafet Y (1976) Magnetic susceptibility of mixed-valence rare-earth compounds. Phys Rev B 13:2950
49. Gunnarsson O, Schönhammer K (1983) Electron spectroscopies for ce compounds in the impurity model. Phys Rev B 28:4315
50. Anderson PW, Yuval G (1996) Exact results in the kondo problem: Equivalence to a classical one-dimensional coulomb gas. Phys Rev Lett 23:89
51. Anderson PW, Yuval G, Hamann DR (1970) Exact results in the kondo problem ii: Scaling theory, qualitatively correct solution and some new results on one-dimensional classical statistical models. Phys Rev B 1:4464
52. Anderson PW (1970) A poor man's derivation of scaling laws for the kondo problem. J Phys C: Solid St Phys 3:2436
53. Mehta P, Andrei N (2006) Nonequilibrium transport in quantum impurity models: The bethe ansatz for open systems. Phys Rev Lett 96:216802
54. Rejec T, Ramšak A (2003) Formulas for zero-temperature conductance through a region with interaction. Phys Rev B 68:035342
55. Žitko R, Bonča J, Ramšak A, Rejec T (2006) Kondo effect in triple quantum dots. Phys Rev B 73:153307
56. Mravlje J, Ramšak A, Rejec T (2006) Kondo effect in double quantum dots with interdot repulsion. Phys Rev B 73:241305(R)
57. Bulla R, Costi T, Pruschke T (2007) The numerical renormalization group method for quantum impurity models. Rev Mod Phys 80:395; cond-mat/0701106
58. Coleman P (2002) Local moment physics in heavy electron systems. cond-mat/0206003
59. Krishna-murthy HR, Wilkins JW, Wilson KG (1980) Renormalization-group approach to the anderson model of dilute magnetic alloys. i. static properties for the symmetric case. Phys Rev B 21:1003
60. Averin DV, Nazarov YV (1990) Virtual electron diffusion during quantum tunneling of the electric charge. Phys Rev Lett 65:2446
61. Imry Y, Landauer R (1999) Conductance viewed as transmission. Rev Mod Phys 71:S306
62. Imry Y (2002) Introduction to mesoscopic physics, 2nd edn. Oxford University Press, Oxford
63. Datta S (1997) Electronic Transport in Mesoscopic Systems. Cambridge University Press, Cambridge
64. Oguri A, Nisikawa Y, Hewson AC (2005) Determination of the phase shifts for interacting electrons connected to reservoirs. J Phys Soc Jpn 74:2554
65. Nisikawa Y, Oguri A (2006) Numerical renormalization group approach to a quartet quantum-dot array connected to reservoirs: Gate-voltage dependence of the conductance. Phys Rev B 73:125108
66. Pustilnik M, Glazman LI (2001) Kondo effect in real quantum dots. Phys Rev Lett 87:216601
67. Meir Y, Wingreen NS (1992) Landauer formula for the current through an interacting electron region. Phys Rev Lett 68: 2512
68. van der Wiel WG, De Franceschi S, Fujisawa T, Elzerman JM, Tarucha S, Kouwenhoven LP (2000) The kodno effect in the unitary limit. Science 289:2105
69. Fabrizio M (2006) Clusters of anderson impurities and beyond: How kondo effect dies and what can we learn about the mott transition. Lecture notes, XI Training Course in the Physics of Strongly Correlated Systems
70. Ramšak A, Mravlje J, Žitko R, Bonča J (2006) Spin qubits in double quantum dots: Entanglement versus the kondo effect. Phys Rev B 74:241305(R)
71. Žitko R, Bonča J (2006) Enhanced conductance through side-coupled double quantum dots. Phys Rev B 73:035332
72. Zaránd G, Chung C-H, Simon P, Vojta M (2006) Quantum criticality in a double quantum-dot system. Phys Rev Lett 97:166802
73. Vojta M, Bulla R, Hofstetter W (2002) Quantum phase transitions in models of coupled magnetic impurities. Phys Rev B 65:140405(R)
74. Cornaglia PS, Grempel DR (2005) Strongly correlated regimes in a double quantum dot device. Phys Rev B 71:075305
75. Žitko R, Bonča J (2007) Fermi-liquid versus non-fermi-liquid behavior in triple quantum dots. Phys Rev Lett 98:047203
76. Jayaprakash C, Krishna-murthy HR, Wilkins JW (1981) Two-impurity kondo problem. Phys Rev Lett 47:737
77. van der Wiel WG, De Franceschi S, Elzerman JM, Tarucha S, Kouwenhoven LP, Motohisa J, Nakajima F, Fukui T (2002) Two-stage kondo effect in a quantum dot at a high magnetic field. Phys Rev Lett 88:126803
78. Hofstetter W, Schoeller H (2002) Quantum phase transition in a multilevel dot. Phys Rev Lett 88:016803
79. Hofstetter W, Zarand G (2004) Singlet-triplet transition in lateral quantum dots: A renormalization group study. Phys Rev B 69:235301
80. Hofstetter W (2000) Generalized numerical renormalization group for dynamical quantities. Phys Rev Lett 85:1508
81. Žitko R, Bonča J (2006) Multi-impurity anderson model for quantum dots coupled in parallel. Phys Rev B 74:045312
82. Galpin MR, Logan DE, Krishnamurthy HR (2005) Quantum phase transition in capacitively coupled double quantum dot. Phys Rev Lett 94:186406
83. Oguri A, Hewson AC (2005) Nrg approach to the transport through a finite hubbard chain connected to reservoirs. J Phys Soc Jpn 74:988
84. Caroli C, Combescot R, Nozieres P, Saint-James D (1971) Calculation of the tunneling current. J Phys C 4:916
85. Aligia AA (2006) Effective kondo model for a trimer on a metallic surface. Phys Rev Lett 96:096804
86. Wahl P, Simon P, Diekhöner L, Stepanyuk VS, Bruno P, Schneider MA, Kern K (2007) Exchange interaction between single magnetic adatoms. Phys Rev Lett 98:056601
87. Krishna-murthy HR, Wilkins JW, Wilson KG (1980) Renormalization-group approach to the anderson model of dilute magnetic alloys. ii. static properties for the asymmetric case. Phys Rev B 21:1044
88. Vojta M (2006) Impurity quantum phase transitions. Phil Mag 86:1807
89. Žitko R, Bonča J (2007) Quantum phase transitions in systems of parallel quantum dots. Phys Rev B 76:241305(R)

90. Anders FB, Schiller A (2005) Real-time dynamics in quantum impurity systems: A time-dependent numerical renormalization group approach. Phys Rev B 74:245113
91. Kotliar G (2005) Quantum impurity models as reference systems for strongly correlated materials: the road from the kondo impurity model to first principles electronic structure calculations with dynamical mean-field theory. J Phys Soc Jpn 74:147

Quantum Information Processing

SETH LLOYD
W.M. Keck Center for Extreme Quantum Information Processing (xQIT), MIT, Cambridge, USA

Article Outline

Glossary

Algorithm A systematic procedure for solving a problem, frequently implemented as a computer program.

Bit The fundamental unit of information, representing the distinction between two possible states, conventionally called 0 and 1. The word 'bit' is also used to refer to a physical system that registers a bit of information.

Boolean algebra The mathematics of manipulating bits using simple operations such as *AND*, *OR*, *NOT*, and *COPY*.

Communication channel A physical system that allows information to be transmitted from one place to another.

Computer A device for processing information. A digital computer uses Boolean algebra (q. v.) to processes information in the form of bits.

Cryptography The science and technique of encoding information in a secret form. The process of encoding is called encryption, and a system for encoding and decoding is called a cipher. A key is a piece of information used for encoding or decoding. Public-key cryptography operates using a public key by which information is encrypted, and a separate private key by which the encrypted message is decoded.

Decoherence A peculiarly quantum form of noise that has no classical analog. Decoherence destroys quantum superpositions and is the most important and ubiquitous form of noise in quantum computers and quantum communication channels.

Error-correcting code A technique for encoding information in a form that is resistant to errors. The syndrome is the part of the code that allows the error to be detected and that specifies how it should be corrected.

Entanglement A peculiarly quantum form of correlation that is responsible for many types of quantum weirdness. Entanglement arises when two or more quantum systems exist in a superposition of correlated states.

Entropy Information registered by the microscopic motion of atoms and molecules. The second law of thermodynamics (q. v.) states that entropy does not decrease over time.

Fault-tolerant computation Computation that uses error-correcting codes to perform algorithms faithfully in the presence of noise and errors. If the rate of errors falls below a certain threshold, then computations of any desired length can be performed in a fault-tolerant fashion. Also known as robust computation.

Information When used in a broad sense, information is data, messages, meaning, knowledge, etc. Used in the more specific sense of information theory, information is a quantity that can be measured in bits.

Logic gate A physical system that performs the operations of Boolean algebra (q. v.) such as *AND*, *OR*, *NOT*, and *COPY*, on bits.

Moore's law The observation, first made by Gordon Moore, that the power of computers increases by a factor of two every year and a half or so.

Quantum algorithm An algorithm designed specifically to be performed by a quantum computer using quantum logic. Quantum algorithms exploit the phenomena of superposition and entanglement to solve problems more rapidly than classical computer algorithms can. Examples of quantum algorithms include Shor's algorithm for factoring large numbers and breaking public-key cryptosystems, Grover's algorithm for searching databases, quantum simulation, the adiabatic algorithm, etc.

Quantum bit A bit registered by a quantum-mechanical system such as an atom, photon, or nuclear spin. A quantum bit, or 'qubit', has the property that it can exist in a quantum superposition of the states 0 and 1.

Qubit A quantum bit.

Quantum communication channel A communication channel that transmits quantum bits. The most common communication channel is the bosonic channel, which transmits information using light, sound, or other substances whose elementary excitations consist of bosons (photons for light, phonons for sound).

Quantum computer A computer that operates on quantum bits to perform quantum algorithms. Quantum computers have the feature that they can preserve quantum superpositions and entanglement.

Quantum cryptography A cryptographic technique that encodes information on quantum bits. Quantum cryptography uses the fact that measuring quantum systems typically disturbs them to implement cryptosystems whose security is guaranteed by the laws of physics. Quantum key distribution (QKD) is a quantum cryptographic technique for distributing secret keys.

Quantum error-correcting code An error-correcting code that corrects for the effects of noise on quantum bits. Quantum error-correcting codes can correct for the effect of decoherence (q. v.) as well as for conventional bit-flip errors.

Quantum information Information that is stored on qubits rather than on classical bits.

Quantum mechanics The branch of physics that describes how matter and energy behave at their most fundamental scales. Quantum mechanics is famously weird and counterintuitive.

Quantum weirdness A catch-all term for the strange and counterintuitive aspects of quantum mechanics. Well-known instances of quantum weirdness include Schrödinger's cat (q. v.), the Einstein–Podolsky–Rosen thought experiment, violations of Bell's inequalities, and the Greenberger–Horne–Zeilinger experiment.

Reversible logic Logical operations that do not discard information. Quantum computers operate using reversible logic.

Schrödinger's cat A famous example of quantum weirdness. A thought experiment proposed by Erwin Schrödinger, in which a cat is put in a quantum superposition of being alive and being dead. Not sanctioned by the Society for Prevention of Cruelty to Animals.

Second law of thermodynamics The second law of thermodynamics states that entropy does not increase. An alternative formulation of the second law states that it is not possible to build an eternal motion machine.

Superposition The defining feature of quantum mechanics which allows particles such as electrons to exist in two or more places at once. Quantum bits can exist in superpositions of 0 and 1 simultaneously.

Teleportation A form of quantum communication that uses pre-existing entanglement and classical communication to send quantum bits from one place to another.

Definition of the Subject

Quantum mechanics is the branch of physics that describes how systems behave at their most fundamental level. The theory of information processing studies how information can be transferred and transformed. Quantum information science, then, is the theory of communication and computation at the most fundamental physical level. Quantum computers store and process information at the level of individual atoms. Quantum communication systems transmit information on individual photons.

Over the past half century, the wires and logic gates in computers have halved in size every year and a half, a phenomenon known as Moore's law. If this exponential rate of miniaturization continues, then the components of computers should reach the atomic scale within a few decades. Even at current (2008) length scales of a little larger than one hundred nanometers, quantum mechanics plays a crucial role in governing the behavior of these wires and gates. As the sizes of computer components press down toward the atomic scale, the theory of quantum information processing becomes increasingly important for characterizing how computers operate. Similarly, as communication systems become more powerful and efficient, the quantum mechanics of information transmission becomes the key element in determining the limits of their power.

Miniaturization and the consequences of Moore's law are not the primary reason for studying quantum information, however. Quantum mechanics is weird: electrons, photons, and atoms behave in strange and counterintuitive ways. A single electron can exist in two places simultaneously. Photons and atoms can exhibit a bizarre form of correlation called entanglement, a phenomenon that Einstein characterized as *spukhafte Fernwirkung*, or 'spooky action at a distance'. Quantum weirdness extends to information processing. Quantum bits can take on the values of 0 and 1 simultaneously. Entangled photons can be used to teleport the states of matter from one place to another. The essential goal of quantum information science is to determine how quantum weirdness can be used to enhance the capabilities of computers and communication systems. For example, even a moderately sized quantum computer, containing a few tens of thousands of bits,

would be able to factor large numbers and thereby break cryptographic systems that have until now resisted the attacks of even the largest classical supercomputers [1]. Quantum computers could search databases faster than classical computers. Quantum communication systems allow information to be transmitted in a manner whose security against eavesdropping is guaranteed by the laws of physics.

Prototype quantum computers that store bits on individual atoms and quantum communication systems that transmit information using individual photons have been built and operated. These prototypes have been used to confirm the predictions of quantum information theory and to explore the behavior of information processing at the most microscopic scales. If larger, more powerful versions of quantum computers and communication systems become readily available, they will offer considerable enhancements over existing computers and communication systems. In the meanwhile, the field of quantum information processing is constructing a unified theory of how information can be registered and transformed at the fundamental limits imposed by physical law.

The remainder of this article is organized as follows:

- Section "Introduction"
  A review of the history of ideas of information, computation, and the role of information in quantum mechanics is presented.
- Section "Quantum Mechanics"
  The formalism of quantum mechanics is introduced and applied to the idea of quantum information.
- Section "Quantum Computation"
  Quantum computers are defined and their properties presented.
- Section "Noise and Errors"
  The effects of noise and errors are explored.
- Section "Quantum Communication"
  The role of quantum mechanics in setting limits to the capacity of communication channels is delineated. Quantum cryptography is explained.
- Section "Implications and Conclusions"
  Implications are discussed.

This review of quantum information theory is mathematically self-contained in the sense that all the necessary mathematics for understanding the quantum effects treated in detail here are contained in the introductory section on quantum mechanics. By necessity, not all topics in quantum information theory can be treated in detail within the confines of this article. We have chosen to treat a few key subjects in more detail: in the case of other topics we supply references to more complete treatments. The standard reference on quantum information theory is the text by Nielsen and Chuang [1], to which the reader may turn for in depth treatments of most of the topics covered here. One topic that is left largely uncovered is the broad field of quantum technologies and techniques for actually building quantum computers and quantum communication systems. Quantum technologies are rapidly changing, and no brief review like the one given here could adequately cover both the theoretical and the experimental aspects of quantum information processing.

Introduction

Information

Quantum information processing as a distinct, widely recognized field of scientific inquiry has arisen only recently, since the early 1990s. The mathematical theory of information and information processing dates to the mid-twentieth century. Ideas of quantum mechanics, information, and the relationships between them, however, date back more than a century. Indeed, the basic formulae of information theory were discovered in the second half of the nineteenth century, by James Clerk Maxwell, Ludwig Boltzmann, and J. Willard Gibbs [2]. These statistical mechanicians were searching for the proper mathematical characterization of the physical quantity known as entropy. Prior to Maxwell, Boltzmann, and Gibbs, entropy was known as a somewhat mysterious quantity that reduced the amount of work that steam engines could perform. After their work established the proper formula for entropy, it became clear that entropy was in fact a form of information — the information required to specify the actual microscopic state of the atoms in a substance such as a gas. If a system has W possible states, then it takes $\log_2 W$ bits to specify one state. Equivalently, any system with distinct states can be thought of as registering information, and a system that can exist in one out of W equally likely states can register $\log_2 W$ bits of information. The formula, $S = k \log W$, engraved on Boltzmann's tomb, means that entropy S is proportional to the number of bits of information registered by the microscopic state of a system such as a gas. (Ironically, this formula was first written down not by Boltzmann, but by Max Planck [3], who also gave the first numerical value $1.38\,10^{-23}$ J/K for the constant k. Consequently, k is called Planck's constant in early works on statistical mechanics [2]. As the fundamental constant of quantum mechanics, $h = 6.63\,10^{-34}$ joule seconds, on which more below, is also called Planck's constant, k was renamed Boltzmann's constant and is now typically written k_B.)

Although the beginning of the information processing revolution was still half a century away, Maxwell, Boltzmann, Gibbs, and their fellow statistical mechanicians were well aware of the connection between information and entropy. These researchers established that if the probability of the ith microscopic state of some system is p_i, then the entropy of the system is $S = k_B(-\sum_i p_i \ln p_i)$. The quantity $\sum_i p_i \ln p_i$ was first introduced by Boltzmann, who called it H. Boltzmann's famous H-theorem declares that H never increases [2]. The H-theorem is an expression of the second law of thermodynamics, which declares that $S = -k_B H$ never decreases. Note that this formula for S reduces to that on Boltzmann's tomb when all the states are equally likely, so that $p_i = 1/W$.

Since the probabilities for the microscopic state of a physical system depend on the knowledge possessed about the system, it is clear that entropy is related to information. The more certain one is about the state of a system – the more information one possesses about the system – the lower its entropy. As early as 1867, Maxwell introduced his famous 'demon' as a hypothetical being that could obtain information about the actual state of a system such as a gas, thereby reducing the number of states W compatible with the information obtained, and so decreasing the entropy [4]. Maxwell's demon therefore apparently contradicts the second law of thermodynamics. The full resolution of the Maxwell's demon paradox was not obtained until the end of the twentieth century, when the theory of the physics of information processing described in this review had been fully developed.

Quantum Mechanics

For the entropy, S, to be finite, a system can only possess a finite number W of possible states. In the context of classical mechanics, this feature is problematic, as even the simplest of classical systems, such as a particle moving along a line, possesses an infinite number of possible states. The continuous nature of classical mechanics frustrated attempts to use the formula for entropy to calculate many physical quantities such as the amount of energy and entropy in the radiation emitted by hot objects, the so-called 'black body radiation'. Calculations based on classical mechanics suggested the amount of energy and entropy emitted by such objects should be infinite, as the number of possible states of a classical oscillator such as a mode of the electromagnetic field was infinite. This problem is known as 'the ultraviolet catastrophe'. In 1901, Planck obtained a resolution to this problem by suggesting that such oscillators could only possess discrete energy levels [3]: the energy of an oscillator that vibrates with frequency ν can only come in multiples of $h\nu$, where h is Planck's constant defined above. Energy is *quantized*. In that same paper, as noted above, Planck first wrote down the formula $S = k \log W$, where W referred to the number of discrete energy states of a collection of oscillators. In other words, the very first paper on quantum mechanics was about information. By introducing quantum mechanics, Planck made information/entropy finite. Quantum information as a distinct field of inquiry may be young, but its origins are old: the origin of quantum information coincides with the origin of quantum mechanics.

Quantum mechanics implies that nature is, at bottom, discrete. Nature is digital. After Planck's advance, Einstein was able to explain the photo-electric effect using quantum mechanics [5]. When light hits the surface of a metal, it kicks off electrons. The energy of the electrons kicked off depends only on the frequency ν of the light, and not on its intensity. Following Planck, Einstein's interpretation of this phenomenon was that the energy in the light comes in chunks, or *quanta*, each of which possesses energy $h\nu$. These quanta, or particles of light, were subsequently termed photons. Following Planck and Einstein, Niels Bohr used quantum mechanics to derive the spectrum of the hydrogen atom [6].

In the mid nineteen-twenties, Erwin Schrödinger and Werner Heisenberg put quantum mechanics on a sound mathematical footing [7,8]. Schrödinger derived a wave equation – the Schrödinger equation – that described the behavior of particles. Heisenberg derived a formulation of quantum mechanics in terms of matrices, matrix mechanics, which was subsequently realized to be equivalent to Schrödinger's formulation. With the precise formulation of quantum mechanics in place, the implications of the theory could now be explored in detail.

It had always been clear that quantum mechanics was strange and counterintuitive: Bohr formulated the phrase 'wave-particle duality' to capture the strange way in which waves, like light, appeared to be made of particles, like photons. Similarly, particles, like electrons, appeared to be associated with waves, which were solutions to Schrödinger's equation. Now that the mathematical underpinnings of quantum mechanics were in place, however, it became clear that quantum mechanics was downright weird. In 1935, Einstein, together with his collaborators Boris Podolsky and Nathan Rosen, came up with a thought experiment (now called the EPR experiment after its originators) involving two photons that are correlated in such a way that a measurement made on one photon appears instantaneously to affect the state of the other photon [9]. Schrödinger called this form of corre-

lation 'entanglement'. Einstein, as noted above, referred to it as 'spooky action at a distance'. Although it became clear that entanglement could not be used to transmit information faster than the speed of light, the implications of the EPR thought experiment were so apparently bizarre that Einstein felt that it demonstrated that quantum mechanics was fundamentally incorrect. The EPR experiment will be discussed in detail below. Unfortunately for Einstein, when the EPR experiment was eventually performed, it confirmed the counterintuitive predictions of quantum mechanics. Indeed, every experiment ever performed so far to test the predictions of quantum mechanics has confirmed them, suggesting that, despite its counterintuitive nature, quantum mechanics is fundamentally correct.

At this point, it is worth noting a curious historical phenomenon, which persists to the present day, in which a famous scientist who received his or her Nobel prize for work in quantum mechanics, publicly expresses distrust or disbelief in quantum mechanics. Einstein is the best known example of this phenomenon, but more recent examples exist, as well. The origin of this phenomenon can be traced to the profoundly counterintuitive nature of quantum mechanics. Human infants, by the age of a few months, are aware that objects – at least, large, classical objects like toys or parents – cannot be in two places simultaneously. Yet in quantum mechanics, this intuition is violated repeatedly. Nobel laureates typically possess a powerful sense of intuition: if Einstein is not allowed to trust his intuition, then who is? Nonetheless, quantum mechanics contradicts their intuition just as it does everyone else's. Einstein's intuition told him that quantum mechanics was wrong, and he trusted that intuition. Meanwhile, scientists who are accustomed to their intuitions being proved wrong may accept quantum mechanics more readily. One of the accomplishments of quantum information processing is that it allows quantum weirdness such as that found in the EPR experiment to be expressed and investigated in precise mathematical terms, so we can discover exactly how and where our intuition goes wrong.

In the 1950's and 60's, physicists such as David Bohm, John Bell, and Yakir Aharonov, among others, investigated the counterintuitive aspects of quantum mechanics and proposed further thought experiments that threw those aspects in high relief [10,11,12]. Whenever those thought experiments have been turned into actual physical experiments, as in the well-known Aspect experiment that realized Bell's version of the EPR experiment [13], the predictions of quantum mechanics have been confirmed. Quantum mechanics is weird and we just have to live with it.

As will be seen below, quantum information processing allows us not only to express the counterintuitive aspects of quantum mechanics in precise terms, it allows us to exploit those strange phenomena to compute and to communicate in ways that our classical intuitions would tell us are impossible. Quantum weirdness is not a bug, but a feature.

Computation

Although rudimentary mechanical calculators had been constructed by Pascal and Leibnitz, amongst others, the first attempts to build a full-blown digital computer also lie in the nineteenth century. In 1822, Charles Babbage conceived the first of a series of mechanical computers, beginning with the fifteen ton Difference Engine, intended to calculate and print out polynomial functions, including logarithmic tables. Despite considerable government funding, Babbage never succeeded in building a working difference. He followed up with a series of designs for an Analytical Engine, which was to have been powered by a steam engine and programmed by punch cards. Had it been constructed, the analytical engine would have been the first modern digital computer. The mathematician Ada Lovelace is frequently credited with writing the first computer program, a set of instructions for the analytical engine to compute Bernoulli numbers.

In 1854, George Boole's *An investigation into the laws of thought* laid the conceptual basis for binary computation. Boole established that any logical relation, no matter how complicated, could be built up out of the repeated application of simple logical operations such as *AND*, *OR*, *NOT*, and *COPY*. The resulting 'Boolean logic' is the basis for the contemporary theory of computation.

While Schrödinger and Heisenberg were working out the modern theory of quantum mechanics, the modern theory of information was coming into being. In 1928, Ralph Hartley published an article, 'The Transmission of Information', in the Bell System Technical Journal [14]. In this article he defined the amount of information in a sequence of n symbols to be $n \log S$, where S is the number of symbols. As the number of such sequences is S^n, this definition clearly coincides with the Planck–Boltzmann formula for entropy, taking $W = S^n$.

At the same time as Einstein, Podolsky, and Rosen were exploring quantum weirdness, the theory of computation was coming into being. In 1936, in his paper "On Computable Numbers, with an Application to the *Entscheidungsproblem*", Alan Turing extended the earlier work of Kurt Gödel on mathematical logic, and introduced

the concept of a Turing machine, an idealized digital computer [15]. Claude Shannon, in his 1937 master's thesis, "A Symbolic Analysis of Relay and Switching Circuits", showed how digital computers could be constructed out of electronic components [16]. (Howard Gardner called this work, "possibly the most important, and also the most famous, master's thesis of the century".)

The Second World War provided impetus for the development of electronic digital computers. Konrad Zuse's Z3, built in 1941, was the first digital computer capable of performing the same computational tasks as a Turing machine. The Z3 was followed by the British Colossus, the Harvard Mark I, and the ENIAC. By the end of the 1940s, computers had begun to be built with a stored program or 'von Neumann' architecture (named after the pioneer of quantum mechanics and computer science John von Neumann), in which the set of instructions – or program – for the computer were stored in the computer's memory and executed by a central processing unit.

In 1948, Shannon published his groundbreaking article, "A Mathematical Theory of Communication", in the Bell Systems Journal [17]. In this article, perhaps the most influential work of applied mathematics of the twentieth century (following the tradition of his master's thesis), Shannon provided the full mathematical characterization of information. He introduced his colleague, John Tukey's word, 'bit', a contraction of 'binary digit', to describe the fundamental unit of information, a distinction between two possibilities, True or False, Yes or No, 0 or 1. He showed that the amount of information associated with a set of possible states i, each with probability p_i, was uniquely given by formula $-\sum_i p_i \log_2 p_i$. When Shannon asked von Neumann what he should call this quantity, von Neumann is said to have replied that he should call it H, 'because that's what Boltzmann called it'. (Recalling the Boltzmann's original definition of H, given above, we see that von Neumann had evidently forgotten the minus sign.)

It is interesting that von Neumann, who was one of the pioneers both of quantum mechanics and of information processing, apparently did not consider the idea of processing information in a uniquely quantum-mechanical fashion. Von Neumann had many things on his mind, however – game theory, bomb building, the workings of the brain, etc. – and can be forgiven for failing to make the connection. Another reason that von Neumann may not have thought of quantum computation was that, in his research into computational devices, or 'organs', as he called them, he had evidently reached the impression that computation intrinsically involved dissipation, a process that is inimical to quantum information processing [18]. This impression, if von Neumann indeed had it, is false, as will now be seen.

Reversible Computation

The date of Shannon's paper is usually taken to be the beginning of the study of information theory as a distinct field of inquiry. The second half of the twentieth century saw a huge explosion in the study of information, computation, and communication. The next step towards quantum information processing took place in the early 1960s. Until that point, there was an impression, fostered by von Neumann amongst others, that computation was intrinsically irreversible: according to this view, information was necessarily lost or discarded in the course of computation. For example, a logic gate such as an *AND* gate takes in two bits of information as input, and returns only one bit as output: the output of an *AND* gate is 1 if and only if both inputs are 1, otherwise the output is 0. Because the two input bits cannot be reconstructed from the output bits, an *AND* gate is irreversible. Since computations are typically constructed from *AND*, *OR*, and *NOT* gates (or related irreversible gates such as *NAND*, the combination of an *AND* gate and a *NOT* gate), computations were thought to be intrinsically irreversible, discarding bits as they progress.

In 1960, Rolf Landauer showed that because of the intrinsic connection between information and entropy, when information is discarded in the course of a computation, entropy must be created [19]. That is, when an irreversible logic gate such as an *AND* gate is applied, energy must be dissipated. So far, it seems that von Neumann could be correct. In 1963, however, Yves Lecerf showed that Turing Machines could be constructed in such a way that all their operations were logically reversible [20]. The trick for making computation reversible is record-keeping: one sets up logic circuits in such a way that the values of all bits are recorded and kept. To make an *AND* gate reversible, for example, one adds extra circuitry to keep track of the values of the input to the *AND* gate. In 1973, Charles Bennett, unaware of Lecerf's result, rederived it, and, most importantly, constructed physical models of reversible computation based on molecular systems such as DNA [21]. Ed Fredkin, Tommaso Toffoli, Norman Margolus, and Frank Merkle subsequently made significant contributions to the study of reversible computation [22].

Reversible computation is important for quantum information processing because the laws of physics themselves are reversible. It's this underlying reversibility that is responsible for Landauer's principle: whenever a logically irreversible process such as an *AND* gate takes place, the

information that is discarded by the computation has to go somewhere. In the case of an conventional, transistor-based *AND* gate, the lost information goes into entropy: to operate such an *AND* gate, electrical energy must be dissipated and turned into heat. That is, once the *AND* gate has been performed, then even if the logical circuits of the computer no longer record the values of the inputs to the gate, the microscopic motion of atoms and electrons in the circuit effectively 'remember' what the inputs were. If one wants to perform computation in a uniquely quantum-mechanical fashion, it is important to avoid such dissipation: to be effective, quantum computation should be reversible.

Quantum Computation

In 1980, Paul Benioff showed that quantum mechanical systems such as arrays of spins or atoms could perform reversible computation in principle [23]. Benioff mapped the operation of a reversible Turing machine onto the a quantum system and thus exhibited the first quantum-mechanical model of computation. Benioff's quantum computer was no more computationally powerful than a conventional classical Turing machine, however: it did not exploit quantum weirdness. In 1982, Richard Feynman proposed the first non-trivial application of quantum information processing [24]. Noting that quantum weirdness made it hard for conventional, classical digital computers to simulate quantum systems, Feynman proposed a 'universal quantum simulator' that could efficiently simulate other quantum systems. Feynman's device was not a quantum Turing machine, but a sort of quantum analog computer, whose dynamics could be tuned to match the dynamics of the system to be simulated.

The first model of quantum computation truly to embrace and take advantage of quantum weirdness was David Deutsch's quantum Turing machine of 1985 [25]. Deutsch pointed out that a quantum Turing machine could be designed in such a way as to use the strange and counterintuitive aspects of quantum mechanics to perform computations in ways that classical Turing machines or computers could not. In particular, just as in quantum mechanics it is acceptable (and in many circumstances, mandatory) for an electron to be in two places at once, so in a quantum computer, a quantum bit can take on the values 0 and 1 simultaneously. One possible role for a bit in a computer is as part a program, so that 0 instructs the computer to 'do this' and 1 instructs the computer to 'do that'. If a quantum bit that takes on the values 0 and 1 at the same time is fed into the quantum computer as part of a program, then the quantum computer will 'do this' and 'do that' simultaneously, an effect that Deutsch termed 'quantum parallelism'. Although it would be years before applications of quantum parallelism would be presented, Deutsch's paper marks the beginning of the formal theory of quantum computation.

For almost a decade after the work of Benioff, Feynman, and Deutsch, quantum computers remained a curiosity. Despite the development of a few simple algorithms (described in greater detail below) that took advantage of quantum parallelism, no compelling application of quantum computation had been discovered. In addition, the original models of quantum computation were highly abstract: as Feynman noted [24], no one had the slightest notion of how to build a quantum computer. Absent a 'killer ap', and a physical implementation, the field of quantum computation languished.

That languor dissipated rapidly with Peter Shor's discovery in 1994 that quantum computers could be used to factor large numbers [26]. That is, given the product r of two large prime numbers, a quantum computer could find the factors p and q such that $pq = r$. While it might not appear so instantaneously, solving this problem is indeed a 'killer ap'. Solving the factoring problem is the key to breaking 'public-key' cryptosystems. Public-key cryptosystems are a widely used method for secure communication. Suppose that you wish to buy something from me over the internet, for example. I openly send you a public key consisting of the number r. The public key is not a secret: anyone may know it. You use the public key to encrypt your credit card information, and send me that encrypted information. To decrypt that information, I need to employ the 'private keys' p and q. The security of public-key cryptography thus depends on the factoring problem being hard: to obtain the private keys p and q from the public key r, one must factor the public key.

If quantum computers could be built, then public-key cryptography was no longer secure. This fact excited considerable interest among code breakers, and some consternation within organizations, such as security agencies, whose job it is to keep secrets. Compounding this interest and consternation was the fact that the year before, in 1993, Lloyd had shown how quantum computers could be built using techniques of electromagnetic resonance together with 'off-the shelf' components such as atoms, quantum dots, and lasers [27]. In 1994, Ignacio Cirac and Peter Zoller proposed a technique for building quantum computers using ion traps [28]. These designs for quantum computers quickly resulted in small prototype quantum computers and quantum logic gates being constructed by David Wineland [29], and Jeff Kimble [30]. In 1996, Lov Grover discovered that quantum comput-

ers could search databases significantly faster than classical computers, another potentially highly useful application [31]. By 1997, simple quantum algorithms had been performed using nuclear magnetic resonance based quantum information processing [32,33,34]. The field of quantum computation was off and running.

Since 1994, the field of quantum computation has expanded dramatically. The decade between the discovery of quantum computation and the development of the first applications and implementations saw only a dozen or so papers published in the field of quantum computation. As of the date of publication of this article, it is not uncommon for a dozen papers on quantum computation to be posted on the Los Alamos preprint archive (ArXiv) every day.

Quantum Communication

While the idea of quantum computation was not introduced until 1980, and not fully exploited until the mid-1990s, quantum communication has exhibited a longer and steadier advance. By the beginning of the 1960s, J.P. Gordon [35] and Lev Levitin [36] had begun to apply quantum mechanics to the analysis of the capacity of communication channels. In 1973, Alexander Holevo derived the capacity for quantum mechanical channels to transmit classical information [37] (the Holevo–Schumacher–Westmoreland theorem [38,39]). Because of its many practical applications, the so-called 'bosonic' channel has received a great deal of attention over the years [40]. Bosonic channels are quantum communication channels in which the medium of information exchange consists of bosonic quantum particles, such as photons or phonons. That is, bosonic channels include communication channels that use electromagnetic radiation, from radio waves to light, or sound.

Despite many attempts, it was not until 1993 that Horace Yuen and Masanao Ozawa derived the capacity of the bosonic channel, and their result holds only in the absence of noise and loss [41]. The capacity of the bosonic channel in the presence of loss alone was not derived until 2004 [42], and the capacity of this most important of channels in the presence of noise and loss is still unknown [43].

A second use of quantum channels is to transmit *quantum* information, rather than classical information. The requirements for transmitting quantum information are more stringent than those for transmitting classical information. To transmit a classical bit, one must end up sending a 0 or a 1. To transmit a quantum bit, by contrast, one must also faithfully transmit states in which the quantum bit registers 0 and 1 simultaneously. The quantity which governs the capacity of a channel to transmit quantum information is called the coherent information [44,45]. A particularly intriguing method of transmitting quantum information is *teleportation* [46]. Quantum teleportation closely resembles the teleportation process from the television series Star Trek. In Star Trek, entities to be teleported enter a special booth, where they are measured and dematerialized. Information about the composition of the entities is then sent to a distant location, where the entities rematerialize.

Quantum mechanics at first seems to forbid Trekkian teleportation, for the simple reason that it is not possible to make a measurement that reveals an arbitrary unknown quantum state. Worse yet, any attempt to reveal that state is likely to destroy it. Nonetheless, if one adds just one ingredient to the protocol, quantum teleportation is indeed possible. That necessary ingredient is entanglement.

In quantum teleportation, an entity such as a quantum bit is to be teleported from Alice at point A to Bob at point B. For historical reasons, in communication protocols the sender of information is called Alice and the receiver is called Bob; an eavesdropper on the communication process is called Eve. Alice and Bob possess prior entanglement in the form of a pair of Einstein–Podolsky–Rosen particles. Alice performs a suitable measurement (described in detail below) on the qubit to be teleported together with her EPR particle. This measurement destroys the state of the particle to be teleported ('dematerializing' it), and yields two classical bits of information, which Alice sends to Bob over a conventional communication channel. Bob then performs a transformation on his EPR particle. The transformation Bob performs is a function of the information he receives from Alice: there are four possible transformations, one for each of the four possible values of the two bits he has received. After the Bob has performed his transformation of the EPR particle, the state of this particle is now guaranteed to be the same as that of the original qubit that was to be teleported.

Quantum teleportation forms an integral part of quantum communication and of quantum computation. Experimental demonstrations of quantum teleportation have been performed with photons and atoms as the systems whose quantum states are to be teleported [47,48]. At the time of the writing of this article, teleportation of larger entities such as molecules, bacteria, or human beings remains out of reach of current quantum technology.

Quantum Cryptography

A particularly useful application of the counterintuitive features of quantum mechanics is quantum cryptogra-

phy [49,50,51]. Above, it was noted that Shor's algorithm would allow quantum computers to crack public-key cryptosystems. In the context of code breaking, then, quantum information processing is a disruptive technology. Fortunately, however, if quantum computing represents a cryptographic disease, then quantum communication represents a cryptographic cure. The feature of quantum mechanics that no measurement can determine an unknown state, and that almost any measurement will disturb such a state, can be turned into a protocol for performing quantum cryptography, a method of secret communication whose security is guaranteed by the laws of physics.

In the 1970s, Stephen Wiesner developed the concept of quantum conjugate coding, in which information can be stored on two conjugate quantum variables, such as position and momentum, or linear or helical polarization [49]. In 1984, Charles Bennett and Gilles Brassard turned Wiesner's quantum coding concept into a protocol for quantum cryptography [50]: by sending suitable states of light over a quantum communication channel, Alice and Bob can build up a shared secret key. Since any attempt of Eve to listen in on their communication must inevitably disturb the states sent, Alice and Bob can determine whether Eve is listening in, and if so, how much information she has obtained. By suitable privacy amplification protocols, Alice and Bob can distill out secret key that they alone share and which the laws of physics guarantee is shared by no one else. In 1990 Artur Ekert, unaware of Wiesner, Bennett, and Brassard's work, independently derived a protocol for quantum cryptography based on entanglement [51].

Commercial quantum cryptographic systems are now available for purchase by those who desire secrecy based on the laws of physics, rather than on how hard it is to factor large numbers. Such systems represent the application of quantum information processing that is closest to every day use.

The Future

Quantum information processing is currently a thriving scientific field, with many open questions and potential applications. Key open questions include,

- Just what can quantum computers do better than classical computers? They can apparently factor large numbers, search databases, and simulate quantum systems better than classical computers. That list is quite short, however. What is the full list of problems for which quantum computers offer a speed up?
- How can we build large scale quantum computers? Lots of small scale quantum computers, with up to a dozen bits, have been built and operated. Building large scale quantum computers will require substantial technological advances in precision construction and control of complex quantum systems. While advances in this field have been steady, we're still far away from building a quantum computer that could break existing public-key cryptosystems.
- What are the ultimate physical limits to communication channels? Despite many decades of effort, fundamental questions concerning the capacity of quantum communication channels remain unresolved.

Quantum information processing is a rich stream with many tributaries in the fields of engineering, physics, and applied mathematics. Quantum information processing investigates the physical limits of computation and communication, and it devises methods for reaching closer to those limits, and someday perhaps to attain them.

Quantum Mechanics

In order to understand quantum information processing in any non-trivial way, some math is required. As Feynman said, "… it is impossible to explain honestly the beauties of the laws of nature in a way that people can feel, without their having some deep understanding of mathematics. I am sorry, but this seems to be the case" [52]. The counterintuitive character of quantum mechanics makes it even more imperative to use mathematics to understand the subject. The strange consequences of quantum mechanics arise directly out of the underlying mathematical structure of the theory. It is important to note that every bizarre and weird prediction of quantum mechanics that has been experimentally tested has turned out to be true. The mathematics of quantum mechanics is one of the most trustworthy pieces of science we possess.

Luckily, this mathematics is also quite simple. To understand quantum information processing requires only a basic knowledge of linear algebra, that is, of vectors and matrices. No calculus is required. In this section a brief review of the mathematics of quantum mechanics is presented, along with some of its more straightforward consequences. The reader who is familiar with this mathematics can safely skip to the following sections on quantum information. Readers who desire further detail are invited to consult reference [1].

Qubits

The states of a quantum system correspond to vectors. In a quantum bit, the quantum logic state 0 corresponds to a two-dimensional vector, $\binom{1}{0}$, and the quantum logic

state 1 corresponds to the vector $\binom{0}{1}$. It is customary to write these vectors in the so-called 'Dirac bracket' notation:

$$|0\rangle \equiv \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad |1\rangle \equiv \begin{pmatrix} 0 \\ 1 \end{pmatrix}. \tag{1}$$

A general state for a qubit, $|\psi\rangle$, corresponds to a vector $\binom{\alpha}{\beta} = \alpha|0\rangle + \beta|1\rangle$, where α and β are complex numbers such that $|\alpha|^2 + |\beta|^2 = 1$. The requirement that the amplitude squared of the components of a vector sum to one is called 'normalization'. Normalization arises because amplitudes squared in quantum mechanics are related to probabilities. In particular, suppose that one prepares a qubit in the state $|\psi\rangle$, and then performs a measurement whose purpose is to determine whether the qubit takes on the value 0 or 1 (such measurements will be discussed in greater detail below). Such a measurement will give the outcome 0 with probability $|\alpha|^2$, and will give the outcome 1 with probability $|\beta|^2$. These probabilities must sum to one.

The vectors $|0\rangle, |1\rangle, |\psi\rangle$ are column vectors: we can also define the corresponding row vectors,

$$\langle 0| \equiv \begin{pmatrix} 1 & 0 \end{pmatrix}, \quad \langle 1| \equiv \begin{pmatrix} 0 & 1 \end{pmatrix}, \quad \langle \psi| \equiv \begin{pmatrix} \bar{\alpha} & \bar{\beta} \end{pmatrix}. \tag{2}$$

Note that creating the row vector $\langle\psi|$ involves both transposing the vector and taking the complex conjugate of its entries. This process is called Hermitian conjugation, and is denoted by the superscript †, so that $\langle\psi| = |\psi\rangle^\dagger$.

The two-dimensional, complex vector space for a qubit is denoted C^2. The reason for introducing Dirac bracket notation is that this vector space, like all the vector spaces of quantum mechanics, possesses a natural inner product, defined in the usual way by the product of row vectors and column vectors. Suppose $|\psi\rangle = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$ and $|\phi\rangle = \begin{pmatrix} \gamma \\ \delta \end{pmatrix}$, so that $\langle\phi| = \begin{pmatrix} \bar{\gamma} & \bar{\delta} \end{pmatrix}$ The row vector $\langle\phi|$ is called a 'bra' vector, and the column vector $|\psi\rangle$ is called a 'ket' vector. Multiplied together, these vectors form the inner product, or 'bracket',

$$\langle\phi|\psi\rangle \equiv \begin{pmatrix} \bar{\gamma} & \bar{\delta} \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \alpha\bar{\gamma} + \beta\bar{\delta}. \tag{3}$$

Note that $\langle\psi|\psi\rangle = |\alpha|^2 + |\beta^2| = 1$. The definition of the inner product (3) turns the vector space for qubits C^2 into a 'Hilbert space', a complete vector space with inner product. (Completeness means that any convergent sequence of vectors in the space attains a limit that itself lies in the space. Completeness is only an issue for infinite-dimensional Hilbert spaces and will be discussed no further here.)

We can now express probabilities in terms of brackets: $|\langle 0|\psi\rangle|^2 = |\alpha|^2 \equiv p_0$ is the probability that a measurement that distinguishes 0 and 1, made on the state $|\psi\rangle$, yields the output 0. Similarly, $|\langle 1|\psi\rangle|^2 = |\beta|^2 \equiv p_1$ is the probability that the same measurement yields the output 1. Another way to write these probabilities is to define the two 'projectors'

$$\begin{aligned} P_0 &= \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \end{pmatrix} = |0\rangle\langle 0| \\ P_1 &= \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \end{pmatrix} = |1\rangle\langle 1|. \end{aligned} \tag{4}$$

Note that

$$P_0^2 = |0\rangle\langle 0|0\rangle\langle 0| = |0\rangle\langle 0| = P_0. \tag{5}$$

Similarly, $P_1^2 = P_1$. A projection operator or projector P is defined by the condition $P^2 = P$. Written in terms of these projectors, the probabilities p_0, p_1 can be defined as

$$p_0 = \langle\psi|P_0|\psi\rangle, \quad p_1 = \langle\psi|P_1|\psi\rangle. \tag{6}$$

Note that $\langle 0|1\rangle = \langle 1|0\rangle = 0$: the two states $|0\rangle$ and $|1\rangle$ are orthogonal. Since any vector $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$ can be written as a linear combination, or superposition, of $|0\rangle$ and $|1\rangle$, $\{|0\rangle, |1\rangle\}$ make up an orthonormal basis for the Hilbert space C^2. From the probabilistic interpretation of brackets, we see that orthogonality implies that a measurement that distinguishes between 0 and 1, made on the state $|0\rangle$, will yield the output 0 with probability 1 ($p_0 = 1$), and will never yield the output 1 ($p_1 = 0$). In quantum mechanics, orthogonal states are reliably distinguishable.

Higher Dimensions

The discussion above applied to qubits. More complicated quantum systems lie in higher dimensional vector spaces. For example, a 'qutrit' is a quantum system with three distinguishable states $|0\rangle, |1\rangle, |2\rangle$ that live in the three-dimensional complex vector space C^3. All the mechanisms of measurement and definitions of brackets extend to higher dimensional systems as well. For example, the distinguishability of the three states of the qutrit implies $\langle i|j\rangle = \delta_{ij}$. Many of the familiar systems of quantum mechanics, such as a free particle or a harmonic oscillator, have states that live in *infinite* dimensional Hilbert spaces. For example, the state of a free particle corresponds to a complex valued *function* $\psi(x)$ such that $\int_{-\infty}^{\infty} \bar{\psi}(x)\psi(x)\,\mathrm{d}x = 1$. The probability of finding the particle in the interval between $x = a$ and $x = b$ is then $\int_a^b \bar{\psi}(x)\psi(x)\,\mathrm{d}x$. Infinite dimensional Hilbert spaces

involve subtleties that, fortunately, rarely impinge upon quantum information processing except in the use of bosonic systems as in quantum optics [40].

Matrices

Quantum mechanics is an intrinsically *linear* theory: transformations of states are represented by matrix multiplication. (Nonlinear theories of quantum mechanics can be constructed, but there is no experimental evidence for any intrinsic nonlinearity in quantum mechanics.) Consider the set of matrices U such that $U^\dagger U = Id$, where Id is the identity matrix. Such a matrix is said to be 'unitary'. (For matrices on infinite-dimensional Hilbert spaces, i. e., for linear operators, unitarity also requires $UU^\dagger = Id$.) If we take a normalized vector $|\psi\rangle$, $\langle\psi|\psi\rangle = 1$, and transform it by multiplying it by U, so that $|\psi'\rangle = U|\psi\rangle$, then we have

$$\langle\psi'|\psi\rangle = \langle\psi|U^\dagger U|\psi\rangle = \langle\psi|\psi\rangle = 1\,. \tag{7}$$

That is, unitary transformations U preserve the normalization of vectors. Equation (7) can also be used to show that any U that preserves the normalization of all vectors $|\psi\rangle$ is unitary. Since to be given a physical interpretation in terms of probabilities, the vectors of quantum mechanics must be normalized, the set of unitary transformations represents the set of 'legal' transformations of vectors in Hilbert space. (Below, we'll see that when one adds an environment with which qubits can interact, then the set of legal transformations can be extended.) Unitary transformations on a single qubit make up the set of two-by-two unitary matrices $U(2)$.

Spin and Other Observables

A familiar quantum system whose state space is represented by a qubit is the spin 1/2 particle, such as an electron or proton. The spin of such a particle along a given axis can take on only two discrete values, 'spin up', with angular momentum $\hbar/2$ about that axis, or 'spin down', with angular momentum $-\hbar/2$. Here, $\hbar$ is Planck's reduced constant: $\hbar \equiv h/2\pi = 1.05457\,10^{-34}$ joule-sec. It is conventional to identify the state $|\uparrow\rangle$, spin up along the z-axis, with $|0\rangle$, and the state $|\downarrow\rangle$, spin up along the z-axis, with $|1\rangle$. In this way, the spin of an electron or proton can be taken to register a qubit.

Now that we have introduced the notion of spin, we can introduce an operator or matrix that corresponds to the measurement of spin. Let $P_\uparrow = |\uparrow\rangle\langle\uparrow|$ be the projector onto the state $|\uparrow\rangle$, and let $P_\downarrow = |\downarrow\rangle\langle\downarrow|$ be the projector onto the state $|\downarrow\rangle$. The matrix, or 'operator' corresponding to spin 1/2 along the z-axis is then

$$I_z = \frac{\hbar}{2}(P_\uparrow - P_\downarrow) = \frac{\hbar}{2}\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix} = \frac{\hbar}{2}\sigma_z\,, \tag{8}$$

where $\sigma_z \equiv \begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}$ is called the z Pauli matrix. In what way does I_z correspond to spin along the z-axis? Suppose that one starts out in the state $|\psi\rangle = \alpha|\uparrow\rangle + \beta|\downarrow\rangle$ and then measures spin along the z-axis. Just as in the case of measuring 0 or 1, with probability $p_\uparrow = |\alpha|^2$ one obtains the result $\uparrow$, and with probability $p_\downarrow = |\beta|^2$ one obtains the result $\downarrow$. The *expectation value* for the angular momentum along the z-axis is then

$$\langle I_z\rangle = p_\uparrow(\hbar/2) + p_\downarrow(-\hbar/2) = \langle\psi|I_z|\psi\rangle\,. \tag{9}$$

That is, the expectation value of the observable quantity corresponding to spin along the z-axis is given by taking the bracket of the state $|\psi\rangle$ with the operator I_z corresponding to that observable.

In quantum mechanics, every observable quantity corresponds to an operator. The operator corresponding to an observable with possible outcome values $\{a\}$ is $A = \sum_a a|a\rangle\langle a| = \sum_a aP_a$, where $|a\rangle$ is the state with value a and $P_a = |a\rangle\langle a|$ is the projection operator corresponding to the outcome a. Note that since the outcomes of measurements are real numbers, $A^\dagger = A$: the operators corresponding to observables are Hermitian. The states $\{|a\rangle\}$ are, by definition, distinguishable and so make up an orthonormal set. From the definition of A one sees that $A|a\rangle = a|a\rangle$. That is, the different possible outcomes of the measurement are eigenvalues of A, and the different possible outcome states of the measurement are eigenvectors of A.

If more than one state $|a\rangle_i$ corresponds to the outcome a, then $A = \sum_a aP_a$, where $P_a = \sum_i |a\rangle_i\langle a|$ is the projection operator onto the eigenspace corresponding to the 'degenerate' eigenvalue a. Taking, for the moment, the case of non-degenerate eigenvalues, then the expectation value of an observable A in a particular state $|\chi\rangle = \sum_a \chi_a|a\rangle$ is obtained by bracketing the state about the corresponding operator:

$$\langle A\rangle \equiv \langle\chi|A|\chi\rangle = \sum_a |\chi_a|^2 a = \sum_a p_a a\,, \tag{10}$$

where $p_a = |\chi_a|^2$ is the probability that the measurement yields the outcome a.

Above, we saw that the operator corresponding to spin along the z-axis was $I_z = (\hbar/2)\sigma_z$. What then are the operators corresponding to spin along the x- and y-axes? They

are given by $I_x = (\hbar/2)\sigma_x$ and $I_y = (\hbar/2)\sigma_y$, where σ_x and σ_y are the two remaining Pauli spin matrices out of the trio:

$$\sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \sigma_y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} \quad \sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{11}$$

By the prescription for obtaining expectation values (10), for an initial state $|\chi\rangle$ the expectation values of spin along the x-axis and spin along the y-axis are

$$\langle I_x \rangle = \langle \chi | I_x | \chi \rangle \,, \quad \langle I_y \rangle = \langle \chi | I_y | \chi \rangle \,. \tag{12}$$

The eigenvectors of I_x, σ_x and I_y, σ_y are also easily described. The eigenvector of I_x, σ_x corresponding to spin up along the x-axis is $|\rightarrow\rangle = \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}$, while the eigenvector of I_x, σ_x corresponding to spin down along the x-axis is $|\leftarrow\rangle = \begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix}$. Note that these eigenvectors are orthogonal and normalized – they make up an orthonormal set. It's easy to verify that, $\sigma_x|\uparrow\rangle = +1|\uparrow\rangle$, and $\sigma_x|\downarrow\rangle = -1|\downarrow\rangle$, so the eigenvalues of σ_x are ± 1. The eigenvalues of $I_x = (\hbar/2)\sigma_x$ are $\pm\hbar/2$, the two different possible values of angular momentum corresponding to spin up or spin down along the x-axis. Similarly, the eigenvector of I_y, σ_y corresponding to spin up along the y-axis is $|\otimes\rangle = \begin{pmatrix} 1/\sqrt{2} \\ i/\sqrt{2} \end{pmatrix}$, while the eigenvector of I_y, σ_y corresponding to spin down along the y-axis is $|\odot\rangle = \begin{pmatrix} i/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}$. (Here, in deference to the right-handed coordinate system that we are implicitly adopting, $\otimes$ corresponds to an arrow heading away from the viewer, and $\odot$ corresponds to an arrow heading towards the viewer.)

Rotations and SU(2)

The Pauli matrices $\sigma_x, \sigma_y, \sigma_z$ play a crucial role not only in characterizing the measurement of spin, but in generating rotations as well. Because of their central role in describing qubits in general, and spin in particular, several more of their properties are elaborated here. Clearly, $\sigma_i = \sigma_i^\dagger$: Pauli matrices are Hermitian. Next, note that

$$\sigma_x^2 = \sigma_y^2 = \sigma_z^2 = Id = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{13}$$

Since $\sigma_i = \sigma_i^\dagger$, and $\sigma_i^2 = Id$, it's also the case that $\sigma_i^\dagger \sigma_i = Id$: that is, the Pauli matrices are unitary. Next, defining the commutator of two matrices A and B to be $[A, B] = AB - BA$, it is easy to verify that $[\sigma_x, \sigma_y] = 2i\sigma_z$. Cyclic permutations of this identity also hold, e. g., $[\sigma_z, \sigma_x] = 2i\sigma_y$.

Now introduce the concept of a rotation. The operator $e^{-i(\theta/2)\sigma_x}$ corresponds to a rotation by an angle θ about the x-axis. The analogous operators with x replaced by y or z are expressions for rotations about the y- or z-axes. Exponentiating matrices may look at first strange, but exponentiating Pauli matrices is significantly simpler. Using the Taylor expansion for the matrix exponential, $e^A = Id + A + A^2/2! + A^3/3! + \ldots$, and employing the fact that $\sigma_j^2 = Id$, one obtains

$$e^{-i(\theta/2)\sigma_j} = \cos(\theta/2)Id - i\sin(\theta/2)\sigma_j \,. \tag{14}$$

It is useful to verify that the expression for rotations (14) makes sense for the states we have defined. For example, rotation by π about the x-axis should take the state $|\uparrow\rangle$, spin z up, to the state $|\downarrow\rangle$, spin z down. Inserting $\theta = \pi$ and $j = x$ in (14), we find that the operator corresponding to this rotation is the matrix $-i\sigma_x$. Multiplying $|\uparrow\rangle$ by this matrix, we obtain

$$-i\sigma_x|\uparrow\rangle = -i\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1 \\ 0 \end{pmatrix} = -i\begin{pmatrix} 0 \\ 1 \end{pmatrix} = -i|\downarrow\rangle. \tag{15}$$

The rotation does indeed take $|\uparrow\rangle$ to $|\downarrow\rangle$, but it also introduces an overall phase of $-i$.

What does this overall phase do? The answer is Nothing! Or, at least, nothing observable. Overall phases cannot change the expectation value of any observable. Suppose that we compare expectation values for the state $|\chi\rangle$ and for the state $|\chi'\rangle = e^{i\phi}|\chi\rangle$ for some observable corresponding to an operator A. We have

$$\langle \chi | A | \chi \rangle = \langle \chi | e^{-i\phi} A e^{i\phi} | \chi \rangle = \langle \chi' | A | \chi' \rangle \,. \tag{16}$$

Overall phases are undetectable. Keeping the undetectability of overall phases in mind, it is a useful exercise to verify that other rotations perform as expected. For example, a rotation by $\pi/2$ about the x-axis takes $|\otimes\rangle$, spin up along the y-axis, to $|\uparrow\rangle$, together with an overall phase.

Once rotation about the x, $>y$, and z axes have been defined, it is straightforward to construct rotations about any axis. Let $\hat{\iota} = (\iota_x, \iota_y, \iota_z)$, $\iota_x^2 + \iota_y^2 + \iota_z^2 = 1$, be a unit vector along the $\hat{\iota}$ direction in ordinary three-dimensional space. Define $\sigma_{\hat{\iota}} = \iota_x\sigma_x + \iota_y\sigma_y + \iota_z\sigma_z$ to be the *generalized Pauli matrix* associated with the unit vector $\hat{\iota}$. It is easy to verify that $\sigma_{\hat{\iota}}$ behaves like a Pauli matrix, e. g., $\sigma_{\hat{\iota}}^2 = Id$. Rotation by θ about the $\hat{\iota}$ axis then corresponds to an operator $e^{-i(\theta/2)\sigma_{\hat{\iota}}} = \cos(\theta/2)Id - i\sin(\theta/2)\sigma_{\hat{\iota}}$. Once again, it is a useful exercises to verify that such rotations behave as expected. For example, a rotation by π about the

$(1/\sqrt{2}, 0, 1/\sqrt{2})$ axis should 'swap' $|\uparrow\rangle$ and $|\rightarrow\rangle$, up to some phase.

The set of rotations of the form $e^{-i\theta/2\sigma_i}$ forms the *group* $SU(2)$, the set of complex 2 by 2 unitary matrices with determinant equal to 1. It is instructive to compare this group with the 'conventional' group of rotations in three dimensions, $SO(3)$. $SO(3)$ is the set of real 3 by 3 matrices with orthonormal rows/columns and determinant 1. In $SO(3)$, when one rotates a vector by 2π, the vector returns to its original state: a rotation by 2π corresponds to the 3 by 3 identity matrix. In $SU(2)$, rotating a vector by 2π corresponds to the transformation $-Id$: in rotating by 2π, the vector acquires an overall phase of -1. As will be seen below, the phase of -1, while unobservable for single qubit rotations, can be, and has been observed in two-qubit operations. To return to the original state, with no phase, one must rotate by 4π. A macroscopic, classical version of this fact manifests itself when one grasps a glass of water firmly in the palm of one's hand and rotates one's arm and shoulder to rotate the glass without spilling it. A little experimentation with this problem shows that one must rotate glass and hand around twice to return them to their initial orientation.

Why Quantum Mechanics?

Why is the fundamental theory of nature, quantum mechanics, a theory of complex vector spaces? No one knows for sure. One of the most convincing explanations came from Aage Bohr, the son of Niels Bohr and a Nobel laureate in quantum mechanics in his own right [53]. Aage Bohr pointed out that the basic mathematical representation of *symmetry* consists of complex vector spaces. For example, while the apparent symmetry group of rotations in three dimensional space is the real group $SO(3)$, the actual underlying symmetry group of space, as evidenced by rotations of quantum-mechanical spins, is $SU(2)$: to return to the same state, one has to go around not once, but twice. It is a general feature of complex, continuous groups, called 'Lie groups' after Sophus Lie, that their fundamental representations are complex. If quantum mechanics is a manifestation of deep, underlying symmetries of nature, then it should come as no surprise that quantum mechanics is a theory of transformations on complex vector spaces.

Density Matrices

The review of quantum mechanics is almost done. Before moving on to quantum information processing proper, two topics need to be covered. The first topic is how to deal with uncertainty about the underlying state of a quantum system. The second topic is how to treat two or more quantum systems together. These topics turn out to possess a strong connection which is the source of most counterintuitive quantum effects.

Suppose that don't know exactly what state a quantum system is in. Say, for example, it could be in the state $|0\rangle$ with probability p_0 or in the state $|1\rangle$ with probability p_1. Note that this state is not the same as a quantum superposition, $\sqrt{p_0}|0\rangle + \sqrt{p_1}|1\rangle$, which is a definite state with spin oriented in the $x-z$ plane. The expectation value of an operator A when the underlying state possesses the uncertainty described is

$$\langle A\rangle = p_0\langle 0|A|0\rangle + p_1\langle 1|A|1\rangle = \mathrm{tr}\rho A\,, \tag{17}$$

where $\rho = p_0|0\rangle\langle 0| + p_1|1\rangle\langle 1|$ is the *density matrix* corresponding to the uncertain state. The density matrix can be thought of as the quantum mechanical analogue of a probability distribution.

Density matrices were developed to provide a quantum mechanical treatment of statistical mechanics. A famous density matrix is that for the canonical ensemble. Here, the energy state of a system is uncertain, and each energy state $|E_i\rangle$ is weighted by a probability $p_i = e^{-E_i/k_BT}/Z$, where $Z = \sum_i e^{-E_i/k_BT}$ is the partition function. Z is needed to normalize the probabilities $\{p_i\}$ so that $\sum_i p_i = 1$. The density matrix for the canonical ensemble is then $\rho_C = (1/Z)\sum_i e^{-E_i/k_BT}|E_i\rangle\langle E_i|$. The expectation value of any operator, e. g., the energy operator H (for 'Hamiltonian') is then given by $\langle H\rangle = \mathrm{tr}\rho_C H$.

Multiple Systems and Tensor Products

To describe two or more systems requires a formalism called the tensor product. The Hilbert space for two qubits is the space $C^2 \otimes C^2$, where $\otimes$ is the tensor product. $C^2 \otimes C^2$ is a four-dimensional space spanned by the vectors $|0\rangle\otimes|0\rangle, |0\rangle\otimes|1\rangle, |1\rangle\otimes|0\rangle, |1\rangle\otimes|1\rangle$. (To save space these vectors are sometimes written $|0\rangle|0\rangle, |0\rangle|1\rangle, |1\rangle|0\rangle, |1\rangle|1\rangle$, or even more compactly, $|00\rangle, |01\rangle, |10\rangle, |11\rangle$. Care must be taken, however, to make sure that this notation is unambiguous in a particular situation.) The tensor product is *multilinear*: in performing the tensor product, the distributive law holds. That is, if $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$, and $|\phi\rangle = \gamma|0\rangle + \delta|1\rangle$, then

$$\begin{aligned}|\psi\rangle\otimes|\phi\rangle &= (\alpha|0\rangle + \beta|1\rangle)\otimes(\gamma|0\rangle + \delta|1\rangle)\\ &= \alpha\gamma|0\rangle\otimes|0\rangle + \alpha\delta|0\rangle\otimes|1\rangle\\ &\quad + \beta\gamma|1\rangle\otimes|0\rangle + \beta\delta|1\rangle\otimes|1\rangle\,.\end{aligned} \tag{18}$$

A tensor is a thing with slots: the key point to keep track of in tensor analysis is which operator or vector acts on which slot. It is often useful to label the slots, e. g., $|\psi\rangle_1\otimes|\phi\rangle_2$ is

a tensor product vector in which $|\psi\rangle$ occupies slot 1 and $|\phi\rangle$ occupies slot 2.

One can also define the tensor product of operators or matrices. For example, $\sigma_x^1 \otimes \sigma_z^2$ is a tensor product operator with σ_x in slot 1 and σ_z in slot 2. When this operator acts on a tensor product vector such as $|\psi\rangle_1 \otimes |\phi\rangle_2$, the operator in slot 1 acts on the vector in that slot, and the operator in slot 2 acts on the vector in that slot:

$$(\sigma_x^1 \otimes \sigma_z^2)(|\psi\rangle_1 \otimes |\phi\rangle_2) = (\sigma_x^1|\psi\rangle_1) \otimes (\sigma_z^2|\phi\rangle_2)\,. \quad (19)$$

The No-Cloning Theorem

Now that tensor products have been introduced, one of the most famous theorems of quantum information – the no-cloning theorem – can immediately be proved [54]. Classical information has the property that it can be copied, so that $0 \to 00$ and $1 \to 11$. How about quantum information? Does there exist a procedure that allows one to take an arbitrary, unknown state $|\psi\rangle$ to $|\psi\rangle \otimes |\psi\rangle$? Can you clone a quantum? As the title to this section indicates, the answer to this question is No.

Suppose that you could clone a quantum. Then there would exist a unitary operator U_C that would take the state

$$|\psi\rangle \otimes |0\rangle \to U_C|\psi\rangle \otimes |0\rangle = |\psi\rangle \otimes |\psi\rangle\,, \quad (20)$$

for any initial state $|\psi\rangle$. Consider another state $|\phi\rangle$. Since U_C is supposed to clone any state, we have then we would also have $U_C|\phi\rangle \otimes |0\rangle = |\phi\rangle \otimes |\phi\rangle$. If U_C exists, then, the following holds for any states $|\psi\rangle$, $|\phi\rangle$:

$$\begin{aligned}\langle\phi|\psi\rangle &= ({}_1\langle\phi| \otimes {}_2\langle 0|)(|\psi\rangle_1 \otimes |0\rangle_2)\\ &= ({}_1\langle\phi| \otimes {}_2\langle 0|)(U_C^\dagger U_C)(|\psi\rangle_1 \otimes |0\rangle_2)\\ &= ({}_1\langle\phi| \otimes {}_2\langle 0|U_C^\dagger)(U_C|\psi\rangle_1 \otimes |0\rangle_2)\\ &= ({}_1\langle\phi| \otimes {}_2\langle\phi|)(|\psi\rangle_1 \otimes |\psi\rangle_2)\\ &\quad \cdot ({}_1\langle\phi|\psi\rangle_1)({}_2\langle\phi|\psi\rangle_2)\\ &= \langle\phi|\psi\rangle^2\,,\end{aligned} \quad (21)$$

where we have used the fact that U_C is unitary so that $U_C^\dagger U_C = Id$. So if cloning is possible, then $\langle\phi|\psi\rangle = \langle\phi|\psi\rangle^2$ for *any* two vectors $|\psi\rangle$ and $|\phi\rangle$. But this is impossible, as it implies that $\langle\phi|\psi\rangle$ equals either 0 or 1 for all $|\psi\rangle$, $|\phi\rangle$, which is certainly not true. You can't clone a quantum.

The no-cloning theorem has widespread consequences. It is responsible for the efficacy of quantum cryptography, which will be discussed in greater detail below. Suppose that Alice sends a state $|\psi\rangle$ to Bob. Eve wants to discover what state this is, without Alice or Bob uncovering her eavesdropping. That is, she would like to make a copy of $|\psi\rangle$ and send the original state $|\psi\rangle$ to Bob. The no-cloning theorem prevents her from doing so: any attempt to copy $|\psi\rangle$ will necessarily perturb the state. An 'optimal cloner' is a transformation that does the best possible job of cloning, given that cloning is impossible [55].

Reduced Density Matrices

Suppose that one makes a measurement corresponding to an observable A_1 on the state in slot 1. What operator do we take the bracket of to get the expectation value? The answer is $A_1 \otimes Id_2$: we have to put the identity in slot 2. The expectation value for this measurement for the state $|\psi\rangle_1 \otimes |\phi\rangle_2$ is then

$$\begin{aligned}&{}_1\langle\psi| \otimes {}_2\langle\phi|A_1 \otimes Id_2|\psi\rangle_1 \otimes |\phi\rangle_2 =\\ &\quad {}_1\langle\psi|A_1|\psi\rangle_1 \otimes {}_2\langle\phi|Id_2|\phi\rangle_2 = {}_1\langle\psi|A_1|\psi\rangle_1\,.\end{aligned} \quad (22)$$

Here we have used the rule that operators in slot 1 act on vectors in slot 1. Similarly, the operators in slot 2 act on vectors in slot 2. As always, the key to performing tensor manipulations is to keep track of what is in which slot. (Note that the tensor product of two numbers is simply the product of those numbers.)

In ordinary probability theory, the probabilities for two sets of events labeled by i and j is given by a joint probability distribution $p(ij)$. The probabilities for the first set of events on their own is obtained by averaging over the second set: $p(i) = \sum_j p(ij)$ is the marginal distribution for the first set of events labeled by i. In quantum mechanics, the analog of a probability distribution is density matrix. Two systems 1 and 2 are described by a joint density matrix ρ_{12}, and system 1 on its own is described by a 'reduced' density matrix ρ_1.

Suppose that systems 1 and 2 are in a state described by a density matrix

$$\rho_{12} = \sum_{ii'jj'} \rho_{ii'jj'}|i\rangle_1\langle i'| \otimes |j\rangle_2\langle j'|\,, \quad (23)$$

where $\{|i\rangle_1\}$ and $\{|j\rangle_2\}$ are orthonormal bases for systems 1 and 2 respectively. As in the previous paragraph, the expectation value of a measurement made on ρ_{12} alone is given by $\mathrm{tr}\rho_{12}(A_1 \otimes Id_2)$. Another way to write such expectation values is to define the *reduced density matrix*,

$$\begin{aligned}\rho_1 = \mathrm{tr}_2\rho_{12} &\equiv \sum_{ii'jj'} \rho_{ii'jj'}|i\rangle_1\langle i'| \otimes_2 \langle j'|j\rangle_2\\ &= \sum_{ii'j} \rho_{ii'jj}|i\rangle_1\langle i'|\,.\end{aligned} \quad (24)$$

Equation (24) describes the partial trace tr_2 over system 2. In other words, if ρ_{12} has components, $\{\rho_{ii'jj'}\}$,

then reduced density matrix $\rho_1 = \mathrm{tr}_2\rho_{12}$ has components $\{\sum_j \rho_{ii'jj}\}$. The expectation value of a measurement A made on the first system alone is then simply $\langle A\rangle = \mathrm{tr}\rho_1 A$. Just as in ordinary probability theory, where the marginal distribution for system 1 is obtained by averaging over the state of system 2, so in quantum mechanics the reduced density matrix that describes system 1 is obtained by tracing over the state of system 2.

Entanglement

One of the central features of quantum information processing is entanglement. Entanglement is a peculiarly quantum-mechanical form of correlation between quantum systems, that has no classical analogue. Entanglement lies at the heart of the various speedups and enhancements that quantum information processing offers over classical information processing.

A pure state $|\psi\rangle_{12}$ for two systems 1 and 2 is entangled if the reduced density matrix for either system taken on its own has non-zero entropy. In particular, the reduced density matrix for system 1 is $\rho_1 = \mathrm{tr}_2\rho_{12}$, where $\rho_{12} = |\psi\rangle_{12}\langle\psi|$. The entropy of this density matrix is $S(\rho_1) = -\mathrm{tr}\rho_1 \log_2 \rho_1$. For pure states, the entropy of ρ_1 is equal to the entropy of ρ_2 and is a good measure of the degree of entanglement between the two systems. $S(\rho_1) = S(\rho_2)$ measures the number of 'e-bits' of entanglement between systems 1 and 2.

A mixed state ρ_{12} for 1 and 2 is entangled if it is not separable. A density matrix is separable if it can be written $\rho_{12} = \sum_j p_j \rho_1^j \otimes \rho_2^j$. In other words, a separable state is one that can be written as a classical mixture of uncorrelated states. The correlations in a separable state are purely classical.

Entanglement can take a variety of forms and manifestations. The key to understanding those forms is the notion of Local Operations and Classical Communication (LOCC) [56]. Local operations such as unitary transformations and measurement, combined with classical communication, can not, on their own, create entanglement. If one state can be transformed into another via local operations and classical communication, then the first state is 'at least as entangled' as the second. LOCC can then be used to categorize the different forms of entanglement.

Distillable entanglement is a form of entanglement that can be transformed into pure-state entanglement [57]. Systems 1 and 2 posses d qubits worth of distillable entanglement if local operations and classical communication can transform their state into a pure state that contains d e-bits (possibly with some leftover 'junk' in a separate quantum register). Systems that are non-separable, but that possess no distillable entanglement are said to possess bound entanglement [58].

The entanglement of formation for a state ρ_{12} is equal to the minimum number of e-bits of pure-state entanglement that are required to create ρ_{12} using only local operations and classical control [59]. The entanglement of formation of ρ_{12} is greater than or equal to ρ_{12}'s distillable entanglement. A variety of entanglement measures exist. Each one is useful for different purposes. Squashed entanglement, for example, plays an important role in quantum cryptography [60]. (Squashed entanglement is a notion of entanglement based on conditional information.)

One of the most interesting open questions in quantum information theory is the definition of entanglement for marti-partite systems consisting of more than two subsystems. Here, even in the case of pure states, no unique definition of entanglement exists.

Entanglement plays a key role in quantum computation and quantum communication. Before turning to those fields, however, it is worth while investigating the strange and counterintuitive features of entanglement.

Quantum Weirdness

Entanglement is the primary source of what for lack of a better term may be called 'quantum weirdness'. Consider the two-qubit state

$$|\psi\rangle_{12} = \frac{1}{\sqrt{2}}(|0\rangle_1 \otimes |1\rangle_2 - |1\rangle_1|0\rangle_2)\,. \tag{25}$$

This state is called the 'singlet' state: if the two qubits correspond to two spin 1/2 particles, as described above, so that $|0\rangle$ is the spin z up state and $|1\rangle$ is the spin z down state, then the singlet state is the state with zero angular momentum. Indeed, rewriting $|\psi\rangle_{12}$ in terms of spin as

$$|\psi\rangle_{12} = \frac{1}{\sqrt{2}}(|\uparrow\rangle_1 \otimes |\downarrow\rangle_2 - |\downarrow\rangle_1|\uparrow\rangle_2)\,. \tag{26}$$

one sees that if one makes a measurement of spin z, then if the first spin has spin z up, then the second spin has spin z down, and vice versa.

If one decomposes the state in terms of spin along the x-axis, $|\rightarrow\rangle = (1/\sqrt{2})(|\uparrow\rangle + |\downarrow\rangle)$, $|\leftarrow\rangle = (1/\sqrt{2})\cdot(|\uparrow\rangle - |\downarrow\rangle)$, then $|\psi\rangle_{12}$ can be rewritten

$$|\psi\rangle_{12} = \frac{1}{\sqrt{2}}(|\rightarrow\rangle_1 \otimes |\leftarrow\rangle_2 - |\leftarrow\rangle_1 \otimes |\rightarrow\rangle_2)\,. \tag{27}$$

Similarly, rewriting in terms of spin along the y-axis, we obtain

$$|\psi\rangle_{12} = \frac{1}{\sqrt{2}}(|\otimes\rangle_1|\odot\rangle_2 - |\odot\rangle_1|\otimes\rangle_2)\,, \tag{28}$$

where $|\otimes\rangle$ is the state with spin up along the y-axis and $|\odot\rangle$ is the state with spin down along the y-axis. No matter what axis one decomposes the spin about, if the first spin has spin up along that axis then the second spin has spin down along that axis, and vice versa. The singlet state has angular momentum zero about every axis.

So far, this doesn't sound too strange. The singlet simply behaves the way a state with zero angular momentum should: it is not hard to see that it is the unique two-spin state with zero angular momentum about every axis. In fact, the singlet state exhibits lots of quantum weirdness. Look at the reduced density matrix for spin 1:

$$\rho_1 = \mathrm{tr}_2\rho_{12} = \mathrm{tr}_2|\psi\rangle_{12}\langle\psi| = \frac{1}{2}(|\uparrow\rangle_1\langle\uparrow| + |\downarrow\rangle_1\langle\downarrow| = Id/2\,. \tag{29}$$

That is, the density matrix for spin 1 is in a completely indefinite, or 'mixed' state: nothing is known about whether it is spin up or spin down along any axis. Similarly, spin 2 is in a completely mixed state. This is already a little strange. The two spins together are in a definite, 'pure' state, the singlet state. Classically, if two systems are in a definite state, then each of the systems on its own is in a definite state: the only way to have uncertainty about one of the parts is to have uncertainty about the whole. In quantum mechanics this is not the case: two systems can be in a definite, pure state taken together, while each of the systems on its own is in an indefinite, mixed state. Such systems are said to be *entangled* with each other.

Entanglement is a peculiarly quantum form of correlation. Two spins in a singlet state are highly correlated (or, more precisely, anticorrelated): no matter what axis one measures spin along, one spin will be found to have the opposite spin of the other. In itself, that doesn't sound so bad, but when one makes a measurement on one spin, something funny seems to happen. Both spins start out in a completely indefinite state. Now one chooses to make a measurement of spin 1 along the z-axis. Suppose that one gets the result, spin up. As a result of the measurement, spin 2 is now in a definite state, spin down along the z-axis. If one had chosen to make a measurement of spin 1 along the x-axis, then spin 2 would also be put in a definite state along the x-axis. Somehow, it seems as if one can affect the state of spin 2 by making a measurement of spin 1 on its own. This is what Einstein called 'spooky action at a distance'.

In fact, such measurements involve no real action at a distance, spooky or otherwise. If one could really act on spin 2 by making a measurement on spin 1, thereby changing spin 2's state, then one could send information instantaneously from spin 1 to spin 2 by measuring spin 1 alone. Such instantaneous transmission of information would violate special relativity and give rise to all sorts of paradoxical capabilities, such as the ability to travel backwards in time. Luckily, it is easy to see that it is impossible to send information superluminally using entanglement: no matter what one does to spin 1, the outcomes of measurements on spin 2 are unaffected by that action. In particular, operations on spin 1 correspond to operators of the form $A_1 \otimes Id_2$, while operations on spin 2 correspond to operators of the form $Id_1 \otimes B_2$. The commutator between such operators is

$$[A_1 \otimes Id_2, Id_1 \otimes B_2] = A_1 \otimes B_2 - A_1 \otimes B_2 = 0\,. \tag{30}$$

Since they commute, it doesn't matter if one does something to spin 1 first, and then measures spin 2, or if one measures spin 2 first and then does something to spin 1: the results of the measurement will be the same. That is, nothing one does to spin 1 on its own can effect spin 2.

Nonetheless, entanglement is counterintuitive. One's classical intuition would like to believe that before the measurement, the system to be measured is in a definite state, even if that definite state is unknown. Such a definite state would constitute a 'hidden variable', an unknown, classical value for the measured variable. Entanglement implies that such hidden variables can't exist in any remotely satisfactory form. The spin version of the EPR effect described above is due to David Bohm [61]. Subsequently, John Bell proposed a set of relations, the 'Bell inequalities', that a hidden variable theory should obey [62]. Bell's inequalities are expressed in terms of the probabilities for the outcomes of measurements made on the two spins along different axes.

Suppose that each particle indeed has a particular value of spin along each axis before it is measured. Designate a particle that has spin up along the x-axis, spin down along the y-axis, and spin up along the z-axis by $(x+, y-, z+)$. Designate other possible orientations similarly. In a collection of particles, let $N(x+, y-, z+)$ be the number of particles with orientations $(x+, y-, z+)$. Clearly, $N(x+, y-) = N(x+, y-, z+) + N(x+, y-, z-)$. Now, in a collection of measurements made on pairs of particles, originally in a singlet state, let $\#(x_1+, y_2-)$ be the number of measurements that give the result spin up along the x-axis for particle 1, and spin down along the y-axis for particle 2. Bell showed that for classical particles that actually possess definite values of spin along different axes before measurement, $\#(x_1+, y_2+) \leq \#(x_1+, z_2+) + \#(y_1-, z_2-)$, together with inequalities that are obtained by permuting axes and signs.

Quantum mechanics decisively violates these Bell inequalities: in entangled states like the singlet state, particles simply do not possess definite, but unknown, values of spin before they are measured. Bell's inequalities have been verified experimentally on numerous occasions [13], although not all exotic forms of hidden variables have been eliminated. Those that are consistent with experiment are not very aesthetically appealing however (depending, of course, on one's aesthetic ideals). A stronger set of inequalities than Bell's are the CHSH inequalities (Clauser–Horne–Shimony–Holt), which have also been tested in numerous venues, with the predictions of quantum mechanics confirmed each time [63]. One of weirdest violation of classical intuition can be found in the so-called GHZ experiment, named after Daniel Greenberger, Michael Horne, and Anton Zeilinger [64].

To demonstrate the GHZ paradox, begin with the three-qubit state

$$|\chi\rangle = (1/\sqrt{2})(|\uparrow\uparrow\uparrow\rangle - |\downarrow\downarrow\downarrow\rangle) \tag{31}$$

(note that in writing this state we have suppressed the tensor product $\otimes$ signs, as mentioned above). Prepare this state four separate times, and make four distinct measurements. In the first measurement measure σ_x on the first qubit, σ_y on the second qubit, and σ_y on the third qubit. Assign the value $+1$ to the result, spin up along the axis measured, and -1 to spin down. Multiply the outcomes together. Quantum mechanics predicts that the result of this multiplication will always be $+1$, as can be verified by taking the expectation value $\langle\chi|\sigma_x^1 \otimes \sigma_y^2 \otimes \sigma_y^3|\chi\rangle$ of the operator $\sigma_x^1 \otimes \sigma_y^2 \otimes \sigma_y^3$ that corresponds to making the three individual spin measurements and multiplying their results together.

In the second measurement measure σ_y on the first qubit, σ_x on the second qubit, and σ_y on the third qubit. Multiply the results together. Once again, quantum mechanics predicts that the result will be $+1$. Similarly, in the third measurement measure σ_y on the first qubit, σ_y on the second qubit, and σ_x on the third qubit. Multiply the results together to obtain the predicted result $+1$. Finally, in the fourth measurement measure σ_x on all three qubits and multiply the results together. Quantum mechanics predicts that this measurement will give the result $\langle\chi|\sigma_x^1 \otimes \sigma_x^2 \otimes \sigma_x^3|\chi\rangle = -1$.

So far, these predictions may not seem strange. A moment's reflection, however, will reveal that the results of the four GHZ experiments are completely incompatible with any underlying assignment of values of ± 1 to the spin along the x- and y-axes before the measurement. Suppose that such pre-measurement values existed, and that these are the values revealed by the measurements. Looking at the four measurements, each consisting of three individual spin measurements, one sees that each possible spin measurement appears twice in the full sequence of twelve individual spin measurements. For example, measurement of spin 1 along the x-axis occurs in the first of the four three-fold measurements, and in the last one. Similarly, measurement of spin 3 along the 3-axis occurs in the first and second three-fold measurements. The classical consequence of each individual measurement occurring twice is that the product of all twelve measurements should be $+1$. That is, if measurement of σ_x^1 in the first measurement yields the result -1, it should also yield the result -1 in the fourth measurement. The product of the outcomes for σ_x^1 then gives $(-1) \times (-1) = +1$; similarly, if σ_x^1 takes on the value $+1$ in both measurements, it also contributes $(+1) \times (+1) = +1$ to the overall product. So if each spin possesses a definite value before the measurement, classical mechanics unambiguously predicts that the product of all twelve individual measurements should be $+1$.

Quantum mechanics, by contrast, unambiguously predicts that the product of all twelve individual measurements should be -1. The GHZ experiment has been performed in a variety of different quantum-mechanical systems, ranging from nuclear spins to photons [65,66]. The result: the predictions of classical mechanics are wrong and those of quantum mechanics are correct. Quantum weirdness triumphs.

Quantum Computation

Quantum mechanics has now been treated in sufficient detail to allow us to approach the most startling consequence of quantum weirdness: quantum computation. The central counterintuitive feature of quantum mechanics is quantum superposition: unlike a classical bit, which either takes on the value 0 or the value 1, a quantum bit in the superposition state $\alpha|0\rangle + \beta|1\rangle$ takes on the values 0 and 1 simultaneously. A quantum computer is a device that takes advantage of quantum superposition to process information in ways that classical computers can't. A key feature of any quantum computation is the way in which the computation puts entanglement to use: just as entanglement plays a central role in the quantum paradoxes discussed above, it also lies at the heart of quantum computation.

A classical digital computer is a machine that can perform arbitrarily complex logical operations. When you play a computer game, or operate a spread sheet, all that is going on is that your computer takes in the information from your joy stick or keyboard, encodes that information as a sequence of zeros and ones, and then performs

sequences of simple logical operations one that information. Since the work of George Boole in the first half of the nineteenth century, it is known that any logical expression, no matter how involved, can be broken down into sequences of elementary logical operations such as *NOT*, *AND*, *OR* and *COPY*. In the context of computation, these operations are called 'logic gates': a logic gates takes as input one or more bits of information, and produces as output one or more bits of information. The output bits are a function of the input bits. A *NOT* gate, for example, takes as input a single bit, X, and returns as output the flipped bit, *NOT* X, so that $0 \rightarrow 1$ and $1 \rightarrow 0$. Similarly, an *AND* gate takes in two bits X, Y as input, and returns the output X *AND* Y. X *AND* Y is equal to 1 when both X and Y are equal to 1; otherwise it is equal to 0. That is, an *AND* gate takes $00 \rightarrow 0$, $01 \rightarrow 0$, $10 \rightarrow 0$, and $11 \rightarrow 1$. An *OR* gate takes X, Y to 1 if either X or Y is 1, and to 0 if both X and Y are 0, so that $00 \rightarrow 0$, $01 \rightarrow 1$, $10 \rightarrow 1$, and $11 \rightarrow 1$. A *COPY* gate takes a single input, X, and returns as output two bits X that are copies of the input bit, so that $0 \rightarrow 00$ and $1 \rightarrow 11$.

All elementary logical operations can be built up from *NOT, AND, OR* and *COPY*. For example, implication can be written $A \rightarrow B \equiv A\ OR\ (NOT\ B)$, since $A \rightarrow B$ is false if and only if A is true and B is false. Consequently, any logical expression, e. g.,

$$\Big(\big(A\ AND\ (NOT\ B)\big)\ OR\ \big(C\ AND\ (NOT\ A)\big)\ AND\ \big(NOT(C\ OR\ B)\big)\Big), \quad (32)$$

can be evaluated using *NOT*, *AND*, *OR*, and *COPY* gates, where *COPY* gates are used to supply the different copies of A, B and C that occur in different places in the expression. Accordingly, {*NOT*, *AND*, *OR*, *COPY*} is said to form a 'computationally universal' set of logic gates. Simpler computationally universal sets of logic gates also exist, e. g. {*NAND*, *COPY*}, where $X\ NAND\ Y = NOT(X\ AND\ Y)$.

Reversible Logic

A logic gate is said to be *reversible* if its outputs are a one-to-one function of its inputs. *NOT* is reversible, for example: since $X = NOT(NOT\ X)$, *NOT* is its own inverse. *AND* and *OR* are not reversible, as the value of their two input bits cannot be inferred from their single output. *COPY* is reversible, as its input can be inferred from either of its outputs.

Logical reversibility is important because the laws of physics, at bottom, are reversible. Above, we saw that the time evolution of a closed quantum system (i. e., one that is not interacting with any environment) is given by a unitary transformation: $|\psi\rangle \rightarrow |\psi'\rangle = U|\psi\rangle$. All unitary transformations are invertible: $U^{-1} = U^{\dagger}$, so that $|\psi\rangle = U^{\dagger}U|\psi\rangle = U^{\dagger}|\psi'\rangle$. The input to a unitary transformation can always be obtained from its output: the time evolution of quantum mechanical systems is one-to-one. As noted in the introduction, in 1961, Rolf Landauer showed that the underlying reversibility of quantum (and also of classical) mechanics implied that logically irreversible operations such as *AND* necessarily required physical dissipation [19]: any physical device that performs an *AND* operation must possess additional degrees of freedom (i. e., an environment) which retain the information about the actual values of the inputs of the *AND* gate after the irreversible logical operation has discarded those values. In a conventional electronic computer, those additional degrees of freedom consist of the microscopic motions of electrons, which, as Maxwell and Boltzmann told us, register large amounts of information.

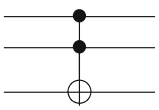
Logic circuits in contemporary electronic circuits consist of field effect transistors, or FETs, wired together to perform *NOT*, *AND*, *OR* and *COPY* operations. Bits are registered by voltages: a FET that is charged at higher voltage registers a 1, and an uncharged FET at Ground voltage registers a 0. Bits are erased by connecting the FET to Ground, discharging them and restoring them to the state 0. When such an erasure or resetting operation occurs, the underlying reversibility of the laws of physics insure that the microscopic motions of the electrons in the Ground still retain the information about whether the FET was charged or not, i. e., whether the bit before the erasure operation registered 1 or 0. In particular, if the bit registered 1 initially, the electrons in Ground will be slightly more energetic than if it registered 0. Landauer argued that any such operation that erased a bit required dissipation of energy $k_B T \ln 2$ to an environment at temperature T, corresponding to an increase in the environment's entropy of $k_B \ln 2$.

Landauer's principle can be seen to be a straightforward consequence of the microscopic reversibility of the laws of physics, together with the fact that entropy is a form of information – information about the microscopic motions of atoms and molecules. Because the laws of physics are reversible, any information that resides in the logical degrees of freedom of a computer at the beginning of a computation (i. e., in the charges and voltages of FETs) must still be present at the end of the computation in some degrees of freedom, either logical or microscopic. Note that physical reversibility also implies that if information can flow from logical degrees of freedom to microscopic degrees of freedom, then it can also flow back again:

the microscopic motions of electrons cause voltage fluctuations in FETs which can give rise to logical errors. Noise is necessary.

Because *AND*, *OR*, *NAND* are not logically reversible, Landauer initially concluded that computation was necessarily dissipative: entropy had to increase. As is often true in the application of the second law of thermodynamics, however, the appearance of irreversibility does not always imply the actual fact of irreversibility. In 1963, Lecerf showed that digital computation could always be performed in a reversible fashion [20]. Unaware of Lecerf's work, in 1973 Bennett rederived the possibility of reversible computation [21]. Most important, because Bennett was Landauer's colleague at IBM Watson laboratories, he realized the physical significance of embedding computation in a logically reversible context. As will be seen, logical reversibility is essential for quantum computation.

A simple derivation of logically reversible computation is due to Fredkin, Toffoli, and Margolus [22]. Unaware of Bennett's work, Fredkin constructed three-input, three-output reversible logic gates that could perform *NOT*, *AND*, *OR*, and *COPY* operations. The best-known example of such a gate is the Toffoli gate. The Toffoli gate takes in three inputs, X, Y, and Z, and returns three outputs, X', Y' and Z'. The first two inputs go through unchanged, so that $X' = X$, $Y' = Y$. The third output is equal to the third input, unless both X and Y are equal to 1, in which case the third output is the *NOT* of the third input. That is, when either X or Y is 0, $Z' = Z$, and when both X and Y are 1, $Z' = NOT\ Z$. (Another way of saying the same thing is to say that $Z' = Z\ XOR\ (X\ AND\ Y)$, where *XOR* is the exclusive *OR* operation whose output is 1 when either one of its inputs is 1, but not both. That is, *XOR* takes $00 \rightarrow 0$, $01 \rightarrow 1$, $10 \rightarrow 1$, $11 \rightarrow 0$.) Because it performs a *NOT* operation on Z controlled on whether both X and Y are 1, a Toffoli gate is often called a controlled-controlled-*NOT* (*CCNOT*) gate.



Quantum Information Processing, Figure 1
A Toffoli gate

To see that *CCNOT* gates can be wired together to perform *NOT*, *AND*, *OR*, and *COPY* operations, note that when one sets the first two inputs X and Y both to the value 1, and allows the input Z to vary, one obtains $Z' = NOT\ Z$. That is, supplying additional inputs allows a *CCNOT* to perform a *NOT* operation. Similarly, setting the input Z to 0 and allowing X and Y to vary yields $Z' = X\ AND\ Y$. *OR* and *COPY* (not to mention *NAND*) can be obtained by similar methods. So the ability to set inputs to predetermined values, together with ability to apply *CCNOT* gates allows one to perform any desired digital computation.

Because reversible computation is intrinsically less dissipative than conventional, irreversible computation, it has been proposed as a paradigm for constructing low power electronic logic circuits, and such low power circuits have been built and demonstrated [67]. Because additional inputs and wires are required to perform computation reversibly, however, such circuits are not yet used for commercial application. As the miniaturization of the components of electronic computers proceeds according to Moore's law, however, dissipation becomes an increasingly hard problem to solve, and reversible logic may become commercially viable.

Quantum Computation

In 1980, Benioff proposed a quantum-mechanical implementation of reversible computation [23]. In Benioff's model, bits corresponded to spins, and the time evolution of those spins was given by a unitary transformation that performed reversible logic operations. (In 1986, Feynman embedded such computation in a local, Hamiltonian dynamics, corresponding to interactions between groups of spins [68].) Benioff's model did not take into account the possibility of putting quantum bits into superpositions as an integral part of the computation, however. In 1985, Deutsch proposed that the ordinary logic gates of reversible computation should be supplemented with intrinsically quantum-mechanical single qubit operations [25]. Suppose that one is using a quantum-mechanical system to implement reversible computations using *CCNOT* gates. Now add to the ability to prepare qubits in desired states, and to perform *CCNOT* gates, the ability to perform single-qubit rotations of the form $e^{-i\theta\hat{\sigma}/2}$ as described above. Deutsch showed that the resulting set of operations allowed *universal quantum computation*. Not only could such a computer perform any desired classical logical transformation on its quantum bits; it could perform any desired unitary transformation U whatsoever.

Deutsch pointed out that a computer endowed with the ability to put quantum bits into superpositions and to perform reversible logic on those superpositions could compute in ways that classical computers could not. In particular, a classical reversible computer can evaluate any desired function of its input bits: $(x_1 \ldots x_n, 0 \ldots 0) \rightarrow (x_1 \ldots x_n, f(x_1 \ldots x_n))$, where x_i represents the logical value, 0 or 1, of the ith bit, and f is the desired function. In

order to preserve reversibility, the computer has been supplied with an 'answer' register, initially in the state $00\ldots0$, into which to place the answer $f(x_1\ldots x_n)$. In a quantum computer, the input bits to any transformation can be in a quantum superposition. For example, if each input bit is in an equal superposition of 0 and 1, $(1/\sqrt{2})(|0\rangle + |1\rangle)$, then all n qubits taken together are in the superposition

$$\frac{1}{2^{n/2}}(|00\ldots0\rangle + |00\ldots1\rangle + \cdots + |11\ldots1\rangle = \frac{1}{2^{n/2}} \sum_{x_1,\ldots,x_n=0,1} |x_1\ldots x_n\rangle\,. \quad (33)$$

If such a superposition is supplied to a quantum computer that performs the transformation $x_1\ldots x_n \rightarrow f(x_1\ldots x_n)$, then the net effect is to take the superposition

$$\frac{1}{2^{n/2}} \sum_{x_1,\ldots,x_n=0,1} |x_1\ldots x_n\rangle|00\ldots0\rangle \rightarrow \frac{1}{2^{n/2}} \sum_{x_1,\ldots,x_n=0,1} |x_1\ldots x_n\rangle|f(x_1\ldots x_n)\rangle\,. \quad (34)$$

That is, even though the quantum computer evaluates the function f only once, it evaluates it on every term in the superposition of inputs simultaneously, an effect which Deutsch termed 'quantum parallelism'.

At first, quantum parallelism might seem to be spectacularly powerful: with only one function 'call', one performs the function on 2^n different inputs. The power of quantum parallelism is not so easy to tease out, however. For example, suppose one makes a measurement on the output state in (33) in the $\{|0\rangle, |1\rangle\}$ basis. The result is a *randomly selected* input-output pair, $(x_1\ldots x_n, f(x_1\ldots x_n))$. One could have just as easily obtained such a pair by feeding a random input string into a classical computer that evaluates f. As will now be seen, the secret to orchestrating quantum computations that are more powerful than classical computations lies in arranging quantum interference between the different states in the superposition of equation (34).

The word 'orchestration' in the previous sentence was used for a reason. In quantum mechanics, states of physical systems correspond to waves. For example, the state of an electron is associated with a wave that is the solution of the Schrödinger equation for that electron. Similarly, in a quantum computer, a state such as $|x_1\ldots x_n\rangle|f(x_1\ldots x_n)\rangle$ is associated with a wave that is the solution of the Schrödinger equation for the underlying quantum degrees of freedom (e. g., electron spins or photon polarizations) that make up the computers quantum bits. The waves of quantum mechanics, like waves of water, light, or sound, can be superposed on each other to construct composite waves. A quantum computer that performs a conventional reversible computation, in which its qubits only take on the values 0 or 1 and are never in superpositions $\alpha|0\rangle + \beta|1\rangle$, can be thought of as an analogue of a piece of music like a Gregorian chant, in which a single, unaccompanied voice follows a prescribed set of notes. A quantum computer that performs many computations in quantum parallel is analogous to a symphony, in which many lines or voices are superposed to create chords, counterpoint, and harmony. The quantum computer programmer is the composer who writes and orchestrates this quantum symphony: her job is to make that counterpoint reveal meaning that is not there in each of the voices taken separately.

Deutsch–Jozsa Algorithm

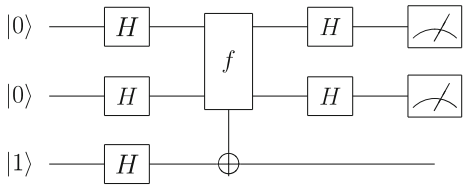
Let's examine a simple example, due to David Deutsch and Richard Jozsa, in which the several 'voices' of a quantum computer can be orchestrated to solve a problem more rapidly than a classical computer [69]. Consider the set of functions f that take one bit of input and produce one bit of output. There are four such functions:

$$f(x) = 0,\quad f(x) = 1,\quad f(x) = x,\quad f(x) = NOT\,x\,. \quad (35)$$

The first two of these functions are constant functions; the second two are 'balanced' in the sense that half of their inputs yield 0 as output, while the other half yield 1. Suppose that one is presented with a 'black box' that implements one of these functions. The problem is to query this black box and to determine whether the function the box contains is constant or balanced.

Classically, it clearly takes exactly two queries to determine whether the function in the box is constant or balanced. Using quantum information processing, however, one query suffices. The following quantum circuit shows how this is accomplished.

Deutsch–Jozsa Circuit



Quantum Information Processing, Figure 2
2-Qubit Deutsch–Jozsa circuit

Quantum circuit diagrams are similar in character to their classical counterparts: qubits enter on the left, un-

dergo a series of transformations effected by quantum logic gates, and exit at the right, where they are measured. In the circuit above, the first gate, represented by H is called a Hadamard gate. The Hadamard gate is a single-qubit quantum logic gate that effects the transformation

$$\begin{aligned} |0\rangle &\rightarrow (1/\sqrt{2})(|0\rangle + |1\rangle)\,, \\ |1\rangle &\rightarrow (1/\sqrt{2})(|0\rangle - |1\rangle)\,. \end{aligned} \tag{36}$$

In other words, the Hadamard performs a unitary transformation $U_H = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix}$ on its single-qubit input. Note that the Hadamard transformation is its own inverse: $U_H^2 = Id$.

The second logic gate implements the unknown, black-box function f. It takes two binary inputs, x, y, and gives two binary outputs. The gate leaves the first input unchanged, and adds $f(x)$ to the second input (modulo 2), so that $x \rightarrow x$ and $y \rightarrow y + f(x)(\mathrm{mod}\ 2)$. Such gates can be implemented using the controlled-*NOT* operation introduced above. Recall that the controlled-*NOT* or *CNOT* leaves its first input bit unchanged, and flips the second if and only if the first input is 1. In the symbol for a controlled-*NOT* operation, the • part represents the control bit and the $\oplus$ part represents the bit that can be flipped. The circuits required to implement the four different functions from one bit to one bit are as follows:

$$f(x) = 0: \quad f(x) = 1: \quad f(x) = x: \quad f(x) = NOT\, x: \tag{37}$$

The black box in the Deutsch–Jozsa algorithm contains one of these circuits. Note that the black-box circuits are 'classical' in the sense that they map input combinations of 0's and 1's to output combinations of 0's and 1's: the circuits of (37) make sense as classical circuits as well as quantum circuits.

Any classical circuit that can determine whether f is constant or balanced requires at least two uses of the f gate. By contrast, the Deutsch–Jozsa circuit above requires only one use of the f gate. Going through the quantum logic circuit, one finds that a constant function yields the output $|0\rangle$ on the first output line, while a balanced function yields the output $|1\rangle$ (up to an overall, unobservable phase). That is, only a single function call is required to reveal whether f is constant or balanced.

Several comments on the Deutsch–Jozsa algorithm are in order. The first is that, when comparing quantum algorithms to classical algorithms, it is important to compare apples to apples: that is, the gates used in the quantum algorithm to implement the black-box circuits should be the same as those used in any classical algorithms. The difference, of course, is that the quantum gates preserve quantum coherence, a concept which is meaningless in the classical context. This requirement has been respected in the Deutsch–Jozsa circuit above.

The second comment is that the Deutsch–Jozsa algorithm is decidedly odd and counterintuitive. The f gates and the controlled-*NOT* gates from which they are constructed both have the property that the first input passes through unchanged $|0\rangle \rightarrow |0\rangle$ and $|1\rangle \rightarrow |1\rangle$. Yet somehow, when the algorithm is identifying balanced functions, the first bit flips. How can this be? This is the part where quantum weirdness enters. Even though the f and controlled-*NOT* gates leave their first input unchanged in the logical basis $\{|0\rangle, |1\rangle\}$, the same property does not hold in other bases. For example, let $|+\rangle = (1/\sqrt{2})(|0\rangle + |1\rangle) = U_H|0\rangle$, and let $|1\rangle = (1/\sqrt{2})(|0\rangle - |1\rangle) = U_H|1\rangle$. Straitforward calculation shows that, when acting on the basis $\{|+\rangle, |-\rangle\}$, the CNOT still behaves like a CNOT, but with roles of its imputs reversed: now the second qubit passes through unchanged, while the first qubit gets flipped.

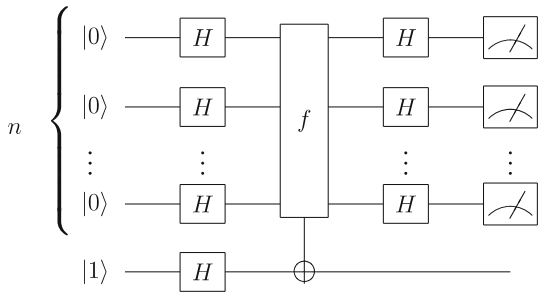
That is, when acting on the basis $\{|+\rangle, |-\rangle\}$, the *CNOT* still behaves like a *CNOT*, but with the roles of its inputs reversed: now the second qubit passes through unchanged, while the first qubit gets flipped. It is this quantum role reversal that underlies the efficacy of the Deutsch–Jozsa algorithm. Pretty weird.

It is important to note that the Deutsch–Jozsa algorithm is not just a theoretical point. The algorithm has been implemented using techniques from nuclear magnetic resonance (NMR) [33]. The results are exactly as predicted by quantum mechanics: a single function call suffices to determine whether that function is constant or balanced.

The two-qubit algorithm was first described by David Deutsch. Later, with Richard Jozsa, he extended this algorithm to a multi-qubit algorithm. Now consider functions f from n qubits to a single qubit. Once again, the problem is to determine whether or not f is constant or balanced. That is, the function f in the black box is either constant: $f(x) = 0$ for all n-bit inputs x, or $f(x) = 1$ for all x, or balanced: $f(x) = 0$ for exactly half of its 2^n possible input strings, and $f(x) = 1$ for the other half. (If this problem statement seems somewhat artificial, note that the algorithm works equally well for distinguishing between constant functions and 'typical' functions, which are *approximately* balanced.)

On average, a classical algorithm takes a little more than two function calls to distinguish between a constant or a balanced function. However, in the worst case, it takes $2^{n-1} + 1$ calls, as more than half the inputs have to

be sampled. As before, the quantum algorithm takes but a single function call, as the following circuit shows.



Quantum Information Processing, Figure 3
Caption: *n*-Qubit Deutsch–Jozsa circuit

To determine whether the f is constant or balanced, one measures the first n output bits: if they are all 0, then the function is constant; if one or more is 1, then the function is balanced.

Other Algorithms: The Quantum Fourier Transform

While it conclusively demonstrates that quantum computers are strictly more powerful than classical computers for certain problems, the Deutsch–Jozsa algorithm does not solve a problem of burning interest to applied computer scientists. Once it was clear that quantum computers could offer a speedup over classical algorithms, however, other algorithms began to be developed. Simon's algorithm [70], for example, determines whether a function f from n bits to n bits is (a) one-to-one, or (b) two-to-one with a large period s, so that $f(x+s) = f(x)$ for all x. (In Simon's algorithm the addition is bitwise modulo 2, with no carry bits.)

Simon's algorithm has a similar 'flavor' to the Deutsch–Jozsa algorithm: it is intriguing but does not obviously admit wide application. A giant step towards constructing more useful algorithms was Coppersmith's introduction [71] of the Quantum Fourier Transform (QFT). The fast Fourier transform maps a function of n bits to its discrete Fourier transform function:

$$f(x) \to g(y) = \sum_{x=0}^{2^n-1} \mathrm{e}^{2\pi i x y/2^n} f(x)\,. \tag{38}$$

The fast Fourier transform takes $O(n2^n)$ steps. The quantum Fourier transform takes a *wave function* over n qubits to a Fourier transformed wave function:

$$\sum_{x=0}^{2^n-1} f(x)|x\rangle \to 2^{-n/2} \sum_{x,y=0}^{2^n-1} \mathrm{e}^{2\pi i x y/2^n} f(x)|y\rangle\,. \tag{39}$$

It is not difficult to show that the quantum Fourier transform is a unitary.

To obtain a quantum logic circuit that accomplishes the QFT, it is convenient to express states in a binary representation. In the equations above, x and y are n-bit numbers. Write x as $x_n \ldots x_1$, where $x_n, \ldots x_1$ are the bits of x. This is just a more concise way of saying that $x = x_1 2^0 + \cdots + x_n 2^{n-1}$. Similarly, the expression $0.y_1 \ldots y_m$ represents the number $y_1/2 + \ldots y_m/2^{-m}$. Using this binary notation, it is not hard to show that the quantum Fourier transform can be written:

$$|x_1 \ldots x_n\rangle \to 2^{-n/2}(|0\rangle + \mathrm{e}^{2\pi i 0.x_1}|1\rangle)(|0\rangle + \mathrm{e}^{2\pi i 0.x_2 x_1}|1\rangle) \ldots (|0\rangle + \mathrm{e}^{2\pi i 0.x_n \ldots x_1}|1\rangle)\,. \tag{40}$$

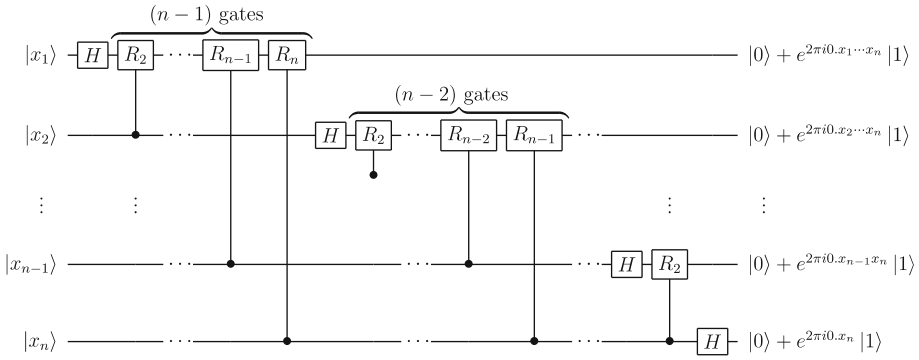
When the quantum Fourier transform is written in this form, it is straightforward to construct a circuit that implements it (see Fig. 4).

Note that the QFT circuit for wave functions over n qubits takes $O(n^2)$ steps: it is exponentially faster than the FFT for functions over n bits, which takes $O(n2^n)$ steps. This exponential speedup of the quantum Fourier transform is what guarantees the efficacy of many quantum algorithms.

The quantum Fourier transform is a potentially powerful tool for obtaining exponential speedups for quantum computers over classical computers. The key is to find a way of posing the problem to be solved in terms of finding periodic structure in a wave function. This step is the essence of the best known quantum algorithm, Shor's algorithm for factoring large numbers [26].

Shor's Algorithm

The factoring problem can be stated as follows: Given $N = pq$, where p, q are prime, find p and q. For large p and q, this problem is apparently hard for classical computers. The fastest known algorithm (the 'number sieve') takes $O(N^{1/3})$ steps. The apparent difficulty of the factoring problem for classical computers is important for cryptography. The commonly used RSA public-key cryptosystem relies on the difficulty of factoring to guarantee security. Public-key cryptography addresses the following societally important situation. Alice wants to send Bob some secure information (e. g., a credit card number). Bob sends Alice the number N, but does not reveal the identity of p or q. Alice then uses N to construct an encrypted version of the message she wishes to send. Anyone who wishes to decrypt this message must know what p and q are. That is,



Quantum Information Processing, Figure 4
Quantum Fourier Transform

encryption can be performed using the public key N, but decryption requires the private key p, q.

In 1994, Peter Shor showed that quantum computers could be used to factor large numbers and so crack public-key cryptosystems that whose security rests on the difficulty of factoring [26]. The algorithm operates by solving the so-called 'discrete logarithm' problem. This problem is, given N and some number x, find the smallest r such that $x^r \equiv 1 (\bmod N)$. Solving the discrete logarithm allows N to be factored by the following procedure. First, pick $x < N$ at random. Use Euclid's algorithm to check that the greatest common divisor of x and N is 1. (Euclid's algorithm is to divide N by x; take the remainder r_1 and divide x by r_1; take the remainder of that division, r_2 and divide r_1 by that, etc. The final remainder in this procedure is the greatest common divisor, or g.c.d., of x and N.) If the g.c.d. of x and N is not 1, then it is either p or q and we are done.

If the greatest common divisor of x and N is 1, suppose that we can solve the discrete logarithm problem to find the smallest r such that $x^r \equiv 1 (\bmod N)$. As will be seen, if r is even, we will be able to find the factors of N easily. If r turns out to be odd, just pick a new x and start again: continue until you obtain an even r (since this occurs half the time, you have to repeat this step no more than twice on average). Once an even r has been found, we have $(x^{r/2} - 1)(x^{r/2} + 1) \equiv 1 (\bmod N)$. In other words, $(x^{r/2} - 1)(x^{r/2} + 1) = bN = bpq$ for some b. Finding the greatest common divisor of $x^{r/2} - 1$, $x^{r/2} + 1$ and N now reveals p and q. The goal of the quantum algorithm, then, is to solve the discrete logarithm problem to find the smallest r such that $x^r \equiv 1 \bmod N$. If r can be found, then N can be factored.

In its discrete logarithm guise, factoring possesses a periodic structure that the quantum Fourier transform can reveal. First, find an x whose g.c.d. with N is 1, as above, and pick n so that $N^2 < 2^n < 2N^2$. The quantum algorithm uses two n-qubit registers. Begin by constructing a uniform superposition $2^{-n/2} \sum_{k=0}^{2^n-1} |k\rangle|0\rangle$. Next, perform exponentiation modulo N to construct the state,

$$2^{-n/2} \sum_{k=0}^{2^n-1} |k\rangle|x^k \bmod N\rangle \,. \tag{41}$$

This modular exponentiation step takes $O(n^3)$ operations (note that $x^{2^k} \bmod N$ can be evaluated by first constructing $x^2 \bmod N$, then constructing $(x^2)^2 \bmod N$, etc.). The periodic structure in (41) arises because if $x^k \equiv a \bmod N$, for some a, then $x^{k+r} \equiv a \bmod N$, $x^{k+2r} \equiv a \bmod N$, $\ldots$, $x^{k+mr} \equiv a \bmod N$, up to the largest m such that $k + mr < N^2$. The same periodicity holds for any a. That is, the wave function (41) is periodic with with period r. So if we apply the quantum Fourier transform to this wave function, we can reveal that period and find r, thereby solving the discrete logarithm and factoring problems.

To reveal the hidden period and find r apply the QFT to the *first* register in the state (41). The result is

$$2^{-n} \sum_{jk=0}^{2^n-1} \mathrm{e}^{2\pi i jk/2^n} |j\rangle|x^k \bmod N\rangle \,. \tag{42}$$

Because of the periodic structure, positive interference takes place when $j(k + \ell r)$ is close to a multiple of 2^n. That is, measuring the first register now yields a number j such that $jr/2^n$ is close to an integer: only for such j does the necessary positive interference take place. In other words, the algorithm reveals a j such that $j/2^n = s/r$ for some integer s. That is, to find r, we need to find fractions s/r that approximate $j/2^n$. Such fractions can be obtained us-

ing a continued fraction approximation. With reasonably high probability, the result of the continued fraction approximation can be shown to yield the desired answer r. (More precisely, repetition of this procedure $O(2 \log N)$ times suffices to identify r.) Once r is known, then the factors p and q of N can be recovered by the reduction of factoring to discrete logarithm given above.

The details of Shor's algorithm reveal considerable subtlety, but the basic idea is straightforward. In its reduction to discrete logarithm, factoring possesses a hidden periodic structure. This periodic structure can be revealed using a quantum Fourier transform, and the period itself in turn reveals the desired factors.

More recent algorithms also put the quantum Fourier transform to use to extract hidden periodicities. Notably, the QFT can be used to find solutions to Pell's equation ($x^2 - ny^2 = 1$, for non-square n) [72]. Generalizations of the QFT to transforms over groups (the dihedral group and the permutation group on n objects S_n) have been applied to other problems such as the shortest vector on a lattice [73] (dihedral group, with some success) and the graph isomorphism problem (S_n, without much success [74]).

The Phase-Estimation Algorithm

One of the most useful applications of the quantum Fourier transform is finding the eigenvectors and eigenvalues of unitary transformations. The resulting algorithm is called the 'phase-estimation' algorithm: its original form is due to Kitaev [75]. Suppose that we have the ability to apply a 'black box' unitary transformation U. U can be written $U = \sum_j e^{i\phi_j} |j\rangle\langle j|$, where $|j\rangle$ are the eigenvectors of U and $e^{i\phi_j}$ are the corresponding eigenvalues. The goal of the algorithm is to estimate the $e^{i\phi_j}$ and the $|j\rangle$. (The goal of the original Kitaev algorithm was only to estimate the eigenvalues $e^{i\phi_j}$. However, Abrams and Lloyd showed that the algorithm could also be used to construct and estimate the eigenvectors $|j\rangle$, as well [76]. The steps of the phase estimation algorithm are as follows.

(0) Begin with the initial state $|0\rangle|\psi\rangle$, where $|0\rangle$ is the n-qubit state $00\ldots0\rangle$, and $|\psi\rangle$ is the state that one wishes to decompose into eigenstates: $|\psi\rangle = \sum_j \psi_j |j\rangle$.

(1) Using Hadamards or a QFT, put the first register into a uniform superposition of all possible states:

$$\rightarrow 2^{-n/2} \sum_{k=0}^{2^n-1} |k\rangle|\psi\rangle \, .$$

(2) In the kth component of the superposition, apply U^k to $|\psi\rangle$:

$$\begin{aligned} &\rightarrow 2^{-n/2} \sum_{k=0}^{2^n-1} |k\rangle U^k |\psi\rangle \\ &= 2^{-n/2} \sum_{j,k=0}^{2^n-1} |k\rangle U^k \psi_j |j\rangle \\ &= 2^{-n/2} \sum_{j,k=0}^{2^n-1} \psi_k e^{ik\phi_j} |k\rangle|j\rangle \, . \end{aligned}$$

(3) Apply inverse QFT to first register:

$$\rightarrow 2^{-n} \sum_{j,k,l=0}^{2^n-1} \psi_k e^{ik\phi_j} e^{-2\pi i k l/2^n} |l\rangle|j\rangle \, .$$

(4) Measure the registers. The second register contains the eigenvector $|j\rangle$. The first register contains $|l\rangle$ where $2\pi l/2^n \approx \phi_j$. That is, the first register contains an n-bit approximation to ϕ_j.

By repeating the phase-estimation algorithm many times, one samples the eigenvectors and eigenvalues of U. Note that to obtain n-bits of accuracy, one must possess the ability to apply U 2^n times. This feature limits the applicability of the phase-estimation algorithm to a relatively small number of bits of accuracy, or to the estimation of eigenvalues of Us that can easily be applied an exponentially large number of times. We've already seen such an example of a process in modular exponentiation. Indeed, Kitaev originally identified the phase estimation algorithm as an alternative method for factoring.

Even when only a relatively small number of applications of U can be performed, however, the phase-estimation algorithm can provide an exponential improvement over classical algorithms for problems such as estimating the ground state of some physical Hamiltonian [76,77], as will now be seen.

Quantum Simulation

One of the earliest uses for a quantum computer was suggested by Richard Feynman [24]. Feynman noted that simulating quantum systems on a classical computer was hard: computer simulations of systems such as lattice gauge theories take up a substantial fraction of all supercomputer time, and, even then, are often far less effective than their programmers could wish them to be. The reason why it's hard to simulate a quantum system on a classical computer is straightforward: in the absence of any sneaky

tricks, the only known way to simulate a quantum system's time evolution is to construct a representation of the full state of the system, and to evolve that state forward using the system's quantum equation of motion. To represent the state of a quantum system on a classical computer is typically exponentially hard, however: an n-spin system requires 2^n complex numbers to represent its state. Evolving that state forward is even harder: it requires exponentiation of a 2^n by 2^n matrix. Even for a small quantum system, for example, one containing fifty spins, this task lies beyond the reach of existing classical supercomputers. True, supercomputers are also improving exponentially in time (Moore's law). No matter how powerful they become, however, they will not be able to simulate more than 300 spins directly, for the simple reason that to record the 2^{300} numbers that characterize the state of the spins would require the use of all 2^{300} particles in the universe within the particle horizon.

Feynman noted that if one used qubits instead of classical bits, the state of an n-spin system can be represented using just n qubits. Feynman proposed a class of systems called 'universal quantum simulators' that could be programmed to simulate any other quantum system. A universal quantum simulator has to possess a flexible dynamics that can be altered at will to mimic the dynamics of the system to be simulated. That is, the dynamics of the universal quantum simulator form an *analog* to the dynamics of the simulated system. Accordingly, one might also call quantum simulators, 'quantum analog computers'.

In 1996, Lloyd showed how Feynman's proposal could be turned into a quantum algorithm [78]. For each degree of freedom of the system to be simulated, allocate a quantum register containing a sufficient number of qubits to approximate the state of that degree of freedom to some desired accuracy. If one wishes to simulate the system's interaction with the environment, a number of registers should also be allocated to simulate the environment (for a d-dimensional system, up to d^2 registers are required to simulate the environment). Now write the Hamiltonian of the system an environment as $H = \sum_{\ell=1}^{m} H_\ell$, where each H_ℓ operates on only a few degrees of freedom. The Trotter formula implies that

$$\mathrm{e}^{-iH\delta t} = \mathrm{e}^{-iH_1\Delta t}\ldots\mathrm{e}^{-iH_m\Delta t} - \frac{1}{2}\sum_{jk}[H_j, H_k]\Delta t^2 + O\left(\Delta t^3\right) . \tag{43}$$

Each $\mathrm{e}^{-iH_\ell\Delta t}$ can be simulated using quantum logic operations on the quantum bits in the registers corresponding to the degrees of freedom on which H_ℓ acts. To simulate the time evolution of the system over time $t = n\Delta t$, we simply apply $\mathrm{e}^{-iH\Delta t}$ n times, yielding

$$\begin{aligned}\mathrm{e}^{-iHt} &= \left(\mathrm{e}^{-iH\Delta t}\right)^n \\ &= \left(\Pi_\ell \mathrm{e}^{-iH_\ell\Delta t}\right)^n - \frac{n}{2}\sum_{jk}[H_j, H_k]\Delta t^2 \\ &\quad + O\left(\Delta t^3\right) .\end{aligned} \tag{44}$$

The quantum simulation takes $O(mn)$ steps, and reproduces the original time evolution to an accuracy $h^2t^2m^2/n$, where h is the average size of $\|[H_j, H_k]\|$ (note that for simulating systems with local interactions, most of these terms are zero, because most of the local interactions commute with each other).

A second algorithm for quantum simulation takes advantage of the quantum Fourier transform [79,80]. Suppose that one wishes to simulate the time evolution of a quantum particle whose Hamiltonian is of the form $H = P^2/2m + V(X)$, where $P = -i\partial/\partial x$ is the momentum operator for the particle, and $V(X)$ is the potential energy operator for the particle expressed as a function of the position operator X. Using an n-bit discretization for the state we identify the x eigenstates with $|x\rangle = |x_n \ldots x_1\rangle$. The momentum eigenstates are then just the quantum Fourier transform of the position eigenstates: $|p\rangle = 2^{-n/2}\sum_{x=0}^{2^n-1}\mathrm{e}^{2\pi ixp/2^n}|x\rangle$. That is, $P = U_{QFT}XU_{QFT}^\dagger$.

By the Trotter formula, the infinitesimal time evolution operator is

$$\mathrm{e}^{-iH\Delta t} = \mathrm{e}^{-iP^2\Delta t/2m}\mathrm{e}^{-iV(X)\Delta t} + O\left(\delta t^2\right) . \tag{45}$$

To enact this time evolution operator one proceeds as above. Write the state of the particle in the x-basis: $|\psi\rangle = \sum_x \psi_x|x\rangle$. First apply the infinitesimal $\mathrm{e}^{-iV(X)\Delta t}$ operator:

$$\sum_x \psi_x|x\rangle \to \sum_x \psi_x \mathrm{e}^{-iV(x)}|x\rangle . \tag{46}$$

To apply the infinitesimal $\mathrm{e}^{-iP^2\delta t/2m}$ operator, first apply an inverse quantum Fourier transform on the state, then apply the unitary transformation $|x\rangle \to \mathrm{e}^{-ix^2\Delta t/2m}|x\rangle$, and finally apply the regular QFT. Because X and P are related by the quantum Fourier transform, these three steps effectively apply the transformation $\mathrm{e}^{-iP^2\Delta t/2m}$. Applying first $\mathrm{e}^{-iV(X)\Delta t}$ then $\mathrm{e}^{-iP^2\Delta t/2m}$ yields the full infinitesimal time evolution (45). The full time evolution operator e^{-iHt} can then be built up by repeating the infinitesimal operator $t/\Delta t$ times. As before, the accuracy of the quantum simulation can be enhanced by slicing time ever more finely.

Quantum simulation represents one of the most powerful uses of quantum computers. It is probably the application of quantum computers that will first give an advantage over classical supercomputers, as only one hundred qubits or fewer are required to simulate, e. g., molecular orbitals or chemical reactions, more accurately than the most powerful classical supercomputer. Indeed, special purpose quantum simulators have already been constructed using nuclear magnetic resonance techniques [81]. These quantum analog computers involve interactions between many hundreds of nuclear spins, and so are already performing computations that could not be performed by any classical computer, even one the size of the entire universe.

Quantum Search

The algorithms described above afford an exponential speedup over the best classical algorithms currently known. Such exponential speedups via quantum computation are hard to find, and are currently limited to a few special problems. There exists a large class of quantum algorithms afford a *polynomial* speedup over the best possible classical algorithms, however. These algorithms are based on Grover's quantum search algorithm.

Grover's algorithm [31] allows a quantum computer to search an unstructured database. Suppose that this database contains N items, one of which is 'marked', and the remainder of which are unmarked. Call the marked item w, for 'winner'. Such a database can be represented by a function $f(x)$ on the items in the database, such that f of the marked item is 1, and f of any unmarked item is 0. That is, $f(w) = 1$, and $f(x \neq w) = 0$. A classical search for the marked item must take $N/2$ database calls, on average. By contrast, a quantum search for the marked item takes $O(\sqrt{N})$ calls, as will now be shown.

Unstructured database search is an 'oracle' problem. In computer science, an oracle is a 'black box' function: one can supply the black box with an input x, and the black box then provides an output $f(x)$, but one has no access to the mechanism inside the box that computes $f(x)$ from x. For the quantum case, the oracle is represented by a function on two registers, one containing x, and the other containing a single qubit. The oracle takes $|x\rangle|y\rangle \rightarrow |x\rangle|y + f(x)\rangle$, where the addition takes place modulo 2.

Grover originally phrased his algorithm in terms of a 'phase' oracle U_w, where $|x\rangle U_w|x\rangle = (-1)^{f(x)}|x\rangle$. In other words, the 'winner' state acquires a phase of -1: $|w\rangle \rightarrow -|w\rangle$, while the other states remain unchanged: $|x \neq w\rangle \rightarrow |x\rangle$. Such a phase oracle can be constructed from the original oracle in several ways. The first way involves two oracle calls. Begin with the state $|x\rangle|0\rangle$ and call the oracle once to construct the state $|x\rangle|f(x)\rangle$. Now apply a σ_z transformation to the second register. The effect of this is to take the state to $(-1)^{f(x)}|x\rangle|f(x)\rangle$. Applying the oracle for a second time yields the desired phase-oracle state $(-1)^{f(x)}|x\rangle|0\rangle$. A second, sneakier way to construct a phase oracle is to initialize the second qubit in the state $(1/\sqrt{2})(|0\rangle - |1\rangle)$. A *single* call of the original oracle on the state $|x\rangle(1/\sqrt{2})(|0\rangle - |1\rangle)$ then transforms this state into $(-1)^{f(x)}|x\rangle((1/\sqrt{2})(|0\rangle - |1\rangle))$. In this way a phase oracle can be constructed from a single application of the original oracle.

Two more ingredients are needed to perform Grover's algorithm. Let's assume that $N = 2^n$ for some n, so that the different states $|j\rangle$ can be written in binary form. Let U_0 be the unitary transformation that takes $|0 \ldots 0\rangle \rightarrow -|0 \ldots 0\rangle$, that takes $|j\rangle \rightarrow |j\rangle$ for $j \neq 0$. That is, U_0 acts in the same way as U_w, but applies a phase of -1 to $|0 \ldots 0\rangle$ rather than to $|w\rangle$. In addition, let H be the transformation that performs Hadamard transformations on all of the qubits individually.

Grover's algorithm is performed as follows. Prepare all qubits in the state $|0\rangle$ and apply the global Hadamard transformation H to create the state $|\psi\rangle = (1/\sqrt{N})\sum_{j=0}^{N-1}|j\rangle$. Apply, in succession, U_w, then H, then U_0, then H again. These four transformations make up the composite transformation $U_G = HU_0HU_w$. Now apply U_G again, and repeat for a total of $\approx (\pi/4)\sqrt{N}$ times (that is, the total number of times U_G is applied is equal to the integer closest to $(\pi/4)\sqrt{N}$). The system is now, with high probability, in the state $|w\rangle$. That is, $U_G^{\sqrt{N}}|0 \ldots 0\rangle \approx |w\rangle$. Since each application of U_G contains a single call to the phase oracle U_w, the winner state $|w\rangle$ has now been identified with $O(\sqrt{N})$ oracle calls, as promised.

The quantum algorithm succeeds because the transformation U_G acts as a *rotation* in the two-dimensional subspace defined by the states $|\psi\rangle$ and $|w\rangle$. The angle of the rotation effected by each application of U_G can be shown to be given by $\sin\theta = 2/\sqrt{N}$. Note that $|\psi\rangle$ and $|w\rangle$ are approximately orthogonal, $\langle\psi|w\rangle = 1/\sqrt{N}$, and that after the initial Hadamard transformation the system begins in the state $|\psi\rangle$. Each successive application of U_G moves it an angle θ closer to $|w\rangle$. Finally, after $\approx (\pi/4)\sqrt{N}$ iterations, the state has rotated the full $\approx \pi/2$ distance to $|w\rangle$.

Grover's algorithm can be shown to be optimal [82]: no black-box algorithm can find $|w\rangle$ with fewer than $O(\sqrt{N})$ iterations of the oracle. The algorithm also works for oracles where there are M winners, so that

$f(x) = 1$ for M distinct inputs. In this case, the angle of rotation for each iteration of U_G is given by $\sin\theta = (2/N)\sqrt{M(N-M)}$, and the algorithm takes $\approx (\pi/4)\sqrt{N/M}$ steps to identify a winner.

The Adiabatic Algorithm

Many classically hard problems take the form of optimization problems. In the well-known traveling salesman problem, for example, one aims to find the shortest route connecting a set of cities. Such optimization problems can be mapped onto a physical system, in which the function to be optimized is mapped onto the energy function of the system. The ground state of the physical system then represents a solution to the optimization problem. A common classical technique for solving such problems is simulated annealing: one simulates the process of gradually cooling the system in order to find its ground state [83]. Simulated annealing is bedeviled by the problem of local minima, states of the system that are close to the optimal states in terms of energy, but very far away in terms of the particular configuration of the degrees of freedom of the state. To avoid getting stuck in such local minima, one must slow the cooling process to a glacial pace in order to insure that the true ground state is reached in the end.

Quantum computing provides a method for getting around the problem of local minima. Rather than trying to reach the ground state of the system by cooling, one uses a purely quantum-mechanical technique for finding the state [84]. One starts the system with a Hamiltonian dynamics whose ground state is simple to prepare (e. g., 'all spins sideways'). Then one gradually deforms the Hamiltonian from the simple dynamics to the more complex dynamics whose ground state encodes the answer to the problem in question. If the deformation is sufficiently gradual, then the adiabatic theorem of quantum mechanics guarantees that the system remains in its ground state throughout the deformation process. When the adiabatic deformation is complete, then the state of the system can be measured to reveal the answer.

Adiabatic quantum computation (also called 'quantum annealing') represents a purely quantum way to find the answer to hard problems. How powerful is adiabatic quantum computation? The answer is, 'nobody knows for sure'. The key question is, what is 'sufficiently gradual' deformation? That is, how slowly does the deformation have to be to guarantee that the transformation is adiabatic? The answer to this question lies deep in the heart of quantum matter. As one performs the transformation from simple to complex dynamics, the adiabatic quantum computer goes through a quantum phase transition. The maximum speed at which the computation can be performed is governed by the size of the minimum energy gap of this quantum phase transition. The smaller the gap, the slower the computation. The scaling of gaps during phase transitions ('Gapology') is one of the key disciplines in the study of quantum matter [85]. While the scaling of the gap has been calculated for many familiar quantum systems such as Ising spin glasses, calculating the gap for adiabatic quantum computers that are solving hard optimization problems seems to be just about as hard as solving the problem itself.

While few quantum computer scientists believe that adiabatic quantum computation can solve the traveling salesman problem, there is good reason to believe that adiabatic quantum computation can outperform simulated annealing on a wide variety of hard optimization problems. In addition, it is known that adiabatic quantum computation is neither more nor less powerful than quantum computation itself: a quantum computer can simulate a physical system undergoing adiabatic time evolution using the quantum simulation techniques described above; in addition, it is possible to construct devices that perform conventional quantum computation in an adiabatic fashion [86].

Quantum Walks

A final, 'physics-based', type of algorithm is the quantum walk [87,88,89,90]. Quantum walks are coherent versions of classical random walks. A classical random walk is a stochastic Markov process, the random walker steps between different states, labeled by j, with a probability w_{ij} for making the transition from state j to state i. Here w_{ij} is a stochastic matrix, $w_{ij} \geq 0$ and $\sum_j w_{ij} = 1$. In a quantum walk, the stochastic, classical process is replaced by a coherent, quantum process: the states $|j\rangle$ are quantum states, and the transition matrix U_{ij} is unitary.

By exploiting quantum coherence, quantum walks can be shown typically to give a square root speed up over classical random walks. For example, in propagation along a line, a classical random walk is purely diffusive, with the expectation value of displacement along the line going as the square root of the number of steps in the walk. By contrast, a quantum walk can be set up as a coherent, propagating wave, so that the expectation value of the displacement is proportional to the number of steps [88]. A particularly elegant example of a square root speed up in a quantum walk is the evaluation of a *NAND* tree [90]. A *NAND* tree is a binary tree containing a *NAND* gate at each vertex. Given inputs on the leaves of the tree, the problem is to evaluate the outcome at the root of the tree: is it zero

or one? *NAND* trees are ubiquitous in, e. g., game theory: the question of who wins at chess, checkers, or Go, is determined by evaluating a suitable *NAND* tree. Classically, a *NAND* tree can be evaluated with a minimum of steps. A quantum walk, by contrast, can evaluate a *NAND* tree using only $2^{n/2}$ steps.

For some specially designed problems, such as propagation along a random tree, quantum walks can give exponential speedups over classical walks [89]. The question of what problems can be evaluated more rapidly using quantum walks than classical walks remains open.

The Future of Quantum Algorithms

The quantum algorithms described above are potentially powerful, and, if large-scale quantum computers can be constructed, could be used to solve a number of important problems for which no efficient classical algorithms exist. Many questions concerning quantum algorithms remain open. While the majority of quantum computer scientists would agree that quantum algorithms are unlikely to provide solutions to NP-complete problems, it is not known whether or not quantum algorithms could provide solutions to such problems as graph isomorphism or shortest vector on a lattice. Such questions are an active field of research in quantum computer science.

Noise and Errors

The picture of quantum computation given in the previous section is an idealized picture that does not take into account the problems that arise when quantum computers are built in practice. Quantum computers can be built using nuclear magnetic resonance, ion traps, trapped atoms in cavities, linear optics with feedback of nonlinear measurements, superconducting systems, quantum dots, electrons on the surface of liquid helium, and a variety of other standard and exotic techniques. Any system that can be controlled in a coherent fashion is a candidate for quantum computation. Whether a coherently controllable system can actually be made to computer depends primarily on whether it is possible to deal effectively with the noise intrinsic to that system. Noise induces errors in computation. Every type of quantum information processor is subject to its own particular form of noise.

A detailed discussion of the various technologies for building quantum computers lies beyond the scope of this article. While the types of noise differ from quantum technology to quantum technology, however, the methods for dealing with that noise are common between technologies. This section presents a general formalism for characterizing noise and errors, and discusses the use of quantum error-correcting codes and other techniques for coping with those errors.

Open-System Operations

The time evolution of a closed quantum-mechanical system is given by unitary transformation: $\rho \rightarrow U\rho U^\dagger$, where U is unitary, $U^\dagger = U^{-1}$. For discussing quantum communications, it is necessary to look at the time evolution of open quantum systems that can exchange quantum information with their environment. The discussion of open quantum systems is straightforward: simply adjoin the system's environment, and consider the coupled system and environment as a closed quantum system. If the joint density matrix for system and environment is

$$\rho_{\mathrm{SE}}(0) \rightarrow \rho_{\mathrm{SE}}(t) = U_{\mathrm{SE}}\rho_{\mathrm{SE}}(0)U_{\mathrm{SE}}^\dagger \,. \tag{47}$$

The state of the system on its own is obtained by taking the partial trace over the environment, as described above: $\rho_{\mathrm{S}}(t) = \mathrm{tr}_{\mathrm{E}}\rho_{\mathrm{SE}}(t)$.

A particularly useful case of system and environmental interaction is one in which the system and environment are initially uncorrelated, so that $\rho_{\mathrm{SE}}(0) = \rho_{\mathrm{S}}(0) \otimes \rho_{\mathrm{E}}(0)$. In this case, the time evolution of the system on its own can always be written as $\rho_{\mathrm{S}}(t) = \sum_k A_k \rho_{\mathrm{S}}(0) A_k^\dagger$. Here the A_k are operators that satisfy the equation $\sum_k A_k^\dagger A_k = Id$: the A_k are called Kraus operators, or effects. Such a time evolution for the system on its own is called a completely positive map. A simple example of such a completely positive map for a qubit is $A_0 = Id/\sqrt{2}$, $A_1 = \sigma_x/\sqrt{2}$. $\{A_0, A_1\}$ can easily be seen to obey $A_0^\dagger A_0 + A_1^\dagger A_1 = Id$. This completely positive map for the qubit corresponds to a time evolution in which the qubit has a 50% chance of being flipped about the x-axis (the effect A_1), and a 50% chance of remaining unchanged (the effect A_0).

The infinitesimal version of any completely positive map can be obtained by taking $\rho_{\mathrm{SE}}(0) = \rho_{\mathrm{S}}(0) \otimes \rho_{\mathrm{E}}(0)$, and by expanding (46) to second order in t. The result is the Lindblad master equation:

$$\frac{\partial \rho_{\mathrm{S}}}{\partial t} = -i[\tilde{H}_{\mathrm{S}}, \rho_{\mathrm{S}}] - \sum_k (L_k^\dagger L_k \rho_{\mathrm{S}} - 2L_k \rho_{\mathrm{S}} Ł_k + \rho_{\mathrm{S}} L_k^\dagger L_k) \,. \tag{48}$$

Here $\tilde{H}_{\mathrm{S}}$ is the effective system Hamiltonian: it is equal to the Hamiltonian H_{S} for the system on its own, plus a perturbation induced by the interaction with the environment (the so-called 'Lamb shift'). The L_k correspond to open system effects such as noise and errors.

Quantum Error-Correcting Codes

One of the primary effects of the environment on quantum information is to cause errors. Such errors can be corrected using quantum error-correcting codes. Quantum error-correcting codes are quantum analogs of classical error-correcting codes such as Hamming codes or Reed–Solomon codes [91]. Like classical error-correcting codes, quantum error-correcting codes involve first encoding quantum information in a redundant fashion; the redundant quantum information is then subjected to noise and errors; then the code is decoded, at which point the information needed to correct the errors lie in the code's syndrome.

More bad things can happen to quantum information than to classical information. The only error that can occur to a classical bit is a bit-flip. By contrast, a quantum bit can either be flipped about the x-axis (the effect σ_x), flipped about the y-axis (the effect σ_y), flipped about the z-axis (the effect σ_z), or some combination of these effects. Indeed, an error on a quantum bit could take the form of a rotation by an unknown angle θ about an unknown axis. Since specifying that angle and axis precisely could take an infinite number of bits of information, it might at first seem impossible to detect and correct such an error.

In 1996, however, Peter Shor [92] and Andrew Steane [93] independently realized that if an error correcting code could detect and correct bit-flip errors (σ_x) and phase-flip errors (σ_z), then such a code would in fact correct any single-qubit error. The reasoning is as follows. First, since $\sigma_y = i\sigma_x\sigma_z$, a code that detects and corrects first σ_x errors, then σ_z errors will also correct σ_y errors. Second, since any single-qubit rotation can be written as a combination of σ_x, σ_y and σ_z rotations, the code will correct arbitrary single qubit errors. The generalization of such quantum error-correcting codes to multiple qubit errors are called Calderbank–Shor–Steane (CSS) codes [94]. A powerful technique for identifying and characterizing quantum codes is Gottesman's stabilizer formalism [95].

Concatenation is a useful method for constructing codes, both classical and quantum. Concatenation combines two codes, with the second code acting on bits that have been encoded using the first code. Quantum error-correcting codes can be combined with quantum computation to perform fault-tolerant quantum computation. Fault-tolerant quantum computation allows quantum computation to be performed accurately even in the presence of noise and errors, as long as those errors occur at a rate below some threshold [96,97,98]. For restricted error models [99], this rate can be as high as $1\% - 3\%$. For realistic error models, however, the rate is closer to $10^{-3} - 10^{-4}$.

Re-focusing

Quantum error-correcting codes are not the only technique available for dealing with noise. If, as is frequently the case, environmentally induced noise possesses some identifiable structure in terms of correlations in space and time, or obeys some set of symmetries, then powerful techniques come into play for coping with noise.

First of all, suppose that noise is correlated in time. The simplest such correlation is a static imperfection: the Hamiltonian of the system is supposed to be H, but the actual Hamiltonian is $H + \Delta H$, where ΔH is some unknown perturbation. For example, an electron spin could have the Hamiltonian $H = -(\hbar/2)(\omega + \Delta\omega)\sigma_z$, where $\Delta\omega$ is an unknown frequency shift. If not attended to, such a frequency shift will introduce unknown phases in a quantum computation, which will in turn cause errors.

Such an unknown perturbation can be dealt with quite effectively simply by flipping the electron back and forth. Let the electron evolve for time T; flip it about the x-axis; let it evolve for time T; finally, flip it back about the x-axis. The total time-evolution operator for the system is then

$$\sigma_x e^{i(\omega+\Delta\omega)T\sigma_z}\sigma_x e^{i(\omega+\Delta\omega)T\sigma_z} = Id\,. \tag{49}$$

That is, this simple refocusing technique cancels out the effect of the unknown frequency shift, along with the time evolution of the unperturbed Hamiltonian.

Even if the environmental perturbation varies in time, refocusing can be used significantly to reduce the effects of such noise. For time-varying noise, refocusing effectively acts as a filter, suppressing the effects of noise with a correlation time longer than the refocusing timescale T. More elaborate refocusing techniques can be used to cope with the effect of couplings between qubits. Refocusing requires no additional qubits or syndromes, and so is a simpler (and typically much more effective) technique for dealing with errors than quantum error-correcting codes. For existing experimental systems, refocusing typically makes up the 'first line of defence' against environmental noise. Once refocusing has dealt with time-correlated noise, quantum error correction can then be used to deal with any residual noise and errors.

Decoherence-free Subspaces and Noiseless Subsystems

If the noise has correlations in space, then quantum information can often be encoded in such a way as to be

resistant to the noise even in the absence of active error correction. A common version of such spatial correlation occurs when each qubit is subjected to the *same* error. For example, suppose that two qubits are subjected to noise of the form of a fluctuating Hamiltonian $H(t) = (\hbar/2)\gamma(t)(\sigma_z^1 + \sigma_z^2)$. This Hamiltonian introduces a time-varying phase $\gamma(t)$ between the states $|\uparrow\rangle_i, |\downarrow\rangle_i$. The key point to note here is that this phase is the *same* for both qubits. A simple way to compensate for such a phase is to encode the logical state $|0\rangle$ as the two-qubit state $|\uparrow\rangle_1|\downarrow\rangle_2$, and the logical state $|1\rangle$ as the two-qubit state $|\downarrow\rangle_1|\uparrow\rangle_2$. It is simple to verify that the two-qubit encoded states are now invariant under the action of the noise: any phase acquired by the first qubit is canceled out by the equal and opposite phase acquired by the second qubit. The subspace spanned by the two-qubit states $|0\rangle, |1\rangle$ is called a decoherence-free subspace: it is invariant under the action of the noise.

Decoherence-free subspaces were first discovered by Zanardi [100] and later popularized by Lidar [101]. Such subspaces can be found essentially whenever the generators of the noise possess some symmetry. The general form that decoherence-free subspaces take arises from the following observation concerning the relationship between noise and symmetry.

Let $\{E_k\}$ be the effects that generate the noise, so that the noise takes $\rho \to \sum_k E_k \rho E_k^\dagger$, and let $\mathcal{E}$ be the algebra generated by the $\{E_k\}$. Let G be a symmetry of this algebra, so that $[g, E] = 0$ for all $g \in G, E \in \mathcal{E}$. The Hilbert space for the system then decomposes into irreducible representation of $\mathcal{E}$ and G in the following well-known way:

$$\mathcal{H} = \sum_j \mathcal{H}_{\mathrm{E}}^j \otimes \mathcal{H}_G^j \,, \tag{50}$$

where $\mathcal{H}_{\mathrm{E}}^j$ are the irreducible representations of $\mathcal{E}$, and $\mathcal{H}_G^j$ are the irreducible representations of G.

The decomposition (49) immediately suggests a simple way of encoding quantum information in a way that is immune to the effects of the noise. Look at the effect of the noise on states of the form $|\phi\rangle_j \otimes |\psi\rangle_j$ where $|\phi\rangle_j \in \mathcal{H}_{\mathrm{E}}^j$, and $|\psi\rangle_j \in \mathcal{H}_G^j$ for some j. The effect E_k acts on this state as $(E_k^j|\phi\rangle_j) \otimes |\psi\rangle_j$, where $E_k{}^j$ is the effect corresponding to E_k within the representation $\mathcal{H}_{\mathrm{E}}^j$. In other words, if we encode quantum information in the state $|\psi\rangle_j$, then the noise has *no effect* on $|\psi\rangle_j$. A decoherence-free subspace corresponds to an $\mathcal{H}_G^j$ where the corresponding representation of $\mathcal{E}$, $\mathcal{H}_{\mathrm{E}}^j$, is one-dimensional. The case where $\mathcal{H}_{\mathrm{E}}^j$, is higher dimensional is called a noiseless subsystem [102].

Decoherence-free subspaces and noiseless subsystems represent highly effective methods for dealing with the presence of noise. Like refocusing, these methods exploit symmetry to encode quantum information in a form that is immune to noise that possesses that symmetry. Where refocusing exploits temporal symmetry, decoherence-free subspaces and noiseless subsystems exploit spatial symmetry. All such symmetry-based techniques have the advantage that no error-correcting process is required. Like refocusing, therefore, decoherence-free subspaces and noiseless subsystems form the first line of defense against noise and errors.

The tensor product decomposition of irreducible representations in (49) lies behind all known error-correcting codes [103]. A general quantum-error correcting code begins with a state $|00\ldots0\rangle_{\mathrm{A}}|\psi\rangle$, where $|00\ldots0\rangle_{\mathrm{A}}$ is the initial state of the ancilla. An encoding transformation U_{en} is then applied; an error E_k occurs; finally a decoding transformation U_{de} is applied to obtain the state

$$|\mathrm{e_k}\rangle_{\mathrm{A}}|\psi\rangle = U_{\mathrm{de}} E_k U_{\mathrm{en}} |00\ldots0\rangle_{\mathrm{A}}|\psi\rangle \,. \tag{51}$$

Here, $|e_k\rangle_{\mathrm{A}}$ is the state of the ancilla that tells us that the error corresponding to the effect E_k has occurred. Equation. (50) shows that an error-correcting code is just a noiseless subsystem for the 'dressed errors' $\{U_{\mathrm{de}} E_k U_{\mathrm{en}}\}$. At bottom, all quantum error-correcting codes are based on symmetry.

Topological Quantum Computing

A particularly interesting form of quantum error correction arises when the underlying symmetry is a topological one. Kitaev [104] has shown how quantum computation can be embedded in a topological context. Two-dimensional systems with the proper symmetries exhibit topological excitations called anyons. The name, 'anyon', comes from the properties of these excitations under exchange. Bosons, when exchanged, obtain a phase of 1; fermions, when exchanged, obtain a phase of -1. Anyons, by contrast, when exchanged, can obtain an arbitrary phase $e^{i\phi}$. For example, the anyons that underlie the fractional quantum Hall effect obtain a phase $e^{2\pi i/3}$ when exchanged. Fractional quantum Hall anyons can be used for quantum computation in a way that makes two-qubit quantum logic gates intrinsically resistant to noise [105].

The most interesting topological effects in quantum computation arise when one employs non-abelian anyons [104]. Non-abelian anyons are topological excitations that possess internal degrees of freedom. When two non-abelian anyons are exchanged, those internal degrees of freedom are subjected not merely to an additional

phase, but to a general unitary transformation U. Kitaev has shown how in systems with the proper symmetries, quantum computation can be effected simply by exchanging anyons. The actual computation takes place by dragging anyons around each other in the two-dimensional space. The resulting transformation can be visualized as a braid in two dimensional space plus the additional dimension of time.

Topological quantum computation is intrinsically fault tolerant. The topological excitations that carry quantum information are impervious to locally occurring noise: only a global transformation that changes the topology of the full system can create a error. Because of their potential for fault tolerance, two-dimensional systems that possess the exotic symmetries required for topological quantum computation are being actively sought out.

Quantum Communication

Quantum mechanics provides the fundamental limits to information processing. Above, quantum limits to computation were investigated. Quantum mechanics also provides the fundamental limits to communication. This section discusses those limits. The session closes with a section on quantum cryptography, a set of techniques by which quantum mechanics guarantees the privacy and security of cryptographic protocols.

Multiple Uses of Channels

Each quantum communication channel is characterized by its own open-system dynamics. Quantum communication channels can possess memory, or be memoryless, depending on their interaction with their environment. Quantum channels with memory are a difficult topic, which will be discussed briefly below. Most of the discussion that follows concerns the memoryless quantum channel. A single use of such a channel corresponds to a completely positive map, $\rho \to \sum_k A_k \rho A_k^\dagger$, and n uses of the channel corresponds to a transformation

$$\begin{aligned}\rho_{1\ldots n} &\to \sum_{k_1\ldots k_n} A_{k_n} \otimes \cdots \otimes A_{k_1} \rho_{1\ldots n} A_{k_1}^\dagger \otimes \cdots \otimes A_{k_n}^\dagger \\ &\equiv \sum_K A_K \rho_{1\ldots n} A_K^\dagger ,\end{aligned} \tag{52}$$

where we have used the capital letter K to indicate the n uses of the channel $k_1 \ldots k_n$. In general, the input state $\rho_{1\ldots n}$ may be entangled from use to use of the channel. Many outstanding questions in quantum communication theory remain unsolved, including, for example, the question of whether entangling inputs of the channel helps for communicating classical information.

Sending Quantum Information

Let's begin with using quantum channels to send quantum information. That is, we wish to send some quantum state $|\psi\rangle$ from the input of the channel to the output. To do this, we encode the state as some state of n inputs to the channel, send the encoded state down the channel, and then apply a decoding procedure at the output to the channel. It is immediately seen that such a procedure is equivalent to employing a quantum error-correcting code.

The general formula for the capacity of such quantum channels is known [44,45]. Take some input or 'signal' state $\rho_{1\ldots n}$ for the channel. First, construct a purification of this state. A purification of a density matrix ρ for the signal is a pure state $|\psi\rangle_{\rm AS}$ for the signal together with an ancilla, such that the state $\rho_{\rm S} = {\rm tr}_{\rm A}|\psi\rangle_{\rm AS}\langle\psi|$ is equal to the original density matrix ρ. There are many different ways to purify a state: a simple, explicit way is to write $\rho = \sum_j p_j |j\rangle\langle j|$ in diagonal form, where $\{|j\rangle\}$ is the eigenbasis for ρ. The state $|\psi\rangle_{\rm AS} = \sum_j \sqrt{p_j}|j\rangle_{\rm A}|j\rangle_{\rm S}$, where $\{|j\rangle_{\rm A}\}$ is an orthonormal set of states for the ancilla, then yields a purification of ρ.

To obtain the capacity of the channel for sending quantum information, proceed as follows. Construct a purification for the signal $\rho_{1\ldots n}$: $|\psi_n\rangle = \sum_J \sqrt{p_J}|J\rangle_{\rm A}^n|J\rangle_{\rm S}^n$, where we have used an index J instead of j to indicate that these states are summed over n uses of the channel. Now send the signal state down the channel, yielding the state

$$\rho_{\rm AS} = \sum_{JJ'} \sqrt{p_J p_{J'}}|J\rangle_{\rm A}^n\langle J'| \otimes \sum_K A_K |J\rangle_{\rm S}^n \langle J'|A_K^\dagger , \tag{53}$$

where as above $K = k_1 \ldots k_n$ indicates k uses of the channel. $\rho_{\rm AS}$ is the state of output signal state together with the ancilla. Similarly, $\rho_{\rm S} = {\rm tr}_{\rm A}\rho_{\rm AS}$ is the state of the output signal state on its own.

Let $I(AS) = -{\rm tr}\rho_{\rm AS}\log_2\rho_{\rm AS}$ be the entropy of $\rho_{\rm AS}$, measured in bits. Similarly, let $I(S) = -{\rm tr}\rho_{\rm S}\log_2\rho_{\rm S}$ be the entropy of the output state $\rho_{\rm S}$, taken on its own. Define $I(S/A) \equiv I(S) - I(AS)$ if this quantity is positive, and $I(S/A) \equiv 0$ otherwise. The quantity $I(S/A)$ is a measure of the capacity of the channel to send quantum information if the signals being sent down the channel are described by the density matrix $\rho_{1\ldots n}$. It can be shown using either CSS codes [106] or random codes [45,107] that encodings exist that allow quantum information to be sent down the channel and properly decoded at the output at a rate of $I(S/A)/n$ qubits per use.

$I(S/A)$ is a function only of the properties of the channel and the input signal state $\rho_{1\ldots n}$. The bigger $I(S/A)$ is, the less coherence the channel has managed to destroy. For example, if the channel is just a unitary transformation of the input, which destroys no quantum information, then $I(AS) = 0$ and $I(S/A) = I(S)$: the state of the signal and ancilla after the signal has passed through the channel is pure, and all quantum information passes down the channel unscathed. By contrast, a completely decohering channel takes an the input $\sum_j \sqrt{p_j}|j\rangle_A|j\rangle_S$ to the output $\sum_j p_j|j\rangle_A\langle j| \otimes |j\rangle_S\langle j|$. In this case, $I(AS) = I(S)$ and $I(S/A) = 0$: the channel has completely destroyed all quantum information sent down the channel.

In order to find the absolute capacity of the channel to transmit quantum information, we must maximize the quantity $I(S/A)/n$ over all n-state inputs $\rho_{1\ldots n}$ to the channel and take the limit as $n \to \infty$. More precisely, define

$$I_C = \lim_{n\to\infty} \min \sup I(S/A)/n \,, \tag{54}$$

where the supremum (sup) is taken over all n-state inputs $\rho_{1\ldots n}$. I_C is called the coherent information [44,45]: it is the capacity of the channel to transmit quantum information reliably.

Because the coherent information is defined only in the limit that the length of the input state goes to infinity, it has been calculated exactly in only a few cases. One might hope, in analogue to Shannon's theory of classical communication, that for memoryless channels one need only optimize over single inputs. That hope is mistaken, however: entangling the input states typically increases the quantum channel capacity even for memoryless channels [108].

Capacity of Quantum Channels to Transmit Classical Information

One of the most important questions in quantum communications is the capacity of quantum channels to transmit classical information. All of our classical communication channels – voice, free space electromagnetic, fiber optic, etc. – are at bottom quantum mechanical, and their capacities are set using the laws of quantum mechanics. If quantum information theory can discover those limits, and devise ways of attaining them, it will have done humanity a considerable service.

The general picture of classical communication using quantum channels is as follows. The conventional discussion of communication channels, both quantum and classical, designates the sender of information as Alice, and the receiver of information as Bob. Alice selects an ensemble of input states ρ_J over n uses of the channel, and send the Jth input ρ_J with probability p_J. The channel takes the n-state input ρ_J to the output $\tilde{\rho}_J = \sum_K A_K \rho_J A_K^\dagger$. Bob then performs a generalized measurement $\{B_\ell\}$ with outcomes $\{\ell\}$ to try to reveal which state Alice sent. A generalized measurement is simply a specific form an open-system transformation. The $\{B_\ell\}$ are effects for a completely positive map: $\sum_\ell B_\ell^\dagger B_\ell = Id$. After making the generalized measurement on an output state $\tilde{\rho}_J$, Bob obtains the outcome ℓ with probability $p_{\ell|J} = \mathrm{tr} B_\ell \tilde{\rho}_J B_\ell^\dagger$, and the system is left in the state $(1/p_{\ell|J}) B_\ell \tilde{\rho}_J B_\ell^\dagger$.

Once Alice has chosen a particular ensemble of signal states $\{\rho_J, p_J\}$, and Bob has chosen a particular generalized measurement, then the amount of information that can be sent along the channel is determined by the input probabilities p_J and the output probabilities $p_{\ell|J}$ and $p_\ell = \sum_J p_J p_{\ell|J}$. In particular, the rate at which information can be sent through the channel and reliably decoded at the output is given by the mutual information $I(\text{in} : \text{out}) = I(\text{out}) - I(\text{out}|\text{in})$, where $I(\text{out}) = -\sum_\ell p_\ell \log_2 p_\ell$ is the entropy of the output and $I(\text{out}|\text{in}) = \sum_J p_J(-\sum_\ell p_{\ell|J} \log_2 p_{\ell|J})$ is the average entropy of the output conditioned on the state of the input.

To maximize the amount of information that can be sent down the channel, Alice and Bob need to maximize over both input states and over Bob's measurement at the output. The Schumacher–Holevo–Westmoreland theorem, however, considerably simplifies the problem of maximizing the information transmission rate of the channel by obviating the need to maximize over Bob's measurements [37,38,39]. Define the quantity

$$\chi = S\left(\sum_J p_J \tilde{\rho}_J\right) - \sum_J p_J S(\tilde{\rho}_J) \,, \tag{55}$$

where $S(\rho) \equiv -\mathrm{tr}\rho \log_2 \rho$. χ is the difference between the entropy of the average output state and the average entropy of the output states. The Schumacher–Holevo–Westmoreland theorem then states that the capacity of the quantum channel for transmitting classical information is given by the limit as $\lim_{n\to\infty} \min \sup \chi/n$, where the supremum is taken over all possible ensembles of input states $\{\rho_J, p_J\}$ over n uses of the channel.

For Bob to attain the channel capacity given by χ, he must in general make entangling measurements over the channel outputs, even when the channel is memoryless and when Alice does not entangle her inputs. (An entangling measurement is one the leaves the outputs in an entangled state after the measurement is made.) It would simplify the process of finding the channel capacity still further if the optimization over input states could be performed over a single use of the channel for memoryless

channels, as is the case for classical communication channels, rather than having to take the limit as the number of inputs goes to infinity. If this were the case, then the channel capacity for memoryless channels would be attained for Alice sending unentangled states down the channel. Whether or not one is allowed to optimize over a single use for memoryless channels was for many years one of the primary unsolved conjectures of quantum information theory.

Let's state this conjecture precisely. Let χ_n be the maximum of χ over n uses of a memoryless channel. We then have the

Channel Additivity Conjecture: $\chi_n = n\chi_1$ Shor showed that the channel additivity conjecture is equivalent to two other additivity conjectures, the additivity of minimum output entropy and the additivity of entanglement of formation [109]. Entanglement of formation was discussed in the section on entanglement above. The minimum output entropy for n uses of a memoryless channel is simply the minimum over input states ρ_n, for n uses of the channel, of $S(\tilde{\rho}_n)$, where $\tilde{\rho}_n$ is the output state arising from the input ρ_n. We then have the

Minimum Output Entropy Additivity Conjecture: The minimum over ρ_n of $S(\tilde{\rho}_n)$ is equal to n times the minimum over ρ_1 of $S(\tilde{\rho}_1)$.

Shor's result shows that the channel additivity conjecture and the minimum output entropy additivity conjecture are equivalent: each one implies the other. If these additivity conjectures could have been proved to be true, that would have resolved some of the primary outstanding problems in quantum channel capacity theory. Remarkably, however, Hastings recently showed that the minimum output entropy conjecture is *false*, by exhibiting a channel whose minimum output entropy for multiple uses is achieved for entangled inputs [110]. As a result, the question of just how much classical information can be sent down a quantum channel, and just which quantum channels are additive and which are not, remains wide open.

Bosonic Channels

The most commonly used quantum communication channel is the so-called bosonic channel with Gaussian noise and loss [40]. Bosonic channels are ones that use bosons such as photons or phonons to communicate. Gaussian noise and loss is the most common type of noise and loss for such channels, it includes the effect of thermal noise, noise from linear amplification, and leakage of photons or phonons out of the channel. It has been shown that the capacity for bosonic channels with loss alone is attained by sending coherent states down the channel [42]. Coherent states are the sort of states produced by lasers and are the states that are currently used in most bosonic channels.

It has been conjectured that coherent states also maximize the capacity of quantum communication channels with Gaussian noise as well as loss [43]. This conjecture, if true, would establish the quantum-mechanical equivalent of Shannon's theorem for the capacity of classical channels with Gaussian noise and loss. The resolution of this conjecture can be shown to be equivalent to the following, simpler conjecture:

Gaussian Minimum Output Entropy Conjecture: Coherent states minimize the output entropy of bosonic channels with Gaussian noise and no loss.

The Gaussian minimum output entropy is intuitively appealing: an equivalent statement is that the *vacuum* input state minimizes the output entropy for a channel with Gaussian noise. In other words, to minimize the output entropy of the channel, send *nothing*.

Despite its intuitive appeal, the Gaussian minimum output entropy conjecture has steadfastly resisted proof for decades. This conjecture is related to the additivity conjectures above: in particular, if the additivity conjectures can be shown to be true, then the Gaussian minimum output entropy conjecture is also true [110]. It is not known whether the converse implication also holds. Proving the additivity conjectures and the Gaussian minimum output entropy conjecture is one of the primary goals of quantum information theory.

Entanglement Assisted Capacity

Just as quantum bits possess greater mathematical structure than classical bits, so quantum channels possess greater variety than their classical counterparts. A classical channel has but a single capacity. A quantum channel has one capacity for transmitting quantum information (the coherent information), and another capacity for transmitting classical information (the Holevo quantity χ). We can also ask about the capacity of a quantum channel in the presence of prior entanglement.

The entanglement assisted capacity of a channel arises in the following situation. Suppose that Alice and Bob have used their quantum channel to build up a supply of entangled qubits, where Alice possesses half of the entangled pairs of qubits, and Bob possesses the other half of the pairs. Now Alice sends Bob some qubits over the channel. How much classical information can these qubits convey?

At first one might think that the existence of shared prior entanglement should have no effect on the amount of information that Alice can send to Bob. After all, entanglement is a form of correlation, and the existence of prior correlation between Alice and Bob in a classical setting has no effect on the amount of information sent. In the quantum setting, however, the situation is different.

Consider, for example, the case where Alice and Bob have a perfect, noiseless channel. When Alice and Bob share no prior entanglement, then a single qubit sent down the channel conveys exactly one bit of classical information. When Alice and Bob share prior entanglement, however, a single quantum bit can convey more than one bit of classical information. Suppose that Alice and Bob share an entangled pair in the singlet state $(1/\sqrt{2})(|0\rangle_A|1\rangle_B - |1\rangle_A|0\rangle_B)$. Alice then performs one of four actions on her qubit: either she does nothing (performs the identity Id on the qubit), or she flips the qubit around the x-axis (performs σ_x), or she flips the qubit around the y-axis (performs σ_y), she flips the qubit around the z-axis (performs σ_z).

Now Alice sends her qubit to Bob. Bob now possesses one of the four orthogonal states, $(1/\sqrt{2}) \cdot (|0\rangle_A|1\rangle_B - |1\rangle_A|0\rangle_B)$, $(1/\sqrt{2})(|1\rangle_A|1\rangle_B - |0\rangle_A|0\rangle_B)$, $(i/\sqrt{2}) \cdot (|1\rangle_A|1\rangle_B + |0\rangle_A|0\rangle_B)$, $(1/\sqrt{2})(|0\rangle_A|1\rangle_B + |1\rangle_A|0\rangle_B)$. By measuring which of these states he possesses, Bob can determine which of the four actions Alice performed. That is, when Alice and Bob share prior entanglement, Alice can send *two* classical bits for each quantum bit she sends. This phenomenon is known as superdense coding [111].

In general, the quantum channel connecting Alice to Bob is noisy. We can then ask, given the form of the quantum channel, how much does the existence of prior entanglement help Alice in sending classical information to Bob? The answer to this question is given by the following theorem, due to Shor et al. The entanglement assisted capacity of a quantum channel is equal to the maximum of the quantum mutual information between the input and output of the channel [112]. The quantum mutual information is defined as follows. Prepare a purification $|\psi\rangle_{AS}$ of an input state ρ and send the signal state S down the channel, resulting the state ρ_{AS} as in (52) above. Defining $\rho_S = \mathrm{tr}_A \rho_{AS}$, $\rho_A = \mathrm{tr}_S \rho_{AS}$, as before, the quantum mutual information is defined to be $I_Q(A:S) = S(\rho_A) + S(\rho_S) - S(\rho_{AS})$. The entanglement assisted capacity of the channel is obtained by maximizing the quantum mutual information $I_Q(A:S)$ over input states ρ.

The entanglement assisted capacity of a quantum channel is greater than or equal to the channel's Holevo quantity, which is in turn greater than or equal to the channel's coherent information. Unlike the coherent information, which is known not to be additive over many uses of the channel, or the Holevo quantity, which is suspected to be additive but which has not been proved to be so, the entanglement assisted capacity is known to be additive and so can readily be calculated for memoryless channels.

Teleportation

As mentioned in the introduction, one of the most strange and useful effects in quantum computation is teleportation [46]. The traditional, science fiction picture of teleportation works as follows.

An object such as an atom or a human being is placed in a device called a teleporter. The teleporter makes complete measurements of the physical state of the object, destroying it in the process. The detailed information about that physical state is sent to a distant location, where a second teleporter uses that information to reconstruct an exact copy of the original object.

At first, quantum mechanics would seem to make teleportation impossible. Quantum measurements tend to disturb the object measured. Many identical copies of the object are required to obtain even a rough picture of the underlying quantum state of the object. In the presence of shared, prior entanglement, however, teleportation is in fact possible in principle, and simple instances of teleportation have been demonstrated experimentally.

A hint to the possibility of teleportation comes from the phenomenon of superdense coding described in the previous section. If one qubit can be used to convey two classical bits using prior entanglement, then maybe two classical bits might be used to convey one qubit. This hope turns out to be true. Suppose that Alice and Bob each possess one qubit out of an entangled pair of qubits (that is, they mutually possess one 'e-bit'). Alice desires to teleport the state $|\psi\rangle$ of another qubit. The teleportation protocol goes as follows.

First, Alice makes a Bell-state measurement on the qubit to be teleported together with her half of the entangled pair. A Bell-state measurement on two qubits is one that determines whether the two qubits are in one of the four states $|\phi_{00}\rangle = (1/\sqrt{2})(|01\rangle - |10\rangle)$, $|\phi_{01}\rangle = (1/\sqrt{2})(|00\rangle - |11\rangle)$, $|\phi_{10}\rangle = (1/\sqrt{2})(|00\rangle + |11\rangle)$, or $|\phi_{11}\rangle = (1/\sqrt{2})(|01\rangle + |10\rangle)$. Alice obtains two classical bits of information as a result of her measurement, depending on which $|\phi_{ij}\rangle$ the measurement revealed. She sends these two bits to Bob. Bob now performs a unitary transformation on his half of the entangled qubit pair. If he receives 00, then he does nothing. If he receives 01, then he applies σ_x to flip his bit about the x-axis. If he receives

10, then he applies σ_y to flip his bit about the y-axis. If he receives 11, then he applies σ_z to flip his bit about the z-axis. The result? After Bob has performed his transformation conditioned on the two classical bits he received from Alice, his qubit is now in the state $|\psi\rangle$, up to an overall phase. Alice's state has been teleported to Bob.

It might seem at first somewhat mysterious how this sequence of operations can teleport Alice's state to Bob. The mechanism of teleportation can be elucidated as follows. Write $|\psi\rangle = \alpha|0\rangle_i + \beta|1\rangle_i$. Alice and Bob's entangled pair is originally in the state $|\phi_{00}\rangle_{AB} = (1/\sqrt{2})(|0\rangle_A|1\rangle_B - |1\rangle_A|0\rangle_B)$. The full initial state of qubit to be teleported together with the entangled pair can then be written as

$$\begin{aligned}
|\psi\rangle|\phi_{00}\rangle_{AB} &= (\alpha|0\rangle_i + \beta|1\rangle_i)\frac{1}{\sqrt{2}}(|0\rangle_A|1\rangle_B - |1\rangle_A|0\rangle_B) \\
&= \frac{1}{2\sqrt{2}}(|0\rangle_i|1\rangle_A - |1\rangle_i|0\rangle_A) \otimes (\alpha|0\rangle_B + \beta|1\rangle_B) \\
&\quad + \frac{1}{2\sqrt{2}}(|0\rangle_i|0\rangle_A - |1\rangle_i|1\rangle_A) \otimes (\alpha|1\rangle_B + \beta|0\rangle_B) \\
&\quad + \frac{1}{2\sqrt{2}}(|0\rangle_i|0\rangle_A + |1\rangle_i|1\rangle_A) \otimes (\alpha|1\rangle_B - \beta|0\rangle_B) \\
&\quad + \frac{1}{2\sqrt{2}}(|0\rangle_i|1\rangle_A + |1\rangle_i|0\rangle_A) \otimes (\alpha|0\rangle_B - \beta|1\rangle_B) \\
&= \frac{1}{2}(|\phi_{00}\rangle_{iA} \otimes |\psi\rangle_B + |\phi_{01}\rangle_{iA} \otimes \sigma_x|\psi\rangle_B \\
&\quad + |\phi_{10}\rangle_{iA} \otimes i\sigma_y|\psi\rangle_B + |\phi_{11}\rangle_{iA} \otimes \sigma_z|\psi\rangle_B) \,.
\end{aligned} \tag{56}$$

When the initial state is written in this form, one sees immediately how the protocol works: the measurement that Alice makes contains exactly the right information that Bob needs to reproduce the state $|\psi\rangle$ by performing the appropriate transformation on his qubit.

Teleportation is a highly useful protocol that lies at the center of quantum communication and fault tolerant quantum computation. There are several interesting features to note. The two bits of information that Alice obtains are completely random: 00, 01, 10, 11 all occur with equal probability. These bits contain no information about $|\psi\rangle$ taken on their own: it is only when combined with Bob's qubit that those bits suffice to recreate $|\psi\rangle$. During the teleportation process, it is difficult to say just where the state $|\psi\rangle$ 'exists'. After Alice has made her measurement, the state $|\psi\rangle$ is in some sense 'spread out' between her two classical bits and Bob's qubit. The proliferation of quotation marks in this paragraph is a symptom of quantum weirdness: classical ways of describing things are inadequate to capture the behavior of quantum things. The only way to see what happens to a quantum system during a process like teleportation is to apply the mathematical rules of quantum mechanics.

Quantum Cryptography

A common problem in communication is security. Suppose that Alice and Bob wish to communicate with each other with the secure knowledge that no eavesdropper (Eve) is listening in. The study of secure communication is commonly called cryptography, since to attain security Alice must encrypt her messages and Bob must decrypt them. The no-cloning theorem together with the fact that if one measures a quantum system, one typically disturbs it, implies that quantum mechanics can play a unique role in constructing cryptosystems. There are a wide variety of quantum cryptographic protocols [49,50,51]. The most common of these fall under the heading of quantum key distribution (QKD).

The most secure form of classical cryptographic protocols is the one-time pad. Here, Alice and Bob each possess a random string of bits. This string is called the key. If no one else possesses the key, then Alice and Bob can send messages securely as follows. Suppose that Alice's message has been encoded in bits in some conventional way (e. g., mapping characters to ASCII bit strings). Alice encrypts the message by adding the bits of the key to the bits of her message one by one, modulo 2 (i. e., without carrying). Quantum information theory is a rich and fundamental field. Its origins lie with the origins of quantum mechanics itself a century ago. The field has expanded dramatically since the mid 1990s, due to the discovery of practical applications of quantum information processing such as factoring and quantum cryptography, and because of the rapid development of technologies for manipulating systems in a way that preserves quantum coherence.

As an example of the rapid pace of development in the field of quantum information, while this article was in proof, a new algorithm for solving linear sets of equations was discovered [116]. Based on the quantum phase algorithm, this algorithm solves the following problem: given a sparse matrix A and a vector $\vec{b}$, find a vector $\vec{x}$ such that $A\vec{x} = \vec{b}$. That is, construct $\vec{x} = A^{-1}\vec{b}$. If A is an n by n matrix, the best classical algorithms for solving this problem run in time $O(n)$. Remarkably, the quantum matrix inversion algorithm runs in time $O(\log n)$, an exponential improvement: a problem that could take 10^{12} – 10^{15} operations to solve on a classical computer could be solved on a quantum computer in fewer than one hundred steps.

When they were developed in the mid twentieth century, the fields of classical computation and communica-

tion provided unifying methods and themes for all of engineering and science. So at the beginning of the twenty first century, quantum information is providing unifying concepts such as entanglement, and unifying techniques such as coherent information processing and quantum error correction, that have the potential to transform and bind together currently disparate fields in science and engineering.

The idea of quantum cryptography was proposed, in embryonic form, by Stephen Wiesner in [49]. The first quantum cryptographic protocol was proposed by Bennett and Brassard in 1984 and is commonly called BB84 [50]. The BB84 protocol together with its variants is the one most commonly used by existing quantum cryptosystems.

In BB84, Alice sends Bob a sequence of qubits. The protocol is most commonly described in terms of qubits encoded on photon polarization. Here, we will describe the qubits in terms of spin, so that we can use the notation developed in Sect. "Quantum Mechanics". Spin 1/2 is isomorphic to photon polarization and so the quantum mechanics of the protocol remains the same.

Alice chooses a sequence of qubits from the set $\{|\uparrow\rangle, |\downarrow\rangle, |\leftarrow\rangle, |\rightarrow\rangle\}$ at random, and sends that sequence to Bob. As he receives each qubit in turn, Bob picks at random either the z-axis or the x-axis and measures the received qubit along that axis. Half of the time, on average, Bob measures the qubit along the same axis along which it was prepared by Alice.

Alice and Bob now check to see if Eve is listening in. Eve can intercept the qubits Alice sends, make a measurement on them, and then send them on to Bob. Because she does not know the axis along which any individual qubit has been prepared, however, here measurement will inevitably disturb the qubits. Alice and Bob can then detect Eve's intervention by the following protocol.

Using an ordinary, insecure form of transmission, e. g., the telephone, Alice reveals to Bob the state of some of the qubits that she sent. On half of those qubits, on average, Bob measured them along the same axis along which they were sent. Bob then checks to see if he measured those qubits to be in the same state that Alice sent them. If he finds them all to be in the proper state, then he and Alice can be sure that Eve is not listening in. If Bob finds that some fraction of the qubits are not in their proper state, then he and Alice know that either the qubits have been corrupted by the environment in transit, or Eve is listening in. The degree of corruption is related to the amount of information that Eve can have obtained: the greater the corruption, the more information Eve may have. From monitoring the degree of corruption of the received qubits, Alice and Bob can determine just how many bits of information Eve has obtained about their transmission.

Alice now reveals to Bob the axis along which she prepared the remainder of her qubits. On half of those, on average, Bob measured using the same axis. If Eve is not listening, those qubits on which Bob measured using the same axis along which Eve prepared them now constitute a string of random bits that is shared by Alice and Bob and by them only. This shared random string can then be used as a key for a one-time pad.

If Eve is listening in, then from their checking stage, Alice and Bob know just how many bits out of their shared random string are also known by Eve. Alice and Bob can now perform classical privacy amplification protocols [113] to turn their somewhat insecure string of shared bits into a shorter string of shared bits that is more secure. Once privacy amplification has been performed, Alice and Bob now share a key whose secrecy is guaranteed by the laws of quantum mechanics.

Eve could, of course, intercept *all* the bits sent, measure them, and send them on. Such a 'denial of service' attack prevents Alice and Bob from establishing a shared secret key. No cryptographic system, not even a quantum one, is immune to denial of service attacks: if Alice and Bob can exchange no information then they can exchange no secret information! If Eve lets enough information through, however, then Alice and Bob can always establish a secret key.

A variety of quantum key distribution schemes have been proposed [50,51]. Ekert suggested using entangled photons to distribute keys to Alice and Bob. In practical quantum key distribution schemes, the states sent are attenuated coherent states, consisting of mostly vacuum with a small amplitude of single photon states, and an even smaller amplitude of states with more than one photon. It is also possible to use continuous quantum variables such as the amplitudes of the electric and magnetic fields to distribute quantum keys [114,115]. To guarantee the full security of a quantum key distribution scheme requires a careful examination of all possible attacks given the actual physical implementation of the scheme.

Implications and Conclusions

Quantum information theory is a rich and fundamental field. Its origins lie with the origins of quantum mechanics itself a century ago. The field has expanded dramatically since the mid 1990s, due to the discovery of practical applications of quantum information processing such as factoring and quantum cryptography, and because of the rapid development of technologies for manipulating systems in

a way that preserves quantum coherence. When they were developed in the mid twentieth century, the fields of classical computation and communication provided unifying methods and themes for all of engineering and science. So at the beginning of the twenty first century, quantum information is providing unifying concepts such as entanglement, and unifying techniques such as coherent information processing and quantum error correction, that have the potential to transform and bind together currently disparate fields in science and engineering.

Indeed, quantum information theory has perhaps even a greater potential to transform the world than classical information theory. Classical information theory finds its greatest application in the man-made systems such as electronic computers. Quantum information theory applies not only to man-made systems, but to all physical systems at their most fundamental level. For example, entanglement is a characteristic of virtually all physical systems at their most microscopic levels. Quantum coherence and the relationship between symmetries and the conservation and protection of information underlie not only quantum information, but the behavior of elementary particles, atoms, and molecules.

When or whether techniques of quantum information processing will become tools of mainstream technology is an open question. The technologies of precision measurement are already fully quantum mechanical: for example, the atomic clocks that lie at the heart of the global positioning system (GPS) rely fundamentally on quantum coherence. Ubiquitous devices such as the laser and the transistor have their roots in quantum mechanics. Quantum coherence is relatively fragile, however: until such a time as we can construct robust, easily manufactured coherent systems, quantum information processing may have its greatest implications at the extremes of physics and technology.

Quantum information processing analyzes the universe in terms of information: at bottom, the universe is composed not just of photons, electrons, neutrinos and quarks, but of quantum bits or qubits. Many aspects of the behavior of those elemental qubits are independent of the particular physical system that registers them. By understanding how information behaves at the quantum mechanical level, we understand the fundamental behavior of the universe itself.

Bibliography

1. Nielsen MA, Chuang IL (2000) Quantum Computation and Quantum Information. Cambridge University Press, Cambridge
2. Ehrenfest P, Ehrenfest T (1912) The Conceptual Foundations of the Statistical Approach in Mechanics. Cornell University Press, Ithaca, NY
3. Planck M (1901) Ann Phys 4:553
4. Maxwell JC (1871) Theory of Heat. Appleton, London
5. Einstein A (1905) Ann Phys 17:132
6. Bohr N (1913) Phil Mag 26:1–25, 476–502, 857–875
7. Schrödinger E (1926) Ann Phys 79:361–376, 489–527; (1926) Ann Phys 80:437–490; (1926) Ann Phys 81:109–139
8. Heisenberg W (1925) Z Phys 33:879–893; (1925) Z Phys 34:858–888; (1925) Z Phys 35:557–615
9. Einstein A, Podolsky B, Rosen N (1935) Phys Rev 47:777
10. Bohm D (1952) Phys Rev 85:166–179, 180–193
11. Bell JS (1964) Physics 1:195
12. Aharonov Y, Bohm D (1959) Phys Rev 115:485–491; (1961) Phys Rev 123:1511–1524
13. Aspect A, Grangier P, Roger G (1982) Phys Rev Lett 49:91–94; Aspect A, Dalibard J, Roger G (1982) Phys Rev Lett 49:1804–1807 ; Aspect A (1999) Nature 398:189
14. Hartley RVL (1928) Bell Syst Tech J 535–563
15. Turing AM (1936) Proc Lond Math Soc 42:230–265
16. Shannon CE (1937) Symbolic Analysis of Relay and Switching Circuits. Master's Thesis, MIT
17. Shannon CE (1948) A Math theory of commun. Bell Syst Tech J 27:379–423, 623–656
18. von Neumann J (1966) In: Burks AW (ed) Theory of Self-Reproducing Automata. University of Illinois Press, Urbana
19. Landauer R (1961) IBM J Res Develop 5:183–191
20. Lecerf Y (1963) Compte Rendus 257:2597–2600
21. Bennett CH (1973) IBM J Res Develop 17:525–532; (1982) Int J Theor Phys 21:905–940
22. Fredkin E, Toffoli T (1982) Int J Theor Phys 21:219–253
23. Benioff P (1980) J Stat Phys 22:563–591; (1982) Phys Rev Lett 48:1581–1585; (1982) J Stat Phys 29:515–546; Ann NY (1986) Acad Sci 480:475–486
24. Feynman RP (1982) Int J Th Ph 21:467
25. Deutsch D (1985) Proc Roy Soc Lond A 400:97–117; (1989) Proc Roy Soc Lond A 425:73–90
26. Shor P (1994) In: Goldwasser S (ed) Proceedings of the 35th Annual Symposium on Foundations of Computer Science. IEEE Computer Society, Los Alamitos, pp 124–134
27. Lloyd S (1993) Science 261:1569; (1994) Science 263:695
28. Cirac JI, Zoller P (1995) Phys Rev Lett 74:4091
29. Monroe C, Meekhof DM, King BE, Itano WM, Wineland DJ (1995) Phys Rev Lett 75:4714
30. Turchette QA, Hood CJ, Lange W, Mabuchi H, Kimble HJ (1995) Phys Rev Lett 75:4710
31. Grover LK (1996) In: Proceedings, 28th Annual ACM Symposium on the Theory of Computing, p 212
32. Cory D, Fahmy AF, Havel TF (1997) Proc Nat Acad Sci USA 94:1634–1639
33. Chuang IL, Vandersypen LMK, Zhou X, Leung DW, Lloyd S (1998) Nature 393:143–146
34. Chuang IL, Gershenfeld N, Kubinec M (1998) Phys Rev Lett 80:3408–3411
35. Gordon JP (1962) Proc IRE 50:1898–1908
36. Lebedev DS, Levitin LB (1963) Sov Phys Dok 8:377
37. Holevo AS (1973) Prob Per Inf 9:3; Prob Inf Trans (USSR) 9:110
38. Schumacher B, Westmoreland M (1997) Phys Rev A 55:2738
39. Holevo AS (1998) IEEE Trans Inf Th 44:69

40. Caves CM, Drummond PD (1994) Rev Mod Phys 66:481–537
41. Yuen HP, Ozawa M (1993) Phys Rev Lett 70:363–366
42. Giovannetti V, Guha S, Lloyd S, Maccone L, Shapiro JH, Yuen HP (2004) Phys Rev Lett 92:027902
43. Giovannetti V, Guha S, Lloyd S, Maccone L, Shapiro JH (2004) Phys Rev A 70:032315
44. Nielsen M, Schumacher B (1996) Phys Rev A 54:2629–2635
45. Lloyd S (1997) Phys Rev A 55:1613–1622
46. Bennett CH, Brassard G, Cripeau C, Jozsa R, Peres A, Wootters WK (1993) Phys Rev Lett 70:1895–1899
47. Pan JW, Daniell M, Gasparoni S, Weihs G, Zeilinger A (2001) Phys Rev Lett 86:4435–4438
48. Zhang TC, Goh KW, Chou CW, Lodahl P, Kimble HJ (2003) Phys Rev A 67:033802
49. Wiesner S (1983) SIGACT News 15:78–88
50. Bennett CH, Brassard G (1984) Proc IEEE Int Conf Comput 10–12:175–179
51. Ekert A (1991) Phys Rev Lett 67:661–663
52. Feynman RP (1965) The Character of Physical Law. MIT Press, Cambridge
53. Bohr A, Ulfbeck O (1995) Rev Mod Phys 67:1–35
54. Wootters WK, Zurek WH (1982) Nature 299:802–803
55. Werner RF (1998) Phys Rev A 58:1827–1832
56. Bennett CH, Bernstein HJ, Popescu S, Schumacher B (1996) Phys Rev A 53:2046–2052
57. Horodecki M, Horodecki P, Horodecki R (1998) Phys Rev Lett 80:5239–5242
58. Horodecki M, Horodecki P, Horodecki R (2000) Phys Rev Lett 84:2014–2017
59. Wootters WK (1998) Phys Rev Lett 80:2245–2248
60. Christandl M, Winter A (2004) J Math Phys 45:829–840
61. Bohm D (1952) Am J Phys 20:522–523
62. Bell JS (2004) Aspect A (ed) Speakable and Unspeakable in Quantum Mechanics. Cambridge University Press, Cambridge
63. Clauser JF, Horne MA, Shimony A, Holt RA (1969) Phys Rev Lett 23:880–884
64. Greenberger DM, Horne MA, Zeilinger A (1989) In: Kafatos M (ed) Bell's Theorem, Quantum Theory, and Conceptions of the Universe. Kluwer, Dordrecht
65. Pan JW, Daniell M, Weinfurter H, Zeilinger A (1999) Phys Rev Lett 82:1345–1349
66. Nelson RJ, Cory DG, Lloyd S (2000) Phys Rev A 61:022106
67. Frank MP, Knight TF (1998) Nanotechnology 9:162–176
68. Feynman RP (1985) Opt News 11:11; (1986) Found Phys 16:507
69. Deutsch D, Jozsa R (1992) Proc R Soc London A 439:553
70. Simon DR (1994) In: Goldwasser S (ed) Proceedings of the 35th Annual Symposium on Foundations of Computer Science. IEEE Computer Society, Los Alamitos, pp 116–123
71. Coppersmith D (1994) IBM Research Report RC19642
72. Hallgren S (2007) J ACM 54:1
73. Kuperberg G (2003) arXiv:quant-ph/0302112
74. Hallgren S, Moore C, Rötteler M, Russell A, Sen P (2006) In: Proc 38th ann ACM symposium theory comput, pp 604–617
75. Kitaev AY, Shen A, Vyalyi MN (2002) Classical and Quantum Computation. American Mathematical Society, Providence
76. Abrams DS, Lloyd S (1997) Phys Rev Lett 79:2586–2589
77. Aspuru-Guzik A, Dutoi AD, Love PJ, Head-Gordon M (2005) Science 309:1704–1707
78. Lloyd S (1996) Science 273:1073–8
79. Wiesner S (1996) arXiv:quant-ph/9603028
80. Zalka C (1998) Proc Roy Soc Lond A 454:313–322
81. Ramanathan C, Sinha S, Baugh J, Havel TF, Cory DG (2005) Phys Rev A 71:020303(R)
82. Boyer M, Brassard G, Hoyer P, Tapp A (1998) Fortsch Phys 46:493–506
83. Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Science 220: 671–680
84. Farhi E, Goldstone J, Gutmann S (2001) Science 292:472–476
85. Sachdev S (1999) Quantum Phase Transitions. Cambridge University Press, Cambridge
86. Aharonov D, van Dam W, Kempe J, Landau Z, Lloyd S, Regev O (2004) In: Proceedings of the 45th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2004), pp 42–51. quant-ph/0405098
87. Farhi E, Gutmann S (1998) Phys Rev A 58:915–928
88. Aharonov D, Ambainis A, Kempe J, Vazirani U (2001) In: Proc 33th ACM Symposium on Theory of Computing (STOC 2001), pp 50–59
89. Childs AM, Cleve R, Deotto E, Farhi E, Gutmann S, Spielman DA (2003) Proc 35th ACM Symposium on Theory of Computing (STOC 2003), pp 59–68
90. Farhi E, Goldstone J, Gutmann S (2007) A quantum algorithm for the Hamiltonian NAND tree. arXiv:quant-ph/0702144
91. Hamming R (1980) Coding and Information Theory. Prentice Hall, Upper Saddle River
92. Shor PW (1995) Physical Review A, 52:R2493–R2496
93. Steane AM (1996) Phys Rev Lett 77:793–797
94. Calderbank AR, Shor PW (1996) Phys Rev A 54:1098–1106
95. Gottesman D (1996) Phys Rev A 54:1862
96. Aharonov D, Ben-Or M (1997) In: Proc 29th ann ACM symposium theory comput, pp 176–188
97. Knill E, Laflamme R, Zurek W (1996) arXiv:quant-ph/9610011
98. Gottesman D (1998) Phys Rev A 57:127–137
99. Knill E (2005) Nature 434:39–44
100. Zanardi P, Rasetti M (1997) Phys Rev Lett 79:3306–3309
101. Lidar DA, Chuang IL, Whaley KB (1998) Phys Rev Lett 81: 2594–2597
102. Knill E, Laflamme R, Viola L (2000) Phys Rev Lett 84:2525–2528
103. Zanardi P, Lloyd S (2003) Phys Rev Lett 90:067902
104. Kitaev AY (2003) Ann Phys 30:2–30
105. Lloyd S (2004) Quant Inf Proc 1:15–18
106. Devetak I (2005) IEEE Trans Inf Th 51:44–55
107. Shor PW (2002) Lecture Notes, MSRI Workshop on Quantum Computation. Mathematical Sciences Research Institute, Berkley
108. DiVincenzo DP, Shor PW, Smolin JA (1998) Phys Rev A 57: 830–839
109. Shor PW (2004) Comm Math Phys 246:453–472
110. Hastings MB (2008) A counterexample to additivity of minimum output entropy. arXiv:0809.3972
111. Bennett CH, Wiesner SJ (1992) Phys Rev Lett 69:2881
112. Bennett CH, Shor PW, Smolin JA, Thapliyal AV (1999) Phys Rev Lett 83:3081–3084
113. Bennett CH, Brassard G, Robert JM (1988) SIAM Journal on Computing 17:210–229
114. Ralph TC (1999) Phys Rev A 61:010303
115. Braunstein S, Pati AK (2003) Quantum Information With Continuous Variables. Springer, New York
116. Harrow AW, Hassidim A, Lloyd S (2008) Quantum algorithm for solving linear sets of equations. arXiv:0811.3171

Quantum Information Science, Introduction to

Joseph F. Traub
Computer Science Department, Columbia University, New York, USA

Quantum information science is a hugely exciting fairly new field that has attracted many researchers from diverse fields such as physics, chemistry, computer science, and mathematics. A driving force is that Moore's law will soon end due to a number of factors. These include tunneling, the heat problem, approaching atomic size, difficulty of design, and cost of fabrication facilities. Multicores and manycores may provide a near term fix. But if we are to continue Moore's Law exponential trajectory long term entirely new technologies are needed. Possible future technologies include molecular, biological, photonic, or quantum computers. Quantum computers are based on the laws of quantum mechanics.

We've seen progress on quantum algorithms and complexity. But can quantum computers be built? Two of the impediments include the small number of qubits to date and the short decoherence times. Will there be enough qubits to do new science especially with the requirements of error correction and fault tolerant computing? Will there be enough time to do new science before decoherence sets in? In contrast to quantum computation some quantum communication applications have already been realized.

This section begins with a broad overview of quantum information science. The articles that follow discuss quantum algorithms, quantum algorithms and complexity for continuous problems, quantum computational complexity, quantum error correction and fault tolerant computing, quantum computing with trapped ions, quantum computing using optics, and finally cryptography.

Lloyd (see ► Quantum Information Processing) argues that the most important reason for studying quantum information science is to construct a unified theory of how information can be registered and transformed at the fundamental limits imposed by physical law. He introduces the formalism of quantum mechanics and applies it to the idea of quantum information. He then identifies quantum superposition and entanglement as being at the heart of quantum computation. He discusses the Deutsch–Josza algorithm which solves a problem faster than could be done by any classical computer. This is an artificial example. Decidably not artificial is Shor's algorithm for factoring large integers in polynomial time. The fastest classical algorithm known is exponential in the number of bits. Shor's algorithm is important for cryptography since the commonly used RSA public-key system relies on the difficulty of factoring to guarantee security. The importance of the quantum Fourier transform for algorithm design is discussed.

Lloyd does not describe the various technologies that can be used to build quantum computers. Every one of them is subject to its own particular form of noise. He does present a general formalism for characterizing noise and errors and discusses techniques for coping with them.

Mosca (see ► Quantum Algorithms) defines quantum algorithms as algorithms that run on any realistic model of quantum computation. The most commonly used model of quantum computation is the circuit model (more strictly, the model of uniform families of acyclic quantum circuits), and the quantum Strong Church–Turing thesis states that the quantum circuit model can efficiently simulate any realistic model of computation. Several other models of quantum computation have been developed, and indeed they can be efficiently simulated by quantum circuits. Quantum circuits closely resemble most of the currently pursued approaches for attempting to construct scalable quantum computers.

Papageorgiou and Traub (see ► Quantum Algorithms and Complexity for Continuous Problems) point out that most continuous mathematical formulations arising in science and engineering can only be solved numerically and therefore approximately. There are two major motivations for studying quantum algorithms and complexity for continuous problems.

1. Are quantum computers more powerful than classical computers for important scientific problems? How much more powerful?
2. Many important scientific and engineering problems have continuous formulations. These problems occur in fields such as physics, chemistry, engineering and finance.

The continuous formulations include path integration, partial differential equations (in particular, the Schrödinger equation) and continuous optimization.

To answer the first question the classical computational complexity of the problem must be known. There have been decades of research on the classical complexity of continuous problems in the field of information-based complexity. The reason the complexity of many continuous problems is known is that adversary arguments can be used to obtain their query complexity. Regarding the second motivation, in this article they report on high-dimensional integration, path integration, Feynman path in-

tegration, the smallest eigenvalue of a differential equation, approximation, partial differential equations, ordinary differential equations and gradient estimation. They also briefly report on the simulation of quantum systems on a quantum computer.

Watrous (see ► Quantum Computational Complexity) provides a survey of quantum computational complexity, with a focus on three fundamental notions: polynomial-time quantum computations, the efficient verification of quantum proofs, and quantum interactive proof systems. Based on these notions he defines quantum complexity classes that contain computational problems of varying hardness. Properties of these complexity classes, and the relationships among these classes and classical complexity classes, are presented. As these notions and complexity classes are typically defined within the quantum circuit model, this article includes a section that focuses on basic properties of quantum circuits that are important in the setting of quantum complexity. A selection of other topics in quantum complexity, including quantum advice, space-bounded quantum computation, and bounded-depth quantum circuits, is also presented.

Grassl and Rotteler (see ► Quantum Error Correction and Fault Tolerant Quantum Computing) point out that it has been shown that even with imperfect quantum memory and imperfect quantum operations it is possible to implement arbitrary long quantum computation, provided that the failure probability of each element is below a certain threshold. They provide an overview of the ingredients leading to fault tolerant quantum computation (FTQC). In the first part, they present the theory of quantum error-correcting codes (QECCs) and in particular two important classes of QECCs, namely the so-called CSS codes and stabilizer codes. Both are related to classical error-correcting codes, so they start with some basics from this area. In the second part of the article, they present a high-level view of the main ideas of FTQC and the threshold theorem.

Lange (see ► Quantum Computing with Trapped Ions) summarizes the state-of-the-art of quantum computing with trapped ions. All the necessary components of a trapped ion quantum computer have been demonstrated, from quantum memory and fundamental quantum logic gates to simple quantum algorithms. Current experimental efforts are directed towards scaling up the small systems investigated so far and enhancing the fidelity of operations to a level where error correction can be applied efficiently. The first task at which a quantum computer is expected to outperform a classical one is the efficient simulation of quantum systems too complex for classical treatment.

Milburn and White (see ► Quantum Computing Using Optics) point out that optical implementations of quantum computing have largely focused on encoding quantum information using single photon states of light. For example, a single photon could be excited to one of two carefully defined orthogonal mode functions of the field with different momentum directions. However, as optical photons do not interact with each other directly, physical devices that enable one encoded bit of information to unitarily change another are hard to implement. In principle it can be done using a Kerr nonlinearity, but Kerr nonlinear phase shifts are too small to be useful. Knill et al. discovered another way in which the state of one photon could be made to act conditionally on the state of another using a measurement based scheme. They discuss this approach in some detail as it has led to experiments that have already demonstrated many of the key elements required for quantum computation with optics.

Lo and Zhao (see ► Quantum Cryptography) point out that the goal of quantum cryptography is to perform tasks that are impossible or intractable with conventional cryptography. Quantum cryptography makes use of the subtle properties of quantum mechanics such as the quantum no-cloning theorem and the Heisenberg uncertainty principle. Unlike conventional cryptography, whose security is often based on unproven computational assumptions, quantum cryptography has an important advantage in that its security is often based on the laws of physics. Thus far, proposed applications of quantum cryptography include quantum key distribution (abbreviated QKD), quantum bit commitment and quantum coin tossing. These applications have varying degrees of success. The most successful and important application (QKD) has been proven to be unconditionally secure. Moreover, experimental QKD has now been performed over hundreds of kilometers over both standard commercial telecom optical fibers and open-air. In fact, commercial QKD systems are currently available on the market.

Quantum Phenomena in Semiconductor Nanostructures

Umberto Ravaioli
University of Illinois at Urbana-Champaign, Urbana, USA

Article Outline

Glossary

Ballistic transport A regime of conduction where charge carriers do not exhibit appreciable scattering while moving in a certain material region.

Band structure The effective energy-momentum dispersion followed by valence and conduction electrons in a crystal, when represented as quasi-particles moving in a uniform medium equivalent to the actual crystal environment characterized by a periodic distribution of atomic potential wells.

Hot carriers Electrons or holes in a semiconductor, with energy and velocity appreciably exceeding the average values for thermal equilibrium with the lattice, typically in regions of high electric fields.

Quantum correction Addition of terms to a semi-classical model to cause particles obeying classical motion laws to follow collectively quantum behavior, for the purpose to simulate correctly quantum transport effects with considerably cheaper computations.

Scattering Collisions experienced by charge carriers moving through a semiconductor with vibrational modes of the crystal lattice (phonon scattering), with charged or neutral impurities (impurity scattering), with other charge carriers in proximity (short-range charge-charge scattering) or with collective charge vibrational modes of other carriers over a long-range (plasmon scattering).

Quantum sub-band Discrete projection of the band structure in the directions of unrestricted motion, corresponding to a specific discrete state created by size quantization.

Size quantization Separation of electronic states into discrete energy levels, in a region of restricted dimensionality, typically at interfaces between different materials which create a quantum potential well. In semiconductor devices, size quantization is normally present in the transverse cross-section of narrow conduction channels delimited by heterojunction interfaces.

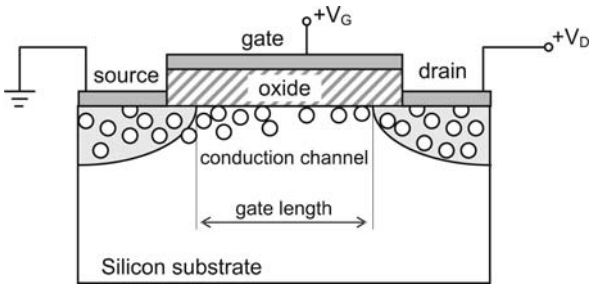
Definition of the Subject

Semiconductor devices are the building blocks of integrated circuits at the heart of computer chips and other electronics systems which are ubiquitous in modern society. While quantum physics is essential to describe the properties of semiconductor materials, transport of electronic carriers can be treated with classical laws of motion, as long as particles experience random scattering events with the crystal lattice along their path. However, in many practical devices, carriers are confined in narrow channels, so that size quantization effects should be included in the direction normal to the conduction path to model accurately particle density. As the device size is reduced to pack more and more functionality in integrated circuits, quantum coherence may eventually become important in the direction of propagation if there is a sufficient decrease of scattering events approaching quasi-ballistic transport conditions. If the device structure presents potential barriers along the direction of propagation, coherence may also manifest itself via quantum tunneling. It is essential to have a detailed understanding of quantum effects in electronic transport to design effectively devices at the nanoscale, sustain the miniaturization trends of integrated circuits, and create new engineered nanostructures.

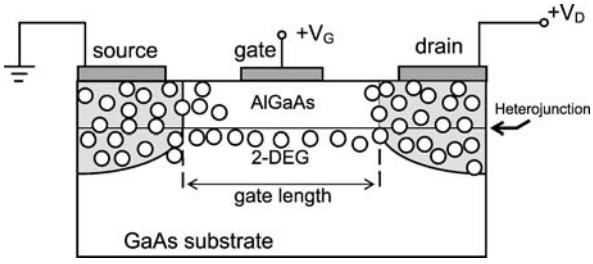
Introduction

In the second half of the 20th century, semiconductor devices have progressively made their way into nearly every aspect of technology and everyday life. The introduction of integrated circuits and microprocessors ushered the present era of computers, internet, and mobile communications. This unprecedented growth has been made possible by the ability to scale down the size of devices and to increase their speed of operation, so that the complexity of integrated circuits has steadily climbed, leading to new and more powerful applications. The pace of innovation is exemplified by Moore's law [49], an empirical observation which has predicted the doubling of transistors on a chip every two years. This has meant that over the last thirty five years the number of transistors on a commercial microprocessor has increased from slightly over two thousand to nearly one billion.

The most common transistor found in integrated circuits is the metal-oxide-semiconductor field-effect transistor (MOSFET) [50,67] based on a thin layer of silicon dioxide (SiO_2) sandwiched between a silicon conducting channel and a gate electrode, which forms a capacitive structure controlling the flow of charge in the channel when one varies the voltage applied to the gate. A schematic diagram is shown in Fig. 1. In modern com-



Quantum Phenomena in Semiconductor Nanostructures, Figure 1
Schematic diagram of the metal-oxide-semiconductor field-effect transistor (MOSFET)



Quantum Phenomena in Semiconductor Nanostructures, Figure 2
Schematic diagram of a high electron mobility transistor (HEMT)

mercial devices, the effective length of the channel has shrunk to about 50 nm and the thickness of the oxide layer can be of the order of just 1 nm.

The use of silicon MOSFETs has become so widespread for a number of reasons. Silicon is a readily available material and it is very easy to grow layers of oxide on a silicon surface. When compared to other semiconductor materials, however, silicon does not seem to possess particularly outstanding properties. Electrons move somewhat slowly under the application of an electric field, because of a relatively heavy effective mass and the high rate of scattering events with the vibrations (phonons) of the atoms forming the crystal structure. Other III–V semiconductor materials like GaAs have a smaller effective mass, so that electrons can move much faster in a certain range of electric fields with a higher drift velocity. It was widely thought at some point that for high-performance applications it would be better to adopt such semiconductor materials with an intrinsically faster mechanism for transistor switching. Also, if a heterojunction is formed between materials like GaAs and the AlGaAs alloy, electron mobility could be increased by orders of magnitudes for motion on the interface plane, leading to high electron mobility transistors (HEMT) [11,67] shown schematically in Fig. 2.

Despite their potential, III–V compound materials have not been widely employed for computing chips and have been mainly relegated to niche applications in microwaves and optoelectronics. Silicon has had of course the advantage of several more decades of development and the corresponding lower production costs and higher reliability, but in many regards the perceived weaknesses of silicon have been the reason for its success in integrated circuits applications. For miniaturization, it is very important for a device to retain its behavior essentially unchanged when scaled, so that fabrication approaches and system architectures do not need to change radically from one generation to the next. The carrier transport behavior in GaAs and other III–V materials is strongly dependent on the actual electric fields established inside the conduction channel. While electron mobility is very high at low fields, the saturation velocity, at fields established in a practical device, is actually slightly lower than in silicon, due to the intervalley scattering mechanisms to upper conduction valleys which have a higher effective mass and lower mobility [26].

To realize even higher mobility, devices need to be sufficiently small so that they can operate in the so-called overshoot regime, where electrons traverse the channel before intervalley transfer events can take place and never reach the saturation velocity in a bulk material at a comparable electric field. However, by scaling further the device, a regime of quasi-ballistic transport can be established more easily where quantum effects and coherence of the electron wave need to be dealt with. Silicon, with its heavier effective mass, has essentially retained until today most of its classical behavior as far as transport along the channel axis is concerned. Size quantization effects in the cross-section of the channel are becoming more influential but the nature of electronic transport has not changed substantially even in the nanometer size range. Generations of device designers have been able to gradually adapt their approaches to make new devices behave acceptably as they have been scaled, but the process has been an evolutionary rather than a revolutionary one. From the point of view of circuit realization, the broad adoption of silicon MOSFETs has been aided by the availability of the C-MOS architecture [50,61] based on a basic inverter structure which only conducts current when the switching between two logic states takes place. If millions of transistors are packed in the same integrated circuit it is crucial to minimize power consumption to avoid the possibility of thermal failure. The ready availability of SiO_2 oxide for the gate insulator has been of fundamental importance, since realization of comparable gate oxides in alternative material systems is in general costly, impractical or not possi-

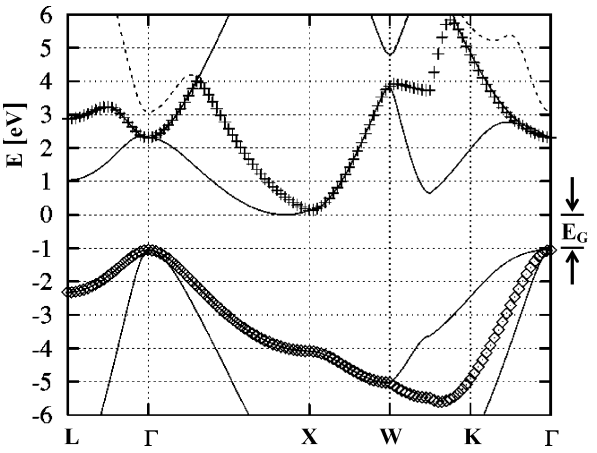
ble at all. To-date, there is simply no other technology that can approach, even remotely, what can be achieved with present C-MOS silicon technology for large-scale device integration.

This contribution deals with quantum models and simulation techniques used for the analysis of semiconductor structures within the effective mass approximation. These are applicable to devices with standard topology which are scaled down to the nanometer range. At even smaller scale, one has to deal with small clusters of materials and atomistic models are necessary to capture the full physics of the problem, since a bulk description is no longer sufficient to characterize the properties which may be strongly dependent on the cluster size. A detailed treatment of atomistic approaches is beyond the scope of this work.

Quantum Effects in Semiconductors

Quantum effects manifest themselves in a number of ways in semiconductor structures. First of all, the underlying model of transport in bulk material is based on the semiconductor band structure, which is derived from a quantum mechanical description of the crystal lattice [30]. The band structure describes the extended valence and conduction band states originating from superposition of the individual atomic states of the lattice nodes. The motion of actual electrons can then be studied in terms of quasi-particles obeying the energy-momentum dispersion relation expressed by the band structure. Negative electron quasi-particles are considered for the excited states in the conduction band and positive hole quasi-particles (vacancies of electrons) are considered in the valence band. Semiconductors normally have a band of forbidden energies (band gap) separating conduction and valence band states, as one can see in Fig. 3 showing the band structure of silicon.

Although the realistic band structure for a 3-D bulk semiconductor has in general a rather complicated shape and multiple eigenvalue branches, for simple considerations a two-band model can be used to describe hole motion at the top of the valence band and electron motion at the bottom of the conduction band, adopting an approximately parabolic energy-momentum dispersion relation. The curvatures of these parabolic dispersions define effective masses to express the kinetic energy of the quasi-particles. The effective masses for electrons and holes in a semiconductor normally differ significantly from the actual mass of an electron in vacuum. The essential quantum effects of the crystal are encapsulated in the effective mass, which is adequate to study particle transport at least in conditions not very far from equilibrium. In a bulk region,



Quantum Phenomena in Semiconductor Nanostructures, Figure 3

Simplified band structure of silicon in the standard textbook representation, drawn along crystal momentum directions connecting the main symmetry points, with the Γ point corresponding to zero crystal momentum. Calculations were performed with an empirical pseudo-potential approach. The band gap E_G between the top of the valence band and the bottom of the conduction band is slightly larger than 1 eV in silicon

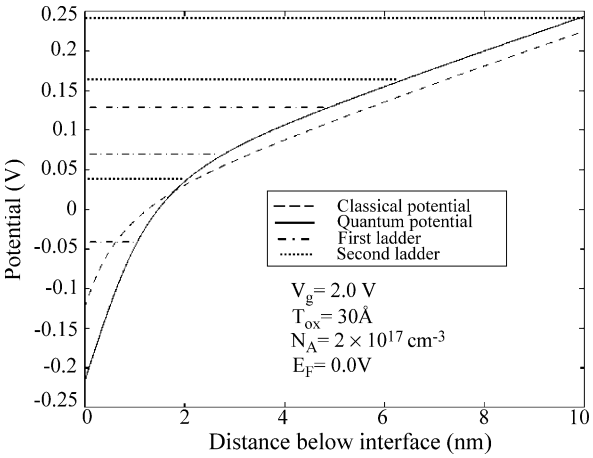
classical laws of motions can be used by applying the effective mass to relate momentum and energy. In conditions of reduced dimensionality at interfaces, the wave nature of quasi-particles may have to be considered and a simple quantum model also adopts the effective mass, for instance to express the kinetic energy term in the Schrödinger equation. The effective mass parabolic band model is widely used in a quantum model because of its simplicity. When the system is driven sufficiently out of equilibrium, a more complete band structure should be included but the complexity of the model increases rapidly resulting in considerable numerical complications for the solution of transport equations. Often, an intermediate model is acceptable for practical applications, where a nonparabolic relation is introduced to express the deviation from parabolic behavior of the dispersion relation at higher kinetic energies, still retaining the effective mass of the parabolic model as a parameter that calibrates the model.

Significant quantum effects are first encountered when an interface between dissimilar materials (heterojunction) is formed. The two materials will have in general different band gap and different positioning of conduction and valence band edges with respect to a common energy reference. As a result, at the interface there is a potential energy discontinuity between the bands. For the Si/SiO_2 interface of MOS systems, the typical discontinuity of the conduction band is on the order of 3.0 eV because the oxide is

a dielectric medium with a large band gap, while heterojunctions between usual III–V materials is typically only a fraction of eV. Carriers may be concentrated electrostatically at the heterojunction interface, creating a thin potential well, where electronic states are quantized. In the MOS system, the space charge in the semiconductor side is created by applying an appropriate potential bias on a gate electrode placed on top of the oxide layer [50,61,67]. In III–V compound systems, the technique known as modulation doping is used instead, where doping atoms are placed in the layer of semiconductor with a wider band gap [10,67]. Considering for illustration the case of *n*-type doping, the excess electrons provided by the doping migrate to the semiconductor layer with a smaller band gap where the conduction band is at lower potential energy. Thus, a layer of negative charge is formed at the interface, leaving behind a positive layer due to the ionized donor atoms in the wide band gap semiconductor. A metallic gate is also used in this case to control the flow of carriers in the channel and shut the device off.

The mobile charge at the interface forms a thin sheet, confined by the self-consistent potential well which has a nearly triangular profile close to the interface. Geometric confinement may also be added, by placing a second interface underneath, so that the narrow band gap material layer is sandwiched between layers of oxide or wider band gap semiconductor. The mobile charge sheet is usually called a two-dimensional electron gas (2-DEG). Because carriers reside in the potential well, transport normal to the interface is restricted, but they are free to move on a 2-D plane parallel to the interface. The Hamiltonian of the carriers may be decomposed into normal and parallel components. The confined quantum energy states in the potential well define sub-bands [11], for transport in the parallel plane, which are 2-D projections of the band structure. Each transverse quantum level becomes the reference zero of kinetic energy in the corresponding 2-DEG sub-band. When an electric field parallel to the interface is applied, motion of the carriers is two-dimensional and scattering is also restricted to the plane, so that the momentum of the final states cannot acquire a component normal to the interface. MOSFET and HEMT devices are realized by placing a source and a drain electrode at the two ends of a conduction channel realized with a 2-DEG. The gate electrode in the middle regulates the flow of carriers and can be biased with a repulsive potential to shut the device off.

For illustration, Fig. 4 shows an example of self-consistent equilibrium calculations of electron energy states in MOS structure at the Si/SO_2 interface aligned on the [100] direction of the crystal. Silicon has six degenerate

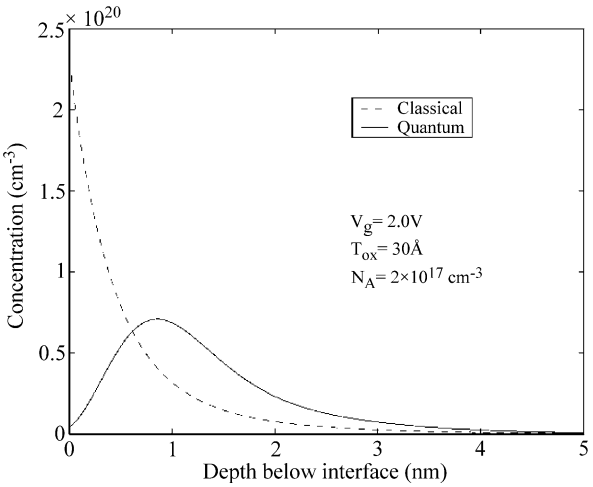


Quantum Phenomena in Semiconductor Nanostructures, Figure 4

Example of potential distribution and energy states obtained on the silicon layer of an MOS capacitor structure. The reference zero energy is the Fermi level in the semiconductor

equivalent valleys, close to the X point of the band structure which, at relatively low energy values, are associated with ellipsoidal iso-energy surfaces aligned two by two along the three principal axes of symmetry perpendicular to equivalent <100> planes. On the interface, two ellipsoids project as degenerate energy circles, corresponding to a first set of quantum state solutions (first ladder). The other four ellipsoids project sideways as ellipses, corresponding to a second set of solution at higher eigenenergies (second ladder). Figure 4 shows the first three energy solutions for each ladder and the corresponding self-consistent potential energy profile obtained from the quantum calculation as well as the profile obtained with a classical model neglecting the quantum states. The corresponding quantum and classical charge distributions are shown in Fig. 5. One can see that the classical model predicts a maximum of charge density exactly at the interface. The quantum density, obtained from the detail of energy states and wave functions, shows nearly zero density at the interface, while the maximum of charge has shifted to a depth of almost 1.0 nm.

As a further step, motion can be restricted in one of the parallel plane directions, either by geometrical confinement, for instance by etching the material, electrical confinement by placing additional electrodes or both. One could obtain in this way a one-dimensional electron gas, confining the carriers to a quantum wire [54]. A 1-D subband structure is associated with the motion along the channel direction corresponding to the only remaining degree of freedom for kinetic energy. In such a structure,



Quantum Phenomena in Semiconductor Nanostructures, Figure 5

Example of inversion electron charge concentration under the gate of an MOS structure as a function of depth below the oxide interface

scattering events can only result in a final state with momentum pointing forward or backwards, with reference to the axis of the conduction channel. Motion in a section of quantum wire can also be further restricted by placing geometric or electrostatic barriers, to create a 3-D cavity called a quantum dot, where states are fully quantized [58]. There are no sub-bands associated with these discrete quantum states because motion is completely restricted. Coupling of the states with the environment outside the cavity may take place, for instance by penetration of evanescent waves through potential barriers delimiting the cavity.

The quantum effects discussed above are the result of size quantization. Restriction of motion in a specific direction causes the energy to be quantized in that direction, in terms of discrete eigenvalues, but quantum effects may also appear along a direction of unrestricted motion where energy belongs to a continuum distribution, if there are sufficiently abrupt potential or geometric discontinuities. A step in potential energy, for instance, creates quantum mechanical reflection, while motion against a thin potential barrier may result in tunneling with evanescent propagation inside the barrier region. Interesting phenomena of mode coupling take place in a quantum wire undergoing changes in the geometric cross-section, because of the relative shifts of the transverse energy states [44].

A much-studied quantum problem in semiconductor devices is resonant tunneling [22,36,56,64]. Here, two narrowly spaced layers of a wide band gap material are embedded in a semiconducting medium of lower band gap, creating two potential barriers. The region between the barriers is a potential well with quantum states which are resonantly coupled with the continuum states on the two sides of the structure. Tunneling is mostly favorable at incident energies corresponding to these resonant states, because the multiple reflections at the interface discontinuities create a constructive interference for transmission and destructive interference for reflection. The current-voltage characteristics peak in correspondence of the resonances, therefore exhibiting negative differential resistance behavior. One can extend the concept and create a large number of wide band gap material layers, thus creating an artificial material called a superlattice [72]. In the direction normal to the interfaces, the coupling between multiple quantum reflections creates ranges of energies with favorable transmission (minibands) and forbidden energy gaps that create an effective artificial band structure.

Size Quantization

In the size quantization problem, discrete stationary states are associated with mobile carriers, obeying the time-independent Schrödinger equation for the coordinates restricted to motion with a general form

$$\hat{H}\psi_j = E_j\psi_j \,, \tag{1}$$

where $\hat{H}$ is the Hamiltonian, E_j the energy eigenvalues of the confined states and ψ_j the corresponding wave functions. For simplicity, the following consideration will be for electrons in the conduction band. Similar considerations hold in principle for holes in the valence band, although the details of the band structure may require some more involved formalism.

For 1-D confinement, where the motion in the quantum channel is delimited by a planar interface, e. g. a heterojunction, the electron states can be characterized by an envelope wave function

$$\psi(\vec{r}) = \xi_j(z)\exp(i\vec{k}\cdot\vec{r}) \,, \tag{2}$$

where z is the direction perpendicular to the heterojunction, and $\vec{r}$ and $\vec{k}$ are the 2-D position vector and wave vector for motion parallel to the interface. In the plane parallel to the interface the wave function component is simply a plane wave, while confinement is described by the function $\xi_j(z)$ which corresponds to the jth confined sub-band and satisfies the 1-D time-independent Schrödinger equation in the z direction

$$-\frac{\hbar^2}{2m^*}\frac{\mathrm{d}}{\mathrm{d}z}\left(\frac{1}{m^*}\frac{\mathrm{d}\xi_j(z)}{\mathrm{d}z}\right) + U(z)\xi_j(z) = E_j\xi_j(z) \,. \tag{3}$$

The potential energy $U(z)$ is given by

$$U(z) = -e\phi(z) + V_h(z) + V_{\text{ex}}(z)\,, \tag{4}$$

where $\phi(z)$ is the electrostatic potential, $V_h(z)$ is a step function describing the interface barrier, and $V_{\text{ex}}(z)$ is the local exchange-correlation potential. Assuming the case of an electron channel obtained on a p-type substrate, the electrostatic potential is obtained from a solution of Poisson's equation

$$\frac{\mathrm{d}}{\mathrm{d}z}\left(\varepsilon_d \frac{\mathrm{d}\phi(z)}{\mathrm{d}z}\right) = -e\left[\sum\nolimits_j N_j \left|\xi_j(z)\right|^2 - p\,(z,\phi) + N_{\text{A}}^-(z) - N_{\text{D}}^+(z)\right] = -\rho\,. \tag{5}$$

$N_{\text{A}}(z)$ and $N_{\text{D}}(z)$ are the concentrations of ionized acceptor and donor dopants, N_j represents the sub-band electron concentration, p is the hole concentration in the substrate, ε_d is the space dependent permittivity and ρ represents the net charge density. The hole concentration may be approximated as a classical equilibrium distribution with a space profile that is nonlinearly dependent on the electrostatic potential. The formulation for N_j involves the integral of the Fermi function which, in the case of a 2-DEG, can be integrated analytically, giving [11]

$$N_j = \frac{D}{e} g_{\text{v}} k_{\text{B}} T \ln\left[1 + \exp\left(\frac{E_F - E_j}{k_{\text{B}} T}\right)\right]. \tag{6}$$

Here, g_{v} is the valley degeneracy for the specific eigenvalue ladder, related to the band structure symmetry and the projection of the conduction band valleys on the interface plane. The 2-DEG density of states is a constant as a function of energy and it is given by

$$D = \frac{em^*}{\pi\hbar^2}\,. \tag{7}$$

If a qualitative approximation is sought, starting from a fixed potential profile, solutions of the Schrödinger equation for the wave functions may just be available analytically in some cases. For instance, for a potential approximated by an infinite barrier at a heterojunction interface ($z = 0$) and by a triangular distribution $V(z) = eF_{\text{s}}z$ for $z > 0$ in the semiconductor (where F_{s} is an effective electric field assumed to be constant throughout the layer) solutions for wave functions and energy eigenvalues can be expressed in terms of Airy functions [1]

$$\xi_j(z) = Ai\left\{\left(\frac{2m^* eF_{\text{s}}}{\hbar^2}\right)^{\frac{1}{3}}\left[z - \left(\frac{E_j}{eF_{\text{s}}}\right)\right]\right\}, \tag{8}$$

$$E_j(z) = \left(\frac{\hbar^2}{2m^*}\right)^{\frac{1}{3}}\left[\frac{3}{2}\pi eF_{\text{s}}\left(j + \frac{3}{4}\right)\right]^{\frac{2}{3}}. \tag{9}$$

For double confinement, the mathematical formulation of the problem is very similar with more general differential operators [38]

$$-\frac{\hbar^2}{2m^*}\nabla\cdot\left(\frac{1}{m^*}\nabla\psi_j\right) + U(z)\psi_j = E_j\psi_j\,, \tag{10}$$

$$\nabla\cdot(\varepsilon_d\nabla\phi) = -e\left[\sum\nolimits_j N_j\left|\psi_j\right|^2 - p\,(\phi) + N_{\text{A}}^- - N_{\text{D}}^+\right]. \tag{11}$$

The main formal difference is in the expression for the sub-band carrier concentration since, except for the 2-DEG case, the Fermi function cannot be integrated exactly. This quantity is now expressed in terms of the Fermi integral of order $-1/2$

$$N_j = \frac{g_v}{\pi}\left(\frac{2m^* k_{\text{B}} T}{\hbar^2}\right)^{\frac{1}{2}} F_{\frac{-1}{2}}\left(\frac{E_F - E_j}{k_{\text{B}} T}\right), \tag{12}$$

which needs to be evaluated numerically. An efficient and very accurate approach is based on rational function approximations [2]. For practical applications, the most interesting quantity is the quantum electron density, which is obtained from combined information given by the wave function and the energy eigenstates as

$$n_{\text{q}}(\phi) = \frac{g_v}{\pi}\left(\frac{2m^* k_{\text{B}} T}{\hbar^2}\right)^{\frac{1}{2}} \sum_j \psi_j^2 F_{\frac{-1}{2}}\left(\frac{E_F - E_j}{k_{\text{B}} T}\right). \tag{13}$$

The time-independent Schrödinger equation specifies an eigenvalue problem and the Poisson equation is an elliptical partial differential equation. Because of the different mathematical nature, the two equations cannot be solved simultaneously and an iterative procedure is necessary instead. A simple iteration by itself does not converge and usually an underrelaxation in the electron density is applied with a relaxation parameter $\omega^{(m)}$ which may have to be adaptively modified as the iteration progresses. The algorithm flow of this simple approach can be summarized as follows:

(1) Solve the nonlinear Poisson equation using the quantum electron density from the last iteration

$$\nabla\cdot\left(\varepsilon\nabla\phi^{(m+1)}\right) = -\rho\left[n_{\text{q}}^{(m)}, \phi^{(m+1)}\right].$$

(2) Solve Schrödinger equation using the old $n_{\text{q}}^{(m)}$

$$\hat{H}\left[\phi^{(m+1)}, V_{\text{xc}}\left(n_{\text{q}}^{(m)}\right)\right]\psi_i^{(m+1)} = E_i^{(m+1)}\psi_i^{(m+1)}.$$

(3) Calculate an intermediate new quantum electron density $n_{\mathrm{q,int}}^{(m+1)}$

$$n_{\mathrm{q,int}}^{(m+1)} = n_{\mathrm{q}}\left[E_i^{(m+1)}, \psi_i^{(m+1)}\right] .$$

(4) Underrelax in n_{q} to achieve convergence

$$n_{\mathrm{q}}^{(m+1)} = \omega^{(m+1)} n_{\mathrm{q,int}}^{(m+1)} + \left(1 - \omega^{(m+1)}\right) n_{\mathrm{q}}^{(m)} .$$

(5) Repeat the iteration until n_{q} becomes stationary

$$\left\| n_{\mathrm{q}}^{(m+1)} - n_{\mathrm{q}}^{(m)} \right\|_2 \leq \text{tolerance} .$$

The problem with this method is the inherent instability of the outer iteration which is controlled by the underrelaxation procedure only. The necessary relaxation parameter $\omega^{(m)}$ is not known in advance and needs to be readjusted during the iteration. If $\omega^{(m)}$ is too large, the integral of the quantized charge oscillates without converging or if it is too small, convergence may be unbearably slow.

The algorithm may be modified to address these shortcomings, in a way that partially decouples the differential equations and dampens the electric charge oscillation. If one knew the exact dependence of the quantum electron density n_{q} on the electrostatic potential, it would be sufficient to solve the nonlinear Poisson equation, without the need for coupling with the Schrödinger equation by an outer iteration, but what one can do practically is to find a suitable approximate expression for the quantum electron density $\tilde{n}_{\mathrm{q}}(\phi)$ and use such an expression in a predictor-corrector type of approach [71]. This procedure would approximately decouple both equations and move most of the nonlinearities in a Poisson equation of the type

$$\nabla \cdot (\varepsilon \nabla \phi) = -\rho\left[\tilde{n}_{\mathrm{q}}(\phi), \phi\right] . \tag{14}$$

The predicted result for n_{q} and ϕ from this equation would then be corrected in an outer iteration step by an exact solution of Schrödinger equation. The electrostatic potential enters the quantum electron density $n_{\mathrm{q}}(\phi)$ through the potential dependence of sub-band energy levels and wave functions, following the form of Eq. (13)

$$n_{\mathrm{q}}(\phi) = \frac{g_v}{\pi}\left(\frac{2m^* k_{\mathrm{B}} T}{\hbar^2}\right)^{\frac{1}{2}} \cdot \sum_j \psi_j^2(\phi) F_{\frac{-1}{2}}\left(\frac{E_F - E_j(\phi)}{k_{\mathrm{B}} T}\right) . \tag{15}$$

By using the derivative property of the Fermi–Dirac integral,

$$\frac{\mathrm{d}}{\mathrm{d}x} F_k(x) = F_{k-1}(x) , \tag{16}$$

one can show [71] that, under a perturbation $\delta\phi$ of the potential, the corresponding variation in quantum density can be approximated as

$$\delta\tilde{n}_{\mathrm{q}}(\phi, \delta\phi) = \frac{g_v}{\pi}\left(\frac{2m^* k_{\mathrm{B}} T}{\hbar^2}\right)^{\frac{1}{2}} \cdot \sum_j \psi_j^2(\phi) F_{\frac{-3}{2}}\left(\frac{E_F - E_j(\phi)}{k_{\mathrm{B}} T}\right) \frac{q\delta\phi}{k_{\mathrm{B}} T} . \tag{17}$$

Since for a perturbation one can write

$$\tilde{n}_{\mathrm{q}}(\phi + \delta\phi) = \tilde{n}_{\mathrm{q}}(\phi) + \delta\tilde{n}_{\mathrm{q}}(\phi, \delta\phi) , \tag{18}$$

using once more the derivative properties of the Fermi–Dirac integrals, one obtains

$$\tilde{n}_{\mathrm{q}}(\phi + \delta\phi) = \frac{g_v}{\pi}\left(\frac{2m^* k_{\mathrm{B}} T}{\hbar^2}\right)^{\frac{1}{2}} \cdot \sum_j \psi_j^2(\phi) F_{\frac{-1}{2}}\left(\frac{E_F - E_j(\phi) + q\delta\phi}{k_{\mathrm{B}} T}\right) . \tag{19}$$

Compared with the expression for $n_{\mathrm{q}}(\phi)$ one can see that the effect of the potential perturbation is simply translated into a variation of the energy levels by a corresponding potential amount

$$E_n(\phi) \to E_n(\phi) - q\delta\phi . \tag{20}$$

Using this result, the predictor-corrector procedure starts with the solution of a nonlinear Poisson equation

$$\nabla \cdot (\varepsilon \nabla \phi) = -q\left[\tilde{n}_{\mathrm{q}}(\phi) - p(\phi) - N_{\mathrm{D}}^{+}(\phi) + N_{\mathrm{A}}^{-}(\phi)\right] , \tag{21}$$

where we use the potential dependent solution for the quantum electron density (predictor)

$$\tilde{n}_{\mathrm{q}}(\phi) = \frac{g_v}{\pi}\left(\frac{2m^* k_{\mathrm{B}} T}{\hbar^2}\right)^{\frac{1}{2}} \cdot \sum_j \left(\psi_j^{(m)}\right)^2 F_{\frac{-1}{2}}\left(\frac{E_F - E_n^{(m)} + q\left(\phi - \phi^{(m)}\right)}{k_{\mathrm{B}} T}\right) . \tag{22}$$

The superscripts (m) denote quantities obtained at the previous outer iteration step. Solution of the nonlinear Poisson equation is conveniently accomplished with a Newton–Raphson procedure, since terms of the Jacobian matrix are easily obtained in terms of $F_{-3/2}$. The electrostatic

potential $\phi^{(m+1)}$ obtained from solving the Poisson equation is then used within the Schrödinger equation (corrector) formulated as

$$-\frac{\hbar^2}{2}\nabla\cdot\left(\frac{1}{m^*}\nabla\psi_j^{(m+1)}\right)+\left[V_h-q\phi^{(m+1)}+V_{\rm xc}\left(\tilde{n}_{\rm q}^{(m+1)}\right)\right]\psi_j^{(m+1)}=E_j^{(m+1)}\psi_j^{(m+1)}\,, \quad (23)$$

to calculate the update of the corrected quantum density

$$n_{\rm q}^{(m+1)}=\frac{g_v}{\pi}\left(\frac{2m^*k_{\rm B}T}{\hbar^2}\right)^{\frac{1}{2}}\cdot\sum_j\left(\psi_j^{(m+1)}\right)^2F_{\frac{-1}{2}}\left(\frac{E_F-E_j^{(m+1)}}{k_{\rm B}T}\right). \quad (24)$$

The predictor-corrector algorithm steps can be summarized as follows:

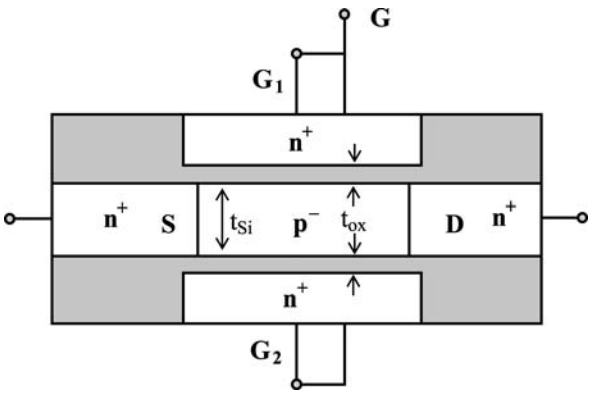
(1) Solve the nonlinear Poisson equation using the approximation $\tilde{n}_{\rm q}(\phi)$.
(2) Solve Schrödinger equation using the latest predictor value $\tilde{n}_{\rm q}^{(m+1)}$.
(3) Calculate the exact quantum electron density $n_{\rm q}^{(m+1)}$.
(4) Repeat iteration until the electron density becomes stationary

$$\left\|n_{\rm q}^{(m+1)}-n_{\rm q}^{(m)}\right\|_2\le\text{tolerance}\,.$$

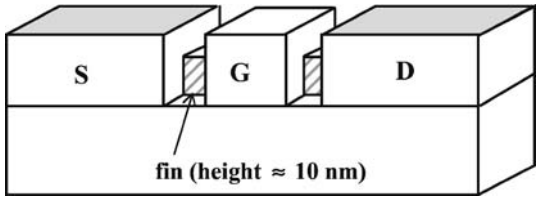
Note that at step (2) the latest predictor value $\tilde{n}_{\rm q}^{(m+1)}$ is also used to evaluate the exchange correlation term in the Schrödinger equation, since practical observation suggests that convergence would be much worse if the previous value $\tilde{n}_{\rm q}^{(m)}$ was used. A range of numerical tests have also shown that the predictor-corrector algorithm does not need an additional underrelaxation and that the simple outer iteration is inherently stable and convergent.

Size quantization is important in narrow channel devices. As the channel is shortened, the bulk region under the interface makes it difficult to shut the device off completely because stray carriers may transfer from source to drain causing a parasitic current. The situation is in principle improved by using a double gate MOSFET structure [21], as shown in Fig. 6, which reduces the silicon region to a thin slab which may emphasize the effects of size quantization since two interfaces are now present. Realization of a double gate structure with horizontal interfaces is technologically difficult, because a layer of silicon would need to be grown on oxide.

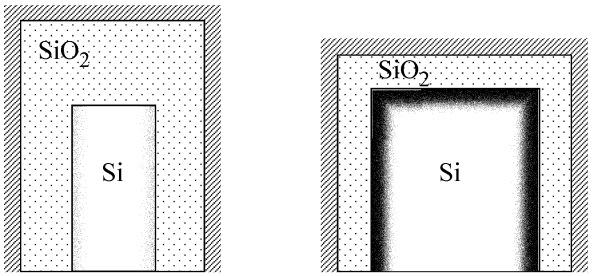
An approximate practical realization is made with a fin structure with oxide and gate contact wrapped around as



Quantum Phenomena in Semiconductor Nanostructures, Figure 6
Schematic structure of a double gate MOSFET

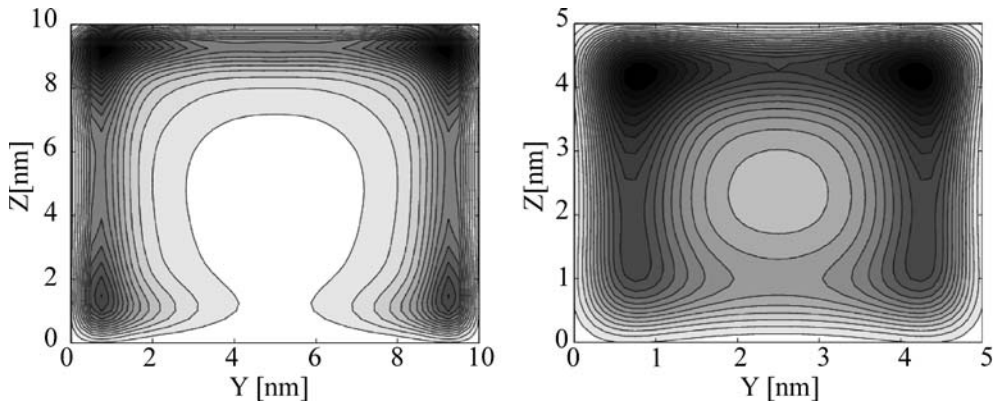


Quantum Phenomena in Semiconductor Nanostructures, Figure 7
Structure of a finFET to realize approximately a double gate MOSFET from the two vertical oxide interfaces with the silicon fin

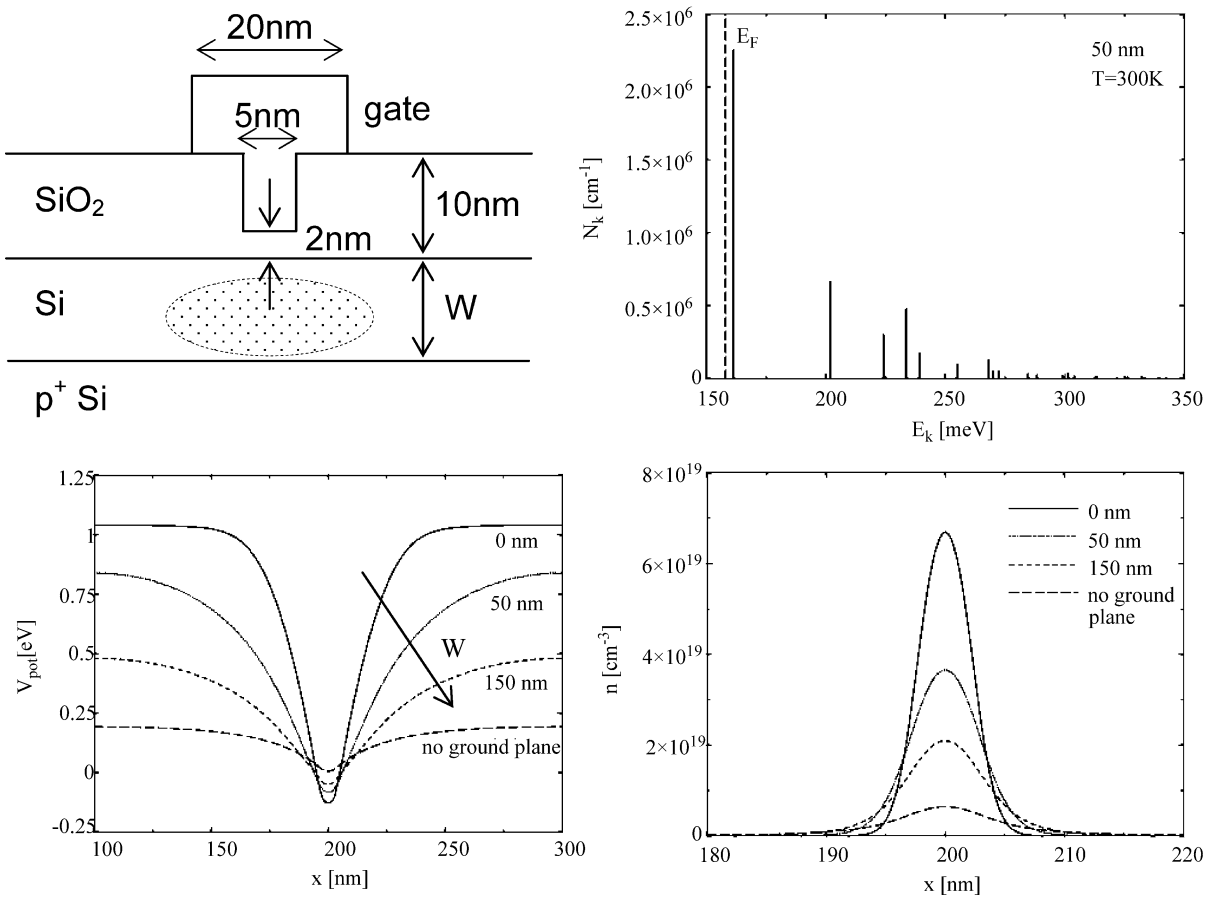


Quantum Phenomena in Semiconductor Nanostructures, Figure 8
Cross-section of conduction channels in finFET (a) and tri-gate MOSFET (b)

in the finFET of Fig. 7 [12,37,52], so that on the cross-section of the channel in Fig. 8a the electronic charge is concentrated at the two vertical interfaces where the oxide is thinner. To increase the total charge in the channel, the channel cross-section may be widened as in Fig. 8b to realized the so-called tri-gate MOSFET [69] where charge accumulates at the three interfaces controlled by the wrap-



Quantum Phenomena in Semiconductor Nanostructures, Figure 9
Example of self-consistent calculation of quantum electron distribution on the square cross-section of tri-gate MOS channels with sides 10 nm (*left*) and 5 nm (*right*). One may notice the concentration peaks at the *top* corners of the structure



Quantum Phenomena in Semiconductor Nanostructures, Figure 10
***Top left*: quantum wire formation under a trench in the gate oxide for a structure formed with a layer of undoped silicon on top of a highly doped ground plane; *top right*: energy states in the structure with thickness of the undoped silicon layer W = 50 nm; *bottom left*: trench confining potential as function of layer of thickness W; *bottom right*: corresponding total charge densities in the quantum well. Gate voltages are 1.6 V, 1.0 V, 0.7 V and 0.3 V for increasing W. The metal gate is assumed to be aluminum**

around gate structure. Examples of charge distribution obtained on the cross-section by 2-D self-consistent solution of the Schrödinger/Poisson problem are shown in Fig. 9, for tri-gate square cross-sections with sides of 10 nm and 5 nm. Peak concentration of charge close to the top corners of the structure is noticeable.

A simpler planar structure may be realized instead with a gate contact embedded in a trench of the oxide (Fig. 10) to form a quasi-1-D quantum wire, as studied in [24]. This structure may be advantageous because of reduced interface scattering and simpler fabrication, but carriers are confined in the channel by an electrostatic potential, rather than physical interface walls, so that there is less channel isolation. Figure 10 shows several examples of the potential well formed by the trench structure of the gate for various thicknesses W of the undoped silicon layer placed above a highly p-doped ground plane.

Inclusion of Realistic Band Structure

For high carrier concentrations or high confinement barriers in a conduction channel, the range of energies necessary to account for size quantization may be much above the conduction band edge, so that a simple parabolic description of the band structure is questionable. The next approximation level is to describe the deviation from the simple behavior with the addition of non-parabolic terms. There are various possible approaches which can be followed to improve the description of the band structure. A general expression provides the energy as a series expansion

$$E = a_0 + a_2 k^2 + a_4 k^4 + a_6 k^6 + \cdots = \sum_{i=0}^{\infty} a_{2i} k^{2i} \,. \quad (25)$$

Examples in the literature simply add the fourth order term to include nonparabolic effects in the Schrödinger equation [15,53]. Other attempts have tried to define an energy-dependent effective mass, but such a formulation is in general not a sound approach and may lead to erroneous results [53]. Another commonly followed approach is to represent the wave vector instead as a series expansion of energy

$$\frac{\hbar^2 k^2}{2m^*} = E\left(1 + \alpha E + \beta E^2 + \gamma E^3 + \cdots\right) \,. \quad (26)$$

Most commonly, the expression is truncated to include only up to the quadratic energy term, leading to the standard nonparabolic approximation widely used in the literature

$$\frac{\hbar^2 k^2}{2m^*} = E\left(1 + \alpha E\right) \,, \quad (27)$$

where α is the coefficient of nonparabolicity with dimensions of an inverse energy. Application of this simple nonparabolic form into the Schrödinger equation for the case of a 2-DEG system was presented in [43].

The nonparabolic dispersion relation can be easily inverted by solving the quadratic equation for energy and then applying a series to the resulting square root

$$\begin{aligned} E &= \frac{1}{2\alpha}\left[\sqrt{1 + 4\alpha\frac{\hbar^2 k^2}{2m^*}} - 1\right] \\ &= \frac{1}{2\alpha}\sum_{n=1}^{\infty}\binom{\frac{1}{2}}{n}\left[4\alpha\frac{\hbar^2 k^2}{2m^*}\right]^2 \,. \end{aligned} \quad (28)$$

If we then substitute the crystal momentum k by the momentum operator $i\nabla$ we may write the Schrödinger equation in the form [23,43]

$$\begin{aligned} \frac{1}{2\alpha}\sum_{n=1}^{\infty}\binom{\frac{1}{2}}{n}\left[-4\alpha\frac{\hbar^2}{2}\right]^n\left[\frac{\nabla^2}{m^*}\right]\psi(\vec{r}) + U(\vec{r})\psi(\vec{r}) \\ = \varepsilon\psi(\vec{r}) \,. \end{aligned} \quad (29)$$

One should realize that if the dispersion relation is not strictly parabolic, the effective mass does not contain complete information on the lattice periodic potential effects and that the eigenfunctions are not strictly approximated by plane waves. However, rather than modifying the eigenfunctions, this procedure chooses to modify the form of the Schrödinger equation. In order to proceed, an ansatz is necessary on the form of the wave function, by associating plane waves to the coordinates of unrestricted motion. For instance, if a 1-D confining potential is present along the z-direction, one can write explicitly

$$\psi(\vec{r}) = \mathrm{e}^{i(k_x x + k_y y)}\xi(z) \,, \quad (30)$$

or for double confinement on the (y, z) plane

$$\psi(\vec{r}) = \mathrm{e}^{i\,k_x x}\zeta(y)\xi(z) \,. \quad (31)$$

The forms used for the wave function contain an implicit assumption that the symmetry axes of the effective-mass tensor are along the Cartesian coordinate axes. The energy can be split into components that individually still verify the nonparabolic dispersion relation along the directions of unrestricted motion. For single confinement one can use $\varepsilon = \varepsilon_z + \varepsilon_\parallel$ with

$$\varepsilon_\parallel(1 + \alpha\varepsilon_\parallel) = \frac{\hbar^2}{2}\left[\frac{k_x^2}{m_x} + \frac{k_y^2}{m_y}\right] \,, \quad (32)$$

leading, after several manipulations, to the modified Schrödinger equation [43]

$$\frac{1+2\alpha\varepsilon_{\|}}{2\alpha}\sum_{l=1}^{\infty}\binom{\frac{1}{2}}{l}\left(-4\alpha\frac{\hbar^2}{2m_z\left(1+2\alpha\varepsilon_{\|}\right)^2}\right)^l \cdot\frac{\partial^{2l}\xi(z)}{\partial z^{2l}}+U(z)\xi(z)=\varepsilon_z\xi(z)\,. \quad (33)$$

For a quantum wire system which is doubly confined, there is no straightforward way to develop a general model for nonparabolicity. An approximate solution is to extend the procedure above for 2-DEG, by decomposing the solution space into two transverse directions and solving two 1-D quantization problems. Following a similar procedure as for 1-D confinement, one obtains the analogous modified equation [23]

$$\frac{1+2\alpha\varepsilon_{\|}}{2\alpha}\sum_{l=1}^{\infty}\binom{\frac{1}{2}}{l}\left(-4\alpha\frac{\hbar^2}{2\left(1+2\alpha\varepsilon_x\right)^2}\right)^l \cdot\sum_{q=0}^{l}\binom{l}{q}\left[\frac{1}{m_y^{l-q}}\frac{\partial^{2(l-q)}\zeta(y)}{\partial y^{2(l-q)}}\frac{1}{m_z^q}\frac{\partial^{2q}\xi(z)}{\partial z^{2q}}\right] +U(y,z)\zeta(y)\xi(z)=\varepsilon_{yz}\zeta(y)\xi(z)\,. \quad (34)$$

Both equations recover the usual parabolic form in the limit $\alpha\to 0$. It is of course very difficult to use directly the modified forms of the Schrödinger equation because they involve infinite series. However, one can derive a dispersion relationship if appropriate test functions are formulated for given boundary conditions. For example, considering double confinement, we have for $\varepsilon_{yz} > U(y,z)$ (oscillatory condition for electron energy above the conduction band edge)

$$\zeta(y)=Ae^{ik_yy}+Be^{-ik_yy}\,,\quad \xi(z)=Ce^{ik_zz}+De^{-ik_zz}\,, \quad (35)$$

and for $\varepsilon_{yz} < U(y,z)$ (evanescent condition with decaying behavior)

$$\zeta(y)=Ae^{k_yy}+Be^{-k_yy}\,,\quad \xi(z)=Ce^{k_zz}+De^{-k_zz}\,, \quad (36)$$

with dispersion relationship (choosing "+" for oscillatory and "−" for evanescent)

$$\pm\left[\varepsilon_{yz}-U(y,z)\right]\left\{1+2\alpha\varepsilon_x\pm\alpha\left[\varepsilon_{yz}-U(y,z)\right]\right\} =\frac{\hbar^2}{2}\left(\frac{k_y^2}{m_y}+\frac{k_z^2}{m_z}\right)\,. \quad (37)$$

In order to get a numerical solution, one can use a shooting method. First, a value of the transverse energy ε_{yz} is selected as a guess. Then the dispersion relation is applied to test whether the corresponding wave function satisfies the necessary conditions as number of zeros and derivative sign, within a specified accuracy for satisfying the boundary conditions considering continuity and probability current of the envelope wave function.

The full numerical band structure may be included in a 3-D treatment of the Schrödinger equation [70]. This method is useful for relatively large structures for which an atomistic model is impractical while at the same time a simple analytical band approximation is inadequate. The kinetic energy of carriers can be described by a function $E(\vec{p})$ which may not be available in analytical form but only as an interpolated table, as is common for the output of band structure calculations. The usual quantization rules still apply and one may replace the momentum $\vec{p}$ by the momentum operator $\hat{p}$ to obtain the Hamiltonian. The Schrödinger equation for envelope wave functions, describing carrier motion under the influence of an external potential $V(\vec{x})$, can be written as

$$i\hbar\frac{\partial\psi(\vec{x},t)}{\partial t}=E(\hat{p})\psi(\vec{x},t)+V(\vec{x})\psi(\vec{x},t) \qquad \hat{p}=\frac{\hbar}{i}\nabla\,. \quad (38)$$

This model is reasonable as long as $V(\vec{x})$ does not vary too rapidly within a lattice constant a and the dimensions of the device structure are much larger than a as well.

It is not straightforward to solve an equation of this kind, because the kinetic energy operator may contain high order powers of the momentum. Since $\hat{p}$ is proportional to the gradient vector in the position representation, this means that one has to solve an equally high-order differential equation. Numerical solutions based on finite differences or finite elements require a Taylor expansion of $E(\vec{p})$, for instance, at least up to an order $n+1$ in $\vec{p}$ to compute $E(\vec{p})$ to order n. While it is numerically possible to implement such a high order scheme, the approach becomes inefficient for large n, and depending on the functional form of $E(\vec{p})$ the convergence radius of its Taylor series may be too small for polynomial approximations.

Since the kinetic energy operator $E(\vec{p})$ is diagonal in momentum space, a spectral solution approach using Fourier transforms seems a natural choice for this problem. Starting from the equation in position space, we may insert Fourier transforms around the kinetic energy operator, which then reduces to simple multiplication in momentum space. The Schrödinger equation has the form

$$i\hbar\frac{\partial\psi(\vec{x},t)}{\partial t}=\mathrm{FT}^{-1}\left[E(\vec{p})\mathrm{FT}\psi(\vec{x},t)\right]+V(\vec{x})\psi(\vec{x},t), \quad (39)$$

where one could view

$$E(\hat{p})\psi(\vec{x},t) = \mathrm{FT}^{-1}\left[E(\vec{p})\mathrm{FT}\psi(\vec{x},t)\right] , \tag{40}$$

as the definition for the operator $E(\hat{p})$ in position space. An alternative approach could be to express the Schrödinger equation in momentum space by inserting Fourier transforms around the potential energy as

$$i\hbar\frac{\partial\psi(\vec{p},t)}{\partial t} = E(\vec{p})\psi(\vec{p},t) + \mathrm{FT}\left[V(\vec{x})\mathrm{FT}^{-1}\psi(\vec{p},t)\right] , \tag{41}$$

which provides a Hamiltonian similar to those employed in density functional calculation. In both position and momentum formulation the numerically difficult problem of applying a high-order kinetic energy operator on a wave function has now been reduced to a more manageable calculation of Fourier transforms.

Electronic structure calculations are usually formulated in momentum space and device simulation in position space. It is not clear which of the two formulations might lead to a numerically superior algorithm for the problem at hand, but position space seems to be a more natural choice. In any case, fast Fourier transforms must be used for an efficient numerical implementation, involving only $O(N\log N)$ computational steps on a grid with N nodes, as compared to $O(N^2)$ for other methods.

The use of fast Fourier transforms imposes restrictions on the choice of computational grid. Grid lines should be equidistant in each coordinate direction and for most available fast Fourier transforms the number of grid lines should be a power of 2 in each direction as well. If the grid is too coarse, only parts of momentum space are sampled providing insufficient spatial resolution, as one would readily expect. However, also an excessively fine grid creates problems as well, since one would obtain a computational first Brillouin zone larger than the lattice one, opening the question of how to deal numerically with the states at high momentum values outside the lattice Brillouin zone, which would create spurious solutions if mapped inside the zone. Although this approach allows one, in principle, to account for general band structures, if a parabolic band is acceptable in the quantum problem at hand, it is always more efficient to use a traditional finite difference solver for large problems. Model calculations using this generalized band approach in 3D silicon quantum cavity structures were demonstrated in [70] using a numerical tabulation of a nonparabolic ellipsoidal valley.

Brief Survey of Quantum Transport Models

While size quantization models have been well developed, the treatment of transport remains a complicated quantum problem even in the assumption of envelope wave function under the effective mass approximation. In the most elementary applications, the single particle time-dependent Schrödinger equation as given in Eq. (38) can be used to probe ballistic transport in a nanostructure, assuming a parabolic band [7,25]. Coupled with appropriate conditions to simulate open contacts, one can study transients and switching between ON and OFF states in a quantum structure [59]. The standard approaches that can be used apply finite differences with a direct discretization in time [7,25,41,45,47,59,62] or the split-operator technique [79] based on a spectral method requiring Fast Fourier Transform at each step. The main limitation of the time-dependent Schrödinger equation is that it can treat a single energy at a time and that is naturally suitable for ballistic transport since scattering cannot be easily included, unless a set of coupled equations is set up. Simulation of contacts for an open system requires the numerical development of absorbing boundary conditions, which is not a trivial problem [41,45,47,59,62,68,79].

A more sophisticated transport model that includes a set of states in the formulation is based on the density matrix [48], the evolution of which is governed by the Liouville–von Neumann equation. For a simple parabolic band single-particle Hamiltonian, the evolution equation has the form

$$i\hbar\frac{\partial\rho}{\partial t} = -\frac{\hbar^2}{2m^*}\left(\frac{\partial^2}{\partial x^2} - \frac{\partial^2}{\partial x'^2}\right)\rho + \left[U(x) - U(x')\right]\rho , \tag{42}$$

with the density matrix $\rho(x,x')$ defined as a summation of the real-valued probabilities p_i over a complete set of states i in the system to be studied

$$\rho\left(x,x'\right) = \sum_j p_j \left\langle x|j\right\rangle \left\langle j|x'\right\rangle . \tag{43}$$

Note that, for transport through an open system, a reasonable normalization of the density matrix is not to equate the trace to one, but to relate it to the actual particle density in the system, giving also an intuitive picture of the density matrix. Another approach uses instead the Wigner distribution function [5,46] which is a mathematical transform (Wigner–Weyl transformation) of the density matrix. Starting from the density matrix the Wigner function

is defined as

$$f_W(r,p) = \int_{-\infty}^{\infty} \mathrm{d}r' \rho\left(\frac{r+r'}{2}, \frac{r-r'}{2}\right) \exp\left(\frac{-ipr'}{\hbar}\right), \tag{44}$$

where the coordinate transformations $r = (x + x')/2$ and $r' = x - x'$. The corresponding transport equation is obtained by applying the same transformation to the Liouville equation, obtaining the Wigner transport equation

$$\frac{\partial f_W}{\partial t} = \frac{p}{m^*}\frac{\partial f_W}{\partial r} - \frac{1}{\hbar}\int_{-\infty}^{\infty} \frac{\mathrm{d}p'}{2\pi\hbar} V_W(r, p-p')\, f_W(r', p'). \tag{45}$$

The equation above has similarities with the semi-classical Boltzmann transport equation and the Wigner distribution approaches the classical distribution function as the system tends to a classical state at slowly varying potentials or sufficiently high temperatures. The kernel of the potential operator is provided by the relation

$$V_W(r, p-p') = 2\int_{-\infty}^{\infty} \mathrm{d}r' \sin\left(\frac{pr'}{\hbar}\right)\left[V\left(\frac{r+r'}{2}\right) - V\left(\frac{r-r'}{2}\right)\right]. \tag{46}$$

Numerical implementation of the Wigner equation presents a number of challenges. First of all, it is a partial differential equation with hyperbolic behavior and it requires special solution techniques. Because of the wave nature of the equation, spurious numerical dispersion is difficult to contain since a range of momentum components are solved for simultaneously. Second, both position and momentum and coordinates are included in the Wigner function model, as is the case for the semi-classical Boltzmann equation, so the numerical cost increases quite rapidly for a multi-dimensional problem. On the other hand, fairly detailed scattering models can be implemented [27]. In recent applications, the Wigner function approach has also been combined with particle simulation for resonant tunneling structures [63].

An approach which has gained popularity in recent times is the nonequilibrium Green's function (NEGF) which is governed by Dyson's equation. This approach looks at a nano-system with a unified picture of a quantum channel for charge transport connecting two reservoirs held at different electro-chemical potentials [9,60,74]. The availability of numerical approaches to study a variety of systems, from nanoscale MOSFETs to molecular devices, has greatly contributed to the increase in interest in NEGF. Even an introductory treatment of NEGF would be rather cumbersome and beyond the scope of this work. The interested readers are directed to consult the relevant literature and the material available on line, as detailed in a section below. The main issue with a Green's function approach is that practical applications are possible under carefully defined assumptions. While the approach is potentially the most general and powerful to treat dissipative transport, the formalism and the computational cost may quickly become unmanageable in general conditions.

For a nanoscale system relatively close to equilibrium, a ballistic Green's function model is equivalent to the Schrödinger equation and it is quite useful to readily obtain the transmission and reflection properties of quantum wire and electron wave guide structures requiring 2-D or 3-D simulation [44,65,66]. For a nonrectilinear system with single input and output leads the formulation is very flexible, since it can be split into an equivalent 1-D Green's function problem for the transport and a size quantization problem to resolve transverse states in the cross-section of linear channel elements or cavity states for more general nonrectilinear elements. Another advantage is that application of open-boundary conditions in a Green's function problem is very straightforward, because it is very easy to define the Green's function of a semi-infinite lead. Inclusion of detailed scattering models is possible, but typically complications increase very rapidly, particularly if self-consistent treatment is applied. Reasonable simplified models are possible, for instance the approximate Büttiker probe treatment of scattering [6], which makes a NEGF formulation much less costly.

Quantum Corrections in Particle Simulation

For practical device simulation, it is still desirable to retain the mature and computationally inexpensive semi-classical techniques, extending their validity by the introduction of quantum corrections, to account for size quantization and tunneling effects. In many applications this is a reasonable alternative, since quantum coherence along the transport path is not expected to become significant or dominant for room temperature operation in MOS devices, until channel lengths are scaled down to the sub-10 nm range. The particle Monte Carlo method solves stochastically the Boltzmann Transport Equation [16,29,33,34], by a computer experiment where particle motion is explicitly simulated as sequences of random free flights interrupted by scattering events. The complete details of the band structure may be included in tabular form and the scattering model is very detailed,

so that the description of semi-classical transport provided by a Monte Carlo simulation can be very accurate [13,19,29,42].

Physically based models, such as particle simulation based on Monte Carlo methods, are a useful starting point to add quantum corrections because they adequately describe the nonequilibrium and hot carrier phenomena which dominate present devices, and can be used to calibrate simpler models in the simulation hierarchy, such as drift-diffusion. The computationally expensive alternative to add quantum effects in Monte Carlo simulation is to resolve in detail the transverse sub-bands in a conduction channel and describe particle dynamics within the sub-bands. This requires a detailed model for intra-sub-band and inter-sub-band scattering rates [20,80], as well as 2-D to 3-D scattering mechanisms if a continuum of states is used at high energies [55]. When a complete device simulation is assembled, with contacts characterized by continuum states, one also needs to implement a model that resolves the discontinuity between quantum and continuum description at the channel entrance. Because of the considerable cost associated with a detailed sub-band Monte Carlo model, applications have been limited mainly to uniform 2-DEG simulation under constant electric field conditions [20,80]. In device simulation typical approaches use the global information from the details of quantum states to formulate a potential correction that congregates otherwise semi-classical particles in a configuration that mimics the quantum distributions. The goal is not to simulate individual particles to behave as quantum ones, but to obtain a particle distribution that globally approximates the quantum charge flow, resolving noncoherent quantum effects, like size quantization and non-resonant tunneling [17,73,76].

The motivation to pursue quantum corrections in semi-classical simulation is that full quantum transport remains impractical for many cases. Also, in most realistic device structures one may identify regions where either quantum or classical features of the transport dominate, therefore, a quantum-corrected semi-classical model is useful to treat the whole structure in a unified way. In addition, while a complete quantum simulation may be extremely expensive, there is little computational cost added by quantum corrections to the standard semi-classical Monte Carlo simulators. Quantum corrections to semi-classical simulation may follow various approaches by coupling different quantum formalisms: Feynman effective potential, Wigner transport equation, Bohm potential, Schrödinger equation.

The simplest approach to quantum correction follows the effective potential idea introduced originally by Feynman for statistical mechanics [18]. In this model, particles feel the nearby potential due to quantum fluctuations around the classical path of least action, via a nonlocal function effective potential V_{eff} obtained by convoluting a Gaussian function with the electrostatic potential

$$V_{\text{eff}}(x) = \int V(x) \mathrm{e}^{-\frac{(x-x')^2}{2a^2}} \mathrm{d}x' , \quad a = \frac{\hbar^2}{12 m k_{\text{B}} T} . \quad (47)$$

The effective potential is very simple to implement and computationally inexpensive. In addition it is not sensitive to the intrinsic statistical noise of Monte Carlo simulation [17]. This approach works best for smooth, symmetric potentials. The parameter a describes the effective "size" of the particle and can be treated as a fitting parameter. The detailed solution next to a heterointerface is typically incorrect. Fitting may be applied to obtain the correct average displacement of the charge from an abrupt interface to correspond to the maximum of quantum charge, but the space distribution is not very accurate, because the Gaussian weight function is not suitable to resolve asymmetries. A more general effective potential is possible, which depends explicitly on the wave vector k and does not contain fitting parameters [39]. The Feynman effective potential is a particular case of this general potential correction which has the form

$$V^{\text{Q}}(x,k) = \frac{1}{(2\pi)^3} \int \frac{2m^*}{\beta\hbar^2 k \cdot \xi} \sinh\left(\frac{\beta\hbar^2 k \cdot \xi}{2m^*}\right) \times \exp\left(-\frac{\beta\hbar^2}{8m^*} |\xi|^2\right) V(y) \, \mathrm{e}^{i\xi\cdot(x-y)} \, \mathrm{d}y \, \mathrm{d}\xi . \quad (48)$$

The cost of applying this correction is much higher if done self-consistently. For specific structures like the double gate MOSFET, one can show that much better agreement is obtained by using a Pearson distribution instead of a Gaussian one for the formulation of the effective potential [35].

The Wigner method was also developed as a correction to statistical dynamics [75], by writing the Wigner transport equation for the distribution function in a form that casts it as the standard Boltzmann transport equation plus a correction containing the quantum terms [73]

$$\frac{\partial f}{\partial t} + \vec{v} \cdot \nabla_{\vec{r}} f - \frac{1}{\hbar} \nabla_{\vec{r}} U \cdot \nabla_{\vec{k}} f + \sum_{\alpha=1}^{\infty} \frac{(-1)^{\alpha+1}}{\hbar 4^{\alpha} (2\alpha+1)!} \left(\nabla_{\vec{r}} U \cdot \nabla_{\vec{k}} f\right)^{2\alpha+1} = \left(\frac{\partial f}{\partial t}\right)_c . \quad (49)$$

At first order, one simply stops the first summation term for $\alpha = 1$ and the formula can be rewritten in a form that

closely resembles the standard Boltzmann equation [73]

$$\frac{\partial f}{\partial t} + \vec{v} \cdot \nabla_{\vec{r}} f - \frac{1}{\hbar} F_{\mathrm{qc}} \cdot \nabla_{\vec{k}} f = \left(\frac{\partial f}{\partial t} \right)_c , \tag{50}$$

where the term F_{qc} is the quantum corrected force. The particles move as if under the influence of a classical potential but following approximately equivalent quantum trajectories. By making an assumption on the distribution function, like a displaced Maxwellian, and a band structure, like the simple parabolic dispersion, one can formulate expressions for the quantum corrected force that depend explicitly on momentum. These correction forces still have some problems near sharp interfaces, where quantum effects are prominent. To alleviate this, a smooth potential approximation can be obtained by integrating the displaced Maxwellian distribution with momentum [73]. Simplified versions of the force can be obtained by assuming a thermal energy to specify the momentum. Although this formulation has solid physical foundations, the result for the corrected force depends on the mobile charge density as $\nabla^2 \ln(n)$. Since in a particle simulation the density is recovered from a temporal average of particle occupation in space, convergence is slow, particularly in small structures with few particles and the overall procedure is sensitive to simulation noise. Corrections of this type work instead very well in continuum models like drift-diffusion, where noise is not an issue. From simulation experiments it was found that for a MOS interface no fitting is required in the semiconductor region, but at the oxide interface one has to specify a finite value of the carrier concentration which becomes the fitting parameter in the procedure.

An approximate correction can also be based on Bohm's potential, deriving an effective conduction band edge [78]. If we express the wave function ψ and the eigenenergy E as

$$\begin{aligned} \psi &= R \cdot \exp\left(\frac{iS}{\hbar} \right) E = V\left(\vec{r}\right) - \frac{\hbar^2}{4m} \left[\frac{\nabla^2 P}{P} - \frac{(\nabla P)^2}{2P^2} \right] \\ &= V\left(\vec{r}\right) - \frac{\hbar^2}{2m} \frac{\nabla^2 R}{R} , \end{aligned} \tag{51}$$

the effective conduction band edge is defined as

$$V^*\left(\vec{r}\right) \approx V\left(\vec{r}\right) - \frac{\hbar^2}{2m} \frac{\nabla^2 \sqrt{n}}{\sqrt{n}} = V\left(\vec{r}\right) + V_{\mathrm{B}}^{Q}\left(\vec{r}\right) . \tag{52}$$

Using the approximate relation for carrier density $n \propto \exp\left(-V^*/k_{\mathrm{B}}T\right)$ one can obtain a 1-D effective conduction band equation, valid for conditions close to equilibrium

$$\frac{\mathrm{d}^2 V^*}{\mathrm{d}y^2} - \frac{1}{2k_{\mathrm{B}}T} \left(\frac{\mathrm{d}V^*}{\mathrm{d}y} \right)^2 = \frac{4m^* k_{\mathrm{B}} T}{\hbar^2} \left(V^* - V(y) \right) . \tag{53}$$

The equation represents a first-order quantum correction to the semi-classical BTE by taking into account the effect of carriers only occupying the quantized ground state. This is reasonable in many situations since usually, where the quantum correction is needed, density is not too high and the Fermi level is below the first excited state.

Another technique for quantum correction of particle simulation is based on a direct application of the Schrödinger equation [76]. This technique is strictly suitable for size quantization in the direction perpendicular to transport, while the other techniques can be used in the direction of transport to simulate barrier lowering in tunneling, for instance. The Schrödinger correction, however, does not have the drawbacks of the other approximations, since it is very accurate, requires no fitting parameters and is not sensitive to particle simulation noise. The procedure requires explicit solutions of 1-D or 2-D Schrödinger equations along the channel, but this can be done very efficiently on the regular grids used for particle simulation and it does not affect appreciably the overall computational time. What is lost in cost per iteration is gained in excellent convergence properties.

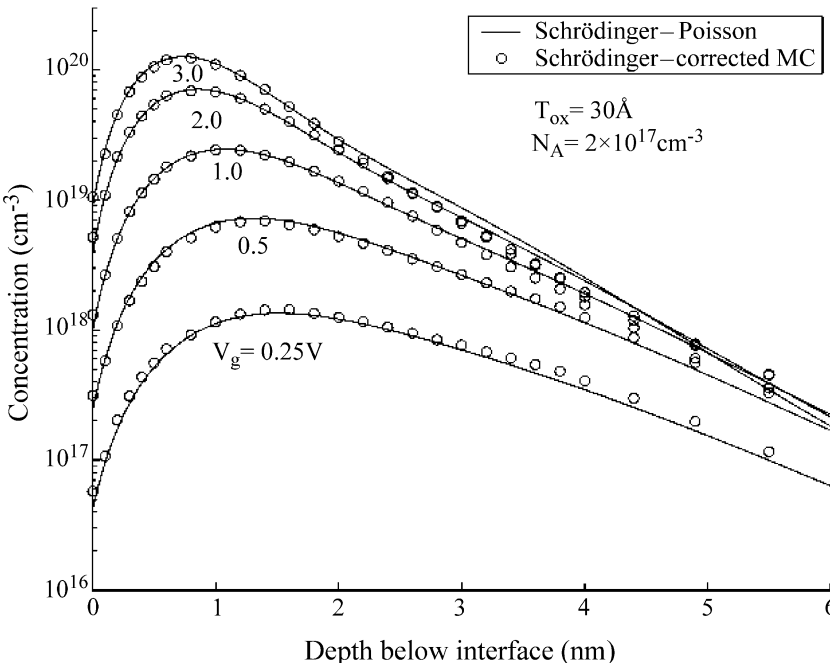
When the Schrödinger equation is solved at different channel locations, the self-consistent Monte Carlo potential, already available from the overall simulation, is the input while the output is a quantum density, n_{q}, obtained from the detailed sub-band structure and an assumption of quasi-equilibrium distribution on the transverse cross section. If one assumes a simple Maxwellian behavior for the quantum carrier distribution

$$n_{\mathrm{q}}(z) \propto \exp\left[\frac{-\{Vp(z) + Vqc(z)\}}{kT} \right] , \tag{54}$$

the potential quantum correction has a shape as

$$Vqc(z) = -kT \log\left(n_{\mathrm{q}}(z)\right) - V_p(z) + V_0 , \tag{55}$$

where V_0 is a reference potential where the quantum charge is zero. When the Monte Carlo simulation runs, application of the potential correction forces the shape of the quantum density onto the semi-classical particles. The actual simulated density does not need to be evaluated to formulate the correction, since only the potential solution from solving the Poisson equation is needed. The potential is normally very smooth and therefore the iteration is not affected appreciably by numerical noise.

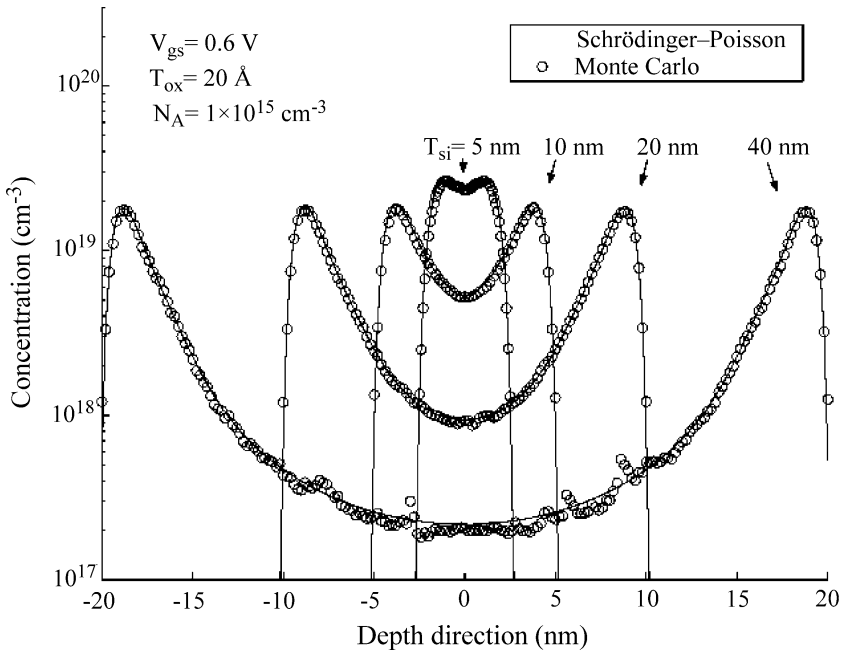


Quantum Phenomena in Semiconductor Nanostructures, Figure 11
Electron concentration in an inverted MOS capacitor calculated with the Schrödinger-corrected Monte Carlo and self-consistent Schrödinger–Poisson methods over a range of gate bias

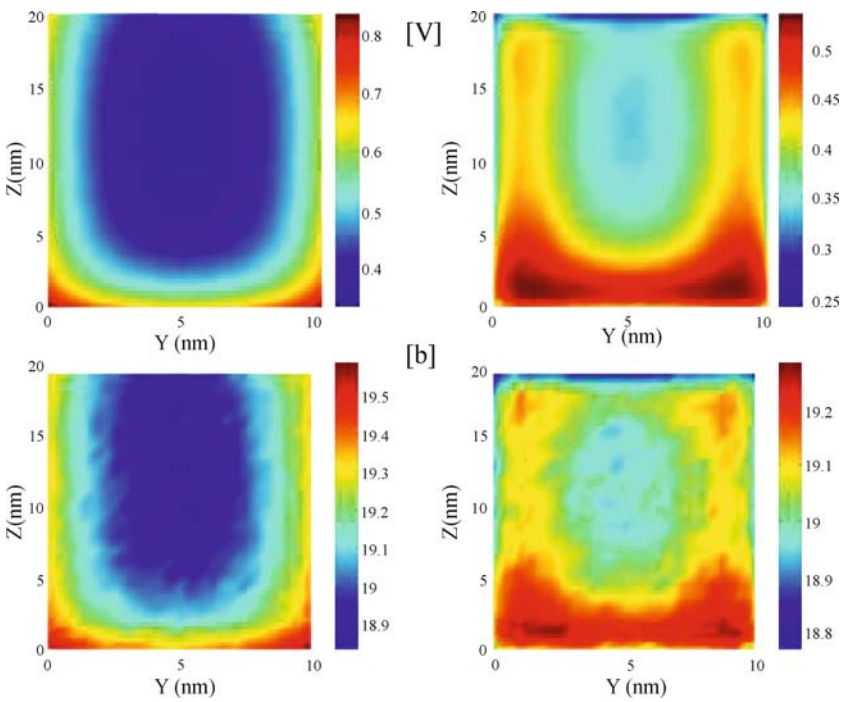
Monte Carlo simulations of MOS capacitor structures match very closely the result of the self-consistent 1-D Schrödinger/Poisson solution, without the need for any calibration or tuning parameter, proving the overall robustness of the procedure. In the case of device simulation transport occurs in the direction of the channel, therefore, energy is acquired by particles. In this case, the actual lattice temperature should not be used to evaluate the shape of the quantum density profile on the transverse section. One can define an effective parameter with dimension of temperature which has been dubbed transverse temperature, to account for heating of the distribution. The transverse temperature is in general smaller than the temperature obtained from the average carrier energy along the channel. It should be stressed that temperature is really a tensorial quantity and that the parameter called transverse temperature cannot be identified with a well-defined physical quantity. In a structure without a clearly defined substrate reference for the potential, like in a double gate structure, the definition of transverse temperature may become ambiguous. A more detailed analysis can be carried out invoking the stress tensor along each transverse direction, to account for the variation of the electron temperature along the longitudinal direction [77]. Comparisons have been carried out with a NEGF simulator for double gate MOSFET structures. Results are in good agreement as long as the thickness of the channel layer is larger than 3 nm. For thinner layers, a quantum correction does not seem to be adequate any longer and more sub-band details should be included [57].

Monte Carlo simulations were conducted with the Schrödinger correction on MOS capacitor structures to verify the quality of the approach. Figure 11 shows a set of simulations conducted at several gate voltages, showing a consistently good agreement. With no adjustable parameter involved, the interface concentration is always accurately resolved. Figure 12 shows a set of simulation for double gate structures with separation between the oxide interfaces from 5 to 40 nm, again with excellent agreement.

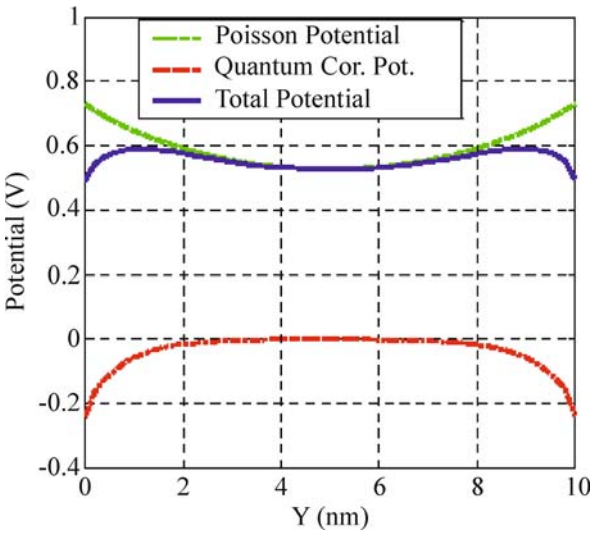
Finally, representative results from a complete 3-D finFET simulation, including a 2-D quantum correction approach, are shown for a channel cross-section in Fig. 13. Potential and carrier distribution are shown both for a classical transport model (left) and with the addition of quantum corrections (right). In this particular simulation both gate and drain bias are 1 V, the side oxide thickness is 1 nm and the background acceptor concentration is 10^{16} cm^{-3}. Figure 14 shows how the quantum correction potential is added to the electrostatic potential to cause charge repulsion from the interfaces.



Quantum Phenomena in Semiconductor Nanostructures, Figure 12
Electron concentration in a double-gate MOS capacitor calculated with the Schrödinger-corrected Monte Carlo and self-consistent Schrödinger–Poisson methods over a range of body thicknesses



Quantum Phenomena in Semiconductor Nanostructures, Figure 13
Potential (*top row*) and electron concentration (*bottom row*) in the cross-section of a finFET as simulated by 3-D Monte Carlo simulation. Results on the left correspond to a classical simulation without quantum correction. The application of the Schrödinger quantum correction, in the results on the *right*, causes a shift of the charge concentration maximum away from the interfaces. Note that the figures are oriented upside down so that the top oxide interface corresponds to $z = 0$



Quantum Phenomena in Semiconductor Nanostructures, Figure 14

The quantum correction potential (*red curve*) added to the electrostatic potential solution from the Poisson equation resulted in a quantum corrected potential that repels electrons from the interface

Online Resources

Over the last several years, the availability of advanced material on internet services has increased dramatically. The readers of this article will be particularly interested in exploring the many relevant resources accessible at http://www.nanohub.org, the portal of the Network for Computational Nanotechnology, an initiative funded by the National Science Foundation. The nanohub provides an ever increasing portfolio of online simulations that allow users to access fully functional research codes through comprehensive interactive interfaces. Many of the available applications in the nanoelectronics section of the nanohub relate directly to topics outlined in this work, including solvers for band structure, size quantization, quantum-corrected Monte Carlo, NEGF applied to nanodevice structures. Besides nanoelectronics, the site provides tools for nano-electro-mechanical system and nanobiology simulation. The unique approach of this organization has been to embed mature and working research codes, virtually unmodified, within a powerful infrastructure which provides at the same time a complete graphical user interface with intuitive selection menus and controls to initialize and start a simulation, access to large computational resources on efficient computer clusters, visualization of the results in the interface window with a range of facilities for data exploration. Because of this arrangement, deployment time of simulation codes has been drastically shortened, leading to availability of a vast array of applications which is expected to keep growing considerably in the future. With integration of all simulation aspects under one single interface, the nanohub services are appealing not only to computational experts in the field, but also to experimentalists and students. Beyond online computation, the nanohub is also the depository of a large number of tutorials, lectures and seminars which are regularly augmented with new contributions. The identification of specific material of interest is facilitated by flexible search tools.

Future Directions

The saturation in our ability to scale further down the dimensions of standard silicon devices for integrated circuits, which is expected to occur in the near future, opens many questions which make the formulation of future directions quite difficult. The main issue is that a replacement for standard silicon MOSFET technology has not been clearly identified, despite a flurry of activities that have explored many possibilities. Molecular systems have been studied as possible candidates for new nanotechnologies. One can find in nature a variety of stable and robust molecules that could be produced or harvested with nearly perfect uniformity of shapes and properties and which exhibit device-like behavior under external stimuli. In certain areas of electronics, organic materials have already made promising inroads, particularly for the realization of optical devices and displays [28]. Conduction through organic molecules is being studied with the goal to realize elementary computational or memory elements while a great deal of attention is also being paid to carbon nanotubes, which yield solid-state structures dominated in their behavior by the specific molecular footprint achieved during growth and which may allow a rich variety of new applications [3,40]. There is also interest in biomolecular systems. DNA molecules, which have long wire shapes, have been suggested as possible structures for computing applications [51]. Even biological ionic channels [31], that act as natural nanoscale devices in every cell and display a range of behaviors, have been examined for possible applications as switches, sensors and actuators [8]. Finally, semiconductor materials with magnetic properties are investigated to provide the ability to select the spin quantum number of carriers injected into a device [4]. Since the time decay constant of a given spin state can be much longer than characteristic device operation times, the possibility to encode information in spin states provides a new computational paradigm full of intriguing possibilities. Fun-

damental research on quantum computation in general is currently receiving considerable attention [32] but practical applications are still a matter of considerable speculation and the main limitation today is due to the fact that extremely low temperatures are still necessary to demonstrate working quantum computing applications of any kind.

While molecular systems are attractive to attempt the realization of more advanced nanoscale device elements, realistically one cannot expect a rapid technological development that may quickly provide an alternative to present semiconductor devices. One should expect that even when intrinsic scaling limits are reached, silicon technology would continue to be refined, still playing a fundamental role in any electronics application. Even if totally new devices become practical, we are likely to see for a long period of time the emergence of hybrid technologies where new systems, made of nanoscale building blocks that realize specialized tasks, will be implanted onto host systems based on traditional semiconductor technology approaches. Systems completely realized with molecular technologies are still difficult to envision. Many efforts are focused on finding a one-to-one replacement for the MOSFET but, arguably, completely new architecture and computational paradigms are necessary to realize the promise of new potential molecular approaches. Biological or artificial (biomimetic) nano-channels may find uses as single molecule detectors in novel sensing applications, possibly operating as a combination of analog and digital computing elements.

This preamble provides some grounds to venture into a prediction of future directions. It is unquestionable that miniaturization will continue and that elementary devices will continue to seek the true ultimate physical limit, which is arguably on the order of the size of an atom. Quantum and first principle calculations will be indispensable to provide the necessary predictive power for the design of reliable systems. Today, computer aided design (CAD) is already an essential stage in the design and fabrication loop. Design of large systems would be inconceivable without the use of CAD tools for circuit simulation [14]. While it was possible in the past to fine tune a device structure by trial-and-error in the laboratory, at the nanoscale regime of today one has to rely more and more on process and device simulators to optimize and calibrate design and fabrication stages to achieve the necessary performance and reliability. As the scale of device elements decreases, so is our ability to observe and measure directly anything related to device behavior. In the future, the exploration of new device concepts will have to rely even more on new generations of multi-scale and multi-physics computational approaches, while development teams will be by necessity increasingly interdisciplinary. One may envision that the computational approaches described earlier will continue to be used and improved, but increasingly in the context of coupled physical effects, where electrical, mechanical and thermal behavior are intimately connected to determine the behavior of a complex nanoscale system. For all of these aspects, quantum mechanics will play a pivotal role, but one should expect that the standard envelope-function effective-mass picture, adopted so successfully in most semiconductor device applications until now, will have to be complemented often by an atomistic approach requiring new practical simulation tools able to perform first principle calculations at various levels of complexity, from electronic structure to molecular dynamics.

This outlook opens up more general issues for which the science community needs to be prepared. Quantum mechanics is not going to be any longer the domain of a limited number of specialists, but it will permeate a much broader range of disciplines and applications. At the same time, computational requirements will be much greater and the complexity of simulations will make it impractical for single individuals to manage complete simulation codes alone. The computational community is already actively addressing the issues of Peta-scale computing technology which is expected to debut in the next five years. This unprecedented computing power will create new difficulties in data management, visualization, and numerical strategies for efficient use of the resources, while new educational challenges will have to be addressed in order to prepare professionals who can deal with the complexity of the simulation models and of the computational environments. This is why this author firmly believes that large cyber-infrastructure initiatives in nanotechnology will be crucial to create virtual collaborative communities and to develop the necessary educational and computational resources that an academic institution or industrial laboratory may not be able to provide individually to solve the electronics problems of the future.

Acknowledgments

This work was supported by the National Science Foundation through the Network for Computational Nanotechnology grant EEC-0228390.

Bibliography

Primary Literature

1. Ando T, Fowler AB, Stern F (1982) Electronic properties of two-dimensional systems. Rev Mod Phys 54:437–669

2. Antia HM (1993) Rational function approximations for Fermi–Dirac integrals. Astrophys J Suppl 84:101–106
3. Appenzeller J, Knoch J, Martel R, Derycke V, Wind SJ, Avouris P (2002) Carbon nanotube electronics. IEEE Trans Nanotech 1:184–189
4. Awschalom DD, Samarth N, Loss D (eds) (2002) Semiconductor spintronics and quantum computation. Springer, Berlin
5. Bertoni A, Bordone P, Brunetti R, Jacoboni C (1999) The Wigner function for electron transport in mesoscopic systems. J Phys Cond Matt 11:5999–6012
6. Büttiker M (1986) Four-terminal phase-coherent conductance. Phys Rev Lett 57:1761–1764
7. Cahay M, Kreskovsky JP, Grubin HL (1989) Electron diffraction through an aperture in a potential wall. Solid-State Electron 32:1185–1989
8. Cornell BA, Braach-Maksvytis, King LG, Osman PDJ, Raguse B, Wieczorek L, Pace RJ (1997) A biosensor that uses ion-channel switches. Nature 387:580–583
9. Datta S (2000) Nanoscale device simulation: the Green's function formalism. Superlatt Microstruc 28:253–278
10. Delagebeaudeuf D, Linh NT (1982) Metal-(n) AlGaAs-GaAs two-dimensional electron gas FET. IEEE Trans Electron Dev 29: 955–960
11. Dingle R, Störmer HL, Gossard AC, Wiegmann W (1978) Electron mobilities in modulation-doped semiconductor heterojunction superlattices. Appl Phys Lett 33:665–667
12. Doyle BS, Datta S, Doczy M, Hareland S, Jin B, Kavalieros J, Linton T, Murthy A, Rios R, Chau R (2003) High performance fully-depleted tri-gate CMOS transistors. IEEE Trans Electron Dev 24:263–265
13. Duncan A, Ravaioli U, Jakumeit J (1998) Full-band Monte Carlo investigation of hot carrier trends in the scaling of metal-oxide-semiconductor field-effect transistors. IEEE Trans Electron Dev 45:867–876
14. Dutton RW, Yu Z (1993) Technology CAD: computer simulation of IC processes and devices. Kluwer Academic Publishers, Norwell
15. Ekenberg U (1989) Nonparabolicity effects in a quantum well: sublevel shift, parallel mass and Landau levels. Phys Rev B 40:7714–7726
16. Fawcett W, Boardman AD, Swain S (1970) Monte Carlo determination of electron transport properties in gallium arsenide. J Phys Chem Sol 31:1963–1990
17. Ferry DK (2000) The onset of quantization in ultra-submicron semiconductor devices. Superlatt Microstruct 27:61–66
18. Feynmann R, Kleinert H (1986) Effective classical partition functions. Phys Rev A 34:5080–5084
19. Fischetti MV, Laux S (1988) Monte Carlo analysis of electron transport in small semiconductor devices including band-structure and space-charge effects. Phys Rev B 38: 9721–9745
20. Fischetti MV, Laux SE (1993) Monte Carlo study of electron transport in silicon inversion layers. Phys Rev B 48: 2244–2274
21. Frank DJ, Laux SE, Fischetti MV (1992) Monte Carlo simulation of 30 nm dual-gate MOSFET: how short can Si go? IEDM Tech Dig 1992:553–556
22. Frensley WR (1987) Wigner function model of a resonant-tunneling semiconductor device. Phys Rev B 36:1570–1580
23. Godoy A, Yang Z, Ravaioli U, Gamiz F (2005) Effects of non-parabolic bands in quantum wires. J Appl Phys 98:013702
24. Godoy A, Ruiz-Gallardo A, Sampedro C, Gámiz F (2007) Quantum-mechanical effects in multiple-gate MOSFETs. J Comput Electron 6:145–148
25. Goldberg A, Schey HM, Schwartz JL (1966) Computer-generated motion pictures of one-dimensional quantum-mechanical transmission and reflection phenomena. Am J Phys 35: 177–186
26. Gunn JB (1963) Microwave oscillations of current in II–V semiconductors. Solid-State Commun 1:88–91
27. Han Z, Goldsman N, Lin CK (2005) Incorporation of quantum correction to semiclassical two-dimensional device modeling with the Wigner–Boltzmann equation. Solid-State Electron 49:145–154
28. Heeger AJ (1998) Light emission from semiconducting polymers: light-emitting diodes, light-emitting electrochemical cells, lasers and white light for the future. Solid State Communic 107:673–679
29. Hess K (1991) Monte Carlo Device Simulation: full band and beyond. Kluwer Academic Publishers, Norwell
30. Hess K (2000) Advanced theory of semiconductors devices. IEEE Press, Piscataway
31. Hille B (1992) Ionic channels of excitable membranes. Sinauer Associates Inc, Sunderland
32. Hirvensalo M (2004) Quantum computing. Springer, New York
33. Hockney R, Eastwood J (1981) Computer simulation using particles. McGraw-Hill, New York
34. Jacoboni C, Lugli P (1989) The Monte Carlo method for semiconductor device simulation. Springer, Wien
35. Jaud MA, Barraud S, Dollfus P, Jaouen H, Le Carval G (2007) Pearson versus Gaussian effective potentials for quantum-corrected Monte-Carlo simulation. J Comput Electron 6:19–22
36. Jensen KL, Buot FA (1989) Numerical simulation of transient response and resonant-tunneling characteristics of double-barrier semiconductor structures as a function of experimental parameters. J Appl Phys 65:5248–5250
37. Kathawala GA, Ravaioli U (2003) 3-D Monte Carlo simulations of FinFETs. IEDM Tech Dig 2003:29.2.1–29.2.4
38. Kerkhoven T, Galick AT, Ravaioli U, Arends JH, Saad Y (1990) Efficient numerical simulation of electron states in quantum wires. J Appl Phys 74:1199–1204
39. Knezevic I, Vasileska DZ, Ferry DK (2002) Impact of strong quantum confinement on the performance of a highly asymmetric device structure: Monte Carlo particle-based simulation of a focused-ion-beam MOSFET. IEEE Trans Electron Dev 49:1019–1026
40. Kong J, Franklin NR, Zhou C, Chapline MG, Peng S, Cho K, Dai H (2000) Nanotube molecular wires as chemical sensors. Science 287:662–625
41. Kuska JP (1992) Absorbing boundary conditions for the Schrödinger equation. Phys Rev B 46:5000–5003
42. Laux SE, Fischetti MV, Frank DJ (1990) Monte Carlo analysis of semiconductor devices: the DAMOCLES program. IBM J Res Develop 34:466–494
43. Lopez-Villanueva JA, Melchor I, Cartujo P, Carceller JE (1993) Phys Rev B 48:1626–1631
44. Macucci M, Galick AT, Ravaioli U (1995) Quasi-three-dimensional Green's function simulation of coupled electron waveguides. Phys Rev B 52:5210–5220
45. Mains RK, Haddad GI (1988) Time-dependent modeling of resonant-tunneling diodes from a direct solution of the Schrödinger equation. J Appl Phys 64:3564–3569

46. Mains RK, Haddad GI (1988) Wigner function modeling of resonant tunneling diodes with high peak-to-valley ratios. J Appl Phys 64:5041–5044
47. Mains RK, Haddad GI (1990) Improved boundary conditions for the time-dependent Schrödinger equation. J Appl Phys 67:591–593
48. Mizuta H, Goodings CJ (1991) Transient quantum transport simulation based on the statistical density matrix. J Phys Cond Matt 3:3739–3756
49. Moore GE (1965) Cramming more components onto integrated circuits. Electronics 38:116–119
50. Nicollian EH, Brews JR (1982) MOS (Metal Oxide Semiconductor) physics and technology. Wiley, New York
51. Paun G, Rozenberg G, Salomaa A (1998) DNA computing: new computing paradigms. Springer, Berlin
52. Pei G, Kedzierski J, Oldiges P, Ieong M, Kan ECC (2002) FinFET design considerations based on 3-D simulation and analytical modeling. IEEE Trans Electron Dev 49:1411–1419
53. Persson A, Cohen RM (1988) Reformulated Hamiltonian for nonparabolic bands in semiconductor quantum wells. Phys Rev B 38:5568–5575
54. Petroff PM, Gossard AC, Logan RA, Wiegmann W (1982) Toward quantum well wires: fabrication and optical properties. Appl Phys Lett 41:635–638
55. Ravaioli U, Ferry DK (1986) MODFET ensemble Monte Carlo model including the quasi-two-dimensional electron gas. IEEE Trans Electron Dev 33:677–681
56. Ravaioli U, Osman MA, Pötz W, Kluksdahl N, Ferry DK (1985) Investigation of ballistic transport through resonant-tunneling quantum wells using Wigner function approach. Physica B 134:36–40
57. Ravishankar R, Kathawala G, Ravaioli U, Hasan S, Lundstrom M (2005) Comparison of Monte Carlo and NEGF simulations of double gate MOSFETs. J Comput Electron 4:39–43
58. Reed MA, Randall JN, Aggarwal RJ, Matyi RJ, Moore TM, Wetsel AE (1988) Observation of discrete electronic states in a zero-dimensional semiconductor nanostructure. Phys Rev Lett 60:535–537
59. Register LF, Ravaioli U, Hess K (1991) Numerical simulation of mesoscopic systems with open boundaries using the multidimensional time-dependent Schrödinger equation. J Appl Phys 69:7153–7158
60. Ren Z, Venugopal R, Goasguen S, Datta S, Lundstrom MS (2003) nanoMOS 2.5: a two-dimensional simulator for quantum transport in double-gate MOSFET's. IEEE Trans Electron Dev 50:1853–1864
61. Sah CT (1991) Fundamentals of solid state electronics. World Scientific
62. Shibata T (1991) Absorbing boundary conditions for the finite-difference time-domain calculation of the one-dimensional Schrödinger equation. Phys Rev B 42:6760–6763
63. Shifren L, Ringhofer C, Ferry DK (2003) A Wigner function-based quantum ensemble Monte Carlo study of a resonant tunneling diode. IEEE Trans Electron Dev 50:769–773
64. Sollner TCLG, Goodhue WD, Tannenwald PE, Parker CD, Peck DD (1983) Resonant tunneling through quantum wells at frequencies up to 2.5 THz. Appl Phys Lett 43:588–590
65. Sols F, Macucci M, Ravaioli U, Hess K (1989) On the possibility of transistor action based on quantum interference. Appl Phys Lett 54:350–352
66. Sols F, Macucci M, Ravaioli U, Hess K (1989) Theory for a quantum modulated transistor. J of Appl Phys 66:3892–3906
67. Sze SM (1981) Physics of semiconductor devices. Wiley, New York
68. Talebian MA, Pötz MA (1996) Open boundary conditions for a time-dependent analysis of the resonant tunneling structure. Appl Phys Lett 69:1148–1150
69. Trellakis A, Ravaioli U (1999) Lateral scalability limits of silicon conduction channels. J Appl Phys 86:3911–3916
70. Trellakis A, Ravaioli U (2002) Three-dimensional spectral solution of the Schrödinger equation for arbitrary band structures. J Appl Phys 92:3711–3716
71. Trellakis A, Galick AT, Pacelli A, Ravaioli U (1997) Iteration scheme for the solution of the two-dimensional Schrödinger–Poisson equation in quantum structures. J Appl Phys 81:7880–7884
72. Tsu R, Esaki L (1973) Tunneling in a finite superlattice. Appl Phys Lett 22:563–564
73. Tsuchiya H, Ravaioli U (2001) Particle Monte Carlo simulation of quantum phenomena in semiconductor nanostructures. J Appl Phys 89:4023–4029
74. Venugopal R, Ren Z, Datta S, Lundstrom MS, Jovanovic D (2000) Simulating quantum transport in nanoscale transistors: real versus mode-space approaches. J Appl Phys 93: 3730–3739
75. Wigner E (1932) On the quantum correction for thermodynamic equilibrium. Phys Rev 40:749–759
76. Winstead B, Ravaioli U (2003) A quantum correction based on Schrödinger equation applied to Monte Carlo device simulation. IEEE Trans Electron Dev 50:440–446
77. Wu B, Tang TW (2004) The effective conduction band edge method of quantum correction to the Monte Carlo device simulation. J Comput Electron 3:347–350
78. Wu B, Tang TW, Nam J, Tsai JH (2003) Monte Carlo simulation of symmetric and asymmetric double-gate MOSFETs using Bohm-based quantum correction. IEEE Trans Nanotech 2:291–294
79. Yalabik MC, Ecemiş MI (1995) Numerical implementation of absorbing and injecting boundary conditions for the time-dependent Schrödinger equation. Phys Rev B 51:2082–2086
80. Yamakawa S, Ueno H, Taniguchi K, Hamaguchi C, Miyatsuji K, Masaki K, Ravaioli U (1996) Study of interface roughness dependence of electron mobility in Si inversion layers using the Monte Carlo method. J Appl Phys 79:911–916

Books and Reviews

Datta S (1995) Electronic transport in mesoscopic systems. Cambridge University Press, New York

Ferry DK, Goodnick SM (1997) Transport in nanostructures. Cambridge University Press, New York

Frensley WR (1990) Boundary conditions for open quantum systems driven far from equilibrium. Rev Mod Phys 62:745–791

Fu Y, Willander M (1999) Physical models of semiconductor quantum devices. Kluwer Academic Publishers, Norwell

Hamaguchi C (2001) Basic semiconductor physics. Springer, Berlin

Lundstrom M (2000) Fundamentals of carrier transport. Addison-Wesley, Reading

Nag BR (2000) Physics of quantum well devices. Springer, Berlin

Paiella R (2006) Intersubband transitions in quantum structures. Mc-Graw Hill, New York

Quantum Similarity and Quantum Quantitative Structure-Properties Relationships (QQSPR)

Ramon Carbó-Dorca, Ana Gallegos
Institut de Química Computacional,
Universitat de Girona, Girona, Spain

Article Outline

Glossary

There is a brief description of the terms used herein. The defined items appear in alphabetical order. When in a definition a term already defined in the glossary is mentioned, it is written in **bold face**; then, the reader has to refer to the corresponding glossary item, where more information is given.

Carbó (similarity) index The Carbó (Similarity) Index is a **QS** index, which corresponds to the cosine of the angle subtended by the density function (**DF**) tags of any pair of quantum objects. The values of the Carbó index lie within the interval $(0, 1]$. The lower bound corresponds to a complete dissimilarity, while the unit value is encountered when comparing a quantum object with itself. *See also:* **Euclidian Distance (Similarity) Index**.

Core set A **QOS** with the additional information of a known and well-defined property value for every **QO**.

(First order) density function (DF) The first order DF is a quantum theory concept, associated to a known electronic submicroscopic system. To be understood by this term is a **wave function** squared module or the **full DF**, reduced by integration over the space and spin coordinates to a function of the space coordinates of a unique electron. As it is customary in the literature, the name will be shortened to DF. Such a function can be also obtained via direct computation within Density Functional Theory (DFT). DF can be employed as tags associated to well-defined **quantum objects**. DF are non-negative functions. According to the usual quantum mechanical interpretation, within the DF collection there is contained all the information one can extract from the system. This last statement is the basic postulate which appears in first place when developing **QQSPR** theory.

(Full) density function A quantum theoretical concept, associated to a known electronic (or other particle sets) system. It corresponds to the system's **wave function** squared module.

Density function (DF) tags The tag set, (see: **tagged set**) collecting the quantum mechanical density functions used as descriptors in a **QOS**.

DQOS *Discrete Quantum Object Set*. A **QOS** whose tags are finite dimensional vectors, whose elements, in turn, are computed as **quantum similarity measures**.

Euclidean distance (similarity) index Like the **Carbó Index** the Euclidean Distance Index is a **quantum similarity index**, involving the Euclidean distance between two **density functions**. The range of this index is $[0, \infty)$, thus their minimal values correspond to two identical **quantum objects**, while the index grows in relation to the difference in the compared **quantum objects**.

FQQSPR equation *Fundamental Quantum Quantitative Structure-Properties Relationships equation*. The subject of study and analysis of the present contribution. A non-empirical equation, which can be deduced by quantum mechanical theoretical means, serving as the basic tool to obtain **QQSAR** or *Quantum QSAR*. Here, only the linear **FQQSPR** equations are deeply discussed, as they are the main source of **QQSPR** studies, but the **FQQSPR** equation can easily be extended to any order.

Molecular descriptors Parameters of varied origin: empirical, theoretical or experimental, or functions obtained from quantum mechanical manipulations like the **Density Function** or the Electrostatic Molecular Potential, or by empirical considerations, associated to a given molecule, which represent the molecular structure environment and can be employed to obtain **QSPR**. Any set of parameters or functions, which can be used as tags in a **tagged set**, whose objects are molecules.

QO *Quantum Object*. An element of a **QOS**, a **tagged set** where every element is constructed as a composite of a submicroscopic system description (the object) and a density function (the tag). The possibility to represent all the information contained within any quan-

tum system by the full or reduced **DF**, permits one to describe such an entity formed by the structure of the system and its **DF** together; this can be called a **QO**. Plural: **QO's**.

QOS *Quantum Object Set*. A **tagged set** made of **QO's**.

QS *Quantum Similarity*: A discipline dealing with similarity measures between submicroscopic systems. Such kinds of quantum similarity measures (**QSM**) can be computed using the quantum theoretical description of such kinds of objects. According to quantum theory all the information one can obtain about a quantum system is contained in the state **DF** and the set of possible reduced forms, hence **quantum similarity** is part of this information contained in the **DF**. Usually, for computational convenience, the (first order reduced) **DF** has been employed as a universal descriptor for comparison purposes and it is the employed tag of a **QO**. Similarity comparisons become possible by means of comparing the **QO DF** tags.

QSAC *Quantum Similarity Aufbau Condition*: A condition that the **QS Matrix** has to comply with in order to be positive definite and admit Choleski decomposition.

QSAR & QSTR QSAR are **QSPR** employed to estimate biological molecular activity values via **molecular descriptors**. When **QSPR** are employed to estimate molecular toxicity, they can be called QSTR.

Quantum similarity matrices (QS matrices) [Depending on the context so as not to cause confusion with the term **QS Measures** the abbreviation **QSM** can also be used to denote **QS Matrices**]: Any matrix which contains computed **QSM** results involving several **QO**. Usually the **QS Matrix** is symmetric and square when the **QSM** on the involved elements of a unique **QOS** are ordered in pairs. In this case the **QSAC** must hold. The **QS Matrix** can be rectangular when computing and ordering into a matrix the **QSM** of two different **QOS**. **QS Hypermatrices** appear when higher order **QSM** associated to more than two **DF** are involved.

Quantum similarity measures (QSM) and indices (QSI) A positive definite integral, computed employing the **QO DF tags** as integrands and, if necessary, including a positive definite operator which can be supposed to act as a weight. Possessing the structure of a measure, such an integral can be interpreted as a generalized volume. **QSM** between two or several **QO** correspond to integrals, which can be constructed primarily with integrands made with the product of the density tags of the compared **quantum objects** plus a positive definite operator. Such a definition ensures the positive definite nature of the **QS** integrals. A quantum *self-similarity* measure corresponds to the **QSM** computed with the **density function tags** of a unique **QO**. **QSI** are manipulations of the **QSM** in order to obtain **QS** comparisons within an adequate range.

QSPR *(Classical) Quantitative Structure-Properties Relationships*. This term refers to any empirical relationship permitting the connection between molecular structure, represented by a set of parameters (**molecular descriptors**) of any origin and molecular properties. Usually, by a classical QSPR can be understood a non-causal multilinear relationship, obtained via statistical reasoning and procedures. This is the generic name given to any functional (usually linear) connecting the properties of a molecule with the attached **molecular descriptors**. The QSPR functionals are empirically obtained by statistical analysis, usually employing (non-)linear regression techniques or any variant of it.

QQSPR *Quantum Quantitative Structure-Properties Relationships*. These are *non-empirical* functionals derived from the structure of a **FQQSPR equation**. Thus, this kind of relationship, if it exists, to some extent can be considered *causal*. A QQSPR permits us to compute the molecular properties of **U-molecules**, just employing non-empirical considerations and parameters or **molecular descriptors** of quantum mechanical origin. In this sense, these relationships can be considered *universal*, applicable to any molecular structure. However, due to the quantum mechanical nature of the QQSPR, one can obtain structure-properties relationships between **QO's** of any kind: nuclei, atoms, molecules. By QQSPR are here understood **QSPR** obtained by means of the application of canonical quantum theoretical methods to obtain expectation values of a Hermitian operator, which within QQSPR theory is described by functionals of the **density function tags** of a given **QOS**, acting as a basis set. QQSPR are more general than classical **QSPR** as they can be obtained for any **QO**, incorporating information difficult to take into account with classical **molecular descriptors**. The descriptors in the QQSPR framework are the **DF tags** of the **QO**.

QS matrix *Quantum Similarity Matrix*. The matrix possessing positive definite elements, constructed by **quantum similarity measures** using the **DF tags** of the elements of a **QOS**. *See:* **Quantum Similarity Matrices**.

Tagged set A set constructed as the Cartesian product of two separate sets. One of them, the *object set* is composed of well-defined elements of any nature, called

objects. The other set in the tagged set is the *tag set*, made of some *descriptors* attached to every object, which is supposed to contain information about the objects; tags can be mathematical elements made by bit strings, vectors, matrices, functions…

U-m *Unknown (properties) molecule*. A molecule which can be described as a **QO**, but lacks the property information associated to the molecules belonging to the **Core Set**. U-m properties act as the unknowns to be evaluated in the **QQSPR** theory developed here.

U-molecule A **U-m**.

U-m set A **QOS** containing **U-m's** as elements.

Wave function A by-product of solving the Schrödinger equation. When studying stationary submicroscopic systems by means of classical quantum mechanics, the pair energy – wave function correspond to the eigenvalue – eigenfunction pairs of the Hermitian Hamiltonian operator constructed for the system, which substitutes the classical Hamilton function. The squared module of the stationary wave functions for each system state is customarily interpreted, since Born times, as the probability density function to find the system as a whole in some space infinitesimal element of volume.

Definition of the Subject

The concept of Quantum Similarity (QS) was introduced for the first time in 1980 in a paper by Carbó et al. [1] entitled: *How Similar is a Molecule to Another?* There the basic aspects of the theory were set up. The backbone of QS was constituted by the conceptual support of the QSM.

QSM. A QSM between two quantum objects (QO): associated to the density function (DF) tags: $\{\rho_A, \rho_B\}$ was defined as the density overlap integral:

$$Z_{AB} = \int_D \rho_A(\mathbf{r})\rho_B(\mathbf{r})\,\mathrm{d}\mathbf{r} = \langle \rho_A \rho_B \rangle = \langle \rho_B \rho_A \rangle = Z_{BA}\,,$$

which is always a positive real number, because the involved DF are non-negative definite real functions. Self-similarity measure integrals were defined in the same manner:

$$I = A, B\colon Z_{II} = \int_D \rho_I(\mathbf{r})\rho_I(\mathbf{r})\,\mathrm{d}\mathbf{r} = \left\langle \rho_I^2 \right\rangle .$$

QS Matrix. A simple example, which remains formally valid for larger QO sets, can illustrate the basic procedures of QS. For a set of two QO a symmetric QS matrix can be set up:

$$\mathbf{Z} = \begin{pmatrix} Z_{AA} & Z_{AB} \\ Z_{BA} & Z_{BB} \end{pmatrix} = \mathbf{Z}^{\mathrm{T}} \leftarrow Z_{AB} = Z_{BA}\,.$$

The columns (or rows) of the QS matrix:

$$|z_A\rangle = \begin{pmatrix} Z_{AA} \\ Z_{BA} \end{pmatrix} \wedge |z_B\rangle = \begin{pmatrix} Z_{AB} \\ Z_{BB} \end{pmatrix}$$

can be interpreted here as the construction of a two dimensional discrete representation of the DF tags pair, quantum mechanically associated to the two involved QO. In this way one can write the association:

$$\forall I = A, B\colon \rho_I \leftrightarrow \||z_I\rangle\,.$$

QSI. From the QS matrix elements several kinds of QS indices (QSI) can also be described. Two indices were described in the seminal paper:

A) Carbó Index; defined as a cosine of the angle subtended by the pair of DF tags $\{\rho_A, \rho_B\}$:

$$R_{AB} = Z_{AB}(Z_{AA}Z_{BB})^{-\frac{1}{2}} \in [0;1] \rightarrow R_{AB} = R_{BA}\,.$$

B) Euclidean Distance Index; constructed with the well-known form:

$$\begin{aligned} D_{AB}^2 &= \int_D |\rho_A(\mathbf{r}) - \rho_B(\mathbf{r})|^2\,\mathrm{d}\mathbf{r} \\ &= Z_{AA} + Z_{BB} - 2Z_{AB} \in [0;+\infty] \\ &\rightarrow D_{AB} = D_{BA}\,. \end{aligned}$$

Ordering a QOS and Mendeleyev Conjecture. QSM or QSI, computed as previously described on the elements of a Quantum Object Set (QOS), are sufficient to permit ordering the elements of the set; thus, opening the way to construct non-artisan periodic tables of molecular sets, for instance.

The possibility to order QOS by means of their DF tags opened the way to estimate unknown properties of some QO; provided a set of QO with known properties was previously ordered. This has led to the path towards Mendeleyev conjecture, described by Carbó and Besalú [2] in 1996.

QQSPR. Mendeleyev conjecture opened the way to employ QS techniques in order to obtain QSPR (Quantitative Structure-Properties Relationships), providing the necessary background for the description of quantum QSPR (QQSPR); a new kind of QSPR functionals that possess the properties to be: (1) *Universal*: as it can be employed to study any QOS. (2) *Unbiased*: as the user cannot choose any other QO descriptor than the DF tag. (3) *Causal*: as the QQSPR functionals are based on a non-empirical equation, derived from the application of quantum theoretical methodology.

Aims. This contribution pretends to provide a mathematical basis for the understanding of the quantum QSPR problem, which tries to find out how to construct universal, unbiased and causal QSPR models. In turn these QSPR models can be employed to predict complex (in the sense of complicated observables) molecular properties. The ultimate purpose of such a theoretical framework is aimed at overcoming the fact that in previous applications molecular quantum similarity numerical values, ordered in the form of similarity matrices, have been employed just like molecular descriptors within a classical QSPR computational way. The future of molecular quantum similarity must be foreseen within the description and further development of an autonomous QQSPR set of computational procedures.

Introduction

Quantum Object Sets and Core Sets. QSPR studies are based on some set of molecules: M, attached to a collection of descriptors and properties; the whole is the *core set*, symbolized as: C. For all the elements of the set M, via the Schrödinger equation for every molecule in the *core set* a wave function can be computed, providing in turn a set of density functions: $P = \rho_I$, which can be considered as unique continuous molecular descriptors, by means of the quantum mechanical interpretation [3,4].

According to this it can be considered that:

$$\forall m_I \in M \rightarrow \exists \rho_I \in P \wedge \forall I\colon m_I \leftrightarrow \rho_I \,.$$

Therefore, quantum similarity theory [5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22], permits us to perform a Cartesian product of the M and P sets, which is used to build up a *tagged set* [23] $Q = M \times P$, named a *quantum object set* (QOS). In a QOS the molecules constitute the *object set* and the density functions the *tag set* [23,24,25]. The elements of a QOS are *quantum objects* (QO). QO ordered pairs, are constructed in the following way:

$$\forall m_I \in M \wedge \forall \rho_I \in P \rightarrow \forall \omega_I \in Q\colon \omega_I = (m_I; \rho_I)\,. \quad (1)$$

Then the *core set*, C is a well-defined QOS. Because of the QOS definition in Eq. (1), C can have the form:

$$\forall \omega_I \in Q \wedge \exists \pi_I \in \Pi \rightarrow \forall c_I \in C = Q \times \Pi\colon c_I = (\omega_I; \pi_I) \equiv (m_I; \rho_I; \pi_I),$$

Π contains the properties of some elements of M. Hence, C elements are triples made of molecular structures, density functions and properties: $C = M \times P \times \Pi$.

In classical QSPR density functions are replaced by finite dimensional vectors, whose elements are the so-called *molecular descriptors*. Construction of the *core set* with discrete vector spaces, substituting P, will also appear within the QQSPR. The substitution of the continuous density tags by discrete vectors in QQSPR has a mathematical-theoretical meaning, while in empirical QSPR this remains arbitrary. The elements of the *core set* C are core molecules, C-molecules or briefly C-m.

QQSPR Operators, Quantum Similarity Measures and the Fundamental QQSPR Equation. Correspondence principle provides the rules to construct Hermitian operators, with expectation values associated to the experimental outcomes of submicroscopic systems observables [3,4]. For some complex (complicated) observables, like biological activities, the correspondence principle cannot be applied, as Hermitian operators are unavailable or difficult to be obtained. The QQSPR operators and the attached fundamental QQSPR equation, create an approximate quantum mechanical computational environment in order to estimate the expectation values of complicated observables.

QQSPR Operators. The fundamental QQSPR equation arises when density function tags: $\{\rho_I(\mathbf{r})\}$ of some QOS are used to construct a QQSPR operator as:

$$\Omega(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3, \ldots) = x_0\Theta_0(\mathbf{r}_1) + \sum_I x_I\rho_I(\mathbf{r}_2)\Theta_1(\mathbf{r}_1, \mathbf{r}_2) + \sum_I \sum_J x_I x_J \rho_I(\mathbf{r}_2)\rho_J(\mathbf{r}_3)\Theta_2(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3) + O(3) \quad (2)$$

in Eq. (2), x_0 is an arbitrary constant; $\{\Theta_\omega(\mathbf{R})|\omega = 0, 1, 2, \ldots\}$ is a known positive definite operator set, acting as a weight set; and $\{x_I\}$ is a set of parameters, determined through the fundamental QQSPR equation.

The structure of a QQSPR operator (2) has to be seen as a first step algorithm permitting us to define approximate quantum mechanical operators. The QQSPR operators can be employed afterwards to evaluate their quantum mechanical expectation values.

Expectation Values of the QQSPR Operator. To determine the parameter set $\{x_I\}$, it is necessary to compute the set of expectation values over the elements, the core molecules or C-m, of the core set C. Besides a well-defined structure and a known density function, as members of a QOS, the C-m possess a known property value of the set: $\Pi = \{x_I\}$, attached to each one.

Then, every known property of the C-m elements can be expressed as an expectation value of a QQSPR operator:

$$\forall m_K \in \mathrm{C}\colon \pi_K \approx \langle \Omega\rho_K\rangle = x_0\langle\Theta_0\rho_K\rangle + \sum_I x_I\langle\rho_I\Theta_1\rho_K\rangle + \sum_I x_I x_J\langle\rho_I\rho_J\Theta_2\rho_K\rangle + O(3)\,. \tag{3}$$

Zero-th Order Term. In the expectation values (3) of the elements of C, the Zero-th order term is:

$$\theta_K[\Theta_0] = x_0\langle\Theta_0\rho_K\rangle = x_0\int_D \Theta_0(\mathbf{r}_1)\rho_K(\mathbf{r}_1)\,\mathrm{d}\mathbf{r}_1$$

being a constant for each C-m, the Zero-th order term: $x_0\Theta_0(\mathbf{r})$ acts as an origin shift. Choosing: $\Theta_0(\mathbf{r}) = I$, this term becomes proportional to the number of electrons of the C-m considered:

$$\theta_K[I] = x_0\langle\rho_K\rangle = x_0\int_D \rho_K(\mathbf{r}_1)\,\mathrm{d}\mathbf{r}_1 = x_0 N_K\,.$$

The Zero-th order term can be omitted if it is no longer necessary to shift the property values of the C-m.

Quantum Similarity Measures in First and Second-Order Expectation Value Terms. The first-order term in Eq. (3) contains QSM integrals between pairs of C-m density function tags, long time known [15]:

$$z_{IK}[\Theta_1] = \langle\rho_I\Theta_1\rho_K\rangle = \int_D\int_D \rho_I(\mathbf{r}_2)\Theta_1(\mathbf{r}_1,\mathbf{r}_2)\rho_K(\mathbf{r}_1)\,\mathrm{d}\mathbf{r}_1\,\mathrm{d}\mathbf{r}_2\,,$$

and the second-order term is made of triple-density quantum similarity measures [16]

$$z_{IJK}[\Theta_2] = \langle\rho_I\rho_J\Theta_2\rho_K\rangle = \int_D\int_D\int_D \rho_I(\mathbf{r}_2)\rho_J(\mathbf{r}_3)\Theta_2(\mathbf{r}_1,\mathbf{r}_2,\mathbf{r}_3)\rho_K(\mathbf{r}_1)\,\mathrm{d}\mathbf{r}_1\,\mathrm{d}\mathbf{r}_2\,\mathrm{d}\mathbf{r}_3\,.$$

The matrix symbol $\mathbf{Z}$ will be used to represent any collection of QSM: $\{z_{IJ}\}$, independently of the nature of the weighting operator Θ_1.

Quantum Similarity Matrices (QSM) in the Construction of First-Order QSPR Operators and the Definition of Discrete QOS (DQOS). The first-order approach of the QSPR operator [13,14,15,16,17], applied to the *core set* with the known property set: $\Pi = \{\pi_I\}$, produces the equation collection:

$$\forall I\colon p_I = \pi_I - \langle\Theta_0[\rho_I]\rangle \approx \sum_J x_J\langle\rho_J\Theta_1[\rho_I]\rangle = \sum_J x_J z_{JI}\,. \tag{4}$$

If $\Theta_1 = I$ is used, the first-order integrals (4) are:

$$\left\{\langle\rho_J[\rho_I]\rangle = \int_D \rho_J\rho_I\,\mathrm{d}V = z_{JI} = z_{IJ} = \int_D \rho_I\rho_J\,\mathrm{d}V = \langle\rho_I[\rho_J]\rangle\right\},$$

and can be ordered into a $(n\times n)$ symmetric array, constructing in this way the so-called *quantum similarity matrix*: $\mathbf{Z} = \{z_{IJ}\}$ (QS Matrix) [18]. The property set form a column vector: $|\mathbf{p}\rangle = \{p_I\}$. Equations (4) are a linear system, which can be used to evaluate unknown molecular properties for some QOS members of the *core set*.

Every column of the QSM [19]: $\mathbf{Z} = \{|\mathbf{z}_I\rangle = \{z_{JI}\}\}$, is a discrete representation of each QO density function in: $P = \{\rho_I\}$. A one-to-one correspondence exists between the density tag set and the QSM column submatrices:

$$\forall m_I \in M\colon \rho_I \leftrightarrow |\mathbf{z}_I\rangle \Rightarrow \mathrm{P} \Leftrightarrow \mathbf{Z}\,.$$

The QSM column set can be used as a n-dimensional vector tag set, attached to the molecular set, building up a tagged set, called *discrete quantum object set* (DQOS) [19,20,21,22,26]:

$$Q_Z = \mathrm{M}\times\mathbf{Z}\,. \tag{5}$$

In DQOS, the density function tags of the original QOS, Q, belonging to the tag set P, are substituted by the columns of the QSM. There also exists a one-to-one correspondence between both QOS: $Q \leftrightarrow Q_Z$.

Fundamental QQSPR Equation Setup. Expectation values of the QQSPR operator (3) can be collected in a column vector, providing the fundamental QQSPR equation:

$$|\mathbf{p}\rangle \approx \mathbf{Z}_1|\mathbf{x}\rangle + \langle\mathbf{x}|\mathbf{Z}_2|\mathbf{x}\rangle + O(3)\,. \tag{6}$$

In (6), $|\mathbf{p}\rangle = \{p_K\}$ is the shifted C-m properties vector: $|\mathbf{p}\rangle = |\pi\rangle - |\theta\rangle$, where $|\pi\rangle = \{\pi_I\}$ is the original property vector and $|\theta\rangle = \{\theta_K\}$ is the completely determined Zero-th order origin shift vector, $\{\mathbf{Z}_\omega|\omega = 1,2,\ldots\}$ is a matrix set containing the quantum similarity measures, for instance: $Z_1 = \{z_{IK}\}; Z_2 = \{z_{IJK}\}; \ldots$, and: $|\mathbf{x}\rangle = \{x_I\}$ is a column vector bearing the unknown coefficients, which define explicitly the QQSPR operator (2).

The unknown coefficients $|\mathbf{x}\rangle = \{x_I\}$ can be obtained solving the linear equation contained in the fundamental QQSPR Eqs. (4):

$$|\pi\rangle - |\theta\rangle = |\mathbf{p}\rangle = \mathbf{Z}_1|\mathbf{x}\rangle \rightarrow |\mathbf{x}\rangle = (\mathbf{Z}_1)^{-1}|\mathbf{p}\rangle\,. \tag{7}$$

Equation (7) has no predictive power. This is so because the first-order similarity matrix $\mathbf{Z}_1$ has to be chosen positive definite by construction. By predictive power it is understood the possible computation of the property value

for an also known quantum object, a U-m, which as such possesses a well-defined structure and density function, but belongs to the *U* set.

In the last years, since the description of QS measures for the first time [1], the predictive power of the information contained in the QS matrices set has been manipulated within the classical QSPR. This is the same as considering the similarity matrices as a source of molecular parameters to construct empirical QSPR. See references [5] and [27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46].

First-Order Fundamental QQSPR Equation. The study of QQSPR predictive potential starts with the first-order fundamental QQSPR equation, involving the *core set*, containing the associated DQOS molecules, possessing known values of some complex property.

The first-order QQSPR fundamental equation in a compact matrix form [60] is written as:

$$\mathbf{Z}|\mathbf{x}\rangle = |\mathbf{p}\rangle\,; \tag{8}$$

Where the matrix $\mathbf{Z}$ is the already described symmetric QSM, $|\mathbf{p}\rangle$ is the known *core set* shifted property vector and $|\mathbf{x}\rangle$ is a $(n \times 1)$ vector, whose coefficients have to be evaluated.

The predictive power of such an equation is a priori null. This is so because the QSM: $\mathbf{Z}$, is by construction non-singular, then exists a QSM inverse $\mathbf{Z}^{-1}$, with the relationships: $\mathbf{Z}^{-1}\mathbf{Z} = \mathbf{Z}\mathbf{Z}^{-1} = \mathbf{I}$. This leads to the trivial result, defining the unknown coefficient vector:

$$|\mathbf{x}\rangle = \mathbf{Z}^{-1}|\mathbf{p}\rangle\,. \tag{9}$$

And exact property values for any molecule of the *core set*, can be reproduced just choosing the scalar products:

$$\forall I\colon p_I = \langle \mathbf{z}_I|\mathbf{x}\rangle\,. \tag{10}$$

QSM for varied *core sets* have been used in a set of prediction studies [27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46], employing up to date statistical tools, typical of classical QSPR studies, [47,48,49,50,51,52,53,54,55,56,57,58]. The use of the fundamental QQSPR equation to construct algorithms, which can be utilized as predictive tools independently of classical QSPR algorithms, has been previously attempted [59], but it has not been continued in practice until recently [60,61,62]. Here will be discussed in several places not only the QQSPR problem itself, but various points of view and a future perspective as well.

Symmetrical Similarity Matrices. The fundamental QQSAR equation has been presented within the particular case where the basis and probe molecular QOS coincide, forming a square symmetric QS Matrix. This choice has the drawback that the fundamental QQSPR linear system becomes well defined, with a unique solution, whenever the similarity matrix is non-singular as no QO coincides with another within the QOS *C*. Even then, there is quite a wide range of solutions to overcome this apparent limitation. Among other procedures, one can use the *C* symmetric QSM, $\mathbf{Z}$, as a source of molecular descriptors and afterwards employ them in classical statistical treatments. This choice, as has been already commented, has been studied in many publications of our laboratory with success [27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46]. Other possible QSM can be constructed bearing rectangular structure, for example using known molecular structures acting as a basis set, which in turn can be compared with the *core set*.

Origin of Hansch QSAR Models. An interesting possibility of the symmetric square representation for the QS Matrices corresponds to its potential to unveil the origin of one parameter classical QSAR models, such as those Hansch [63] described some years ago. A fundamental QQSPR linear equation can be associated to a set of ν equations with the same number of unknowns, and can be rewritten as:

$$\forall J = 1,\nu\colon p_J \simeq \sum_{I=1}^{\nu} x_I z_{IJ} = x_J z_{JJ} + \sum_{I\neq J}^{\nu} x_I z_{IJ}\,. \tag{11}$$

There is no need to attach the QSM elements to any specific QOS, as all of them are computed over a unique basis of density function tags. Considering (11), two terms can be seen. The first one is attached to a self-similarity measure z_{JJ}, while the second term in cases of some not so strongly varying QOS, can be considered almost a constant, that is using:

$$\forall J\colon \alpha = x_I \wedge \beta \simeq \sum_{I\neq J}^{\nu} x_I z_{IJ}\,. \tag{12}$$

Equation (11) takes the final form:

$$\forall J = 1,\nu\colon p_J \simeq \alpha z_{JJ} + \beta \tag{13}$$

which has the required appearance to be considered as possessing a Hansch-like structure.

Equation (13) proves self-similarities can be substitutes of the classical Hansch analysis parameters [63]. They constitute, for co-generic QOS, molecular descriptors with the property to be directly attached to a tri-dimensional molecular structure. Quantum self-similarity

measures vary slowly with conformational changes [10, 64], so their values for the optimal molecular geometry can be safely used.

Mathematical Background of Quantum Similarity

Inward Matrix Product (IMP)

IMP Definition. An essential piece of QSM theory is the matrix operation called the *inward matrix product* (IMP) [65,66,67,68,69], which is based on the structure of the *Hadamard product* [70][1]. Such an operation is an internal composition law, which can be defined within a matrix (or hypermatrix) vector space $M_{(m\times n)}(K)$ of arbitrary dimension $(m \times n)$ and defined over a field K, producing a matrix whose elements are products made in turn by the elements of the matrices appearing in the IMP itself, according to the straightforward algorithm:

$$\forall \mathbf{A} = \{a_{ij}\}, \mathbf{B} = \{b_{ij}\} \in M(K)\colon \mathbf{P} = \mathbf{A} * \mathbf{B} \rightarrow \mathbf{P} = \{p_{ij}\} \in M(K) \wedge \forall i, j\colon p_{ij} = a_{ij} b_{ij} \,. \quad (14)$$

IMP is an operation, which can be applied not only to matrix spaces but over a wide variety of mathematical objects, producing another mathematical object of the same kind as the ones involved in the operation.

IMP Properties. On the other hand, IMP is equivalent to a feature involving arrays, present in high-level computer languages such as Fortran 95 [71], so practical programming of the IMP properties and characteristics is straightforward. IMP is commutative, associative, and distributive with respect to the matrix sum. Moreover, it has a multiplicative neutral element, the *unity matrix*, which has been customarily represented by a bold real unit symbol and formally defined as: $\mathbf{1} = \{1_{ij} = 1\}$.

IMP Powers and Functions. By an IMP power over a matrix $\mathbf{A} = \{a_{ij}\}$, noted as: $\mathbf{A}^{[p]}$ is understood the matrix whose elements are the corresponding powers of the elements of $\mathbf{A}$, that is: $\mathbf{A}^{[p]} = \{a_{ij}^{p}\}$. In the same manner, an IMP function of a matrix, noted as: $f[\mathbf{A}]$, is defined as the matrix whose elements are the functions of the original matrix: $f[\mathbf{A}] = \{f(a_{ij})\}$.

Scalar Product as an IMP Composite Operation. A useful application example of IMP is associated to the *total sum* of the elements of an arbitrary matrix, $\mathbf{A} = \{a_{ij}\} \in M$, by means of the symbol:

$$\langle \mathbf{A} \rangle = \sum_i \sum_j a_{ij} \,. \quad (15)$$

Connecting this definition with IMP, one can easily write:

$$\langle \mathbf{A} * \mathbf{B} \rangle = \langle \mathbf{A} \rangle * \langle \mathbf{B} \rangle \,.$$

Then, it is simple to construct the scalar product of two matrices of the same dimension, symbolized here as:$\langle \mathbf{A}|\mathbf{B} \rangle$, by means of the IMP structure:

$$\langle \mathbf{A}|\mathbf{B} \rangle = \sum_i \sum_j a_{ij} b_{ij} = \langle \mathbf{A} * \mathbf{B} \rangle \,. \quad (16)$$

In this way, the definition of distances and cosines of the angle between two matrices can be also outlined. For instance, the cosine of the angle subtended by two matrices can be written, according to Eq. (16), as:

$$\cos(\alpha) = (\langle \mathbf{A} * \mathbf{A} \rangle \langle \mathbf{B} * \mathbf{B} \rangle)^{-\frac{1}{2}} \langle \mathbf{A} * \mathbf{B} \rangle \,.$$

Vector Semispaces (VSS)

A *vector semispace* (VSS) [23,24,25,72,73] is a vector space, where the additive group has been substituted by an additive *semigroup*. An additive semigroup [74] is an additive group without reciprocal elements, which is the same as to consider negative elements not present in VSS. A matrix VSS will be made by matrices whose elements are positive definite or semi-definite. QS matrix structures belong to positive definite VSS. This is the same as to consider the matrix elements forming a VSS constructed by positive definite real numbers, extracted in turn from the $\mathbf{R}^+$ half-line. All the elements of a matrix VSS are non-singular matrices from the IMP point of view, while any matrix possessing a zero element will be non-existent in a VSS, if this strict sense is adopted. A functional VSS can be constructed by positive definite functions over a given domain and lacking of the null function in order to comply with the strict VSS characteristics.

Minkowski Norms in VSS. Because of the positive definite structure of the components of a VSS, the easiest way to define a norm within such a mathematical configuration is Minkowski's. In a matrix VSS one can write:

$$\forall \mathbf{A} \in M(\mathbf{R}^+) \rightarrow \langle \mathbf{A} \rangle \in \mathbf{R}^+ \,. \quad (17)$$

[1] The *Hadamard product* (sometimes also called *Schur or Kronecker product*) is related to the multiplication result of two sums and constructed by the sum of the resultant diagonal cross-terms only. In this way, the inward (or Hadamard) product of two sums can be specified by the following algorithm:

$$\left(\sum_I^N a_I\right) * \left(\sum_I^N b_I\right) = \sum_I^N a_I b_I \,.$$

Both sums shall have the same number of terms N, for the IMP being feasible.

Meanwhile, in any general functional VSS, an equivalent form can also be defined:

$$\forall \rho(\mathbf{r}) \in F(\mathbf{R}^+) \rightarrow \langle \rho \rangle = \int_D \rho(\mathbf{r})\, \mathrm{d}\mathbf{r} \in \mathbf{R}^+ \,. \tag{18}$$

As final information one can see that a Minkowski norm in $M(\mathbf{R}^+)$, and, thus, the complete sum of a matrix elements, can be considered as a linear operator, that is:

$$\langle \alpha \mathbf{x} + \beta \mathbf{y} \rangle = \alpha \langle \mathbf{x} \rangle + \beta \langle \mathbf{y} \rangle \,.$$

σ-Shell Structure in VSS. Minkowski-like norms classify the VSS elements in subsets, the *σ-shells*, $S(\sigma)$, whose elements are defined by the value of such a norm:

$$\forall x \in S(\sigma) \subset V(\mathbf{R}^+) \rightarrow \langle x \rangle = \sigma \in \mathbf{R}^+ \,. \tag{19}$$

The *unit shell* $S(1)$ is a VSS subset, which can generate all the other VSS shells. The existence of this property can be easily constructed as follows:

$$\forall z \in S(\sigma) \rightarrow \exists x \in S(1) \colon z = \sigma x \,. \tag{20}$$

Convex Conditions. The idea underlying *convex conditions*, associated to a numerical set, a vector, a matrix, or a function, has been described since the initial work on VSS and the related questions [23,24,25,72,73]. By the *convex conditions symbol*: $K(\mathbf{x})$ is meant that the conditions:

$$\langle \mathbf{x} \rangle = 1 \wedge \mathbf{x} \in V(\mathbf{R}^+) \,,$$

hold simultaneously for a given mathematical object $\mathbf{x}$. Convex conditions become the same as considering that the object belongs to the unit shell of some VSS. Then, for such kinds of elements:

$$K(\mathbf{x}) = \{\langle \mathbf{x} \rangle = 1 \wedge \mathbf{x} \in V(\mathbf{R}^+)\} \equiv \{\mathbf{x} \in S(1)\} \,.$$

Conversely, the following property holds over any element of the unit shell: $\forall \mathbf{x} \in S(1) \rightarrow K(\mathbf{x})$.

Convex Linear Combinations Within a σ-Shell. Given an arbitrary σ-shell: $S(\sigma) \subset V(\mathbf{R}^+)$, of some VSS, then convex linear combinations of the elements of the $\sigma-$shell produce a new element of $S(\sigma)$. That is, suppose that the convex conditions:

$$K(\{\gamma_I\}) = \left\{ \sum_I \gamma_I = 1 \wedge \forall I \colon \gamma_I \in \mathbf{R}^+ \right\} \,, \tag{21}$$

hold on a known set of scalars $\{\gamma_I\}$. Then, the following property will be fulfilled for any arbitrary subset of elements belonging to the chosen $\sigma-$-shell:

$$\{\mathbf{x}_I\} \subset S(\sigma) \wedge K(\{\gamma_I\}) \rightarrow \mathbf{x} = \sum_I \gamma_I \mathbf{x}_I \in S(\sigma) \,,$$

owing to the fact that the summation symbol, associated here to a Minkowski norm, is a linear operator, thus:

$$\langle \mathbf{x} \rangle = \left\langle \sum_I \gamma_I \mathbf{x}_I \right\rangle = \sum_I \gamma_I \langle \mathbf{x}_I \rangle$$
$$= \sum_I \gamma_I \sigma = \sigma \sum_I \gamma_I = \sigma \rightarrow \mathbf{x} \in S(\sigma) \,.$$

Such a property is related to the possibility of constructing approximate atomic and molecular DF, by means of convex linear combinations, using a basis set of structurally simpler functions, which shall belong to the same VSS σ-shell. One of the possible technical options has been described in a series of papers, where the choice of the simplified functions, in atomic electronic density fitting, was made by sets of 1s GTO. The approach was termed *atomic shell approximation* (ASA) [75,76,77,78,79,80,81,82,83] and has been successfully employed, among other possibilities, to make the integral computation and molecular superposition inherent in *molecular quantum similarity measures* (MQSM) easier.

Generating Vector Spaces

Generating Symbols. Any VSS σ-shell structure can be supposedly generated by a conventional *vector space* (VS). Such VS can be defined over the complex or real fields. It can be additionally provided by convenience with a positive definite metric structure. Indeed, suppose such a VS, defined for the sake of generality over the complex field: $V(\mathbf{C})$. Then, from a very general point of view, the following algorithm can be envisaged:

$$\forall v \in V(\mathbf{C}) \wedge v \neq \mathbf{0} \rightarrow \langle v|v \rangle = \sigma \in \mathbf{R}^+$$
$$\Rightarrow \exists x = v^* * v \in S(\sigma) \subset V(\mathbf{R}^+) \,. \tag{22}$$

Where the IMP: $x = v^* * v$ has been used to construct the VSS vector. Then the following sequence:

$$\langle x \rangle = \langle v^* * v \rangle = \langle v|v \rangle = \sigma$$

holds and has been employed to set the form of Eq. (22).

The quantum mechanical image of the density function construction appears as a particular case of the definition attached to Eq. (22). In addition, from a complementary point of view, a symbol to briefly summarize Eq. (22) could be described. One can say compactly that the vector *generates* a VSS vector x, using a *generating symbol*: $R(v \rightarrow x)$, whenever the sequence of relationships in Eq. (22) holds [23,24,25,72,73].

Probability Density Distributions. From the quantum mechanical point of view, when a wave function $\Psi(\mathbf{r})$ is

known, then the attached DF $\rho(\mathbf{r})$ is generated in the following way: $R(\Psi \rightarrow \rho)$ [25].

Another interesting point to be noted is that any probability distribution, discrete or continuous, belongs to some VSS unit shell. Probability distributions can be generated by the conveniently normalized elements of some normed or metric space, in order that the resultant VSS element belongs to the unit shell, $S(1)$. Potentially, in this way they can belong to any other σ-shell: $\forall \rho(\mathbf{r}) \in S(1) \wedge \sigma \in \mathbf{R}^+ \rightarrow p(\mathbf{r}) = \sigma\rho(\mathbf{r}) \in S(\sigma)$.

Thus, in VSS one can consider that the unit shell resumes every other σ-shell. In this manner, probability distributions of any kind can be transformed into any other element of the associated VSS. One can also say that any VSS σ-shell, $S(\sigma)$, can be considered like a homothetic construct of the unit shell, $S(1)$ [72]. Because the elements of the unit shell comply with the adequate form of a probability distribution, then they also fulfil the convenient convex conditions. That is: $\forall x \in S(1) \rightarrow K(x)$. In other words, any probability distribution can be considered as an element of the unit shell forming part of a VSS.

Scalar Products and Measures in VSS. Because of these possible connections attached to probability vectors, scalar products of two distinct compatible probability distributions are always positive definite, as one has:

$$\forall x, y \in S(1) \rightarrow \langle x|y\rangle = \langle x * y\rangle \in \mathbf{R}^+ . \tag{23}$$

Therefore, the cosine of the angle subtended by two probability distributions has to be contained in the open interval: $(0, 1]$, due to Eq. (23). This is so, because the cosine of the subtended angle between two elements can be computed as:

$$\forall x, y \in S(1): \cos(\alpha) = (\langle x*x\rangle\langle y*y\rangle)^{-\frac{1}{2}}\langle x*y\rangle \in (0, 1].$$

Furthermore, Eq. (23) shows that the scalar product between two, or more, elements of the unit shell of a given metric VSS, constitute in any case a *measure*. In this way, such a scalar product can be considered a generalized volume.

Quantum Similarity

QSM over the Unit Shell. A similarity measure over the unit shell of any VSS can be defined through the description of the mathematical elements described so far. *Quantum similarity measures* (QSM) were described a long time ago [1] and have been constantly used since then [84,85,86,87,88,89,90,91,92,93]. In the most simple and at the same time general way, within the easier formalism available, a QSM can be defined knowing the appropriate DF of two quantum systems: $\{\rho_A; \rho_B\}$, adapted to the unit shell of the corresponding VSS, and using as weight some positive definite operator: $\{\Omega\}$, then the integral measure:

$$\begin{aligned} z_{AB}(\Omega) &= \iint_D \rho_A(\mathbf{r}_1)\Omega(\mathbf{r}_1;\mathbf{r}_2)\rho_B(\mathbf{r}_2)\,\mathrm{d}\mathbf{r}_1\,\mathrm{d}\mathbf{r}_2 \\ &= \langle \rho_A \Omega \rho_B\rangle \in \mathbf{R}^+ \end{aligned} \tag{24}$$

will correspond to a weighted scalar product, defined over the unit shell elements, made in turn by the compatible quantum DF. The QSM (24) can be associated to a property very comparable to the one encountered in Eq. (24) and in any instance has to possess a positive definite nature.

Quantum Object Sets. Suppose a set of quantum systems: $S = \{s_I\}$, in a well-defined set of states. Suppose that to every quantum system there is attached a known state DF, forming the set: $P = \{\rho_I\}$, belonging to the unit shell of some functional VSS. A tagged set [23,24,25] can be constructed, using the Cartesian product: $T = S \times P$, where each element, $\tau_I \in T$, is constructed by the ordered pair composition rule: $\tau_I = (s_I; \rho_I)$, forming in this way a *quantum object*. The tagged set T constitutes a *quantum object set*, that is: $T = \{\tau_I\}$. The QSM earlier defined in Eq. (24), can be interpreted, in turn, as a tensor product of the tag part of the QOS.

Quantum Object Sets and Core Sets. The usual problem in all QSPR studies is customarily based on the previous knowledge of some molecular set: M, of cardinality n, such that the structures and properties of the set elements are known beforehand. From now on, one can refer to this collection of molecules, molecular descriptors and properties as the *core set* and name it as: C. With the molecular structures of the elements of the set M known, one can solve the Schrödinger equation associated to, in principle, the ground state of every molecule in the set and compute an attached set of density functions: $\mathrm{P} = \{\rho_I\}$, which can be connected to sole continuous molecular descriptors, according to the quantum mechanical usual custom; such that:

$$\forall m_I \in \mathrm{M} \rightarrow \exists \rho_I \in \mathrm{P} \wedge \forall I: m_I \leftrightarrow \rho_I .$$

In terms of the theoretical settings related with quantum similarity, the Cartesian product of the molecular and the density function sets is used to construct a *tagged set*: $\mathrm{Q} = \mathrm{M} \times \mathrm{P}$. Such a tagged set has been formerly named a *quantum object set*, where the molecules constitute the *object set* and the density functions act as the *tag set*. Thus,

from now on, one can consider the *core set* as a well-defined QOS. One can name the elements of a QOS as *quantum objects*. Therefore, ordered pairs, constructed in the following way, define any QO:

$$\forall m_I \in \mathrm{M} \wedge \forall \rho_I \in \mathrm{P} \rightarrow \forall \omega_I \in \mathrm{Q}\colon \omega_I = (m_I; \rho_I)\,. \quad (25)$$

However, the *core set* shall be structured in an even extended manner. Starting from the QOS definition, the *core set* C has also to be associated with the following characteristic:

$$\forall \omega_I \in \mathrm{Q} \wedge \exists \pi_I \in \Pi \rightarrow \forall c_I \in \mathrm{C} = \mathrm{Q} \times \Pi\colon c_I = (\omega_I \pi_I) \equiv (m_I; \rho_I; \pi_I)\,,$$

Where the set Π contains all the properties of the elements of the molecular set M. Hence, *core set* elements are well-defined triples consisting of molecular structures, density functions, and properties: $\mathrm{C} = \mathrm{M} \times \mathrm{P} \times \Pi$.

In classical QSPR, based on the empirical description of the set M, the set of density functions is replaced by a set of vectors, belonging to a finite dimensional space, whose elements are the chosen *molecular descriptors*. The possible construction of the *core set* within discrete vector spaces, substituting the density tag set P is a characteristic, which will also appear within the QQSPR theoretical development, as will be explained below. It must be said that the QQSPR substitution of the continuous density tags by discrete vectors has a quite well-structured mathematical-theoretical meaning, while in empirical QSPR remains arbitrarily chosen.

Similarity Matrices. Collecting all the QSM computed between the element pairs of a given QOS, a so-called *Quantum Similarity Matrix* (QSM) is obtained, and constructed according to the definition (24) by means of: $\mathbf{Z} = \{z_{ij}\}$. Because of the QS matrix elements structure, the matrix itself can be considered as an element of some VSS of the appropriate dimension. The QS matrix $\mathbf{Z}$ is a symmetric matrix with positive definite elements, whose columns $\{\mathbf{z}_I\}$ (or rows) are also elements of some N-dimensional VSS. As such, there exists a real symmetric matrix, $\mathbf{X}$, such that, in general $R(\mathbf{X} \rightarrow \mathbf{Z})$, that is:

$$\mathbf{Z} = \mathbf{X} * \mathbf{X} = \mathbf{X}^{[2]} \vee \mathbf{X} = \mathbf{Z}^{\left[\frac{1}{2}\right]}\,. \quad (26)$$

As a consequence, any QS matrix belongs to a precise σ-shell of some VSS. That is:

$$\forall \mathbf{Z}\colon \langle \mathbf{Z} \rangle = \sum_i \sum_j z_{ij} = \sigma \rightarrow \mathbf{Z} \in S(\sigma) \subseteq M(\mathbf{R}^+)\,.$$

Stochastic Similarity Matrices. Even if the columns or rows of the QS matrix $\mathbf{Z}$ belong to different σ-shells of some VSS, they can be easily brought to the unit shell, by using a set of simple homothetic transformations, involving a product by a diagonal matrix, with elements constructed by the Minkowski norms of the columns (or rows) of the QS matrix. That is, the diagonal matrix:

$$\mathbf{D} = \mathrm{Diag}(\langle \mathbf{z}_1 \rangle; \langle \mathbf{z}_2 \rangle; \ldots; \langle \mathbf{z}_I \rangle; \ldots)\,, \quad (27)$$

can transform the QS matrix $\mathbf{Z}$ into a column (or row) *stochastic matrix*, simply by multiplying on the right (or the left) of $\mathbf{Z}$ by the inverse of $\mathbf{D}$, respectively [59]. For instance, the stochastic column matrix associated to the QS matrix is:

$$\mathbf{S} = \mathbf{Z}\mathbf{D}^{-1}\,. \quad (28)$$

In this way, the columns $\{s_I\}$ of the stochastic matrix,$\mathbf{S}$, belong to the unit shell of the column vector VSS of the appropriate dimension. That is:

$$\mathbf{S} = \{\mathbf{s}_I\} \rightarrow \forall I\colon \langle \mathbf{s}_I \rangle = \left\langle \langle \mathbf{z}_I \rangle^{-1} \mathbf{z}_I \right\rangle = \langle \mathbf{z}_I \rangle^{-1} \langle \mathbf{z}_I \rangle = 1 \rightarrow \mathbf{s}_I \in S(1)\,.$$

However, the column stochastic QS matrix (28) appears to be no longer symmetric as his originating QS matrix $\mathbf{Z}$ is.

Quantum Similarity Matrix Aufbau Procedure [94]

Suppose a known given Quantum Object Set formed by N molecules, with density tags described as: $\{\rho_I(\mathbf{r})\}$. Up to now, the usual procedure to construct the QSM has been to maximize each of the integrals of type (26) with respect to the translations and rotations of one of the implied QO in relation to the others. This can be expressed formally, for instance, as:

$$\forall I > J\colon z_{IJ} = \langle \rho_I \rho_J \rangle = \int_D \rho_I(\mathbf{r}) \rho_J(\mathbf{r})\, \mathrm{d}\mathbf{r} = z_{JI}\,, \quad (29)$$

However, as has been well known since the first paper on the subject [1] the set of quantum similarity measures $\{z_{IJ}\}$ depend on the relative position in 3D space of the implied Quantum Objects (QO's). As the QO density function labels are positive definite functions, the integrals of type (29) can be considered as measures; thus, they are positive definite too. Up to now the usual procedure to construct the QSM has been to maximize each of the integrals of type (29) with respect to the translations and rotations of one of the implied QO in relation to the other. This can be expressed formally, for instance, as:

$$\forall I > J\colon z_{IJ} = \max_{\mathbf{t};\boldsymbol{\Omega}} \int_D \rho_I(\mathbf{r}) \rho_J(\mathbf{r}|\mathbf{t};\boldsymbol{\Omega})\, \mathrm{d}\mathbf{r} \quad (30)$$

where the pairs: $\{\mathbf{t};\boldsymbol{\Omega}\}$ are translations and rotations respectively, performed on the center of coordinates of the Jth QO. It is irrelevant which one of the QO pair is chosen in order to optimize the integral (29) by means of the algorithm (30), the same result shall be obtained choosing the Ith QO for undertaking translations and rotations.

Apparently, such a procedure, repeated for every non-redundant couple of QO's, shall provide a QSM $\mathbf{Z}$ with appropriate characteristics associated to a metric matrix. The most important one is that the attached QSM has the property to be positive definite; as the density tag set is linearly independent, if the QOS is made of different QO's, then $\mathbf{Z}$ has to be a metric matrix of a pre-Hilbert space [94]. However, in many cases the use of algorithm (30) does not provide a QSM whose whole spectrum is positive definite, but a small amount of the $\mathbf{Z}$ eigenvalues may appear to be negative. This non-definite behavior of the metric matrix $\mathbf{Z}$ can be attributed to the fact that following algorithm (30), when facing the Jth QO to the rest of the QOS elements, then for every distinct QO a different relative position of the Jth QO is found, while reaching the optimal value of the similarity measure (29) for every pair of QO's; that is: the relative position of the Jth QO with respect to the Ith QO, $\forall I\colon J \neq I$, in order to optimize every element z_{IJ}, becomes different, and therefore when optimizing Eq. (30) one will obtain a set of different optimal translations-rotations: $\{\mathbf{t}_I;\boldsymbol{\Omega}_I\}\forall I \neq J$.

When computing any optimal quantum similarity measure by means of algorithm (30), one also must be aware that the final result, can be used to construct the symmetric (2×2) matrix:

$$\mathbf{Z}^{IJ} = \begin{pmatrix} z_{II} & z_{IJ} \\ z_{JI} & z_{JJ} \end{pmatrix} \wedge z_{IJ} = z_{JI}\,, \tag{31}$$

and also has to provide at least a positive definite matrix (31), which is the same as to consider the following property has to be fulfilled:

$$\mathrm{Det}\,|\mathbf{Z}^{IJ}| = z_{II}z_{JJ} - z_{IJ}^2 > 0 \rightarrow z_{II}z_{JJ} > z_{IJ}^2\,. \tag{32}$$

The restriction (32) can also be written as:

$$z_{JJ} > z_{IJ}^2 z_{II}^{-1}\,, \tag{33}$$

and this will provide a form of the (2×2) positive definite restrictions to be easily related to the general analysis which follows. Therefore, the algorithm (30) has to be modified accordingly incorporating the inequality (32) as a restriction:

$$\forall I > J\colon$$

$$z_{IJ} = \max_{\mathbf{t};\boldsymbol{\Omega}} \int_D \rho_I(\mathbf{r})\rho_J(\mathbf{r}|\mathbf{t};\boldsymbol{\Omega})\,\mathrm{d}\mathbf{r} \wedge z_{IJ}^2 < z_{II}z_{JJ} \tag{34}$$

and one can expect that the general QSM $\mathbf{Z}$, can approach in this way the required complete positive definiteness, although this cannot be completely assured. In fact, this (2×2) restriction constitutes an incomplete point of view, as nothing can be said about the positive definiteness of higher dimensional submatrices of the QSM $\mathbf{Z}$. In this sense, the restricted algorithm (34) is more or less similar to the triangle distance relationship coherence, sought by an already published procedure [91].

The Quantum Similarity Matrix Aufbau Recursive Algorithm. Although one can use the Gershgorin theorem to test the positive definiteness of any QSM, a complete QSM calculation algorithm, based on the generalization of property (33) for (2×2) matrices, in order to assure the QSM $\mathbf{Z}$ positive definiteness, shall be based on an Aufbau procedure; that is: starting from any pair of QO, algorithm (34) is put forward. The result will be a positive definite matrix, $\mathbf{Z_0}$ say, with a structure like the matrix (31) defined above. A simple recursive Aufbau algorithm can be described in order to obtain a final positive definite QSM.

Suppose that for some index $P < N$, a $(P\times P)$ positive definite QSM $\mathbf{Z_0}$ has been obtained, using the QO's sequence: $\{I_K; K = 1, P\}$. One can add a new QO to the Aufbau procedure, the Qth QO, say, in such a way that an augmented QSM, $\mathbf{Z_1}$, is obtained possessing the partitioned structure:

$$\mathbf{Z_1} = \begin{pmatrix} \mathbf{Z_0} & |\mathbf{z}\rangle \\ \langle\mathbf{z}| & \theta \end{pmatrix}\,,$$

with the $(1\times P)$ row vector defined as: $\langle\mathbf{z}| = (z_{I_1Q}; z_{I_2Q}; \ldots; z_{I_PQ})$, and the column vector $|\mathbf{z}\rangle$, being just the transpose of the former; finally, $\theta \equiv z_{QQ}$ is the self-similarity of the added QO.

The sufficient relationship, which can be written here as the set of conditions:

$$\theta > \langle|\mathbf{z}|\rangle \wedge \forall K = 1, P\colon z_{I_KI_K} > \sum_{L\neq K} z_{I_KI_L} + z_{I_KQ}\,, \tag{35}$$

assuring that the augmented matrix $\mathbf{Z_1}$ has a positive definite structure, can be alternatively rewritten via a recursive Cholesky decomposition algorithm, described in several places [17,96].

The necessary and sufficient condition for the positive definiteness of the augmented QSM $\mathbf{Z_1}$ can be stated as:

$$\theta - \langle\mathbf{z}|\mathbf{Z_0^{-1}}|\mathbf{z}\rangle > 0$$

$$\rightarrow z_{QQ} > \sum_K \sum_L z_{I_KQ} z_{I_LQ} Z^{(-1)}_{0;I_KI_L}\,. \tag{36}$$

The Cholesky decomposition condition, which can be called here *Quantum Similarity Aufbau Condition*

(QSAC), means that it cannot be reliable to use a pair of QO's every time that a new element of the QSM has to be computed, but that the added QO density function: $\rho_Q(\mathbf{r}|\mathbf{t};\boldsymbol{\Omega})$ has to be translated-rotated with the same values of the pair: $\{\mathbf{t};\boldsymbol{\Omega}\}$, for every computed element of the vector $|\mathbf{z}\rangle$, connecting recursively the QO Q with all the ones previously employed in constructing the QS submatrix $\mathbf{Z_0}$. When the QS submatrix $\mathbf{Z_0}$ has scalar (1×1) dimension as occurs in the submatrix (31) case, then the QSAC (36) becomes the relationship (33). Moreover, the QSAC condition is a stronger positive definiteness condition than the diagonal dominance, as QSAC becomes the necessary and sufficient condition for constructing a positive definite augmented matrix.

The maximal pair condition (30) can be substituted in the general $(P \times P)$ case, for instance, by maximizing the sum of the whole vector $|\mathbf{z}\rangle$, which due to the positive definiteness of its elements is coincident with the search of a maximal Minkowski norm:

$$\max_{\mathbf{t};\boldsymbol{\Omega}}[\langle|\mathbf{z}\rangle\rangle] = \max_{\mathbf{t};\boldsymbol{\Omega}}\left[\sum_{K=1}^{P}\int_D \rho_{I_K}(\mathbf{r})\rho_Q(\mathbf{r}|\mathbf{t};\boldsymbol{\Omega})\,\mathrm{d}\mathbf{r}\right] = \max_{\mathbf{t};\boldsymbol{\Omega}}\left[\sum_{K=1}^{P} z_{I_K Q}\right]. \tag{37}$$

This can be done admitting the same translation-rotation sequence performed on every term of the vector $|\mathbf{z}\rangle$ in Eq. (37), whenever such transformation increases the Minkowski norm. However, while the maximal value of the sum leading to the Gershgorin radius is searched as in the condition (37) of the previous sentence, the QSAC relationship (36) has to be equally tested and if not fulfilled the pair $\{\mathbf{t};\boldsymbol{\Omega}\}$ rejected. Such a procedure will assure the positive definiteness of the QSM $\mathbf{Z}$ at the final step of the recursion and will provide the same relative position in the calculation of the quantum similarity measures for every recursively added QO.

Geometrical Interpretation of the QSAC. Leaving apart the linear algebra concept of diagonal dominance which similarity matrices usually do not fulfill, the alternative Cholesky decomposition condition property leading to the QSAC, assuring in this manner the positive definite structure of the final QSM form and written as in Eq. (36), has a clear geometrical meaning. The positive definite quadratic form: $\langle \mathbf{z}|\mathbf{Z_0^{-1}}|\mathbf{z}\rangle \in \mathbf{R}^+$, is nothing else than the Euclidean norm of the vector $|\mathbf{z}\rangle$ in the reciprocal metric space defined by the density tags: $\{\rho_{I_K}(\mathbf{r}); K = 1, P\}$. The QO's tags are employed to form the QSM $\mathbf{Z_0}$, which because of the QSAC construction has been structured positive definite and acts accordingly as a metric matrix of a P-dimensional pre-Hilbert space. Since in the quadratic form appearing in Eq. (36), the inverse of the metric $\mathbf{Z_0}$ appears, the implicit Euclidean norm equivalent to the aforementioned quadratic form is computed in the metric reciprocal space with the matrix $\mathbf{Z_0^{-1}}$, acting as a positive definite metric matrix, because: $Sp[\mathbf{Z_0}] \in \mathbf{R}^+ \rightarrow Sp\left[\mathbf{Z_0^{-1}}\right] \in \mathbf{R}^+$. Accordingly, the QSAC forces this Euclidean norm in the reciprocal P-dimensional pre-Hilbert space to be less than the self-similarity of the recursively added Qth QO.

This permits us to associate the described Quantum Similarity Aufbau procedure as an algorithm maximizing the Minkowski norm of each recursive column $|\mathbf{z}\rangle$ of the QSM, submitted to the QSAC restriction which means that its Euclidean norm, computed in the recursive reciprocal pre-Hilbert space, remains less than the recursive QSM diagonal self-similarity elements.

Finally, the following points must be taken into account:

1. Because it is not necessary to start the recursive QSAC with any a priori chosen QO, the final QSM will certainly depend on the QO recursive order chosen. Thus, there are just $N!$ possible choices, each one producing an equally positive definite QSM. However, the ordering imposed by the self-similarity measures can be chosen as a way to reach a systematic QSM Aufbau. That is, if one calls the QSM diagonal self-similarity measures set computed on the QOS elements: $D(\mathbf{Z}) = \{z_{II}\}$, then the obvious choices are defined by the maximal ordering:

$$z_{11} = \max_I[D(\mathbf{Z})] \rightarrow z_{22} = \max_I[D(\mathbf{Z}) - z_{11}] \ldots$$

or by the minimal:

$$z_{11} = \min_I[D(\mathbf{Z})] \rightarrow z_{22} = \min_I[D(\mathbf{Z}) - z_{11}] \ldots$$

This ensures that the QO's will be ordered in decreasing or increasing complexity, while providing a generic reproducible way of computing QSM under QSAC premises.

2. When constructing the QSM according to the proposed Aufbau procedure, it is well known that the overlap quantum similarity measures, as defined in Eq. (29), can be substituted by a more general form involving a positive definite operator: $\Omega(\mathbf{r_1};\mathbf{r_2})$; so, in general, the similarity measures can be described as the integral:

$$z_{IJ}(\Omega) = \iint_D \rho_I(\mathbf{r_1})\Omega(\mathbf{r_1};\mathbf{r_2})\rho_J(\mathbf{r_2})\,\mathrm{d}\mathbf{r_1}\,\mathrm{d}\mathbf{r_2}\,;$$

while the positive definite operator choice ensures that the QSM, when constructed according to the equiva-

lent QSAC, like the one depicted previously for overlap quantum similarity measures in Eq. (36), is positive definite. One just shall make the substitution: $z_{IJ} \leftarrow z_{IJ}(\Omega)$.
3. The QSAC is also valid for quantum similarity measures involving the off-diagonal terms of the density matrix.

Linear Quantum QSPR Fundamental Equation

Expectation Values

In quantum mechanics, the expectation value of some QO observable, associated in turn to some Hermitian operator W, is measured in the usual statistical way [97,98,99], using the tag part ρ_A of the corresponding QO:

$$\langle \pi_A \rangle = \langle W|\rho_A \rangle = \int_D W \rho_A \, \mathrm{d}V \, . \tag{38}$$

At the same time, in general, the operator W can be decomposed as follows:

$$W = Q\Omega \, .$$

Ω bears a positive definite and known form. On the other hand, the operator: Q, can be approximately expressed in terms of an appropriate linear (or multilinear) combination of a known density function set $\{\rho_I\}$, provided with the adequate variable count, in order to match the one of ρ_A, that is:

$$Q \simeq \sum_I w_I \rho_I \, . \tag{39}$$

The structure associated to an operator like: W, permits us to construct expectation values of entangled or complicated observables of submicroscopic systems. Such entangled observables can be considered connected to experimental outcomes, like biological activity, whose Hermitian operator cannot be completely well defined. So, this way to proceed appears quite appropriate for cases where the complexity of the observed phenomenon does not possess a straightforward association with any known or easily describable Hermitian operator. One shall stress the fact that the set up (39) is not appropriate for well-defined observables like kinetic energy, Coulomb energy, dipole and multipole moments,...

Quantum QSPR Fundamental Equation

Substituting the expression of the operator, described by Eq. (39) into expectation value expression (38), one arrives at the following result, related to the corresponding QO:

$$\langle \pi_A \rangle = \langle Q\Omega|\rho_A \rangle = \langle Q|\Omega\rho_A \rangle \simeq \sum_I w_I \langle \rho_I|\Omega\rho_A \rangle = \sum_I w_I z_{IA}(\Omega)$$

which, after supposing that several QO or the whole elements of a QOS are considered, this result can be brought into the matrix form of a linear equation:

$$\mathbf{Zw} \simeq |\pi\rangle \, . \tag{40}$$

Where the column vector $|\pi\rangle$ contains the collection of expectation values of the considered QOS, $\mathbf{Z} = \{z_{IA}(\Omega)\}$ is a quantum similarity matrix and, finally, the column vector $\mathbf{w}$ collects the coefficients by which the operator Q is approximately expressed by means of Eq. (39).

Interpretation and Characteristics of the Quantum QSPR Fundamental Equation

The interpretation of the linear system (40) can proceed as follows. The vector $|\pi\rangle$ can supposedly contain known values of a well-defined, but arbitrarily complicated observable property of the chosen QOS. The quantum similarity matrix $\mathbf{Z}$ can be computed, once the elements of the QOS are supposedly known. The coefficient vector $\mathbf{w}$ has to be determined.

Put in such terms, Eq. (40) has the same well-known structure as the usual classical QSPR problems. However, this fact constitutes a very important and crucial result: Because, it permits us to interpret the columns (or rows) of the quantum similarity matrix, $\mathbf{Z}$, as being the QSM finite-dimensional, discrete, *descriptors* of every QO used in the study. These considerations are sufficient to allow us to name Eq. (40) the *quantum QSPR* (QQSPR) *fundamental equation*.

Characteristics of the QQSPR Fundamental Equation. Unlike the problems present in classical QSPR models, the QQSPR fundamental equation has several characteristics lacking in the former usual equations, these are:

1. *Universal applicability*, because Eq. (40) can be used to model any kind of QOS: nuclei, atoms, molecules ...
2. *Unbiased background* descriptor structures, because the QSM elements, forming the quantum similarity matrix, $\mathbf{Z}$, appearing in Eq. (40), are not arbitrarily chosen by the user, among those belonging to a given descriptor pool, but appear as a consequence of the theory.
3. *Causal character*, as the QQSPR models obtained are the result of solving a well-defined equation, as shown through the set up of Eq. (40), and are deducible

from the general theoretical structure of quantum mechanics.

Quantum Similarity Matrices (QSM) in the Construction of First-Order QSPR Operators and the Definition of Discrete QOS. The first-order approach of the QSPR operator, for the *core set* known molecular property tag set: $\Pi = \{\pi_I\}$ generates the following equation collection:

$$\forall I = 1, n:$$
$$p_I = \pi_I - \langle \sigma[\rho_I] \rangle \approx \sum_J x_J \langle \rho_J[\rho_I] \rangle = \sum_J x_J z_{JI} \,. \tag{41}$$

The set of integrals:

$$\left\{ \langle \rho_J[\rho_I] \rangle = \int_D \rho_J \rho_I \, \mathrm{d}V = z_{JI} \right.$$
$$\left. = z_{IJ} = \int_D \rho_I \rho_J \, \mathrm{d}V = \langle \rho_I[\rho_J] \rangle \right\} ,$$

appearing in Eqs. (197) can be ordered into a $(n \times n)$ symmetric array, constructing in this way the so-called *quantum similarity matrix*: $\mathbf{Z} = \{z_{IJ}\}$ (QSM). In turn, the ordered set of shifted properties: $\{p_I\}$ can form a $(n \times 1)$ column vector: $|\mathbf{p}\rangle = \{p_I\}$. Therefore, the equation set (197) is simply a linear system, which will be discussed next, in order to describe its possible use for evaluating *U-m* unknown molecular properties.

Discrete QOS. Every column of the QSM: $\mathbf{Z} = \{|\mathbf{z}_I\rangle = \{z_{JI}\}\}$, can be interpreted as a discrete matrix representation of each QO density matrix, present within the density function tag set: $\mathrm{P} = \{\rho_I\}$. In this way a one-to-one correspondence can be established between the density tag set and the QSM column submatrices, which can be written as:

$$\forall m_I \in \mathrm{M}: \rho_I \leftrightarrow |\mathbf{z}_I\rangle \Rightarrow \mathrm{P} \Leftrightarrow \mathbf{Z} \,.$$

In other words, the QSM column set can be used as a new *n*-dimensional vector tag set, attached to the molecular set M, in order to build up a new tagged set, namely a *discrete quantum object set*:

$$\mathrm{Q_Z} = \mathrm{M} \times \mathbf{Z} \,. \tag{42}$$

In this DQOS, the density function tags of the original QOS, Q, belonging to the tag set P, are substituted by the columns of the QSM. Therefore, there also exists a one-to-one correspondence between both QOS: $\mathrm{Q} \leftrightarrow \mathrm{Q_Z}$.

The Nature of the QSM Descriptors. In both the quantum similarity matrix **Z**, or its stochastic column transformation **S**, the involved columns forming both matrices possess a very special character, besides the fact that they belong to some VSS of the appropriate dimension.

Starting from the QOS, where each QO is defined by the ordered pair of submicroscopic systems and state density functions:

$$\tau_A = (s_A; \rho_A) \in T = S \times P \,,$$

then, when dealing with the construction of the QSM or the stochastic transformation (28), which one can consider expressed through the decomposition of its columns as: $\mathbf{Z} = \{\mathbf{z}_I\}$ or $\mathbf{S} = \{\mathbf{s}_I\}$, it can be deduced that both matrices induce a new possible QOS, made with discrete *N*-dimensional tags, instead of the infinite dimensional density function ones, namely:

$$\theta_A = (s_A; \mathbf{z}_A) \in \Theta = S \times \mathbf{Z} \,, \tag{43}$$

or one can see the equivalent structure, which can also be considered alternatively as:

$$\sigma_A = (s_A; \mathbf{s}_A) \in \Sigma = S \times \mathbf{S} \,. \tag{44}$$

$$\mathbf{Sw} \simeq \mathbf{p} \,.$$

The *discrete* QOS, represented by the definitions (43) or (44), can be admittedly considered, without doubt, as finite dimensional representations of the original QOS, based on density function tags. This can be so, as both $\mathbf{z}_A$ and $\mathbf{s}_A$ discrete tags, essentially are elements of some VSS. Perhaps the representation (44), with tags belonging to the unit shell, corresponds to the most adequate of such discrete forms and, at the same time, the one which is more connected to the original infinite dimensional unit shell made by the collection of density functions.

Even if the choice to build up the problem is the DQOS, represented by Eq. (44), the fundamental QQSPR Eq. (40) can be transformed conveniently into the new row stochastic system:

$$\mathbf{Sw} \simeq \mathbf{p} \,, \tag{45}$$

simply by multiplying on the left by the inverse of the diagonal matrix (27), and using accordingly the transformed properties vector: $\mathbf{p} = \mathbf{D}^{-1}|\pi\rangle$. The column stochastic transformation can be used straightforwardly too, and so it will not be discussed here anymore, as it has been exhaustively studied within several papers [59,100].

First-Order Fundamental QQSPR (FQQSPR) Equation

The analysis of the QQSPR problem can start with the first order or linear fundamental QQSPR equation, involving the *core set*, formed with the molecules of the associated DQOS, which are also linked with known values of some property.

One can write the first-order QQSPR fundamental equation in a compact matrix form:

$$\mathbf{Z}|\mathbf{x}\rangle = |\mathbf{p}\rangle \,; \tag{46}$$

Where the matrix $\mathbf{Z}$ is the already described symmetric QSM, $|\mathbf{p}\rangle$ is the known *core set* property vector and $|\mathbf{x}\rangle$ is a $(n \times 1)$ vector, whose coefficients have to be evaluated.

The predictive power of such an equation is a priori null, because being the QSM: $\mathbf{Z}$, by construction non-singular (otherwise two density functions will be exactly the same), then there always can be computed a QSM inverse: $\mathbf{Z^{-1}}$, obeying the usual relationships: $\mathbf{Z}^{-1}\mathbf{Z} = \mathbf{Z}\mathbf{Z}^{-1} = \mathbf{I}$, in such a way that the trivial result, defining the unknown coefficient vector:

$$|\mathbf{x}\rangle = \mathbf{Z}^{-1}|\mathbf{p}\rangle \,, \tag{47}$$

will be always obtained within a *core set* scenario. Furthermore, one can retrieve the exact value of the property for any molecule of the *core set* QOS choosing the scalar products:

$$\forall I\colon\, p_I = \langle \mathbf{z}_I|\mathbf{x}\rangle \,. \tag{48}$$

The QSM for diverse *core sets* has been used in a quite large set of prediction studies, in every case employing up-to-date statistical tools, the usual procedures currently available in classical QSPR studies. In the present study, the reader can find in the following sections new theoretical developments of the fundamental QQSPR equation prediction ability. However, a reminder of some simple linear algebra for the FQQSPR equation is needed first in order to understand the following arguments; therefore this will be described in the following sections.

Partitioning the FQQSPR Equation and the QSM Inverse. Supposing now one can organize the QSM in the fundamental Eq. (40) in such a way that the last column and row correspond to a *U-m*, then, the unknown property element will be supposedly stored in the last position, the $(n+1)$th, of the vector $|\mathbf{p}\rangle$ and will be symbolized by an a priori undefined parameter: π. With this in mind, one can design a partition of the QSM and the entire FQQSPR Eq. (40) in the following way:

$$\begin{pmatrix} \mathbf{Z_0} & |\mathbf{z}\rangle \\ \langle \mathbf{z}| & \theta \end{pmatrix} \begin{pmatrix} |\mathbf{x_0}\rangle \\ x \end{pmatrix} = \begin{pmatrix} |\mathbf{p_0}\rangle \\ \pi \end{pmatrix} \,. \tag{49}$$

Where, the $(n \times 1)$ column vector $|\mathbf{z}\rangle$ corresponds to the representation of the *U-m* in terms of the density tags of the *core set* and θ is the *U-m* self-similarity measure, which according to the simplified formalism of the expectation values can be defined by means of a simple overlap quantum similarity measure, as the Euclidean norm:

$$\theta = \int_D \rho_U^2(\mathbf{r})\, \mathrm{d}\mathbf{r} \,.$$

One can find the solution of the partitioned linear system by using the following symbols for the partitioned QSM inverse:

$$\mathbf{Z}^{-1} = \begin{pmatrix} \mathbf{Z_0^{(-1)}} & |\mathbf{z}^{(-1)}\rangle \\ \langle \mathbf{z}^{(-1)}| & \theta^{(-1)} \end{pmatrix} \,, \tag{50}$$

and one can evaluate the inverse elements for partitioned QSM matrices in the usual way.

Remarks on the Structure of the Fundamental QQSPR Equation

The following remarks relate to the result given by the fundamental QQSPR equation that was discussed in the previous section. Each of these remarks poses new problems that will be studied separately in subsequent sections.

Symmetrical Similarity Matrices. In the first place, it must be said that the fundamental QQSAR equation has been usually presented in previous literature within the particular case where the basis and probe molecular quantum object tagged sets coincide, providing a square symmetric similarity matrix, and thus the equality: $\mathbf{A} = \mathbf{Z}$, between the involved similarity matrices holds. This choice has the drawback that the fundamental QQSPR linear system becomes well defined, with a unique solution, whenever the similarity matrix is non-singular, which shall be the usual case, as far as no quantum object coincides with another within the quantum object set.

But even then, there is quite a wide range of solutions to overcome this apparent limitation. Among other procedures, one can use the symmetric similarity matrix as a source of molecular descriptors and afterwards employ them in classical statistical treatments. This choice, as was already commented, has been studied in many publications of our laboratory with success. In the same way, the similarity matrix can be transformed into a column or row stochastic matrix and, as a consequence, this form suggests several possibilities, which still are far from being exploited. Some analysis of the stochastic issue will be developed in a forthcoming section of this paper.

Origin of Hansch QSAR Models. An interesting possibility of the symmetric square representation of the quantum similarity matrices corresponds to its potential to unveil the origin of one parameter classical QSAR models, such as those Hansch described some years ago. Indeed, under the equivalence of both the basis B and probe P quantum object sets, the FQQSPR linear equation corresponds to a set of N equations with the same number of unknowns, and can be rewritten as:

$$\forall J = 1, N\colon p_J \simeq \sum_{I=1}^{N} \omega_I z_{IJ} = \omega_J z_{JJ} + \sum_{I \neq J}^{N} \omega_I z_{IJ} \,, \quad (51)$$

where there is no need to attach the similarity matrix elements to any specific quantum object set, as all of them are computed over a unique basis of density function tags. Considering the two terms at the end of the previous equation, it can be seen that the first one, with a diagonal value of the similarity matrix, is attached to a self-similarity measure z_{JJ}, while the second term in cases of a not so strongly varying family of quantum objects, can be considered almost a constant, that is using:

$$\forall J\colon \alpha = \omega_J \wedge \beta \simeq \sum_{I \neq J}^{N} \omega_I z_{IJ} \,, \quad (52)$$

the above equation takes the final form:

$$\forall J = 1, N\colon p_J \simeq \alpha z_{JJ} + \beta \,, \quad (53)$$

which has the required appearance to be considered as possessing a Hansch structure.

Besides this last deduction, it must be said that self-similarity measures of different kinds have been used to test this simple linear equation with quite a large series of quantum objects, yielding usually good results. Self-similarities can be sound substitutes of the classical Hansch analysis parameters. They constitute for co-generic molecular sets molecular descriptors with the property to be directly attached to a tri-dimensional molecular structure. Self-similarity measures vary slowly with conformational changes, so their values for the optimal geometry can be safely used, knowing that the magnitude of the descriptor will differ not very much from the one which is attached to the active conformation associated to the observable property.

Solutions of the Linear QQSPR Fundamental Equation

Equation (40) can be solved choosing the customary methods, as in the usual algorithms employed within the QSPR field [20]. They have been manipulated in this fashion, since its first description on a great deal of cases, as well as for a large variety of problems and subjects [26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,101]. This means that in all these studies the QQSPR fundamental equation has been solved with some algorithm, based on the least squares or similar technique [102,103,104,105,106,107,108,109,110,111,112,113,114,115,116,117,118,119,120,121,122,123,124,125,126,127,128,129,130,131,132,133,134,135,136,137,138,139,140,141,142,143,144,145,146].

However, the characteristic features of the QQSPR fundamental equation, and its definition within the VSS formalism, show that several alternative possibilities can be described, which will be studied next. Such non-classical solutions have in turn provided a collection of many new properties, concepts and application examples related to the tagged sets, VSS and IMP definitions.

Therefore, in order to exploit the QQSPR a plausible alternative to the principal components analysis will be proposed and then, the use of IMP and other techniques to obtain approximate solutions of the QQSPR fundamental equations will also be discussed.

Similarity Matrix Eigenvectors as Basis Sets to Construct the Solutions of the QQSPR Fundamental Equations. Suppose we set the fundamental QQSPR Eq. (40) for a given problem. The secular equation of the involved quantum similarity matrix, $\mathbf{Z}$, can be written as:

$$\mathbf{ZC} = \mathbf{C}\Theta \,, \quad (54)$$

where: $\mathbf{C} = (\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_N)$ is the matrix collection of the eigenvectors of the similarity matrix and: $\Theta = \mathrm{Diag}(\theta_1; \theta_2; \ldots; \theta_N)$ is a diagonal matrix made by the ordered eigenvalues in descending order. The eigenvector matrix can be considered orthogonal, that is, the following property holds: $\mathbf{CC}^{\mathrm{T}} = \mathbf{C}^{\mathrm{T}}\mathbf{C} = \mathbf{I}_N$, with the symbol: $\mathbf{C}^{\mathrm{T}}$ indicating matrix transposition. The eigenvector associated to the greater eigenvalue has their entire elements positive definite, according to the Perron–Frobenius theorems [147]. The spectral decomposition:

$$\mathbf{Z} = \mathbf{C}\Theta\mathbf{C}^{\mathrm{T}} = \sum_{I} \theta_I \mathbf{c}_I \mathbf{c}_I^{\mathrm{T}} \quad (55)$$

can be used in Eq. (40), to obtain, after straightforward rearrangements, the equation:

$$\sum_{I} \theta_I (\mathbf{c}_I^{\mathrm{T}} \mathbf{w}) \mathbf{c}_I \simeq |\pi\rangle \,, \quad (56)$$

so, renaming the set of scalars in Eq. (56) as:

$$\gamma_I = \theta_I (\mathbf{c}_I^{\mathrm{T}} \mathbf{w}) \rightarrow |\gamma\rangle = \{\gamma_I\} \,,$$

then, the new equation could be written by means of a linear combination of the similarity matrix eigenvectors:

$$\sum_I \gamma_I \mathbf{c}_I = \mathbf{C}|\gamma\rangle \simeq |\pi\rangle \,. \tag{57}$$

Equation (57) permits us to compute the new coefficients $|\gamma\rangle$ in an obvious way, by using the orthogonal nature of the eigenvector matrix:

$$|\gamma\rangle \simeq \mathbf{C}^{\mathrm{T}}|\pi\rangle \,.$$

In fact, the original fundamental QQSPR equation coefficient vector can be obtained taking into account the spectral decomposition of the similarity matrix inverse, that is:

$$\begin{aligned}\mathbf{w} \simeq \mathbf{C}\Theta^{-1}\mathbf{C}^{\mathrm{T}}|\pi\rangle &= \sum_I \theta_I^{-1}\mathbf{c}_I\mathbf{c}_I^{\mathrm{T}}|\pi\rangle \\ &= \sum_I \left[\theta_I^{-1}\left(\mathbf{c}_I^{\mathrm{T}}|\pi\rangle\right)\right]\mathbf{c}_I = \sum_I \omega_I \mathbf{c}_I \,. \end{aligned} \tag{58}$$

Therefore, Eq. (58) indicates that the coefficient vector, the solution of the previous Eq. (40), may be expressed by a linear combination of the eigenvectors of the similarity matrix too. Then, to every eigenvector, $\mathbf{c}_I$, there is associated a well-defined scalar coefficient, ω_I, which may be used as a reordering rule in order to obtain approximate solutions of the fundamental QQSPR equation. That is, suppose the eigenvectors are now ordered by the decreasing values of the set: $|\omega\rangle = \{\omega_I\}$, then one can write:

$$\begin{aligned}\mathbf{w} &\simeq \sum_I \delta(\omega_I > \varepsilon)\omega_I\mathbf{c}_I + \sum_J \delta(\omega_J \le \varepsilon)\omega_J\mathbf{c}_J \\ &= \mathbf{w}_{\mathrm{a}} + \mathbf{w}_{\mathrm{error}} \,,\end{aligned}$$

where ε is a given threshold splitting the vector construction in an approximate vector, $\mathbf{w}_{\mathrm{a}}$, and the remaining one, $\mathbf{w}_{\mathrm{error}}$, which can be interpreted or used as a residual error vector.

Stochastic Matrix Eigenvectors as Basis Sets to Construct the Solutions of the Fundamental QQSPR Equations

Similar treatments can be designed for the QQSPR fundamental equation of the stochastic matrices, like the one employed for Eq. (50). The problem is that the stochastic matrix **S** is no longer symmetric and the attached eigensystem, apparently appears to be more laboriously solved, than in the case of the symmetric quantum similarity matrix **Z**. The problem has been already discussed in a general manner [59], so here only some simplified discussion will be given.

Suppose the secular equation, attached to the stochastic matrix **S** is written as:

$$\mathbf{SX} = \mathbf{X}\Sigma \tag{59}$$

substituting the matrix **S** by the expression of the row stochastic transformation as in Eq. (50), then:

$$\mathbf{D}^{-1}\mathbf{ZX} = \mathbf{X}\Sigma \,,$$

which can readily be transformed by simple matrix manipulations and by using the square root of the diagonal matrix **D**, into the new secular equation:

$$\mathbf{D}^{-\frac{1}{2}}\mathbf{ZD}^{-\frac{1}{2}}\mathbf{D}^{\frac{1}{2}}\mathbf{X} = \mathbf{X}\Sigma \,,$$

which from here, calling: $\mathbf{A} = \mathbf{D}^{-\frac{1}{2}}\mathbf{ZD}^{-\frac{1}{2}}$ and $\mathbf{Y} = \mathbf{D}^{\frac{1}{2}}\mathbf{X}$, a new equivalent secular equation is readily made:

$$\mathbf{AY} = \mathbf{Y}\Sigma \,. \tag{60}$$

Equation (60) has the advantage that the matrix A is symmetric, hence the eigenvector matrix Y appears to be orthogonal: $\mathbf{Y}^{\mathrm{T}}\mathbf{Y} = \mathbf{YY}^{\mathrm{T}} = \mathbf{I}$. Then, the sought eigenvectors of the stochastic matrix can be obtained by using the relationship between **Y**and **X**, that is:

$$\mathbf{X} = \mathbf{D}^{-\frac{1}{2}}\mathbf{Y} \,.$$

The matrix **D**, in these circumstances acts as a metric with respect to the eigenvectors of **S**, as one can see that the orthogonality relationships:

$$\mathbf{X}^{\mathrm{T}}\mathbf{DX} = \mathbf{XDX}^{\mathrm{T}} = \mathbf{I} \,,$$

hold, due to the orthogonality of the eigenvector matrix **Y**.

This proves finally that, in any case, Eq. (50) can be solved in the same way as previously commented for Eq. (40), simply by using the appropriate spectral decomposition:

$$\mathbf{S} = \mathbf{X}\Sigma\mathbf{X}^{\mathrm{T}} = \sum_I \sigma_I \mathbf{x}_I\mathbf{x}_I^{\mathrm{T}} \,.$$

Similarity Matrix IMP Decomposition in Order to Construct Approximate Solutions of the QQSPR Fundamental Equations

QQSPR Fundamental Equation over VSS. Going back to the QQSPR fundamental Eq. (40), now one can consider the positive definite nature of the elements, which appear to build the quantum similarity matrix. This can be expressed by means of the symbol: $\mathbf{Z}* > 0$[2]. The structure

[2]By the symbol: $\mathbf{A}* > 0$, applied to an arbitrary matrix $\mathbf{A} = \{a_{ij}\}$, is meant that $\forall i, j\colon a_{ij} > 0$.

of the property vectors can also be taken into account. In the case when the following characteristic also holds: $|\pi\rangle * > 0$, for the involved QOS property or activity vector, then Eq. (40) can be associated to some linear transformation occurring on a VSS, for it can be written:

$$\mathbf{Zw} = \sum_I w_I \mathbf{z}_I \simeq |\pi\rangle \,, \tag{61}$$

showing that a linear combination of vectors, belonging to a VSS, has to be brought into another vector belonging to a VSS.

This situation could only be *generally* achieved by using the condition: $\mathbf{w}* > 0$, which means that the linear system solution also shall belong to the VSS. Equ. (61) above becomes a constrained linear system of equations, since one is seeking solutions for which:

$$\mathbf{Zw} \simeq |\pi\rangle \wedge \mathbf{w}* > \mathbf{0} \,.$$

Approximate Restricted Solutions of Fundamental QQSPR Equation in VSS. An approximate solution of the QQSPR fundamental equation can be obtained in the following way. As all the involved columns of the problem belong to a VSS, they can be decomposed by means of an IMP in terms of some IMP square powers of real matrices. If the treatment has to be more general the squared module of some complex matrices can be alternatively employed, but the treatment becomes slightly more difficult and the needed set of symbols heavier, so just real matrices will be supposed in this discussion. Therefore, owing to these considerations one can write:

$$\mathbf{Z} = \mathbf{A} * \mathbf{A} \wedge \mathbf{w} = \mathbf{x} * \mathbf{x} \wedge |\pi\rangle = \mathbf{p} * \mathbf{p} \,.$$

So Eq. (61) can now be rewritten as:

$$(\mathbf{A} * \mathbf{A})(\mathbf{x} * \mathbf{x}) \simeq \mathbf{p} * \mathbf{p} \,, \tag{62}$$

suggesting an alternative approximate equation, which may be written in the following terms:

$$(\mathbf{Ax}) * (\mathbf{Ax}) \simeq \mathbf{p} * \mathbf{p} \,, \tag{63}$$

which has been obtained in turn, simply using the plausible approximation:

$$(\mathbf{A} * \mathbf{A})(\mathbf{x} * \mathbf{x}) \approx (\mathbf{Ax}) * (\mathbf{Ax}) \,.$$

However, the approximate Eq. (63), suggests that the new linear system:

$$\mathbf{Ax} \simeq \mathbf{p} \tag{64}$$

can now be solved, as it does not have to be submitted to any restriction at all, then:

$$\mathbf{x} \simeq \mathbf{A}^{-1}\mathbf{p} \,,$$

and finally, the approximate solutions of the original system can be written as:

$${}^{a}\mathbf{w} = \mathbf{x} * \mathbf{x} \simeq (\mathbf{A}^{-1}\mathbf{p}) * (\mathbf{A}^{-1}\mathbf{p}) \,,$$

however this is sufficient to ensure:

$$\mathbf{w} \approx {}^{a}\mathbf{w}* > 0 \,.$$

The only problem, which now arises, is the existence of an inverse of the IMP square root of a non-singular matrix. Since the system (64) furnishes approximate solutions to the original problem (61) has to be found, there will be no major problem then to use approximate solutions in the least squares sense of the Eq. (64), as a way to obtain the approximate solution of the QQSPR fundamental equation, restricted to belonging to a VSS.

Convex Conditions Imposed on the Solution Vector of the QQSPR Fundamental Equations

Generating Vector Considerations. The associated problem, to a form like Eq. (61), can also be solved, for instance, as in the well-known ASA fitting procedure [75,76,77,78,79,80,81,82,83]. That is, by using a convex condition on the solution vector: $K(\mathbf{w})$, with the additional meaning that the solution is now forced to belong not only to a VSS, but also to the unit shell.

If the solution of the linear Eq. (61) has to be found as an element of a VSS, $\mathbf{w}* > 0$, necessarily it has to be expressible as an IMP power of some generating real vector like: $\mathbf{w} = \mathbf{x} * \mathbf{x}$. In choosing the convex conditions over the solution: $K(\mathbf{w})$, then the additional restriction is admitted to hold too:

$$\langle \mathbf{w} \rangle = \langle \mathbf{x} * \mathbf{x} \rangle = 1 \,. \tag{65}$$

However, this becomes the same as to consider that the generating vector $R(\mathbf{x} \rightarrow \mathbf{w})$ is normalized. Orthogonal transformations on the generating vector leaves the vector norm invariant, that is: whenever the condition (65) holds, and an orthogonal transformation $\mathbf{U}$ is performed on the generating vector, still the generating rule and the associated convex conditions apply: $R(\mathbf{Ux} \rightarrow {}^{U}\mathbf{w}) \wedge K({}^{U}\mathbf{w})$. Such an idea has been applied to obtain the ASA approximate density functions, using *elementary Jacobi rotations* [148] as a source of orthogonal transformations.

Stochastic Transformations. Still more interesting appears the structure of the fundamental QQSPR equation, when the stochastic transform of the similarity matrix is considered. Equation (40) can thus be multiplied by the inverse of the diagonal matrix **D** on the right as defined in (27), providing:

$$\mathbf{Z}\mathbf{D}^{-1}\mathbf{D}\mathbf{w} = |\pi\rangle\,,$$

which can be transformed into:

$$\mathbf{S}\mathbf{v} = |\pi\rangle\,, \tag{66}$$

whenever it is considered that the following equalities hold:

$$\mathbf{S} = \mathbf{Z}\mathbf{D}^{-1} \wedge \mathbf{v} = \mathbf{D}\mathbf{w}\,. \tag{67}$$

Equation (66) can be also written as a linear combination of the columns of the stochastic matrix $\mathbf{S} = \{s_I\}$:

$$\sum_I v_I \mathbf{s}_I = |\pi\rangle\,.$$

Therefore, this is the same considering the σ-shells of the vectors **v** and $|\pi\rangle$ as being almost the same, as:

$$\langle|\pi\rangle\rangle = \left\langle \sum_I v_I \mathbf{s}_I \right\rangle = \sum_I v_I \langle \mathbf{s}_I\rangle = \sum_I v_I = \langle \mathbf{v}\rangle\,.$$

Thus, if the vector $|\pi\rangle$ is transformed so as to become a unit shell element, this will be completely equivalent to applying the same transformation into the transformed unknown vector **v**. Therefore, the following implications are straightforwardly deduced:

$$|\pi\rangle \in S(1) \rightarrow \mathbf{v} \in S(1) \rightarrow K(\mathbf{v})\,.$$

Demonstrating that in the stochastic QQSPR fundamental Eq. (66) case, the solution contained within a given VSS amounts to the same as obtaining a convex combination of the stochastic matrix columns.

Stochastic QQSPR Least Squares Solution via Jacobi Rotations. After all the previous discussions, there appears to be another possibility which has remained unexplored. Starting from the transformed Eq. (66), with the appropriate definitions (67) in mind, one can seek an approximate solution of the stochastic equation in the least squares sense, defining the quadratic error function by means of the difference vector:

$$|\delta\rangle = \mathbf{S}\mathbf{v} - |\pi\rangle\,,$$

whose Euclidean norm furnishes the quadratic error function, expressible in terms of a scalar product or the inward product sum:

$$\begin{aligned}\varepsilon^{(2)} = \langle\delta|\delta\rangle &= (\mathbf{S}\mathbf{v} - |\pi\rangle)^{\mathrm{T}}(\mathbf{S}\mathbf{v} - |\pi\rangle) \\ &= \langle(\mathbf{S}\mathbf{v} - |\pi\rangle) * (\mathbf{S}\mathbf{v} - |\pi\rangle)\rangle\,.\end{aligned}$$

One easily arrives at the quadratic form:

$$\varepsilon^{(2)}(\mathbf{v}) = \mathbf{v}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{S}\mathbf{v} - \mathbf{v}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}|\pi\rangle - \langle\pi|\mathbf{S}\mathbf{v} + \langle\pi|\pi\rangle\,, \tag{68}$$

however, the quadratic error function optimization has to be carried out preserving the condition of convexity $K(\mathbf{v})$ on the solution vector, otherwise one risks obtaining solutions that do not belong to the unit shell. In order to obtain an appropriate algorithm to perform this task, the following considerations can be taken into account.

The vector v can be expressed as an IMP of an auxiliary vector a, that is:

$$\mathbf{v} = \mathbf{a} * \mathbf{a} \rightarrow \forall I\colon \mathrm{v}_I = a_I^2\,, \tag{69}$$

then the belonging of v to the unit shell is equivalent to the Euclidean normalization of a:

$$\langle \mathbf{v}\rangle = \sum_I \mathrm{v}_I = \sum_I a_I^2 = \langle \mathbf{a}|\mathbf{a}\rangle\,.$$

The quadratic error function (68) can be expressed in terms of the auxiliary vector:

$$\varepsilon^{(2)}(\mathbf{a}) = \mathbf{a}^{[2]\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{S}\mathbf{a}^{[2]} - {}^{\mathrm{T}}\mathbf{a}^{[2]\mathrm{T}}\mathbf{S}^{\mathrm{T}}|\pi\rangle - \langle\pi|\mathbf{S}\mathbf{a}^{[2]} + \langle\pi|\pi\rangle\,, \tag{70}$$

where the symbol: $\mathbf{a}^{[2]} = \mathbf{v} = \mathbf{a} * \mathbf{a}$ has been used. Also, employing to simplify the notation the following conventional symbols:

$$\begin{aligned}\mathbf{H} = \mathbf{S}^{\mathrm{T}}\mathbf{S} &= \{H_{IJ}\} \\ \wedge\ \mathbf{h} = \mathbf{S}^{\mathrm{T}}|\pi\rangle &= \{h_I\} \wedge \mathbf{h}^{\mathrm{T}} = \langle\pi|\mathbf{S} = \{h_I\} \\ \wedge\ \eta &= \langle\pi|\pi\rangle\,,\end{aligned}$$

so Eq. (70) can be explicitly written as:

$$\varepsilon^{(2)}(\mathbf{a}) = \sum_I \sum_J H_{IJ} a_I^2 a_J^2 - 2\sum_I h_I a_I^2 + \eta\,. \tag{71}$$

Starting with an approximate normalized auxiliary vector, orthogonal transformations can be performed, preserving the norm, thus keeping the condition $K(\mathbf{v}) = K(\mathbf{a}^{[2]})$ constant along the optimization of Eq. (71). Orthogonal transformations can be chosen as elementary Jacobi rotations [148], which at every application over the vector a, change two chosen elements $\{a_P; a_Q\}$ into a pair of new

rotated ones $\{a_P^R; a_Q^R\}$, according to the well-known algorithm:

$$\begin{aligned} a_P^R &\leftarrow c a_P - s a_Q \\ a_Q^R &\leftarrow s a_P + c a_Q \,, \end{aligned} \tag{72}$$

where $\{c, s\}$ are the cosine and the sine of the rotation, with the additional obvious relationship: $c^2 + s^2 = 1$.

Over the generating vector coefficients in Eq. (69) it is easy to apply the EJR represented by Eq. (72), and then, the variation of the quadratic error $\delta\varepsilon^{(2)}$, with respect to the active pair of elements $\{a_P; a_Q\}$ may be easily expressed.

Taking also into account that the quadratic elements, for example, will transform and yield variations like:

$$\begin{aligned} \delta a_P^2 &\rightarrow s^2(a_Q^2 - a_P^2) - 2csa_P a_Q \\ \delta(a_P a_Q) &\rightarrow cs(a_P^2 - a_Q^2) - 2s^2 a_P a_Q \\ \delta a_Q^2 &\rightarrow s^2(a_Q^2 - a_P^2) + 2csa_P a_Q = -\delta a_P^2 \,. \end{aligned}$$

A Jacobi rotation as shown in the expression (72) will produce a variation in the quadratic error (71) in the chosen rotated elements, when taking also into account the symmetric nature of the matrix H, of the form:

$$\delta\varepsilon^{(2)}(\mathbf{a}) = H_{PP}(\delta a_P^2)^2 + H_{QQ}(\delta a_Q^2)^2 + 2H_{PQ}\delta a_P^2 \delta a_Q^2 + 2\sum_{I\neq P,Q} a_I^2 (H_{IP}\delta a_P^2 + H_{IQ}\delta a_Q^2)$$

so, also needed are the quartic variations of the auxiliary vector elements, which can be easily computed as in the second-order case.

Substituting such variations into the corresponding equation and collecting terms one finally arrives at a quartic polynomial on the rotation sine:

$$\delta\varepsilon^{(2)} = E_{04}s^4 + E_{13}cs^3 + E_{02}s^2 + E_{11}cs \,, \tag{73}$$

where the parameters $\{E_{IJ}\}$, appearing in Eq. (73), are described as follows:

$$\begin{aligned} E_{04} &= \Theta\left[\left(a_P^2 - a_Q^2\right)^2 - 4a_P^2 a_Q^2\right] \\ E_{13} &= 4\Theta\left(a_P^2 - a_Q^2\right)a_P a_Q \\ E_{02} &= 4\Theta a_P^2 a_Q^2 - 2\left(a_P^2 - a_Q^2\right)G \\ E_{11} &= -4a_P a_Q G \end{aligned}$$

using the following auxiliary terms:

$$\Theta = H_{PP} + H_{QQ} - 2H_{PQ} \,,$$

and

$$G = \sum_{I\neq P,Q} a_I^2 (H_{IP} - H_{QI}) + a_P^2 H_{PP} - a_Q^2 H_{QQ} - \left(a_P^2 - a_Q^2\right)H_{PQ} - h_P + h_Q \,.$$

The optimal sine can be chosen with the null gradient condition $\mathrm{d}\delta\varepsilon^{(2)}/\mathrm{d}s = 0$, taking into account that: $s/c = t$ and that: $\mathrm{d}c/\mathrm{d}s = -t$, then:

$$\frac{\mathrm{d}\delta\varepsilon^{(2)}}{\mathrm{d}s} = -c\left(T_1 t^2 - 2T_2 t - T_3\right) = 0 \,, \tag{74}$$

holds with the auxiliary definitions:

$$\begin{aligned} T_1 &= E_{13}s^2 + E_{11} \\ T_2 &= 2E_{04}s^2 + E_{02} \\ T_3 &= 3E_{13}s^2 + E_{11} \,. \end{aligned}$$

The best Jacobi rotation angle is found solving the quadratic polynomial equation in the EJR tangent {t}, appearing in expression (74). The optimization can be conducted through an iterative procedure, until the global variation of Jacobi rotation angles or the quadratic error integral function become negligible. The interested reader is conducted to the references [78,79,80] for more details, where a complete account of all the Jacobi rotation techniques can be found. A simplified algorithm can be also used and it will be briefly commented upon here. The procedure is based on the fact that sine and cosine can be written in function of the rotation angle:

$$\begin{aligned} s &= \alpha - \frac{1}{6}\alpha^3 + O(5) \\ c &= 1 - \frac{1}{2}\alpha^2 + O(4) \,, \end{aligned}$$

and for small angles it is only necessary to use, up to second order:

$$s \simeq \alpha \wedge c \simeq 1 - \alpha \,,$$

so Eq. (73) can be transformed into a second-order polynomial in the rotation angle:

$$\delta\varepsilon^{(2)} \simeq (E_{02} - E_{11})\alpha^2 + E_{11}\alpha \,,$$

which submitted to the extremum conditions yields:

$$\alpha \simeq \frac{1}{2}\left(1 - E_{02}E_{11}^{-1}\right)^{-1} \simeq s \,.$$

Non-Linear Terms and Extended Wave Functions

Sobolev Spaces

From early times, quantum mechanics has been emphasizing not only the role of well-behaved wave functions, but also the relevance of their gradients and Laplacian forms. The reason for such requirements, necessarily holding on

the current wave functions, can be simply connected to the presence, in the Schrödinger equation set up, of a second-order derivative, the result of a Laplace operator application, associated to the quantum system kinetic energy term, see for example the references [28,149].

Usually, the adequate quantum mechanical behavior of the wave function is focused, among other simple and obvious mathematical features, to the *compulsive* property that wave functions have to be square summable. In some reference books such a property has been promoted to the category of a postulate [150] and in the very early development times of quantum mechanics [98] has been interpreted by Born as the fact that the square module of the wave function can be associated to a probability density function. It was von Neumann [97], who related such properties, among other crucial quantum mechanical theoretical elements, with the mathematical structures of Hilbert–Banach spaces [94,151,152]. More recently, Landau and Lifshitz [153] described the role of the wave function gradient as a descriptor of infinitesimal translations and rotations. These authors settled as well the use of the square module of the wave functions gradient, in order to obtain an alternative kinetic energy expectation value expression, more likely related to the statistical formalism than the Laplace operator form.

On the other hand no utilization has been reported of the so-called *Sobolev spaces* [154] in applied quantum mechanics, at least to our knowledge. Curiously enough, Sobolev spaces where defined as early as 1938, and apparently have been of practical use in some remotely related theoretical landscape, associated to generalized relativity applications [155]. It was not until recently that Sobolev spaces were proposed by us as a vehicle to take into account the role of the wave function squared module:$|\Psi|^2$, as well as to make *simultaneously relevant* the presence of the wave function gradient squared module:$|\nabla\Psi|^2$ in an extended density function. In all this previous work both terms were also presented as forming part of a new quantum mechanical composite norm. In this way, the classical quantum mechanical Banach space has been transformed into a Sobolev space structure [14,156,157], without losing generality, but gaining flexibility instead.

Sobolev spaces can be defined in several ways, leading all of them to simple forms, ready to be used in reinterpreting the approximate solution of the Schrödinger equation and prone to be included with immediate applications, such as those found among the references [156,157]. They can be constructed in such a way as possessing extended forms even more complex, in order to be used to include arbitrary non-linear terms in the same equation [14]. This can be understood by recognizing the fact that the Banach space can be considered the limiting simplified form of a quite large collection of Sobolev spaces.

Quantum Mechanical Hilbert and Banach Spaces. In order to present the Sobolev spaces step by step, the simplest formalism will be defined first, and other extended possibilities will be described later on. To achieve this objective, suppose a quantum mechanical wave function Hilbert space, which can formally be described as:

$$H_\infty(\mathbf{C}) = \{\Psi(\mathbf{r})|\mathbf{r} \in V_P(\mathbf{R}) \wedge \Psi(\mathbf{r}) \in \mathbf{C}\}\,, \tag{75}$$

where the symbol $\mathbf{r}$, used as the wave function variables, shall be considered as a vector, containing all the necessary particle position coordinates as its components. The number of particles is shortly noted with the dimension P of the coordinates vector space. The wave function elements of the Hilbert space (75), possess as a sine qua non condition, the following well-known property about the existence of a positive definite density function, which is remembered here, just to present the notation that will be hereafter employed:

$$\forall\Psi(\mathbf{r}) \in H_\infty(\mathbf{C}) \rightarrow \exists\rho(\mathbf{r}) = |\Psi(\mathbf{r})|^2 \in H_\infty(\mathbf{R}^+)\,. \tag{76}$$

Besides, the density function attached to every wave function, as proposed in Eq. (76), can be seen as belonging to a Hilbert *semispace*, $H_\infty(\mathbf{R}^+)$, where all function values and coefficients are strictly allowed to be positive real numbers only. That is: in the same way as in the Hilbert space (75), one can write the corresponding formal definition for the density function semispace:

$$H_\infty(\mathbf{R}^+) = \left\{\rho(\mathbf{r})|\mathbf{r} \in V_P(\mathbf{R}) \wedge \rho(\mathbf{r}) \in \mathbf{R}^+\right\}, \tag{77}$$

where the dimension of the coordinates vector space, $V_P(\mathbf{R})$, containing the density function variables, has the same meaning as in definition (75). Moreover, the Hilbert space (75) is a Banach space, as all of their elements shall compulsively fulfill the normalization condition:

$$\forall\Psi(\mathbf{r}) \in H_\infty(\mathbf{C}) \rightarrow \int_D |\Psi(\mathbf{r})|^2\,\mathrm{d}\mathbf{r} = 1\,, \tag{78}$$

which obviously amounts to the same as imposing on every element of the Hilbert semispace (77), a *convexity* condition:

$$\forall\rho(\mathbf{r}) \in H_\infty(\mathbf{R}^+) \rightarrow \int_D \rho(\mathbf{r})\,\mathrm{d}\mathbf{r} = 1\,. \tag{79}$$

Gradient of the Wave Function. As was previously commented, the whole quantum mechanical Hilbert space elements shall present other existence properties, mainly related to their derivatives, for instance:

$$\forall\Psi(\mathbf{r}) \in H_\infty(\mathbf{C}) \rightarrow \exists\nabla\Psi(\mathbf{r})\,, \tag{80}$$

where the *nabla* operator ∇ refers to the gradient with respect to the vector coordinates $\mathbf{r}$. That is, formally it can be also written: $\nabla\Psi \equiv \partial\Psi/\partial\mathbf{r}$. The ordering of the resultant gradient vector elements can be somewhat arbitrary; this means that they can be adapted to the structure of the operating mathematical context. In addition, associated to this mentioned component ordering, the resultant gradient vector can be considered to belong to some appropriate Cartesian product of the initial Hilbert space (75), which can be generally defined and noted in a simplified fashion as:

$$H_{\infty}^{(P)}(\mathbf{C}) = \overset{P}{\underset{I=1}{\times}} H_{\infty}(\mathbf{C})\,, \tag{81}$$

because in any ordering case, the resultant gradient vectors will depend on the particle number P. At the same time, the gradients of type (80), can also be easily associated to squared gradient modules, which shall belong to some Hilbert semispace, very similar to the one defined in Eq. (77):

$$\forall\nabla\Psi(\mathbf{r}) \in H_{\infty}^{(P)}(\mathbf{C}) \rightarrow \exists\kappa(\mathbf{r}) = |\nabla\Psi(\mathbf{r})|^2 \in H_{\infty}(\mathbf{R}^{+})\,, \tag{82}$$

where the positive definite function $\kappa(\mathbf{r})$, will produce, when integrated, twice the quantum mechanical kinetic energy expectation value $\langle K\rangle$ of the attached system:

$$\int_D \kappa(\mathbf{r})\,\mathrm{d}\mathbf{r} = \int_D |\nabla\Psi(\mathbf{r})|^2\,\mathrm{d}\mathbf{r} = 2\langle K\rangle\,, \tag{83}$$

which as is well known, can be described alternatively like the classical quantum mechanical expectation value of the Laplace operator:

$$2\langle K\rangle = -\int_D \Psi^{*}(\mathbf{r})\nabla^2\Psi(\mathbf{r})\,\mathrm{d}\mathbf{r}\,, \tag{84}$$

just employing Green's theorem [158].

The interesting thing to be said now consists in proposing some sentences on the nature of the integrals (83) and (84), which are real and positive definite as kinetic energy shall be, either classically speaking or quantum mechanically, thus providing the integral (83) with a well-defined structure, capable of being interpreted as a norm. It should also be noted that the imaginary unit, usually employed in the quantum mechanical definition of linear momentum does not need to be used here in front of the *nabla* operator. The reason can be found in the fact that the imaginary unit has no active role in the above definitions, unless a Hermitian matrix representation is needed for the ∇ operator. Such an **imaginary** factor has to be present in such a Hermitian representation case, because the matrix associated to the bare *nabla* operator is Skew-Hermitian, that is:

$$\int_D \Psi_I^{*}(\mathbf{r})(\nabla\Psi_J(\mathbf{r}))\,\mathrm{d}\mathbf{r} = -\int_D (\nabla\Psi_I(\mathbf{r}))^{*}\Psi_J(\mathbf{r})\,\mathrm{d}\mathbf{r}\,,$$

a property which can be easily interpreted as a consequence of the application of Green's theorem again.

The Simplest Sobolev Space. In any case, the existence of the wave function norm (78) and the subsequent convexity condition (79), can both be recognized as the parallel properties holding for the gradient of the wave function, and corresponding to Eq. (83), which proves collaterally the positive definite nature of the quantum mechanical kinetic energy expectation value. Thus, if the sequence of equations from (75) up to (83) must hold simultaneously, just to obtain a coherent mathematical structure within the quantum mechanical framework, it is feasible to consider that both Banach spaces (75) and (82), can be supposedly forming a composite norm, in the way of the following definition present within the equation shown below. Simplifying the wave function notation from the variable dependence, in order to ease the form of the subsequent equations, one can define the following norm:

$$\forall\Psi \in H_{\infty}(\mathbf{C}) \rightarrow \exists\|\Psi\|_1^1 = \int_D |\Psi|^2\,\mathrm{d}\mathbf{r} + \int_D |\nabla\Psi|^2\,\mathrm{d}\mathbf{r} = 1 + 2\langle K\rangle\,. \tag{85}$$

Such a composition provides the first definition of the simplest element among the collection of all possible Sobolev spaces, which can be connected to quantum theory. Thus, it can be assumed from now on that the most adequate quantum mechanical wave function space structure is a *Sobolev space*.

Sobolev Spaces. The notation for the norm (85) will be made immediately obvious, by means of defining a general Sobolev space norm as:

$$\forall\Psi \in H_{\infty}(\mathbf{C}) \rightarrow \|\Psi\|_{\alpha}^{\beta} = \sum_{a=1}^{\alpha}\int_D \sum_{b=0}^{\beta} |\nabla^b\Psi|^{2a}\,\mathrm{d}\mathbf{r}\,. \tag{86}$$

The usual Hilbert–Banach space norm can be retrieved from definition (86), simply supposing that the null power of the gradient operator can be substituted by the identity: $\nabla^0 \equiv I$, and using afterwards: $\alpha = 1 \wedge \beta = 0$. In addition, the earlier Sobolev space norm, simplified as in

Eq. (85), is also found employing Eq. (86), but choosing: $\alpha = \beta = 1$.

However, although Eq. (86) contains the classical quantum mechanical Sobolev ∇-norm, it implicitly possesses the restriction consisting in that both wave function and gradient norm powers shall be the same in any circumstance. Consequently, they cannot be monitored as independent terms in the norm definition. An appropriate choice to avoid this situation may be described with the more general formulation:

$$\forall \Psi \in H_\infty(\mathbf{C}) \rightarrow$$

$$\|\Psi\|_\alpha^{\beta,\gamma} = \sum_{a=1}^{\alpha} \int_D |\Psi|^{2a}\, \mathrm{d}\mathbf{r} + \sum_{c=1}^{\gamma} \int_D \sum_{b=1}^{\beta} |\nabla^b \Psi|^{2c}\, \mathrm{d}\mathbf{r} \tag{87}$$

Therefore, Eq. (87) will transform into expression (86), whenever: $\alpha = \gamma$. In order to avoid further interpretation problems, the squared module of the *nabla* powers has to be considered a contraction operation; or has to be considered a scalar product of the corresponding matrix elements, represented by the result of the operation $\nabla^b \Psi$:

$$|\nabla^b \Psi|^{2c} \equiv |\langle \nabla^b \Psi | \nabla^b \Psi \rangle|^c \,. \tag{88}$$

Nested Summation Symbols. This last remark, represented by the Eq. (88), can be alternatively written in a very elegant manner employing the definition of an *inward matrix product* (IMP), already discussed.

Taking the IMP definition into account, then in Eqs. (86) and (87) it can be understood that the present square modules are computed over the resultant wave function derivative hypermatrices as:

$$|\nabla^b \Psi|^{2c} = |\langle (\nabla^b \Psi)^* * (\nabla^b \Psi) \rangle|^c \,,$$

where the symbol $\langle\ \rangle$, associated to any matrix, means a sum of the whole matrix elements, for instance:

$$\forall \mathbf{P} = \{p_{ij}\} \rightarrow \langle \mathbf{P} \rangle = \sum_i \sum_j p_{ij} \,, \tag{89}$$

which constitutes a definition possessing obvious generalization possibilities, within any kind of hypermatrix structure.

This generalization power can be easily seen, taking into account the *nested summation symbol* (NSS) formalism, which was developed several years ago, see references [159,160] for example. Then, using NSS, the expression of the total sum of the elements of an arbitrary hypermatrix can be generally written without any further problem. A NSS is a symbolic device, which has a linear operator nature, and in this way resumes the presence of an undefined number of nested sums and corresponds to an easily programmable algorithm, which generalizes in practice an indefinite number of do loops. In turn, a NSS acts over any kind of complex expression, bearing all the involved indices present within the sums, that is:

$$\sum\nolimits_N (\mathbf{i}) \phi(\mathbf{i}) \equiv \sum_{i_1}^{n_1} \sum_{i_2}^{n_2} \ldots \sum_{i_N}^{n_N} \phi(i_1; i_2; \ldots; i_N) \,,$$

where by the index vector $\mathbf{i}$ it is understood: $\mathbf{i} = (i_1; i_2; \ldots; i_N)$. Thus, if by the definition the following subindex structure is assumed:

$$\mathbf{Z} = \{z_{i_1 i_2 \ldots i_N}\} \equiv \{z(\mathbf{i})\} \,,$$

by which is represented any $(n_1 \times n_2 \times \ldots n_N)$-dimensional hypermatrix element, then the symbolic device associated to the total summation of the elements, particularly defined in Eq. (89), can be generally described by means of the compact NSS expression:

$$\langle \mathbf{Z} \rangle = \sum\nolimits_N (\mathbf{i}) z(\mathbf{i}) \,.$$

Extended Wave and Density Functions

Sobolev spaces appear, after the previous discussion, as a very general kind of extended Hilbert–Banach spaces, which within the quantum mechanical framework are able to put into a unique statement the nature of both the wave function and its gradient. Alternatively, they can produce completely general structures, somehow involving usual quantum mechanical operators, attachable to any system observable. It is a matter of straightforward analysis to translate the subjacent Sobolev mathematical structure into the Hilbert space elements themselves, producing a new breed of spaces, which can be obviously called *Hilbert–Sobolev spaces* [3]. As has been done before, within the previous description of Sobolev spaces, the extension of the wave function and the possible application of the resultant formal structure will be here gradually discussed.

Extended Wave Functions. By a ∇-extended wave function $|\Phi\rangle$ has been understood a composite column vector, or alternatively a diagonal matrix, if one prefers, whose elements are the original wave function Ψ and its gradient $\nabla\Psi$. That is:

$$|\Phi\rangle = \begin{pmatrix} \Psi \\ \nabla\Psi \end{pmatrix} \equiv \mathrm{Diag}(\Psi; \nabla\Psi) = \begin{pmatrix} \Psi & 0 \\ 0 & \nabla\Psi \end{pmatrix} \,. \tag{90}$$

[3] In the case where the nature of the operator or the set of operators, active in the Sobolev norm definition, has to be specified, then the notation Hilbert–Sobolev ∇-, Ω- or Υ-spaces, can be obviously employed.

The column vector form, as will be discussed later on, better represents some applications and the mathematical manipulations one can perform over them; while in other cases the diagonal matrix structure produces more elegant expressions and it is easier to deal with. However, both choices provide equivalent results. This is so because the proposed representations constitute the elements of a pair of isomorphic vector spaces.

The previous discussion on Sobolev spaces permits us to define the extended wave function within a general Hermitian operator scheme, simply as:

$$|\Phi\rangle = \begin{pmatrix} \Psi \\ \Omega\Psi \end{pmatrix} \equiv \mathrm{Diag}(\Psi; \Omega\Psi) = \begin{pmatrix} \Psi & 0 \\ 0 & \Omega\Psi \end{pmatrix} . \quad (91)$$

The definition of Eq. (91) can be called an *Ω-extended wave function*. For example, the quantum mechanical complementary definition of the ∇-extended wave function (90) can be easily written employing the position vector **r**, that is:

$$|\Theta\rangle = \begin{pmatrix} \Psi \\ \mathbf{r}\Psi \end{pmatrix} \equiv \mathrm{Diag}(\Psi; \mathbf{r}\Psi) = \begin{pmatrix} \Psi & 0 \\ 0 & \mathbf{r}\Psi \end{pmatrix} . \quad (92)$$

producing an **r**-extended wave function accordingly.

By inspection of the adopted structure until now, extended wave functions can be also considered as the result of applying some adequate operator over the original Schrödinger wave function. For instance, within the already-mentioned diagonal formalism of the Ω-extended wave function (91), defining the diagonal operator:

$$\Gamma = \mathrm{Diag}(\mathbf{I}; \Omega) ,$$

where **I** is the unit operator, it will be sufficient to see that:

$$\begin{aligned} |\Phi\rangle = \Gamma[\Psi] &= \mathrm{Diag}(\mathbf{I}; \Omega)[\Psi] \\ &= \mathrm{Diag}(\mathbf{I}[\Psi]; \Omega[\Psi]) \\ &= \mathrm{Diag}(\Psi; \Omega\Psi) . \end{aligned} \quad (93)$$

The same can be said if the corresponding vector operator is constructed by means of the vector structure:

$$\Gamma = \begin{pmatrix} \mathbf{I} \\ \Omega \end{pmatrix} ,$$

which, upon application over the original scalar wave function form, permits us to alternatively obtain the isomorphic vector picture of the diagonal expression (93).

Energy Expectation Values. Returning to the ∇-extended wave function in Eq. (90), it is easy to see how the energy expectation value of the associated Schrödinger equation can be expressed, without loosing any information, when writing the final form it takes. For this purpose, the appropriate Hamilton operator, **H**, can be structured by means of a diagonal form, as:

$$\mathbf{H} = \mathrm{Diag}(\mathbf{U}; \tfrac{1}{2}\mathbf{I}) = \begin{pmatrix} \mathbf{U} & 0 \\ 0 & \frac{1}{2}\mathbf{I} \end{pmatrix} , \quad (94)$$

where the symbol: **U**, corresponds to the potential energy operator and **I** is just a unit operator built to fit the adequate dimensions of the extended wave function (90) gradient part. Using Eq. (94) and the ∇-extended wave function (90) in the appropriate way, it is immediate to write:

$$\begin{aligned} E = \langle \mathbf{H} \rangle = \langle \Phi | \mathbf{H} | \Phi \rangle &= \langle \Psi | \mathbf{U} | \Psi \rangle + \tfrac{1}{2} \langle \nabla\Psi | \nabla\Psi \rangle \\ &= \langle U \rangle + \langle K \rangle . \end{aligned} \quad (95)$$

In the same way, whenever the normalization of the wave function holds: $\langle \Psi | \Psi \rangle = 1$, the ∇-extended wave function (90) can be manipulated in order to obtain a positive definite norm like:

$$\langle \Phi | \Phi \rangle = \langle \Psi | \Psi \rangle + \langle \nabla\Psi | \nabla\Psi \rangle = 1 + 2\langle K \rangle , \quad (96)$$

which corresponds to the same result as the one provided by the norm obtained in the definition of the simplest Sobolev space, as presented first in Eq. (85). Obviously, such a Sobolev space can be interpreted as a composite Hilbert space, whose elements are defined by the ∇-extended wave function (90). That is, employing the Cartesian product of the Hilbert spaces (75) and (81), upon reordering the ordered pair in the form of column vector:

$$H_\infty \times H_\infty^{(P)} = \left\{ |\Phi\rangle = \begin{pmatrix} \Psi \\ \nabla\Psi \end{pmatrix} \right\} . \quad (97)$$

Extended Density Functions. The Hilbert semispace corresponding to the Hilbert space (97), can be supposedly formed by the total density functions and computed by using the trace of the extended wave function tensor product, for instance:

$$\begin{aligned} \tau(\mathbf{r}) = \mathrm{Tr}\, |\Phi\rangle\langle\Phi| = \mathrm{Tr} \begin{pmatrix} |\Psi|^2 & \Psi(\nabla\Psi)^* \\ (\nabla\Psi)\Psi^* & |\nabla\Psi|^2 \end{pmatrix} \\ = \rho(\mathbf{r}) + \kappa(\mathbf{r}) ; \end{aligned} \quad (98)$$

that is: as the superposition of the electronic density $\rho(\mathbf{r})$ and the kinetic energy density $\kappa(\mathbf{r}) = |\nabla\Psi|^2$. Such a result appears to be consistent with the previous definitions, as already expressed in Eqs. (85) and (96). An alternative definition of the total density (98) can be also obtained, employing the already defined IMP, in association with the

total sum of elements of a vector, as previously given in Eq. (89) and generalized afterwards:

$$\tau(\mathbf{r}) = \langle |\Phi\rangle * |\Phi^*\rangle\rangle = \left\langle \begin{pmatrix} \Psi \\ \nabla\Psi \end{pmatrix} * \begin{pmatrix} \Psi^* \\ (\nabla\Psi)^* \end{pmatrix} \right\rangle$$
$$= \left\langle \begin{pmatrix} |\Psi|^2 \\ |\nabla\Psi|^2 \end{pmatrix} \right\rangle = \left\langle \begin{pmatrix} \rho(\mathbf{r}) \\ \kappa(\mathbf{r}) \end{pmatrix} \right\rangle = \rho(\mathbf{r}) + \kappa(\mathbf{r}) \,. \quad (99)$$

Such superposition of density functions, producing the total density function $\tau(\mathbf{r})$, can be obviously named as a ∇-extended density function.

It must be noted now, that the total density function, as defined in Eqs. (98) or (99), possesses the statistical interpretation of observing, within a space infinitesimal volume element, both the position of the associated system of particles *or* the same system within a corresponding related infinitesimal kinetic energy range. The conjunction *or*, is a consequence of the obtained statistical expressions, where both densities are summed up, and it is in agreement with *Heisenberg's uncertainty principle*, which will forbid the practical use of the product of both distributions, position and momentum not being simultaneously observable.

Quantum Self-Similarity Measures and Non-linear Schrödinger Equation. The previous experience, crystallized in definitions (90) and (91), about the extended wave functions and the details of their subsequent use in the energy definition (95), as well as the construction of the extended density function in Eqs. (98) or (99) formalisms, shows a plausible way to generalize the presented wave function extensions. This effort can be employed as a way to try, afterwards, to obtain information on the possible utility of such extended general wave function forms.

In the same way as the generalization of Sobolev spaces, starting from the simplest form (85), the Ω-extended expression (91) of the wave function can be generalized accordingly.

In order to do so, a systematic exposition will be followed, in the same way as has been previously done. Thus, the first new breed of extended wave functions will be defined by means of the vector, or diagonal, form:

$$|\Phi\rangle = \begin{pmatrix} \Psi \\ |\Psi|^2 \\ \Omega\Psi \end{pmatrix} \equiv \mathrm{Diag}(\Psi; |\Psi|^2; \Omega\Psi) \,. \quad (100)$$

As the second element of the vector (100), is simply the electronic density function, such a vector can be written equivalently as:

$$|\Phi\rangle = \begin{pmatrix} \Psi \\ \rho \\ \Omega\Psi \end{pmatrix} \equiv \mathrm{Diag}(\Psi; \rho; \Omega\Psi) \,; \quad (101)$$

and the corresponding possible energy expression could be written in turn, employing a Hamilton operator and following the previous experience as presented in Eq. (94) in the form:

$$\mathbf{H} = \begin{pmatrix} \mathbf{U} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \alpha\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \frac{1}{2}\mathbf{I} \end{pmatrix} = \mathrm{Diag}(\mathbf{U}; \alpha\mathbf{I}; \tfrac{1}{2}\mathbf{I}) \,, \quad (102)$$

where the first and the last non null elements are the same as shown before in Eq. (94), and besides α is an arbitrary real parameter. Thus, the energy expectation value equivalent to the expression (95), employing in the extended wave function (101) the substitution $\Omega = \nabla$, gives:

$$E = \langle \Phi|\mathbf{H}|\Phi\rangle = \langle U\rangle + \alpha\langle\rho|\rho\rangle + \langle K\rangle \,. \quad (103)$$

So, calling the classical Schrödinger energy (95):

$$E_0 = \langle U\rangle + \langle K\rangle \,,$$

and owing to the fact that the second term in the expectation value (103), can be manipulated as follows:

$$z = \langle\rho|\rho\rangle = \int_D |\Psi|^2|\Psi|^2 \,\mathrm{d}\mathbf{r} = \int_D \Psi^*|\Psi|^2\Psi \,\mathrm{d}\mathbf{r}$$
$$= \langle\Psi\||\Psi|^2|\Psi\rangle = \langle\Psi|\rho|\Psi\rangle = \langle\rho\rangle \,, \quad (104)$$

then, one can simply write:

$$E = E_0 + \alpha z \,. \quad (105)$$

The nature of the integral (104) is well known in the theoretical formulation and definitions of *quantum similarity measures* (QSM), corresponding to the so-called *quantum self-similarity measure* (QSSM) associated to the density function ρ. As is evident upon inspecting the QSSM appearing in Eq. (104), the integral also corresponds to a scalar product of the density function by itself. That is: a simple Euclidean norm within the associated Hilbert semispace, containing ρ. In addition, it can be quantum mechanically interpreted as the expectation value of the density function over itself. Finally, the integral is also closely related to a relativistic correction appearing in the definition of the Breit Hamiltonian: the term named *spin-spin contact* [161,162,163], although in the present form (104) the two-electron Dirac δ-function is absent. The characteristic non-trivial features of this kind of QSSM integral, as appear in this particular formalism, and more precisely with respect of the spin part of the wave function, have been deeply analyzed in two separate papers [87,164].

Moreover, Eqs. (101) up to (105), tell that, in fact, the new extended wave function produces an energy expectation value, which can be seen to be in correspondence with the non-linear Schrödinger equation.

Upon inspecting definition (100), or Eq. (101), the corresponding extended density function can be deduced, employing the same technique as the one that was previously used in Eq. (99):

$$\gamma(\mathbf{r}) = \langle |\Phi\rangle * |\Phi^*\rangle\rangle = \left\langle \begin{pmatrix} \Psi \\ \rho \\ \nabla\Psi \end{pmatrix} * \begin{pmatrix} \Psi^* \\ \rho \\ (\nabla\Psi)^* \end{pmatrix} \right\rangle$$
$$= \left\langle \begin{pmatrix} |\Psi|^2 \\ \rho^2 \\ |\nabla\Psi|^2 \end{pmatrix} \right\rangle = \left\langle \begin{pmatrix} \rho(\mathbf{r}) \\ \rho^2(\mathbf{r}) \\ \kappa(\mathbf{r}) \end{pmatrix} \right\rangle = \rho(\mathbf{r}) + \rho^2(\mathbf{r}) + \kappa(\mathbf{r}) \,. \tag{106}$$

So, the total density now can be written as the ∇-extended density function (99), with an extra term made of the squared electronic density:

$$\gamma(\mathbf{r}) = \tau(\mathbf{r}) + \rho^2(\mathbf{r}) \,.$$

Finally, the Sobolev norm of the extended wave function (101) will be easily obtained by integrating the extended density function (106):

$$\langle \Phi|\Phi\rangle = 1 + 2\langle K\rangle + z \,, \tag{107}$$

the QSSM integral z, defined in Eq. (104), can be also considered as a norm, associated by construction, to the elements of a Hilbert semispace, then this fact assures the positive definition of the integral (107). It is not difficult to associate the norm in Eq. (107), with a Sobalev norm of type (87), with the parameters chosen accordingly: $\|\Psi\|_2^{1,1}$.

Thus, a naïve generalization of the idea underlying the ∇-extended wave function definition has revealed itself as a powerful tool, which permits the formal description of the non-linear Schrödinger equation. Such formalism allows producing another kind of Hilbert–Sobolev space, and at the same time, within their integral steps, finally puts into evidence the connection of Sobolev spaces and extended wave functions with the concept of QSSM.

Expectation Values Within Extended Density Functions Framework

Landau and Lifshitz proposed the interpretation of expectation values in a statistical formulation, instead of the usual quantum mechanical form. This was anticipated somewhere in the already quoted volume of reference [153], published within a series dedicated to studying the mechanics of particle systems. The same point of view was also masterly described and adopted, later on, by McWeeny and Sutcliffe in the book of reference [165]. In the present paper, employing the concepts associated to the extended wave functions, it will be shown that a similar possibility as the one mentioned in these previous references can be exactly deduced. The difference with the above-mentioned sources consists of the fact that, in the present way, one only needs to base the arguments on the structure of the already described Hilbert–Sobolev spaces. Some related point of view has been found in the same direction, precluding this property, when the deduction of the energy expectation values has been discussed, as can be noticed when observing Eqs. (95) and (103). According to this, the purpose of this section is to deduce a general Hilbert–Sobolev formalism for the expectation values associated to the extended wave functions and provide an application example.

Statistical Form of Expectation Values in the Extended Wave Function Formalism. One can deduce the general composition of an extended density function, corresponding to the extended function (91). For this purpose, an appropriate operator shall be defined. It has to be able to act over the extended wave function structure. Accordingly, it is sufficient to take into account that a diagonal-like operator can be constructed in the following way:

$$\Theta = \mathrm{Diag}(\Gamma; \Lambda) \,, \tag{108}$$

where, in order to be applied along the appropriate extended wave function elements, the involved operators Γ and Λ themselves have to possess an adequate structure. When, within the global expectation value expression, the Ω-extended wave function is used quantum mechanically over the operator (108), in order to obtain the corresponding equation, then owing to the properties of diagonal matrices, both the wave function and the diagonal operator can be manipulated, in the forthcoming manner, employing obvious notation and symbols, to arrive towards a statistical formulation final form:

$$\begin{aligned} \langle\Theta\rangle &= \langle\Phi|\Theta|\Phi\rangle \\ &= \langle \mathrm{Tr}(\mathrm{Diag}(\Psi^*; (\Omega\Psi)^*)\,\mathrm{Diag}(\Gamma; \Lambda)\,\mathrm{Diag}(\Psi; \Omega\Psi))\rangle \\ &= \langle \mathrm{Tr}(\mathrm{Diag}(\Gamma; \Lambda)\,\mathrm{Diag}(\Psi^*; (\Omega\Psi)^*)\,\mathrm{Diag}(\Psi; \Omega\Psi))\rangle \\ &= \langle \mathrm{Tr}(\mathrm{Diag}(\Gamma; \Lambda)\,\mathrm{Diag}(|\Psi|^2; |\Omega\Psi|^2))\rangle \\ &= \langle \mathrm{Tr}(\mathrm{Diag}(\Gamma\rho; \Lambda\omega))\rangle = \langle \Gamma\rho + \Lambda\omega\rangle \\ &= \langle \Gamma\rho\rangle + \langle \Lambda\omega\rangle \equiv \langle \Gamma|\rho\rangle + \langle \Lambda|\omega\rangle \,. \end{aligned} \tag{109}$$

Now it must be taken into account that the external summation symbol employed in the above equation, has to be taken, when appropriate, as an integration procedure.

Energy Expectation Value of a Set of Interacting Quantum Objects. Among the possible uses of the present formalism, it seems worthwhile to consider some theoretical arrangement associated to a previous discussion made by Huzinaga and co-workers [166,167]. This procedure is related to the model potential method, proposed by Bonifacic and Huzinaga [168] in order to study the optimal valence AO, transforming the core electron structure into an electrostatic potential. In the following discussion the structure and final form will be presented, which can take the total energy, when the problem of several interacting quantum objects is studied in a somehow approximate way from the point of view of the wave function. In order to perform such a study, one can suppose known a set of quantum objects [25], whose Hamiltonian operators in this case can be considered constructed with a diagonal structure, similar to the one described in Eq. (102).

However, the second diagonal term has to be transformed necessarily into a new operator and described with the appropriate construction rule, as follows:

$$\mathbf{L} = \mathbf{1} - \mathbf{I} = \{(i \neq j)\} . \tag{110}$$

This operator is in some way the reciprocal mirror image of the well-known unit operator:

$$\mathbf{I} = \{\delta(i = j)\} . \tag{111}$$

In the last definition (111) as well as in the operator previously defined in Eq. (110), a *logical Kronecker symbol* has been utilized [159,160]. Considering the definition of the unit operator (111), the logical Kronecker symbols appear obviously structured, adopting a self-explanatory description. The definition, for instance, can be made clearer, considering a logical expression Λ taken as the Kronecker symbol argument, and then its resultant value can be generally described by means of the logical content of the possible issues of such an argument, that is:

$$\delta(\Lambda) \in \{\delta(\Lambda \equiv .T.) = 1 \wedge \delta(\Lambda \equiv .F.) = 0\} .$$

The operator **1**, the *unity* operator as used in Eq. (110), means the multiplicative unit of the IMP, that is: $\mathbf{1} = \{\mathbf{1}_{ij} = 1\}$.

Considering α as a parameter to be adjusted, according to the nature of the problem, then the Hamiltonian could be written in this case as:

$$\mathbf{H} = \begin{pmatrix} U\mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \alpha\mathbf{L} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \frac{1}{2}\mathbf{I} \end{pmatrix} = \mathrm{Diag}(U\mathbf{I}; \alpha\mathbf{L}; \tfrac{1}{2}\mathbf{I}) , \tag{112}$$

where U is a scalar potential operator, defined by means of the expression:

$$U = \sum_I (V_I + R_I) ,$$

with the sum encompassing all the involved quantum objects. By the symbol V_I is understood an attractive potential; while by R_I, the repulsion terms can be somehow described. Both operator terms have to be associated in turn to the whole set of particles, constituting the Ith quantum object. For instance: when dealing with atoms and molecules, these operators could be associated to the nuclear attraction operator and to the Coulomb-exchange operator terms, respectively.

Then, in this problem the appropriate extended wave function can be taken as the ∇-extended form of Eq. (101), that is:

$$|\Phi\rangle = \begin{pmatrix} \Psi \\ \rho \\ \nabla\Psi \end{pmatrix} \equiv \mathrm{Diag}(\Psi; \rho; \nabla\Psi) , \tag{113}$$

where the wave and the density functions present in the extended function (113) shall be taken as vectors, having as elements the wave and density functions of every quantum object in the considered set, respectively.

Thus, the expectation value of the Hamiltonian (112) under the extended wave function (113), can be easily written, using the technique of Eq. (109) as:

$$\langle\Phi|\mathbf{H}|\Phi\rangle = \sum_I \Big(\sum_J \langle V_J + R_J|\rho_I\rangle + \alpha \sum_{J\neq I} \langle\rho_J|\rho_I\rangle + \langle\kappa_I\rangle \Big) , \tag{114}$$

in the above expression the first term corresponds to the potential energy of the objects plus their interactions; the second term can be associated to the expectation value of the projection operator over each quantum object except itself, and the role of this part of the expectation value is intended to prevent the collapsing tendency of the particles, belonging to each separated quantum object, towards a unique system; finally, the third term corresponds to the global kinetic energy obtained in a way such as the quantum objects were non-interacting.

The second element of the expectation value (114) can be also easily interpreted as a sum of the *overlap* QSM [169] between pairs of quantum objects. In this sense, one can observe Huzinaga's treatment as a procedure, taking into account non-linear terms in the approximate solution of the Schrödinger equation.

Quantum Similarity Measures in Extended Hilbert–Sobolev Spaces

QSM in Hilbert semispaces have been studied from the theoretical point of view as well as considering the potential applications of *quantum similarity* over quantum objects. In this paragraph, the structure of QSM over extended wave and density functions will be analyzed. Before proceeding towards such an analysis, it must be said that, as QSM are essentially defined over density functions, they can be constructed even in the Hilbert–Sobolev spaces framework, provided that the extended density function is known. This is so, because, whenever a total density function can be well defined, like the one present in Eq. (98), for instance, then the construction of any similarity measure can also be put forward. Such a general possibility was analyzed in a particular way several years ago [170], when discussing the extension of the QS concepts into partition functions. Statistical mechanics partition functions can be obviously observed as probability distributions and, thus, they can be considered as elements belonging to a characteristic vector set: a Boltzmann semispace, for example. From such a fact they can be used in the general definitions of QSM, as any probability density function can be used for the same purpose.

The most usual way to produce a QSM, corresponds to the integral constructed as:

$$z_{IJ}(\Omega) = \iint_D \rho_I(\mathbf{r}_1)\Omega(\mathbf{r}_1;\mathbf{r}_2)\rho_J(\mathbf{r}_2)\,\mathrm{d}\mathbf{r}_1\,\mathrm{d}\mathbf{r}_2\,. \tag{115}$$

Where in Eq. (115), $\{\rho_I(\mathbf{r}_1);\rho_J(\mathbf{r}_2)\}$ is a pair of homogeneous order density functions, and $\Omega(\mathbf{r}_1;\mathbf{r}_2)$ is a positive definite operator. The attached properties of the set of integrands ensure that in any case the values of the QSM, defined such as in Eq. (115), will produce a positive real element. For the present purposes, the integral form (115) is sufficient. The already mentioned overlap QSM, which appears in the building up of energy expectation values within non-linear Schrödinger equations, such as those in expressions (104) and (114), are *overlap-like* QSM, and can be deduced from the QSM Eq. (115), by simply using a Dirac delta function as operator, that is: $\Omega(\mathbf{r}_1;\mathbf{r}_2) = \delta(\mathbf{r}_1 - \mathbf{r}_2)$.

The definition of kinetic energy density and other possible density kinds, deducible from the Ω-extended wave function concepts, as discussed earlier, opens the way to produce QSM using the integral (115), upon substitution of the density function pairs by the appropriate extended density function.

So it seems now clear that the QSM integral form, as described in Eq. (115), can be used as it is for extended density functions, just substituting the usual electronic density by the corresponding expression in terms of the chosen extended density functions. Here, the interesting new feature consists of the emergence of QSM integrals, associated to density functions of different origin. For instance, suppose the ∇-extended density function as defined in Eq. (98): the total density defined there, associated to a pair of quantum objects produces a QSM, which can be written in terms of four hybrid QSM integrals as:

$$\begin{aligned} z_{IJ} &= \iint_D \tau_I(\mathbf{r}_1)\Omega(\mathbf{r}_1;\mathbf{r}_2)\tau_J(\mathbf{r}_2)\,\mathrm{d}\mathbf{r}_1\,\mathrm{d}\mathbf{r}_2 \\ &= \iint_D \rho_I(\mathbf{r}_1)\Omega(\mathbf{r}_1;\mathbf{r}_2)\rho_J(\mathbf{r}_2)\,\mathrm{d}\mathbf{r}_1\,\mathrm{d}\mathbf{r}_2 \\ &\quad + \iint_D \rho_I(\mathbf{r}_1)\Omega(\mathbf{r}_1;\mathbf{r}_2)\kappa_J(\mathbf{r}_2)\,\mathrm{d}\mathbf{r}_1\,\mathrm{d}\mathbf{r}_2 \\ &\quad + \iint_D \kappa_I(\mathbf{r}_1)\Omega(\mathbf{r}_1;\mathbf{r}_2)\rho_J(\mathbf{r}_2)\,\mathrm{d}\mathbf{r}_1\,\mathrm{d}\mathbf{r}_2 \\ &\quad + \iint_D \kappa_I(\mathbf{r}_1)\Omega(\mathbf{r}_1;\mathbf{r}_2)\kappa_J(\mathbf{r}_2)\,\mathrm{d}\mathbf{r}_1\,\mathrm{d}\mathbf{r}_2\,. \end{aligned}$$

The first term being exactly the one in Eq. (115), and the last one being a QSM over kinetic energy density distributions, the central terms corresponding to hybrid QSM between electronic and kinetic density functions. Total density QSM integrals of any kind still are waiting to be practically employed in a systematic way. It is important to consider them as potentially interesting quantum object descriptors. Because of their flexible generality, the extended density functions may provide new insights and refinements within QSPR models. In the next sections of this work a discussion of several possible uses of the extended densities technique will be discussed.

Fundamental Quantum QSPR (QQSPR) Equation in Sobolev Spaces

From the previous considerations and the *ad hoc* definitions, it can be said, without doubt, that the general structure of Sobolev spaces can be easily associated to the extended Hilbert wave functions, producing a new mathematical structure, which has been named a Hilbert–Sobolev space.

Hilbert–Sobolev spaces have been used too to obtain up to now, with the appropriate definition of the Hamilton operator, adapted to every circumstance, the corresponding energy expectation values. Elementary reasoning permits us to arrive towards the conclusion that the expectation values of a given observable can be obtained in a similar generalized manner and within the formalism, which is

the most valuable finding, perhaps, of the present discussion: the possibility to write the expectation value expressions using a statistical method.

Thus, if some observable **O**, has attached the Hermitian operator Θ, and the associated quantum object extended density function is τ, then the associated quantum mechanical expectation value, $\langle\vartheta\rangle$, can be written, according to the previous considerations as:

$$\langle\vartheta\rangle = \langle\Theta|\tau\rangle \,.$$

In QQSPR reasoning, the expression above can be further arranged in the following way. The Hermitian operator Θ is usually not known, but it can be expressed as the product of a still unknown operator W by a known one Ω, associated at the same time with a positive definite property, that is:

$$\langle\vartheta\rangle = \langle W\Omega|\rho\rangle \,. \tag{116}$$

Because Ω is chosen as having a positive definite form, one is always assured that knowing Θ, then W can be obtained in turn as:

$$\Omega > 0 \rightarrow \exists\Omega^{-1} \rightarrow W = \Theta\Omega^{-1} \,.$$

However, in practice Θ and W are unknown, and in the expectation value formalism there is always the need to obtain W in an approximate way, employing a least squares procedure. This can be easily done whenever a set of compatible density functions, connected to a certain quantum object set as quantum object tags, is already known: $T = \{\tau_I\}$ and can be used in order to express the operator W as a superposition such that:

$$W(\mathbf{r}) = \sum_I w_I \tau_I(\mathbf{r}) \,. \tag{117}$$

Then, the expectation value (116) becomes expressible as:

$$\langle\vartheta\rangle = \sum_I w_I \langle\tau_I|\Omega|\tau\rangle \,,$$

being the resultant QSM integrals, defined in the usual way in Eq. (115). The well-described procedure to obtain the coefficients: $\mathbf{w} = \{w_I\}$ is the least-squares technique, or anyone of the existing variants, as previously described, by means of an alternative method based on IMP reasoning.

The first step is, in any case, to proceed with the construction of a linear system of equations, whose solution is the coefficient vector **w**. To obtain such a linear system it is necessary first to know, for a given quantum object set, a set of properties: $\{\pi_I\}$, which can be associated to the corresponding expectation values: $\{\vartheta_I\}$. Then, employing the quantum objects density functions, which in particular will coincide with the set $T = \{\tau_I\}$, used to construct the unknown part of the operator, although not necessarily both density function sets shall be the same, it is possible to write:

$$\forall K\colon \langle\vartheta_K\rangle \equiv \pi_K = \sum_I w_I \langle\tau_I|\Omega|\tau_K\rangle \,,$$

so collecting in a vector like $|\pi\rangle$, the property values, and defining the integral matrix elements by means of:

$$\mathbf{Z} = \{z_{IK} = \langle\tau_I|\Omega|\tau_K\rangle = \langle\tau_K|\Omega|\tau_I\rangle = z_{KI}\} \,, \tag{118}$$

it is easy to transform the system into a matrix equation in the form of:

$$\mathbf{Zw} = |\pi\rangle \,. \tag{119}$$

The solution of the above system will make known an approximate form of the implied operator and, in this way, provides a possible path to be followed in order to obtain estimates of the property values of any unknown quantum object, just as in the classical QSPR model procedures.

However, the present kind of quantitative structure (represented by the QSM)-property model is completely based on quantum mechanical propositions. More than this, there are no other suppositions than the usual ones, associated to density function algebra and quantum mechanical basic mathematical background. Hence, the results obtained through the linear system of Eqs. (119), do not depend on user choice, but rely directly on theoretical grounds and because of this are statistically unbiased. Equation (119) can be properly called the fundamental QQSPR equation. Moreover, the models obtained in this way can be interpreted in the light of the quantum mechanical expectation value concept. By this simple fact, contrary to the classical QSPR modeling results, they can be associated to a causal relationship relying on quantum object properties and QSM.

Due to the unavoidable presence of the QSM matrix (118) into the fundamental QQSPR equation, the columns of such a matrix play a fundamental role in the discrete representation of quantum objects. Consequently, the columns of the QSM matrix, involving a given quantum object density function, interacting with the whole basis set of density functions employed to represent the unknown operator W, can be safely considered as natural quantum mechanical discrete descriptors of the associated quantum object.

Non-Linear Terms in QQSPR Models

The usual relationships between structure and properties sometimes needs the presence of non-linear terms. Non-linear terms are needed in order to represent accurately the property as a function of the structural descriptors.

The fundamental QQSPR equation can be deduced to introduce in a natural way these terms, if needed. To see this, it is only necessary to think about the easy path, which was used to introduce the non-linear terms in the Schrödinger equation, just by using simple considerations, associated to the structure of the Hilbert–Sobolev spaces.

Suppose that the extended wave function (106) is employed, along with the corresponding extended density functions. In this case, although the Eq. (117), producing the unknown operator expression, can be supposedly set in the same manner as has been proposed in the usual framework, the kinetic energy distribution as well as the non-linear density terms can be employed separately. That is, the unknown operator can be written now as:

$$W = \sum_I w_I \rho_I + \sum_J k_J \kappa_J + \sum_L l_L |\rho_L|^2 \,. \tag{120}$$

Such an approach will produce a set of linear equations, with an extended number of parameters, but also a matrix representation of the unknown operator with added dimensions. The matrix elements, involving squared density functions, are candidates to be interpreted as the quantum representatives of the possible presence of non-linear terms in the fundamental QQSPR equation.

Non-linearity can be introduced in several alternative ways, due to the flexibility promoted by the ideas around the Hilbert–Sobolev concepts. For example, within the QSM matrix definition (118), the operator Ω can be seen as formed by the expression:

$$\begin{aligned}\Omega = \exp(a\rho) &= \sum_{p=0}^{\infty} \frac{a^p}{p!} |\rho|^p \\ &= I + a\rho + \frac{a^2}{2}|\rho|^2 + O(3) \,, \end{aligned} \tag{121}$$

which is assured to be positive definite whenever the exponent a possesses positive values. Convergence can be assured whenever: $a \in (0, 1)$.

In fact, such an expansion can be generalized by using a set of positive definite coefficients $A = \{a_I\}$, such that:

$$\Omega = \sum_{p=0}^{\infty} a_p |\rho|^p = a_0 I + a_1 \rho + a_2 |\rho|^2 + O(3) \,.$$

In such a general case, the QSM matrix elements will be written as a superposition of terms like:

$$\begin{aligned} z_{IK} = \langle \tau_I | \Omega | \tau_K \rangle &= \left\langle \tau_I \middle| \sum_{p=0}^{\infty} a_p \middle| \rho |^p | \tau_K \right\rangle \\ &= \sum_{p=0}^{\infty} a_p \langle \tau_I \| \rho |^p | \tau_K \rangle \\ &= a_0 \langle \tau_I | \tau_K \rangle + a_1 \langle \tau_I | \rho | \tau_K \rangle \\ &\quad + a_2 \langle \tau_I \| \rho |^2 | \tau_K \rangle + O(3) \,. \end{aligned} \tag{122}$$

Where, in the last line of Eq. (122), it is easy to observe the overlap similarity integrals as the zero-th order term, the triple density similarity integrals as the second element constituting the first-order term, and finally, in the second-order term, the quadruple density integrals appear. Such integrals can be readily defined by means of the expression, chosen among other possible definitions, for example, as:

$$\langle \tau_I \| \rho |^2 | \tau_K \rangle = \int_D \tau_I(\mathbf{r}) \tau_K(\mathbf{r}) \rho^2(\mathbf{r}) \, \mathrm{d}\mathbf{r} \,.$$

This result is still more obvious if the following operator structure is employed: upon substituting in Eq. (121) the density function by a convex superposition like the one in Eq. (120), which to obtain simpler expressions will be written as a convex superposition like:

$$\rho = \sum_A \omega_A \rho_A \,,$$

being the coefficients $\{\omega_A\}$ such that: $\forall A$: $\omega_A \in \mathbf{R}^+ \wedge \sum_A \omega_A = 1$.

In this case, the QSM integral (122) will take the following form:

$$\begin{aligned} z_{IK} = \langle \tau_I | \Omega | \tau_K \rangle &= \left\langle \tau_I \middle| \sum_{p=0}^{\infty} \frac{a^p}{p!} \middle| \rho |^p | \tau_K \right\rangle \\ &= \sum_{p=0}^{\infty} \frac{a^p}{p!} \langle \tau_I \| \rho |^p | \tau_K \rangle \\ &= \langle \tau_I | \tau_K \rangle + a \sum_A \omega_A \langle \tau_I | \rho_A | \tau_K \rangle \\ &\quad + \frac{a^2}{2} \sum_A \sum_B \omega_A \omega_B \langle \tau_I | \rho_A \rho_B | \tau_K \rangle \\ &\quad + O(3) \,. \end{aligned}$$

It can be seen that quadratic or higher order terms can naturally appear in the structure of the fundamental QQSPR equation in this way.

Non-Linear Terms and Variational Approach in Quantum QSPR

Fundamental QQSPR Equation in $(N \times M)$ *Similarity Matrix Spaces.* Suppose a quantum object basis set B composed by M quantum systems, whose homogeneous density functions, acting as quantum object tags, are known: $B = \{\rho_I^B | I = 1, M\}$. Suppose also that a probe quantum object set P is well defined and composed by N quantum systems, which have also known density tags: $\{\rho_I^P\}$, and at least is also known a set of property values: $\{p_J\}$ attached to every quantum object of the set; in this manner: $P = \{\rho_J^P \wedge p_J | J = 1, N\}$.

A general operator Ω can be associated to the expectation value computation of the observable property π, in such a way that, knowing the appropriate quantum state density function tag ρ for a given quantum system, such a quantum object observable property can be evaluated in general by using the integral form:

$$\langle \pi \rangle = \langle \Omega | \rho \rangle = \int_D \Omega(\mathbf{r}) \rho(\mathbf{r}) \, \mathrm{d}\mathbf{r} \,, \tag{123}$$

where D is an appropriate integration domain, where the density and operator variables are defined.

Being the operator Ω, in principle, after the adoption of quantum mechanical rules, a Hermitian operator, without loss of generality can be supposedly decomposed into a product of two commutative operators:

$$\Omega(\mathbf{r}) = W(\mathbf{r}) \Theta(\mathbf{r}) \wedge [W(\mathbf{r}); \Theta(\mathbf{r})] = 0 \,, \tag{124}$$

the operator Θ being a known chosen positive definite one, the remnant Hermitian operator is thus defined as:

$$W(\mathbf{r}) = \Omega(\mathbf{r}) \Theta^{-1}(\mathbf{r}) \,. \tag{125}$$

Using Eq. (123) and the operator composition shown in Eq. (124), then it can be formally written:

$$\langle \pi \rangle = \langle W \Theta | \rho \rangle \equiv \langle W | \Theta \rho \rangle = \langle W | \Theta | \rho \rangle \,, \tag{126}$$

suggesting that the operator W could be approximately obtained, even in the case that it is unknown, due to the nature of the observable attached to the property.

In the case, most usual in QQSPR framework, that an approximate construction of the operator W is needed, if an appropriate quantum object set density function tag set, acting as a basis set, B say, is known, as stated at the beginning, that is: $B = \{\rho_I^B | I = 1, M\}$, then the operator W can be written within a first-order linear approach as:

$$W \simeq \sum_{I=1}^{M} \omega_I \rho_I^B \,, \tag{127}$$

so upon substituting this approximate first-order linear expression into the expectation value in Eq. (126), is obtained:

$$\langle \pi \rangle \simeq \sum_{I=1}^{M} \omega_I \langle \rho_I^B | \Theta | \rho \rangle \,, \tag{128}$$

where the integral in Eq. (128), can be interpreted as a quantum similarity measure, that is:

$$\langle \rho_I^B | \Theta | \rho \rangle \equiv \iint_D \rho_I^B(\mathbf{r}_1) \Theta(\mathbf{r}_1; \mathbf{r}_2) \rho(\mathbf{r}_2) \, \mathrm{d}\mathbf{r}_1 \, \mathrm{d}\mathbf{r}_2 \,. \tag{129}$$

The unknown coefficient set in Eq. (128): $|\omega\rangle = \{\omega_I | I = 1, M\}$, which can be collected into an M-dimensional column (or row) vector, will represent the operator W in terms of the known density function basis set B. This situation, clearly represented by Eq. (128), still has a set of undetermined parameters, associated now to the vector $|\omega\rangle$ components, instead of the operator W.

Equation (128) can be used to obtain the vector $|\omega\rangle$. As is usually the case in classical QSPR, it is only necessary to know, a quantum object tag set, associated to some molecular probe set P of cardinality N, $P = \{\rho_J^P \wedge p_J | J = 1, N\}$, where, as previously commented, every quantum object structure in P has also necessarily to be attached to a known value of the involved observable: $|p\rangle = \{p_J | J = 1, N\}$, which can be also collected in form of a N-dimensional column (or row) vector. Then, Eq. (128) can be rewritten for every element in P, employing the known property values instead of the expectation observable values, that is:

$$\forall J = 1, N \colon p_J \simeq \sum_{I=1}^{M} \omega_I \langle \rho_I^B | \Theta | \rho_J^P \rangle \,, \tag{130}$$

in this way the following set of quantum similarity measures is generated:

$$a_{IJ}^{BP}(\Theta) \equiv a_{IJ}^{BP} = \langle \rho_I^B | \Theta | \rho_J^P \rangle \,, \tag{131}$$

which in turn can be considered, after an appropriate rearrangement, as elements of a $(M \times N)$ similarity matrix, involving the basis and probe quantum object molecular sets respectively: $\mathbf{A} = \{a_{IJ}^{BP}\}$.

With this matrix definition in mind, then Eq. (130) can be rewritten as a linear system in matrix form, connecting the already defined vectors in row space form:

$$\langle p | = \langle \omega | \mathbf{A} \,. \tag{132}$$

Such a linear system can be associated to the most common dual problem in column vector space, just defining

the transpose of the similarity matrix, using the usual definition:

$$\mathbf{Z} = \mathbf{A}^{\mathbf{T}} \rightarrow \forall I = 1, N \wedge J = 1, M\colon z_{JI}^{PB} = a_{IJ}^{BP} \ , \quad (133)$$

and in this manner, the fundamental QQSPR equation is set up, writing a column equivalent dual expression of the former row Eq. (132):

$$\mathbf{Z}|\omega\rangle = |p\rangle \ . \quad (134)$$

As in classical QSPR, the solutions of Eq. (134) may provide the knowledge of the coefficient vector $|\omega\rangle$. However, it must again be stressed that Eq. (134) differs from the classical QSPR setup in the sense that such an equation can be deduced from the quantum mechanical statistical structure, associated to expectation value calculations. In this way, the *causal* connection between molecular structure and molecular properties can be deduced from employing quantum mechanical theoretical fundaments, via the ideas of quantum similarity. The interest of such a relationship lies in the fact that fundamental QQSPR equations can be extended to any quantum object structure and properties. So, obviously, these relationships can be applied to molecular systems as well, provided they can be described as quantum objects, making QQSPR *universal* in the sense that it can be applied, under the same conditions, to any sub-microscopic quantum object set.

Non-Linear QQSPR Equations. In a second remark step, which appears to be sufficiently important as to merit a separate section treatment, the approximate operator linear description (127) may be extended with non-linear terms, which can be easily provided by the nature of the involved quantum object density function tags, which can be founded in turn on the theoretical development of extended wave functions.

In this case, Eq. (127), can be written in a more structured manner as a truncated Taylor series, where only the first two terms are kept for simplicity:

$$W \simeq \sum_{I=1}^{M} \omega_I \rho_I^B + \sum_{P=1}^{M} \sum_{Q \geq P}^{M} \omega_{PQ} \rho_P^B \rho_Q^B + O(3) \ , \quad (135)$$

however, with the potential prospect to add terms up to any order. Equation (135) can be perhaps also considered a simplification of a series involving density functions of growing orders, that is:

$$W \simeq \sum_{I=1}^{M} \omega_I^{(1)} \rho_I^{(1)B} + \sum_{P=1}^{M} \omega_P^{(2)} \rho_P^{(2)B} + O(3) \ . \quad (136)$$

The second-order coefficient set $\{\omega_{PQ}\}$ in Eq. (135), can be also substituted as well, in order to retain a minimal number of unknowns, by products of first-order coefficients, in the following way:

$$\forall P, Q\colon \omega_{PQ} \simeq \omega_P \omega_Q \ . \quad (137)$$

Then, just if this is the case, Eq. (130), transforms into a more computationally convenient form:

$$\forall J = 1, N\colon p_J \simeq \sum_{I=1}^{M} \omega_I \langle \rho_I^B | \Theta | \rho_J^P \rangle + \sum_{P=1}^{M} \sum_{Q \geq P}^{M} \omega_P \omega_Q \langle \rho_P^B \rho_Q^B | \Theta | \rho_J^P \rangle + O(3) \ , \quad (138)$$

Triple Density Quantum Similarity Integrals. The integrals included in the second-order terms of Eq. (138) are *triple density similarity measures*, which can have the form chosen, among many other possibilities, in the following way:

$$\langle \rho_P^B \rho_Q^B | \Theta | \rho_J^P \rangle \equiv \iiint_D \rho_P^B(\mathbf{r}_1) \rho_Q^B(\mathbf{r}_2) \Theta(\mathbf{r}_1; \mathbf{r}_2; \mathbf{r}_3) \rho_J^P(\mathbf{r}_3) \, \mathrm{d}\mathbf{r}_1 \, \mathrm{d}\mathbf{r}_2 \, \mathrm{d}\mathbf{r}_3 \, . \quad (139)$$

Moreover, the usual computational form of the triple density measures can be the one, where the operator becomes unit and all the integrand density functions bear the same variable, so the integral in Eq. (139) acquires a simpler structure, like the triple density overlap integral form:

$$\langle \rho_P^B \rho_Q^B \rho_J^P \rangle \equiv \int_D \rho_P^B(\mathbf{r}) \rho_Q^B(\mathbf{r}) \rho_J^P(\mathbf{r}) \, \mathrm{d}\mathbf{r} \, ; \quad (140)$$

while, first-order similarity measures (129) become, under an equivalent simplification, *overlap-like* integrals:

$$\langle \rho_I^B \rho_J^P \rangle \equiv \int_D \rho_I^B(\mathbf{r}) \rho_J^P(\mathbf{r}) \, \mathrm{d}\mathbf{r} \, . \quad (141)$$

Equations (140) and (141), could be obtained defining the respective weighting operators in terms of an integral operator, involving as many products of Dirac's delta functions as density functions appear into the integrand. For instance, in Eq. (139), the operator $\Theta(\mathbf{r}_1; \mathbf{r}_2; \mathbf{r}_3)$ can be sub-

stituted inside the integral in the following manner:

$$\begin{aligned}\left\langle \rho_P^B \rho_Q^B \rho_J^P \right\rangle \equiv & \int_D \Bigg[\iiint_D \rho_P^B(\mathbf{r}_1)\rho_Q^B(\mathbf{r}_2) \\ & \cdot (\delta(\mathbf{r}_1-\mathbf{r})\delta(\mathbf{r}_2-\mathbf{r})\delta(\mathbf{r}_3-\mathbf{r})) \\ & \cdot \rho_J^P(\mathbf{r}_3)\,\mathrm{d}\mathbf{r}_1\,\mathrm{d}\mathbf{r}_2\,\mathrm{d}\mathbf{r}_3 \Bigg] \mathrm{d}\mathbf{r} \\ = & \int_D \Bigg[\int_D \rho_P^B(\mathbf{r}_1)\delta(\mathbf{r}_1-\mathbf{r})\,\mathrm{d}\mathbf{r}_1 \\ & \cdot \int_D \rho_Q^B(\mathbf{r}_2)\delta(\mathbf{r}_2\mathbf{r})\,\mathrm{d}\mathbf{r}_2 \\ & \cdot \int_D \rho_J^P(\mathbf{r}_3)\delta(\mathbf{r}_3-\mathbf{r})\,\mathrm{d}\mathbf{r}_3 \Bigg] \mathrm{d}\mathbf{r} \\ = & \int_D \rho_P^B(\mathbf{r})\rho_Q^B(\mathbf{r})\rho_J^P(\mathbf{r})\,\mathrm{d}\mathbf{r} \end{aligned} \tag{142}$$

It is, then, straightforward to use the same technique to obtain equations possessing a higher number of density function terms, and so it is easily seen how to take into account and to handle them in the same manner, adding higher order terms within non-linear fundamental QQSPR equations of type (138).

Hansch-Type QQSPR Quadratic Models. In the same manner as above in the linear case, the fundamental quadratic QQSPR Eq. (138) can be simplified, so only the diagonal terms of the initial equation remain. First using just a probe set, taking $B = P$ and then supposing that the remnant equation summation terms are constant under the study of some quantum objects, possessing a great deal of homogeneity. In this case one can write:

$$\begin{aligned} &\forall J = 1, N: \\ &\quad p_J \simeq \beta + \alpha\left\langle \rho_J^P | \Theta | \rho_J^P \right\rangle + \alpha^2 \left\langle \rho_J^P \rho_J^P | \Theta | \rho_J^P \right\rangle + O(3)\,, \end{aligned} \tag{143}$$

which constitutes a quadratic extension of the linear Hansch-type relationships.

Quadratic Fundamental QQSPR Equation in Matrix Form. Having set up in the way outlined above the formal structure of the fundamental QQSPR equations, we now need to discuss its matrix implementation, which is an obligatory step when seeking computational algorithms in practical cases. Two possible equivalent modes will be discussed in this section: the first one corresponds to classical matrix product formalism, while a second part will present an equivalent form just employing inward matrix products. The reason for this second formal presentation is the easiness of setting a general framework up to any approximation order.

a) Classical Form

Equation (138) can be easily written in matrix form. For this purpose it is only necessary to define, besides the column vector of the first-order coefficients:

$$|\omega\rangle = \{\omega_I | I = 1, M\}\,; \tag{144}$$

also, for every quantum object within the probe set, the first-order M-dimensional similarity matrix columns:

$$J = 1, N\colon \left|\mathbf{z}_{IJ}^{(1)}\right\rangle = \left\{ z_{IJ}^{(1)} = \left\langle \rho_I^B | \Theta | \rho_J^P \right\rangle | I = 1, M \right\}, \tag{145}$$

as well as the second-order $(M \times M)$-dimensional similarity matrices:

$$\begin{aligned} &J = 1, N: \\ &\left|\mathbf{Z}_J^{(2)}\right\rangle = \left\{ z_{J;PQ}^{(2)} = \left\langle \rho_J^P | \Theta | \rho_P^B \rho_Q^B \right\rangle | P, Q = 1, M \right\}, \end{aligned} \tag{146}$$

shall be constructed.

Taking the above-defined similarity matrices into account, Eq. (138) can be written as:

$$J = 1, N\colon p_J \simeq \left\langle \mathbf{z}_J^{(1)} | \omega \right\rangle + \left\langle \omega | \mathbf{Z}_J^{(2)} | \omega \right\rangle + O(3)\,, \tag{147}$$

so, collecting the property observable values into a column vector, as already discussed and then, reordering first- and second-order matrix components in the following way:

$$\mathbf{Z}^{(1)} = \left\{ \left|\mathbf{z}_J^{(1)}\right\rangle | J = 1, N \right\}, \tag{148}$$

and

$$\mathbf{Z}^{(2)} = \left\{ \mathbf{Z}_J^{(2)} | J = 1, N \right\}, \tag{149}$$

then the second-order fundamental QQSPR equation becomes a quadratic system of equations in matrix form:

$$|p\rangle \simeq (\mathbf{Z}^{(1)} + [\langle\omega|\mathbf{Z}^{(2)}])|\omega\rangle + O(3)\,. \tag{150}$$

b) Inward Matrix Product Form as a Generalization Device

Alternatively, there is the possibility to express the equations of the previous description by means of *inward matrix products*. The first-order term in Eq. (147) can be expressed within inward product formalism at once, as it is a simple scalar product between the involved vectors, so:

$$J = 1, N\colon \left\langle \mathbf{z}_J^{(1)} | \omega \right\rangle \equiv \left\langle \left|\mathbf{z}_J^{(1)}\right\rangle * |\omega\rangle \right\rangle, \tag{151}$$

while the second-order term may be expressed in inward product form with the aid of the coefficient vector tensor product, forming a square $(N \times N)$ matrix:

$$\mathbf{W} = |\omega\rangle \otimes |\omega\rangle \equiv \{w_{IJ} = \omega_I \omega_J | \forall I, J = 1, N\}, \quad (152)$$

so, one can write then the quadratic term of Eq. (150) as an inward matrix product too:

$$J = 1, N:$$
$$\langle \omega | \mathbf{Z}_J^{(2)} | \omega \rangle \equiv \langle \mathbf{Z}_J^{(2)} * \mathbf{W} \rangle = \langle \mathbf{Z}_J^{(2)} * (|\omega\rangle \otimes |\omega\rangle) \rangle , \quad (153)$$

and consequently Eq. (147), can be rewritten as:

$$J = 1, N:$$
$$p_J \simeq \langle |\mathbf{z}_J^{(1)}\rangle * |\omega\rangle \rangle + \langle \mathbf{Z}_J^{(2)} * (|\omega\rangle \otimes |\omega\rangle) \rangle + O(3) . \quad (154)$$

Inward Matrix Product Formalism of Fundamental QQSPR Equation nth Order Terms. Both, classical and inward product, formalisms are equivalent; however, the inward product Eq. (154), permits one to easily imagine any sequence of corrections into the fundamental QQSPR equation, up to any arbitrarily chosen *n*th order term, just writing:

$$J = 1, N: p_J \simeq \sum_{R=1}^{n} \left\langle \mathbf{Z}_J^{(R)} * \left(\overset{R}{\underset{S=1}{\otimes}} |\omega\rangle \right) \right\rangle + O(n+1), \quad (155)$$

where the leading equation terms are $\{\mathbf{Z}_J^{(R)} | J = 1, N\}$ the *R*th order similarity matrices, which can be constructed as:

$$J = 1, N: \mathbf{Z}_J^{(R)} = \{\mathbf{z}_{J;S(\mathbf{i})}^{(R)} = \langle \rho_J^P | \Theta | \rho_{S_1}^B \rho_{S_2}^B \dots \rho_{S_R}^B \rangle | \forall \alpha = 1, R: L_\alpha \in \{1, 2, \dots, M\}\} , \quad (156)$$

with the index set: $S(\mathbf{i}) = \{S_1, S_2, \dots, S_R\}$ formed by any of the M^R combinations with repetition of *R* elements chosen within the *M* integers and, finally, the *R*th order tensor products of the coefficient vector are noted as: $\otimes_{S=1}^{R} |\omega\rangle$.

Stochastic Transformations. We will now discuss a third remark step, dealing with the stochastic transformation of similarity matrices, because it also merits a separate section. Recently, several studies have dealt with stochastic transformations of the fundamental QQSPR equation in linear symmetric form, that is: using $B = P$.

At the light of the previous manipulation presented in this study, the stochastic structure transformation of the fundamental QQSPR equation has to be performed, at any operator-equation approximation level, using the possibility to compute the sum of the elements of the *R*th order similarity matrices as have been previously defined in Eq. (156), that is:

$$\sigma_J^{(R)} = \langle \mathbf{Z}_J^{(R)} \rangle = \sum (\mathbf{i}) z_{J;S(\mathbf{i})}^{(R)} , \quad (157)$$

where a nested summation symbol $\sum(\mathbf{i})$ has been employed in order to indicate the nested sums over the *R* indices, represented by the index sets: $S(\mathbf{i}) = \{S_1, S_2, \dots, S_R\}$. Using the sum of the similarity matrix elements (157), then the elements of the new matrices scaled by this sum become scaled in turn as follows:

$$\mathbf{S}_J^{(R)} = (\sigma_J^{(R)})^{-1} \mathbf{Z}_J^{(R)} , \quad (158)$$

and the new *R*th order stochastic similarity matrices behave as a discrete probability distribution, as: $\forall S(\mathbf{i})$: $z_{J;S(\mathbf{i})}^{(R)} \in \mathbf{R}^+ \to s_{J;S(\mathbf{i})}^{(R)} \in \mathbf{R}^+$ and besides:

$$\langle \mathbf{S}_J^{(R)} \rangle = 1 . \quad (159)$$

Both properties can be cast into a unique *convex condition* symbol:

$$K(\mathbf{S}_J^{(R)}) = \{\forall S(\mathbf{i}): s_{J;S(\mathbf{i})}^{(R)} \in \mathbf{R}^+ \wedge \langle \mathbf{S}_J^{(R)} \rangle = 1\} . \quad (160)$$

So, in this way, the stochastic matrix set: $\mathbf{S} = \{\mathbf{S}_J^{(R)} | R = 1, n\}$ can be considered, up to *n*th order, as a set of M^R-dimensional *unit shell* elements, belonging to some *vector semispace* with the same dimensions. In these circumstances one can consider the fundamental QQSPR Eq. (155) as to be written:

$$J = 1, N: p_J = \sum_{R=1}^{n} \left\langle \mathbf{S}_J^{(R)} * \left(\overset{R}{\underset{S=1}{\otimes}} |\omega\rangle \right) \right\rangle + O(n+1), \quad (161)$$

where everything is the same as in the former Eq. (155), except for the similarity matrix set, which has been substituted by the stochastic matrices (158).

The coefficient vector has been left unchanged, but evidently its character could be no longer the same as in Eq. (155). However, the nature of the coefficient vector can be more precise in this case of the fundamental QQSPR stochastic Eqs. (161). This is due to the characteristic convex condition properties, which possess the semispace unit shell elements obtained transforming the similarity matrices.

In fact, the stochastic similarity matrix set: $\{\mathbf{S}_J^{(R)}|R = 1, n\}$, so naturally obtained from the original similarity matrix set, can be interpreted as a sequential discrete representation of the continuous normalized density function, associated to the involved Jth quantum object. Then, from the quantum mechanical point of view, the whole stochastic matrix set can be viewed as a discrete quantum object tag collection. Thus, in this case, the tensor products of the coefficient vector can be easily considered as arrays of convex sets, that is:

$$\mathbf{W}^{(R)} = \bigotimes_{S=1}^{R} |\omega\rangle = \{w_{S(\mathbf{i})}^{(R)}\} \rightarrow \langle \mathbf{W}^{(R)} \rangle = \sum (\mathbf{i}) w_{S(\mathbf{i})}^{(R)} = 1 \wedge \forall S(\mathbf{i}) \colon w_{S(\mathbf{i})}^{(R)} \in \mathbf{R}^{+} \tag{162}$$

because, whenever the generating coefficient vector is a convex vector, that is, fulfilling the convex conditions:

$$K(|\omega\rangle) = \left\{ \forall I \colon \omega_I \in \mathbf{R}^{+} \wedge \langle |\omega\rangle \rangle = \sum_I \omega_I = 1 \right\}, \tag{163}$$

then, any tensor product of the convex vector $|\omega\rangle$ fulfils: $K\left(\otimes_{S=1}^{R} |\omega\rangle\right)$. Indeed, if convex conditions (163) hold, then it is easy to see that convex conditions are present within any arbitrary order tensor product of the coefficient vector, as shown in the following deduction:

$$\begin{aligned} \mathbf{W}^{(R)} = \bigotimes_{S=1}^{R} |\omega\rangle = \{w_{S(\mathbf{i})}^{(R)} = \omega_{S_1} \omega_{S_2} \ldots \omega_{S_R} \in \mathbf{R}^{+}\} \\ \wedge \langle \mathbf{W}^{(R)} \rangle = \sum (\mathbf{i}) w_{S(\mathbf{i})}^{(R)} = \left(\sum_I \omega_I \right)^R \\ = (\langle |\omega_I\rangle \rangle)^R = (1)^R = 1 \\ \rightarrow K\left(\bigotimes_{S=1}^{R} |\omega\rangle \right) \equiv K(\mathbf{W}^{(R)}) \,. \end{aligned} \tag{164}$$

Variational QQSPR. So far the fundamental QQSPR equation has been solved using the usual strategy associated to classical QSPR. Equations (134), (150) or (161) as in classical terms, can be solved for the coefficient vector $|\omega\rangle$. As has been previously commented this is done, by substituting in the expectation value expression (138) the vector $|\pi\rangle$ by an experimental property vector $|p\rangle$, associated to the probe quantum object set P. The result will be obtained in the same way as in classical QSPR, but using the quantum similarity matrices as molecular descriptors. However, it can be proven that the fundamental QQSPR equation can be solved within the usual quantum variational procedures.

a) Similarity Matrix Unrestricted Variational Treatment

For such a purpose it is sufficient to rewrite the second-order expectation value Eq. (138) as:

$$\forall J = 1, N \colon \langle \pi_J \rangle \simeq \sum_{P=1}^{M} \omega_P z_{J;P}^{(1)} + \sum_{P=1}^{M} \sum_{Q \geq P}^{M} \omega_P \omega_Q z_{J;PQ}^{(2)} + O(3) \tag{165}$$

then, considering every quantum object expectation value as a variational function of the parameters within the coefficient vector $|\omega\rangle$, the resulting expression can be varied, taking into account that the density functions, supposedly obtained by quantum mechanical procedures, no longer need variation. In this way, every Jth quantum object will have to possess a specific coefficient vector $|\omega\rangle$, which can be thus named as $|\omega_J\rangle$. That is:

$$\forall J = 1, N \colon \delta \langle \pi_J \rangle \simeq \sum_{P=1}^{M} \delta\omega_P z_{J;P}^{(1)} + 2 \sum_{P=1}^{M} \sum_{Q \geq P}^{M} \omega_P \omega_Q z_{J;PQ}^{(2)} + O(3) \,, \tag{166}$$

then, using the variation condition for the Jth quantum object:

$$\delta \langle \pi_J \rangle = 0 \,, \tag{167}$$

is obtained:

$$\forall J = 1, N \colon 0 \simeq \sum_{P=1}^{M} \delta\omega_P z_{J;P}^{(1)} + 2 \sum_{P=1}^{M} \sum_{Q \geq P}^{M} \delta\omega_P \omega_Q z_{J;PQ}^{(2)} + O(3) \,, \tag{168}$$

which can be rewritten as:

$$\forall J = 1, N \wedge P = 1, M \colon 0 \simeq z_{J;P}^{(1)} + 2 \sum_{Q=1}^{M} \omega_Q z_{J;PQ}^{(2)} + O(3) \,. \tag{169}$$

This last equation can be expressed in matrix form, using the appropriate similarity matrices as previously defined in Eqs. (145) and (146):

$$\forall J = 1, N \colon \mathbf{z}_J^{(1)} + 2\mathbf{Z}_J^{(2)} |\omega_J\rangle = 0 \,, \tag{170}$$

thus, the specific coefficients for each quantum object may be computed as:

$$\forall J = 1, N\colon |\omega_J\rangle = -\frac{1}{2}\left[\mathbf{Z}_J^{(2)}\right]^{-1}\mathbf{z}_J^{(1)} . \tag{171}$$

This is the same as associating a particular operator W to each quantum object, and such a result is not too surprising a feature, as the operator W can be easily supposed to vary from one quantum object to another, in the same way as Hamilton operators do. The variational expectation value for the Jth object could be obtained in this case as:

$$\langle \pi_J \rangle \simeq \left\langle \omega_J | \mathbf{z}_J^{(1)} \right\rangle + \left\langle \omega_J | \mathbf{Z}_J^{(2)} | \omega_J \right\rangle + O(3) . \tag{172}$$

Using Eq. (171) into Eq. (172), the following expectation value final optimal form will result:

$$\langle \pi_J \rangle \simeq -\frac{1}{4}\left\langle \mathbf{z}_J^{(1)} \middle| \left[\mathbf{Z}_J^{(2)}\right]^{-1} \middle| \mathbf{z}_J^{(1)} \right\rangle + O(3) . \tag{173}$$

b) Expectation Versus Experimental Values
Then, the set of stationary expectation values $|\pi\rangle$ can be compared with the experimental value vector $|p\rangle$, in such a way as to have:

$$|p\rangle = a + b|\pi\rangle , \tag{174}$$

$\{a, b\}$ being some origin and scale parameters, respectively. They can be obtained by the usual well-known regression techniques.

c) Algorithm for Unrestricted Variational QQSPR
Once the set of coefficients $\{a, b\}$ is obtained by using Eq. (174) for a given probe quantum object set, the property expectation value $\langle \pi_K \rangle$ of any new quantum object K, say, with known density function ρ_K, can be employed to estimate the experimental value ρ_K of the quantum object studied property, by using the following steps:

1. Compute: $\left\{\mathbf{z}_K^{(1)}; \mathbf{Z}_K^{(2)}\right\}$ using the basis set B.
2. Evaluate: $\langle \pi_K \rangle \simeq -\frac{1}{4}\left\langle \mathbf{z}_K^{(1)} \middle| \left[\mathbf{Z}_K^{(2)}\right]^{-1} \middle| \mathbf{z}_K^{(1)} \right\rangle + O(3)$
3. Obtain the estimated property: $p_K = a + b\langle \pi_K \rangle$.

Stochastic Similarity Matrices Restricted Variational Treatment. Of course, all that has been said up to now in this section remains valid for stochastic similarity matrices: $\left\{\mathbf{s}_K^{(1)}; \mathbf{S}_K^{(2)}\right\}$, they just have to be used instead of the similarity matrix pair: $\left\{\mathbf{z}_K^{(1)}; \mathbf{Z}_K^{(2)}\right\}$ in the above algorithm. However, the stochastic case may be interesting if the coefficient set $|\omega\rangle$ can be obtained obeying convex conditions as a restriction, so that the previous unrestricted variation algorithm may no longer be applicable.

Expectation Value Jacobi Rotations Variational Form. To obtain the desired restricted variation over the coefficient vector involved in expectation value expressions, a similar procedure as the one employed in developing the ASA technique [75,76,77,78,79,80,81,82,83] could be easily set up to perform the variational computation over Eq. (165), but taking into account the additional restriction of obtaining a convex vector, as a result of the optimization process.

a) Preliminary Considerations
When this option as discussed above is chosen, it is only necessary to express the operator W variational coefficients with the aid of a new free normalized auxiliary vector; in order to ensure the convex conditions hold throughout the entire optimization process, that is:

$$\begin{aligned} |\omega\rangle = |x\rangle * |x\rangle \wedge \langle x|x\rangle = 1 \rightarrow \langle |\omega\rangle \rangle \\ = \sum_I \omega_I = \sum_I x_I^2 = 1 \wedge \forall I\colon \omega_I = x_I^2 \in \mathbf{R}^+ . \end{aligned} \tag{175}$$

After this consideration, it is only necessary to obtain the variation of Eq. (165), by applying norm conserving, orthogonal elementary Jacobi rotations [148] into the auxiliary vector $|x\rangle$ element pairs, in order to arrive at an expression, depending on the elementary Jacobi rotation angle, which could be easily optimized later on.

An interesting point at this stage is to realize that such a restricted variational procedure can be applied to higher order equations, with orders larger than the ones studied up to now. This is due to the fact that Jacobi rotations over the auxiliary vector just change a couple of the coefficient auxiliary vector elements each time an elementary Jacobi rotation is performed, and the same occurs with the coefficient vector. This knowledge of the coefficient vector variation can be easily brought into the tensor products and worked out up to any tensor order.

The rest becomes a procedure with somehow a growing technical computational complexity, but defined within a well-structured theoretical background algorithm.

b) Elementary Jacobi Rotations Algorithm Scheme
Elementary Jacobi rotations need the cosine, c, and the sine, s, of a rotation angle. These involved trigonometric functions fulfil the usual convex relationship: $c^2 + s^2 = 1$. When acting over a vector, the Jacobi rotations will change two vector components, the Kth

and Lth, say, leaving the remaining components as they are:

$$|x\rangle = \begin{pmatrix} \cdots \\ x_K \\ \cdots \\ x_L \\ \cdots \end{pmatrix} \rightarrow \begin{pmatrix} \cdots \\ cx_K - sx_L \\ \cdots \\ sx_K + cx_L \\ \cdots \end{pmatrix}$$

$$\Rightarrow |w\rangle = \begin{pmatrix} \cdots \\ x_K^2 \\ \cdots \\ x_L^2 \\ \cdots \end{pmatrix} \rightarrow \begin{pmatrix} \cdots \\ (cx_K - sx_L)^2 \\ \cdots \\ (sx_K + cx_L)^2 \\ \cdots \end{pmatrix} . \quad (176)$$

It is easy to obtain the variation in the coefficient vector due to an elementary Jacobi rotation as:

$$|\delta\omega\rangle = v_{KL} \begin{pmatrix} \cdots \\ -1 \\ \cdots \\ +1 \\ \cdots \end{pmatrix} = v_{KL}(|e_L\rangle - |e_K\rangle) , \quad (177)$$

where $\{|e_K\rangle, |e_L\rangle\}$ are the corresponding canonical basis set vectors. The scalar coefficient v_{KL} possesses the form:

$$v_{KL} = s^2\left(x_K^2 - x_L^2\right) + 2csx_Kx_L . \quad (178)$$

Then, employing this result in the equivalent expression of Eq. (147), but written in expectation value matrix form, the following can be deduced:

$$\langle\delta\pi\rangle = \langle\delta\omega|(|\mathbf{z}^{(1)}\rangle + 2\mathbf{Z}^{(2)}|\omega\rangle) + \langle\delta\omega|\mathbf{Z}^{(2)}|\delta\omega\rangle , \quad (179)$$

where the quantum object subindex has been taken out to simplify the notation. Then, upon substituting the coefficient vector variation:

$$\langle\delta\pi\rangle = v_{KL}\left[\left(z_L^{(1)} - z_K^{(1)}\right) + 2\sum_I \omega_I\left(Z_{IL}^{(2)} - Z_{IK}^{(2)}\right)\right] + v_{KL}^2\left(Z_{KK}^{(2)} + Z_{LL}^{(2)} - 2Z_{KL}^{(2)}\right) \quad (180)$$

which, upon equalization to zero and terms rearrangement, can be expressed as a second-order equation on the elementary Jacobi rotation sine and cosine:

$$As^2 + Bsc + \beta = 0 , \quad (181)$$

with the coefficients A and B defined as:

$$\begin{aligned} A &= \alpha(\omega_K - \omega_L) \\ B &= 2\alpha x_Kx_L \end{aligned} \quad (182)$$

and, besides, the parameters are constructed by the elements of the similarity matrices in the following way:

$$\begin{aligned} \alpha &= Z_{KK}^{(2)} + Z_{LL}^{(2)} - 2Z_{KL}^{(2)} \\ \beta &= \left(z_L^{(1)} - z_K^{(1)}\right) + 2\sum_I \omega_I\left(Z_{IL}^{(2)} - Z_{IK}^{(2)}\right) . \end{aligned} \quad (183)$$

Higher Order Stochastic Expectation Value Variational Treatment.

a) General Comments

Whenever Eq. (161) is studied, after being conveniently modified for the expectation values form,

$$\forall J = 1, N\colon \langle\pi_J\rangle = \sum_{R=1}^{n} \left\langle \mathbf{S}_J^{(R)} * \mathbf{W}^{(R)} \right\rangle + O(n+1) \quad (184)$$

the obvious fact appears that the variation will affect just the Rth order tensor products $\mathbf{W}^{(R)}$ of the coefficient vector. So it can be written, dropping the quantum object subindex J just for convenience, as before:

$$\langle\delta\pi\rangle = \sum_{R=1}^{n} \langle \mathbf{S}^{(R)} * \delta\mathbf{W}^{(R)} \rangle + O(n+1) , \quad (185)$$

so the relevant variation will be associated to the terms $\delta\mathbf{W}^{(R)}$, which can be easily written, using a tensor notation as:

$$\begin{aligned} \delta\mathbf{W}^{(R)} &= \delta\left(\overset{R}{\underset{S=1}{\otimes}} |\omega\rangle \right) \\ &= \sum_{S=1}^{R} \binom{R}{S} \left[\left(\overset{R-S}{\underset{P=1}{\otimes}} |\omega\rangle \right) \otimes \left(\overset{S}{\underset{Q=1}{\otimes}} |\delta\omega\rangle \right) \right] , \end{aligned} \quad (186)$$

but being the definition of the coefficient vector variation, upon Jacobi rotations, well known from Eq. (177), it can be written:

$$\delta\mathbf{W}^{(R)} = \sum_{S=1}^{R} \binom{R}{S} (v_{KL})^S \cdot \left[\left(\overset{R-S}{\underset{P=1}{\otimes}} |\omega\rangle \right) \otimes \left(\overset{S}{\underset{Q=1}{\otimes}} [(|e_L\rangle - |e_K\rangle)] \right) \right] . \quad (187)$$

So in this way, the restricted variation of the expectation value QQSPR equations, using elementary Jacobi rotations, is clearly defined up to any order.

b) A Computational Detail Concerning Tensor Products of the Difference of Two Canonical Vectors

The tensor product of the difference between the pair of canonical basis set vectors:

$$|e_L\rangle - |e_K\rangle = \begin{pmatrix} \cdots \\ -1 \\ \cdots \\ +1 \\ \cdots \end{pmatrix} \equiv |L\rangle - |K\rangle \equiv |L-K\rangle\,, \quad (188)$$

which appears in Eq. (187), may be expressed in terms of a nested summation symbol. For example, up to second order the sum of the four tensor terms is readily written as:

$$|L-K\rangle \otimes |L-K\rangle = |L\otimes L\rangle - |L\otimes K\rangle - |K\otimes L\rangle + |K\otimes K\rangle \quad (189)$$

with the obvious meaning for the involved tensors:

$$|L\otimes L\rangle = |e_L\rangle \otimes |e_L\rangle = \mathbf{E}_{LL} = \{e_L L; PQ = \delta_{LP}\delta_{LQ}\} \quad (190)$$

and so on.
In general, up to *S*th order:

$$\bigotimes_{Q=1}^{S} |L-K\rangle = \sum (\mathbf{i})\sigma(Q(\mathbf{i}))|Q(\mathbf{i})\rangle\,, \quad (191)$$

where $Q(\mathbf{i}) = \{Q_1 \otimes Q_2 \ldots \otimes Q_S\}$ is any of the possible 2^n combinations with repetition of the indices K and L, the symbol $|Q(\mathbf{i})\rangle$ meaning a tensor product of the initial canonical basis set vectors with such an index repetition. That is: an object equivalent to a canonical hypermatrix, whose elements are all zero, except the one with indices associated to those entering the set. Also $\sigma(Q(\mathbf{i}))$ corresponds to the sign, associated to the fact that the index K appears in $Q(\mathbf{i})$ an even, $\sigma(Q(\mathbf{i})) = +1$, or odd, $\sigma(Q(\mathbf{i})) = -1$, number of times.

QQSPR Operators, Quantum Similarity Measures and the Fundamental QQSPR Equation

The correspondence principle in quantum theory furnishes the rules to construct Hermitian operators, whose expectation values can be associated with the experimental outcomes of submicroscopic system observables. However, as has been previously commented, for some observables of complex submicroscopic systems, like some biological activities of pharmaceutical interest, the correspondence principle cannot be applied. The construction of the QQSPR operators and the attached fundamental QQSPR equation provide the possibility to attach an approximate quantum mechanical operator to estimate expectation values for these cases.

The QQSPR Operator. The fundamental QQSPR equation arises when from the known quantum objects, belonging to some quantum object set; one realizes that their density function tags: $\{\rho_I(\mathbf{r})\}$ can be used to construct a QQSPR operator in the form:

$$\Omega(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3, \ldots) = w_0\Theta_0(\mathbf{r}_1) + \sum_I w_I\rho_I(\mathbf{r}_2)\Theta_1(\mathbf{r}_1, \mathbf{r}_2) + \sum_I \sum_J w_I w_J \rho_I(\mathbf{r}_2)\rho_J(\mathbf{r}_3)\Theta_2(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3) + O(3)\,, \quad (192)$$

in the Eq. (192) above, w_0 is an arbitrary constant; $\{\Theta_\omega(\mathbf{R})|\omega = 0, 1, 2, \ldots\}$ is a known positive definite operator set, acting as a weight for each term development; finally, $\{w_I\}$ a set of unknown parameters which shall be determined through the fundamental QQSPR equation as will be explained below.

Thus, the structure of a QQSPR operator like the one defined in Eq. (192) has to be seen as the first step of an algorithm permitting the construction of approximate quantum mechanical operators, associated in turn to some observables of complex submicroscopic systems, whose nature do not permit the application of the correspondence principle to construct Hermitian operators for the evaluation of observable values.

The Expectation Values of the QQSPR Operator. In order to determine the parameter set $\{w_I\}$, defining in this way the QQSPR operator as written in Eq. (192), it is just necessary to compute the set of expectation values over the elements of a quantum object set which belong to the core set *C*, constituted by the core molecules or C-m. Besides a well-defined structure and a known density function, as members of a quantum object set, the C-m are supposed to possess an element of a known property set $P = \{p_K\}$, attached to each one.

In this way one can express every known property of the C-m elements as the expectation value of some QQSPR operator:

$$p_K \approx \langle \Omega\rho_K\rangle = w_0\langle\Theta_0\rho_K\rangle + \sum_I w_I\langle\rho_I\Theta_1\rho_K\rangle + \sum_I w_I w_J\langle\rho_I\rho_J\Theta_2\rho_K\rangle + O(3)\,. \quad (193)$$

Zero-th Order Term. When describing the expectation values of the C-m as computed in Eq. (193), one can con-

sider first the Zero-th order term:

$$\theta_K[\Theta_0] = w_0\langle\Theta_0\rho_K\rangle = w_0\int_D \Theta_0(\mathbf{r}_1)\rho_K(\mathbf{r}_1)\,\mathrm{d}\mathbf{r}_1\ ,$$

as being a constant for each C-m, which can be used as an origin shift of the C-m property tags, thus the Zero-th order term: $w_0\Theta_0(\mathbf{r})$ appearing in the above operator definition acts as a gauge. Choosing the Zero-th order operator as the unit, this term becomes proportional to the number of electrons of the C-m considered:

$$\theta_K[I] = w_0\langle\rho_K\rangle = w_0\int_D \rho_K(\mathbf{r}_1)\,\mathrm{d}\mathbf{r}_1 = w_0 N_K\ .$$

In case shape functions, defined as:

$$\sigma_K(\mathbf{r}) = N_K^{-1}\rho_K(\mathbf{r}) \rightarrow \int_D \sigma_K(\mathbf{r})\,\mathrm{d}\mathbf{r} = 1\ ,$$

are employed in the QQSPR operator definition (192) and in the expectation value expression (193), then the Zero-th order contribution to the expectation values $\theta_K[I]$ is a constant for all C-m.

The Zero-th order term can be omitted if it is no longer necessary to shift the property values of the C-m.

First- and Second-Order Expectation Value Terms. The first-order term of the expectation value Eq. (193) contains quantum similarity measure integrals among pairs of density function tags of the C-m, which have been defined a long time ago as:

$$z_{IK}[\Theta_1] = \langle\rho_I\Theta_1\rho_K\rangle = \int_D\int_D \rho_I(\mathbf{r}_2)\Theta_1(\mathbf{r}_1,\mathbf{r}_2)\rho_K(\mathbf{r}_1)\,\mathrm{d}\mathbf{r}_1\,\mathrm{d}\mathbf{r}_2\ ,$$

and in the second-order term the triple density quantum similarity measures appear, defined as well as:

$$Z_{IJK}[\Theta_2] = \langle\rho_I\rho_J\Theta_2\rho_K\rangle = \int_D\int_D\int_D \rho_I(\mathbf{r}_2)\rho_J(\mathbf{r}_3)\cdot\Theta_2(\mathbf{r}_1,\mathbf{r}_2,\mathbf{r}_3,)\rho_K(\mathbf{r}_1)\,\mathrm{d}\mathbf{r}_1\,\mathrm{d}\mathbf{r}_2\,\mathrm{d}\mathbf{r}_3\ .$$

Fundamental QQSPR Equation Setup. The expectation values of the QQSPR operator, as described in Eq. (193), can be collected in a column vector providing the fundamental QQSPR equation:

$$|\mathbf{p}\rangle \approx |\theta\rangle + \mathbf{Z}_1|\mathbf{w}\rangle + \langle\mathbf{w}|\mathbf{Z}_2|\mathbf{w}\rangle + O(3)\ , \qquad (194)$$

where in Eq. (194) the following compact symbols have been used: $|\mathbf{p}\rangle = \{p_K\}$ is the C-m properties vector, $|\theta\rangle = \{\theta_K\}$ is the completely determined gauge shift vector, $\{\mathbf{Z}_\omega|\omega = 1, 2, \ldots\}$ is a matrix set containing the quantum similarity measures, for instance: $\mathbf{Z}_1 = \{z_{IK}\}; \mathbf{Z}_2 = \{z_{IJK}\}; \ldots$, and $|\mathbf{w}\rangle = \{w_I\}$ is a column vector bearing the unknown coefficients, which define explicitly the QQSPR operator.

The easiest way to obtain the unknown coefficients $|\mathbf{w}\rangle = \{w_I\}$ is obviously the linear equation contained in the fundamental QQSPR Eq. (194); that is, they can be evaluated by solving:

$$|\mathbf{p}\rangle = |\theta\rangle + \mathbf{Z}_1|\mathbf{w}\rangle \rightarrow |\mathbf{w}\rangle = (\mathbf{Z}_1)^{-1}(|\mathbf{p}\rangle - |\theta\rangle)\ . \qquad (195)$$

However, Eq. (195) has no predictive power whatsoever. This is so because the first-order similarity matrix $\mathbf{Z}_1$ has to be chosen positive definite by construction, therefore the coefficient vector has a unique determined form.

By predictive power is meant here the possibility to compute the value of the property, which precisely defines the C-m set, for an also known quantum object, which as such possesses well-defined structure and density function, but belongs to the Unknown property molecular set: U, whose elements are made by unknown property molecules or quantum objects, the U-m.

In the last years, since the description of quantum similarity measures, the predictive power of the information contained in the similarity matrices set has been manipulated in the classical QSPR way. For example, using similarity matrices principal components, and finding with them a QSAR model, usually multilinear. This multilinear model can be employed, afterwards, to estimate U-m properties. This amounts to the same as considering the similarity matrices as a source of molecular parameters to construct empirical QSPR.

However, there is a possible way to use the system (195) for predicting properties of U-m without further considerations than the involved algebraic procedures. The possible QQSAPR prediction algorithms will be developed in a separate section.

Evaluation of Unknown Molecular Properties as Expectation Values

In general, one can choose any molecular structure U, possessing an unknown value of the property needed to build up the *core set* triads. Thereafter, one can call such a QO the *U-molecule* or *U-m*, for sake of simplicity. The *U-molecule* can supposedly be associated to a corresponding density function:ρ_U too. Hence, the *U-m* can be certainly considered as a QO. On the other hand, one must keep in mind that, by construction of the QSPR problem, the property lacking in the information about U has to be

already known for all elements of the *core set*. One can easily express an approximate value of the *U-m* unknown property through the simplified Minkowski norm:

$$\langle \Omega[\rho_U]\rangle = \int_D \Omega[\rho_U]\,\mathrm{d}V \approx \langle \sigma[\rho_U]\rangle + \sum_{P=1}^{N}\sum {}_P(\mathbf{i})x(\mathbf{i})\langle \rho(\mathbf{i})[\rho_U]\rangle + O(N+1)\,, \tag{196}$$

provided that the set of coefficients $\{x_I\}$ is well-defined.

The Minkowski norm in Eq. (196) can be computed in more sophisticated ways, using a known positive definite operator, W say, as a weight in the expectation value definition: $\langle \Omega W[\rho_U]\rangle$, producing weighted quantum similarity measures of type: $\langle \rho(\mathbf{i})W[\rho_U]\rangle$ in the right part of the expression (196). In order to simplify the formalism, here the convention: $W = I$, has been adopted.

Within the QQSPR problem settings, the set of coefficients: $\{x_I\}$, in Eq. (196), which can be ordered as a column vector: $|\mathbf{x}\rangle = \{x_I\}$, is unknown beforehand, but can be already computed from the first-order approach using the *core set* known property values, as will be discussed below.

Quantum Similarity Matrices in the Construction of First-Order QSPR Operators and the Definition of Discrete QOS. The first-order approach of the QSPR operator for the *core set* known molecular property tag set: $\Pi = \{\pi_I\}$ generates the following equation collection:

$$\forall I = 1, n\colon p_I = \pi_I - \langle \sigma[\rho_I]\rangle \approx \sum_J x_J\langle \rho_J[\rho_I]\rangle = \sum_J x_J z_{JI}\,. \tag{197}$$

The set of integrals:

$$\left\{\langle \rho_J[\rho_I]\rangle = \int_D \rho_J\rho_I\,\mathrm{d}V = z_{JI} = z_{IJ} = \int_D \rho_I\rho_J\,\mathrm{d}V = \langle \rho_I[\rho_J]\rangle\right\}\,,$$

appearing in Eqs. (197) can be ordered into a $(n \times n)$ symmetric array, constructing in this way the *quantum similarity matrix*: $\mathbf{Z} = \{z_{IJ}\}$ (QSM). In turn, the ordered set of shifted properties: $\{p_I\}$ can form a $(n \times 1)$ column vector: $|\mathbf{p}\rangle = \{p_I\}$. Therefore, the equation set (197) is simply a linear system, which will be discussed next, in order to describe its possible use for evaluating *U-m* unknown molecular properties.

Empirical QSPR. In the empirical QSPR problems, the equivalent matrix to the QSM of the QQSPR framework, as described in Sect. "Quantum Similarity", can be obtained in the following manner. Suppose that every molecular structure of M possesses an arbitrarily chosen empirical descriptor vector, in that way:

$$\forall m_I \in \mathrm{M} \to \exists|\mathbf{d}_I\rangle \in \mathrm{D} \wedge \forall I\colon m_I \leftrightarrow |\mathbf{d}_I\rangle\,,$$

then the descriptor set D acts as a tag set to construct an empirical discrete tagged set:

$$\mathrm{Q_D} = \mathrm{M} \times \mathrm{D}\,,$$

such that:

$$\forall \gamma_I \in \mathrm{Q_D} \to \gamma_I = (m_I; |\mathbf{d}_I\rangle) \wedge m_I \in \mathrm{M}; |\mathbf{d}_I\rangle \in \mathrm{D}\,.$$

The discrete tag set D of molecular descriptors can be considered, in turn, as a linearly independent subset of cardinality n belonging to some real m-dimensional column vector space, that is: $\mathrm{D} \subset V_m(\mathbf{R})$. The linear independence of the set D is strictly necessary to construct a matrix comparable in properties to QSM, and in this way, each molecule becomes independently described from the rest. With this information in mind it is easy to construct, $\mathbf{S}_\mathrm{D}$, a symmetric $(n \times n)$ matrix bearing analogous characteristics as the QSM:

$$\forall\{|\mathbf{d}_I\rangle; |\mathbf{d}_J\rangle\} \in \mathrm{D}\colon \mathbf{S}_\mathrm{D} = \{s_{D;IJ} = \langle \mathbf{d}_I|\mathbf{d}_J\rangle = \mathbf{d}_J|\mathbf{d}_I\rangle = s_{D;JI}\} = \{|\mathbf{s}_{\mathrm{D};I}\rangle = \{s_{D;JI}\}\}\,. \tag{198}$$

In fact, constructing the $(m \times n)$ matrix: $\mathbf{D} = \{|\mathbf{d}_I\rangle\}$, whose columns are the elements of the empirical descriptor set D, then the matrix $\mathbf{S}_\mathrm{D}$ can also be defined as the product:

$$\mathbf{S}_\mathrm{D} = \mathbf{D}^\mathrm{T}\mathbf{D}\,,$$

where $\mathbf{D}^\mathrm{T} = \{\langle \mathbf{d}_I|\}$ is the $(n \times m)$ transpose of matrix $\mathbf{D}$, whose rows are the descriptor vectors ordered in such a way. It is easy to see that matrix $\mathbf{S}_\mathrm{D}$, defined in this manner, is coincident with the Gramian matrix of the tag set D. In order to comply with the same standard properties as the QSM the matrix $\mathbf{S}_\mathrm{D}$ has to fulfil: $\mathrm{Det}\,|\mathbf{S}_\mathrm{D}| > 0$. If this is the case, a discrete *empirical object set* $\mathrm{Q_S}$ can be defined as:

$$\mathrm{Q_S} = \mathrm{M} \times \mathbf{S}_\mathrm{D}\,,$$

in close resemblance to the discrete quantum object set $\mathrm{Q_Z}$ described in Eq. (42).

Finally, one shall comment now that, as a consequence of this definition of the set $\mathrm{Q_S}$, the presentation and discussion about the following procedures, which will be studied

in this paper for QSM, can also be applied to the Gramian matrices, associated to empirical descriptor tag sets and thus to the classical QSPR problem.

However, the different background between quantum and empirical points of view induces the necessary emergence of the following considerations. The QQSPR equations are deductible from the usual quantum theoretical considerations; within the same context, they can be easily generalized to contain higher approximation orders. Therefore, the QQSPR equations of any order can certainly possess in general some causal background; while, except for very particular cases, empirical QSPR equations remain arbitrarily constructed and without a clear causal fundament.

First-Order Fundamental QQSPR (FQQSPR) Equation

The analysis of the QQSPR problem can start with the first order or linear fundamental QQSPR equation, involving the *core set*, formed with the molecules of the associated DQOS, which are also linked with known values of some property, according to the considerations noted above.

One can write Eq. (197) in a compact matrix form:

$$\mathbf{Z}|\mathbf{x}\rangle = |\mathbf{p}\rangle \, ; \tag{199}$$

Where the matrix $\mathbf{Z}$ is the already described symmetric QSM, $|\mathbf{p}\rangle$ is the known *core set* property vector and $|\mathbf{x}\rangle$ is a $(n \times 1)$ vector, whose coefficients have to be evaluated.

The predictive power of such an equation is a priori null, because being the QSM: $\mathbf{Z}$, by construction non-singular (otherwise two density functions will be exactly the same), then there always can be computed a QSM inverse: $\mathbf{Z^{-1}}$, obeying the usual relationships: $\mathbf{Z^{-1}Z} = \mathbf{ZZ^{-1}} = \mathbf{I}$, in such a way that the trivial result, defining the unknown coefficient vector:

$$|\mathbf{x}\rangle = \mathbf{Z}^{-1}|\mathbf{p}\rangle \, , \tag{200}$$

will be always obtained within a *core set* scenario. Furthermore, one can retrieve the exact value of the property for any molecule of the *core set* QOS choosing the scalar products:

$$\forall I \colon p_I = \langle \mathbf{z}_I|\mathbf{x}\rangle \, . \tag{201}$$

The QSM for diverse *core sets* has been used in a quite large set of prediction studies [26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46], in every case employing up to date statistical tools, the usual procedures currently available in classical QSPR studies, see for example reference [171]. The use of the first-order fundamental QQSPR equation to construct algorithms, which can be utilized as predictive tools, has been previously attempted [59], but it has not been continued in practice. In the present study, the reader can find in the following sections new theoretical developments of the prediction ability of the fundamental QQSPR equation. However, a reminder of some simple linear algebra relating to the FQQSPR equation is needed first in order to understand the following arguments; therefore it will be described in the forthcoming final section.

Future Trends

Procedure for Adding One Molecular Structure to a Known Core Set

The Partition of the FQQSPR Linear Equation. The general setup described until now, amounts to the same as virtually considering the *U-m* as forming part of the *core set*, but with a parametrized value of the unknown property. One can refer to this extension of the *core set* as the *parametrized core set*.

The following coefficient vector partitioned expression can be easily written in terms of the inverse partitioned QSM matrix elements:

$$\begin{pmatrix} |\mathbf{x_0}\rangle \\ x \end{pmatrix} = \begin{pmatrix} \mathbf{Z_0^{(-1)}}|\mathbf{p_0}\rangle + \pi|\mathbf{z}^{(-1)}\rangle \\ \langle \mathbf{z}^{(-1)}|\mathbf{p_0}\rangle + \pi\theta^{(-1)} \end{pmatrix} .$$

However, in order to obtain equivalent expressions possessing less entanglement with the elements of the inverse matrix, the most convenient way is to restart the procedure writing explicitly:

$$\begin{pmatrix} \mathbf{Z_0}|\mathbf{x_0}\rangle + x|\mathbf{z}\rangle \\ \langle \mathbf{z}|\mathbf{x_0}\rangle + x\theta \end{pmatrix} = \begin{pmatrix} |\mathbf{p_0}\rangle \\ \pi \end{pmatrix} . \tag{202}$$

From the augmented linear equation first component structure, one can obtain:

$$|\mathbf{x_0}\rangle = \mathbf{Z_0^{-1}}|\mathbf{p_0}\rangle - x\mathbf{Z_0^{-1}}|\mathbf{z}\rangle \, , \tag{203}$$

taking into account that the QSM $\mathbf{Z}_0$, associated to the initial *core set*, is non-singular by construction. Therefore, the first right-hand term is just the solution of the FQQSPR linear equation for the initial *core set*, as shown in Eq. (200). Thus, calling:

$$|\mathbf{q}\rangle = \mathbf{Z_0^{-1}}|\mathbf{p_0}\rangle \wedge |\mathbf{a}\rangle = \mathbf{Z_0^{-1}}|\mathbf{z}\rangle \, , \tag{204}$$

then, Eq. (203) could be rewritten as any of the two following equalities:

$$|\mathbf{x_0}\rangle = |\mathbf{q}\rangle - x|\mathbf{a}\rangle = \mathbf{Z_0^{-1}}(|\mathbf{p_0}\rangle - x|\mathbf{z}\rangle) \, . \tag{205}$$

Taking into account Eq. (205), the second component can be written as:

$$\pi = \langle \mathbf{z}|\mathbf{q}\rangle + (\theta - \langle \mathbf{z}|\mathbf{a}\rangle)x = a_0 + a_1 x \,, \tag{206}$$

where:

$$a_0 = \langle \mathbf{z}|\mathbf{q}\rangle \wedge a_1 = \theta - \langle \mathbf{z}|\mathbf{a}\rangle \,; \tag{207}$$

expression (207) shows the expected trivial result consisting of how the *U-m* property and the coefficient, still not evaluated, obviously are linearly related.

Analysis of the FQQSPR Equation. Only in the case that the *U-m* property can be associated to a concrete numerical value: π_U, say; then the exact linear coefficient: x_U can be written in terms of the quantities appearing in Eq. (206) as:

$$x_U = \frac{\pi_U - a_0}{a_1} = \frac{\pi_U - \langle \mathbf{z}|\mathbf{q}\rangle}{\theta - \langle \mathbf{z}|\mathbf{a}\rangle} = \frac{\pi_U - \langle \mathbf{z}|\mathbf{Z}^{-1}|\mathbf{p}_0\rangle}{\theta - \langle \mathbf{z}|\mathbf{Z}^{-1}|\mathbf{z}\rangle} \,. \tag{208}$$

Equation (208), although it will never hold exactly by construction, tells us about several interesting features.

First, admitting through the previous discussion that the vector $|\mathbf{q}\rangle$ is nothing else than the exact linear coefficient set for the *core set*, then the scalar product: $a_0 = \langle \mathbf{z}|\mathbf{q}\rangle \equiv \pi^{(0)}$, is nothing else than an estimation of the *U-m* property, $\pi^{(0)}$, say, using the discrete representation of the *U-m* with respect to the *core set*. Thus, the numerator of Eq. (208) corresponds to the difference between this rough approximation and the exact property value, if known, of the *U-m*. Obviously enough, if: $\pi_U = \pi^{(0)}$ holds, then the *U-m* coefficient will be null, as the *U-m* property could be solely computed by the descriptor $|\mathbf{z}\rangle$.

Second, the denominator in Eq. (208), tells us about the difference between *U-m* self-similarity and the norm of the vector $|\mathbf{z}\rangle$ computed in the QSM *reciprocal* space, defined as the vector space where the inverse of the QSM acts as a metric matrix. An identity as: $\theta = \langle \mathbf{z}|\mathbf{Z}^{-1}|\mathbf{z}\rangle$, will produce an unacceptable linear algebra result, whichever value the *U-m* property could be. It is plausible to suppose, therefore, that in well-behaved FQQSPR prediction problems, the following inequality shall always hold: $\theta \neq \langle \mathbf{z}|\mathbf{Z}^{-1}|\mathbf{z}\rangle$.

One can conclude, within the settings of the *U-m* prediction problem, that the linear structure of the FQQSPR equation does not permit the evaluation of the *U-m* property in Eq. (206), unless the coefficient x appears defined in some way. The exact coefficient value x_U can be derived, if and only if, a given concrete value of the property is known, but in this case, the prediction problem will not a priori exist as such. Only if the property of the *U-m* appears in a parametrized form, the problem can be handled in an approximate way.

Thus, one arrives at the logical conclusion in that a prediction obstacle is already present in the case of a one-dimensional representation of the QSPR operator Ω, even if the quantum similarity description discrete vector tag: $\begin{pmatrix} |\mathbf{z}\rangle \\ \theta \end{pmatrix}$ is known for the extra added *U-m* structure, but the corresponding property value is not defined, but considered as a parameter.

Analysis of U-m Predicted Property Values. Therefore, the aim of the following discussion will be to find an appropriate way to determine a reasonable optimal approach for the *U-m* coefficient x, by means of manipulating Eq. (205) in order that the unknown property parametrized value π could be estimated using Eq. (206). If some optimal coefficient value $x^{(\mathrm{opt})}$ is found, Eq. (206) can be rewritten as:

$$\pi^{(\mathrm{Estimate})} = \pi^{(0)} + (\theta - \langle \mathbf{z}|\mathbf{a}\rangle)x^{(\mathrm{opt})} \,; \tag{209}$$

in this way the role of the estimated coefficient $x^{(\mathrm{opt})}$ appears with a clear meaning now: it constitutes one of the factors to correct the rough initial estimate of the *U-m* property $\pi^{(0)}$, which can be obtained from the primary information provided by the *core set* by just using Eq. (201). Equation (209) above also enhances the leading role of the *U-m* self-similarity θ for such a property correction-estimation task. In Eq. (209), the *U-m* self-similarity appears shifted, in turn, by the norm of the *U-m* discrete representation vector, $|\mathbf{z}\rangle$, with respect to the *core set*: $\langle \mathbf{z}|\mathbf{a}\rangle = \langle \mathbf{z}|\mathbf{Z}^{-1}|\mathbf{z}\rangle$, computed over *reciprocal* space.

Formulation of the Optimization Problem

In any of both direct and reciprocal space cases, as expected from the linear structure of the fundamental equations used and provided that: $\lambda \in \mathbf{R}$, then it can be written for the unknown sought property:

$$\pi = a + b\lambda \tag{210}$$

also, the equation for the core set unknowns can be written in general as:

$$|\mathbf{u}\rangle = \mathbf{A}(|\pi\rangle - \lambda|a\rangle) \,, \tag{211}$$

where $\mathbf{A}$ is a positive definite matrix.

The unknown property in Eq. (210) will be well-defined whenever, using Eq. (211), one could obtain a well-defined value of the parameter: λ. As the solution of

Eq. (210) corresponds to an infinite collection of real elements, the restricted solution in the case of putting one molecule in, is not unique, as from Eq. (211) one can describe several possible ways to obtain optimal values of the parameter λ. For instance:

a) Defining the difference vector: $|d\rangle = |\pi\rangle - \lambda|a\rangle$, a difference norm can be constructed:

$$\langle d|d\rangle = \langle \pi|\pi\rangle - 2\lambda\langle \pi|a\rangle + \lambda^2\langle a|a\rangle \,, \tag{212}$$

optimizing the expression (212) with respect to the parameter, provides:

$$\lambda^{\text{opt}} = \frac{\langle \pi|a\rangle}{\langle a|a\rangle} \,,$$

besides the optimal value of the difference norm will be a minimum, as the second-order coefficient in Eq. (212) is a Euclidean norm of a non-null vector.

b) One can consider the norm of vector $|\mathbf{u}\rangle$ as defined in Eq. (211) the objective function to be optimized; in this case it can be written:

$$\langle \mathbf{u}|\mathbf{u}\rangle = \langle \pi|\mathbf{A}|\pi\rangle - 2\lambda\langle \pi|\mathbf{A}|a\rangle + \lambda^2\langle a|\mathbf{A}|a\rangle \,,$$

So the optimal value of the parameter is now:

$$\lambda^{\text{opt}} = \frac{\langle \pi|\mathbf{A}|a\rangle}{\langle a|\mathbf{A}|a\rangle} \,,$$

which provides a similar form as in the previous procedure, weighted by the transformation matrix **A**.

c) The scalar product of the vectors $\{\mathbf{u}\rangle; |\pi\rangle\}$ can be optimized, the objective function is now:

$$\begin{aligned} |\langle \pi|\mathbf{u}\rangle|^2 &= |\langle \pi|\mathbf{A}|\pi\rangle - \lambda\langle \pi|\mathbf{A}|a\rangle|^2 \\ &= |\langle \pi|\mathbf{A}|\pi\rangle|^2 - 2\lambda\langle \pi|\mathbf{A}|\pi\rangle\langle \pi|\mathbf{A}|a\rangle \\ &\quad + \lambda^2|\langle a|\mathbf{T}|a\rangle|^2 \,, \end{aligned}$$

producing:

$$\lambda^{\text{opt}} = \frac{\langle \pi|\mathbf{A}|\pi\rangle}{\langle \pi|\mathbf{A}|a\rangle} \,.$$

d) The scalar product of the vectors $\{\mathbf{u}\rangle; |a\rangle\}$ can be now optimized, in an equivalent way as in the previous procedure; that is, using the objective function:

$$\begin{aligned} |\langle t|\mathbf{u}\rangle|^2 &= |\langle a|\mathbf{A}|\pi\rangle - \lambda\langle a|\mathbf{A}|a\rangle|^2 \\ &= |\langle a|\mathbf{A}|\pi\rangle|^2 - 2\lambda\langle a|\mathbf{A}|\pi\rangle\langle a|\mathbf{A}|a\rangle \\ &\quad + \lambda^2|\langle a|\mathbf{A}|a\rangle|^2 \,, \end{aligned}$$

which permits one to obtain the optimal value:

$$\lambda^{\text{opt}} = \frac{\langle a|\mathbf{A}|\pi\rangle}{\langle a|\mathbf{A}|a\rangle} \,;$$

this result, however, corresponds to the same restriction as the one previously studied in procedure II of reference [60]. Thus, optimizing the norm $\langle \mathbf{u}|\mathbf{u}\rangle$ seems to be equivalent to optimizing the squared module: $|\langle a|\mathbf{u}\rangle|^2$.

A Quadratic Error Restricted First-Order (n+1) Estimation

The easiest procedure to overcome the previously mentioned evaluation impasse for the unknown property of the *U-m*, concretely the approximate evaluation of the coefficient x, appears naturally associated to the possibility to introduce a restriction of some sort into the FQQSPR equation solution. Here follows the description of one among some possible restriction procedures. One will discuss a second option and sketch some alternative procedures as well, within a separate section below.

Setting up the Problem. By inspecting Eq. (205), one can define the difference vector:

$$|\Delta\rangle = |\mathbf{p_0}\rangle - x|\mathbf{z}\rangle \,, \tag{213}$$

and compute with it the following associated quadratic error, which in this case describes a second-order polynomial of the unknown coefficient x:

$$\varepsilon^{(2)} = \langle \Delta|\Delta\rangle = \langle \mathbf{p_0}|\mathbf{p_0}\rangle - 2x\langle \mathbf{z}|\mathbf{p_0}\rangle + x^2\langle \mathbf{z}|\mathbf{z}\rangle \tag{214}$$

which, in turn, using the usual null gradient condition, allows us to obtain an optimal value of the coefficient x, obeying the simple quotient expression:

$$x^{(\text{opt})} = \frac{\langle \mathbf{z}|\mathbf{p}_0\rangle}{\langle \mathbf{z}|\mathbf{z}\rangle} \,. \tag{215}$$

The optimal coefficient value (215) produces a minimum of the quadratic error, being the second-order coefficient of the quadratic error polynomial (214), associated to the positive definite Euclidean norm of the *U-m* discrete representation with respect to the *core set* density tags, that is: $\langle \mathbf{z}|\mathbf{z}\rangle > 0$. Such a quadratic error restriction is equivalent to constructing a difference vector (213) with elements as small as possible. In the case, quite unlikely to occur, where the known property vector $|\mathbf{p_0}\rangle$ and the *U-m* quantum similarity vector $|\mathbf{z}\rangle$ are linearly dependent, the present restriction will construct a difference vector (213),

which will be exactly the null vector at the optimal value of the unknown.

Using $x^{(\mathrm{opt})}$, the optimal value obtained with Eq. (215), the corresponding unknown property for the *U-m* can be straightforwardly predicted using Eq. (206):

$$\pi^{(\mathrm{opt})} = \langle \mathbf{z}|\mathbf{c}\rangle + \left(\theta - \langle \mathbf{z}|\mathbf{Z_0^{-1}}|\mathbf{z}\rangle\right)x^{(\mathrm{opt})}$$
$$= \langle \mathbf{z}|\mathbf{Z_0^{-1}}|\mathbf{p}_0\rangle + \left(\theta - \langle \mathbf{z}|\mathbf{Z_0^{-1}}|\mathbf{z}\rangle\right)\frac{\langle \mathbf{z}|\mathbf{p}_0\rangle}{\langle \mathbf{z}|\mathbf{z}\rangle} ,$$

providing an expression, which can be easily rearranged by defining the matrix:

$$\mathbf{A} = \mathbf{Z_0^{-1}} + \alpha\mathbf{I} \wedge \alpha = \langle \mathbf{z}|\mathbf{z}\rangle^{-1}\left(\theta - \langle \mathbf{z}|\mathbf{Z_0^{-1}}|\mathbf{z}\rangle\right) .$$

Therefore, one can compactly write the optimal property equation for the *U-m* as:

$$\pi^{(\mathrm{opt})} = \langle \mathbf{z}|\mathbf{A}|\mathbf{p}_0\rangle ,$$

moreover, defining the $(1 \times n)$ row vector: $\langle \mathbf{b}| = \langle \mathbf{zA}$; then, this new description permits us to write the estimated property by means of the following scalar product form:

$$\pi^{(\mathrm{opt})} = \langle \mathbf{b}|\mathbf{p}_0\rangle = \sum_I b_I p_{0;I} . \tag{216}$$

Hence, within the linear FQQSPR equation under the minimal quadratic error restriction, the result (216) shows that the *estimated* optimal unknown property for any *U-m*, is always expressible as a linear functional of the known molecular properties of the *core set*. Such a result is in agreement with usual classical QSPR treatments.

Additionally, in a very unlikely case, where a linear dependence of the *core set* property vector $|\mathbf{p_0}\rangle$ and the *U-m* quantum similarity vector: $|\mathbf{z}\rangle$ applies, that is whenever: $|\mathbf{p}_0\rangle = \lambda|\mathbf{z}\rangle$; then, the optimal estimated property value will be expressible as a multiple of the *U-m* self-similarity:

$$\pi_{\parallel}^{(\mathrm{opt})} = \theta x^{(\mathrm{opt})} = \theta\frac{\langle \mathbf{zp}_0\rangle}{\langle \mathbf{zz}\rangle} = \lambda\theta . \tag{217}$$

It must be finally noted in any case that the gauge operator expectation value: $\langle\sigma[\rho_U]\rangle$, for the *U-m*, if different from zero, shall be added to the optimal value of the property in Eq. (216) or (217) in order to retrieve the predicted value corresponding to the original property set.

A (2 + 1) Naïve Application Example. In order to illustrate the above procedure one can consider a simple case as follows. Supposing that the *core set* is made of just two molecular structures: $\{A, B\}$ say, with a known shifted property vector:

$$|\mathbf{p}\rangle = \begin{pmatrix} p_A \\ p_B \end{pmatrix} ,$$

and also admitting that the *U-m*, possessing the unknown parametric property, π, could be labeled as $\{U\}$; then, the QSM of the *core set* and the similarity vector of the *U-m* can be respectively written as:

$$\mathbf{Z}_0 = \begin{pmatrix} z_{AA} & z_{AB} \\ z_{AB} & z_{BB} \end{pmatrix} \wedge |\mathbf{z}\rangle = \begin{pmatrix} z_{AU} \\ z_{BU} \end{pmatrix} ,$$

with z_{UU} representing the *U-m* self-similarity measure. One can readily compute the *core set* similarity matrix inverse:

$$\mathbf{Z}_0^{-1} = D^{-1}\begin{pmatrix} z_{BB} & -z_{AB} \\ -z_{AB} & z_{AA} \end{pmatrix} \wedge D = z_{AA}z_{BB} - z_{AB}^2 .$$

It must be now said that when doing this kind of calculation care must be taken with the values of the (2×2) similarity matrix determinant D, because a value approaching zero can render the procedure useless and generate unpredictable computational errors. For all molecular pairs $\{A, B\}$ of the *core set*, the value of the determinant D has to be checked to be significantly greater than a positive definite threshold, that is:

$$\forall\{A, B\}\colon D \geq \varepsilon > 0 ;$$

failure to comply with this condition for any *core set* molecular pair may well represent a computationally unbalanced QSM triplet $\{A, B, U\}$. This test shall be added to the already described coherent calculation procedures, when accurate QSM have to be computed.

However, the positive definite determinant condition can also be rewritten in a positive definite quantum similarity matrix condition, that is:

$$z_{BB} > z_{AA}^{-1}z_{AB}^2 .$$

A $(N \times N)$ positive definite condition problem, which corresponds in general to the positive definite nature of the quantum similarity matrices, can be shown that it can be readily solved for any core set cardinality, but the nature of this subject, although of capital importance for application purposes, appears to be marginal in the present work and hence will be studied elsewhere.

Thus, one can express the needed vector resulting from the product: $\mathbf{Z}_0^{-1}|\mathbf{z}\rangle$ as:

$$\mathbf{Z}_0^{-1}|\mathbf{z}\rangle = D^{-1}\begin{pmatrix} z_{BB}z_{AU} - z_{AB}z_{BU} \\ z_{AA}z_{BU} - z_{AB}z_{AU} \end{pmatrix} ,$$

and the same can be obtained for the scalar products entering the restricted optimal solution:

$$\langle \mathbf{z}|\mathbf{z}\rangle = z_{AU}^2 + z_{BU}^2 \wedge \langle \mathbf{z}|\mathbf{p}\rangle = z_{AU}p_A + z_{BU}p_B .$$

Finally, one can also write:

$$\langle \mathbf{z}|\mathbf{Z}_0^{-1}|\mathbf{z}\rangle = D^{-1}[z_{AU}(z_{BB}z_{AU} - z_{AB}z_{BU}) + z_{BU}(z_{AA}z_{BU} - z_{AB}z_{AU})]$$

and

$$\langle \mathbf{z}|\mathbf{Z}_0^{-1}|\mathbf{p}\rangle = D^{-1}[p_A(z_{BB}z_{AU} - z_{AB}z_{BU}) + p_B(z_{AA}z_{BU} - z_{AB}z_{AU})] .$$

Therefore, one can obtain the unknown optimal property value, after some trivial manipulation of the previous quantities as:

$$\pi_{AB;U}^{\text{opt}} = \left[D\left(z_{AU}^2 + z_{BU}^2\right)\right]^{-1}[(\alpha z_{AU} - \beta z_{BU})p_A + (\alpha z_{BU} + \beta z_{AU})p_B] , \quad (218)$$

where the following symbols are used:

$$\alpha = Dz_{UU}$$
$$\beta = z_{AU}z_{BU}(z_{AA} - z_{BB}) + z_{AB}(z_{BU}^2 - z_{AU}^2) .$$

The expression (218) for the optimal quadratic error restricted property of the *U-m* self-similarity, constitutes an explicit example involving a very limited number of molecular structures. However, it also corresponds to a general equation involving any triad of molecules, where one of them acts as the *U-m*.

This simple way of estimating a property can be structured into a procedure involving all the: $N = 1/2[n(n-1)]$ possible *core set* distinct molecular pairs. Indeed, given a *core set* and a *U-m*, one can compute all the possible property estimates using Eq. (218). Such a process will produce a set of N values of the *U-m* estimated property:

$$\left\{\pi_{IJ;U}^{\text{Opt}} | \forall(I = 1, n-1; J = I+1, n)\right\} ,$$

which can be finally manipulated in the usual statistical way.

Such an example opens the way to other possible choices using as probe core sets three or another number of QO. In order to leave this study within reasonable limits this possibility will not be further investigated here.

The (n + m) Case Under a Quadratic Error Restriction. One can extend the estimation procedure, outlined in the previous section, in order to include a *U-m set* of arbitrary cardinality, *m* say, so a general quadratic error restricted scheme can be also described in this more general case. One may write the partition of the QSM into the *core set*, bearing the label 0 and the *U-m set*, bearing the label 1, then the FQQSPR equation can be written as:

$$\begin{pmatrix} \mathbf{Z}_{00} & \mathbf{Z}_{01} \\ \mathbf{Z}_{01}^{\mathrm{T}} & \mathbf{Z}_{11} \end{pmatrix} \begin{pmatrix} |\mathbf{x}_0\rangle \\ |\mathbf{x}_1\rangle \end{pmatrix} = \begin{pmatrix} |\mathbf{p}_0\rangle \\ |\mathbf{p}_1\rangle \end{pmatrix} , \quad (219)$$

which produce the two matrix equations, as follows:

$$\begin{aligned} \mathbf{Z}_{00}|\mathbf{x}_0\rangle + \mathbf{Z}_{01}|\mathbf{x}_1\rangle &= |\mathbf{p}_0\rangle \\ \mathbf{Z}_{01}^{\mathrm{T}}|\mathbf{x}_0\rangle + \mathbf{Z}_{11}|\mathbf{x}_1\rangle &= |\mathbf{p}_1\rangle . \end{aligned} \quad (220)$$

So, from the first element of Eq. (220), one can deduce:

$$|\mathbf{x}_0\rangle = \mathbf{Z}_{00}^{-1}[|\mathbf{p}_0\rangle - \mathbf{Z}_{01}|\mathbf{x}_1\rangle]$$

with the possibility to construct a difference vector:

$$|\mathbf{d}\rangle = |\mathbf{p}_0\rangle - \mathbf{Z}_{01}|\mathbf{x}_1\rangle . \quad (221)$$

Then, one may immediately use the difference vector (221) to define a quadratic error function like:

$$\begin{aligned} \varepsilon^{(2)} &= \langle \mathbf{d}|\mathbf{d}\rangle \\ &= \langle \mathbf{p}_0|\mathbf{p}_0\rangle - 2\langle \mathbf{p}_0|\mathbf{Z}_{01}|\mathbf{x}_1\rangle + \langle \mathbf{x}_1|\mathbf{Z}_{01}^{\mathrm{T}}\mathbf{Z}_{01}|\mathbf{x}_1\rangle , \end{aligned} \quad (222)$$

which upon derivation and submitted to the extremum condition of null gradient, produces:

$$\frac{\partial \varepsilon^{(2)}}{\partial |\mathbf{x}_1\rangle} = -2\mathbf{Z}_{01}^{\mathrm{T}}|\mathbf{p}_0\rangle + 2\mathbf{Z}_{01}^{\mathrm{T}}\mathbf{Z}_{01}|\mathbf{x}_1\rangle = |\mathbf{0}\rangle ,$$

so, the *U-m set* unknown coefficients, restricted to minimal quadratic error, can be obtained by means of the matrix expression:

$$|\mathbf{x}_1^{\text{Opt}}\rangle = \left(\mathbf{Z}_{01}^{\mathrm{T}}\mathbf{Z}_{01}\right)^{-1}\mathbf{Z}_{01}^{\mathrm{T}}|\mathbf{p}_0\rangle . \quad (223)$$

The solution in Eq. (223) depends only on the circumstance that the matrix:

$$\mathbf{A}_{11} = \mathbf{Z}_{01}^{\mathrm{T}}\mathbf{Z}_{01}$$

shall be non-singular. In fact, the matrix $\mathbf{A}_{11}$ corresponds to the scalar products of the matrix representations of the *U-m* set with respect to the *core set* elements. This condition, if the computed similarities are not faulty, constitutes a metric matrix of the *U-m* space, subtended by the *U-m* QSM columns. Thus, provided that the *U-m* discrete representations with respect to the *core set* are linearly independent, the inverse of $\mathbf{A}$ is guaranteed to exist, as it will be positive definite; moreover, implying the quadratic error (222) is a minimum at the value $|\mathbf{x}_1^{\text{Opt}}\rangle$ given by Eq. (223).

One can easily estimate the unknown parametrized *U-m* property vector $|\mathbf{p}_1\rangle$, submitted to the quadratic error restriction, after defining the auxiliary matrix:

$$\mathbf{X}_{10} = \mathbf{A}_{11}\mathbf{Z}_{01}^{\mathrm{T}} ,$$

and performing some rearrangements using trivial matrix algebra, it is obtained:

$$|\mathbf{p}_1^{\mathrm{Opt}}\rangle = \left[\mathbf{Z}_{01}^{\mathrm{T}}\mathbf{Z}_{00}^{-1} + \left(\mathbf{Z}_{11} - \mathbf{Z}_{01}^{\mathrm{T}}\mathbf{Z}_{00}^{-1}\mathbf{Z}_{01}\right)\mathbf{X}_{10}\right]|\mathbf{p}_0\rangle . \quad (224)$$

Equation (224) shows that the predicted *U-m set* property vector is a linear transformation of the *core set* known properties, a result consistent with the unique *U-m* case, already described in the previous section and coincident with the usual classical QSPR procedures.

Alternative Restrictions and the Associated Prediction Algorithms

The case analyzed in Sect. "Formulation of the Optimization Problem" is not at all unique. One can describe other possible alternative restrictions, which can be imposed to the FQQSPR equation, as follows in this section.

An Alternative Orthogonality Restriction. Here, choosing one of the possible procedures, a deep discussion will be carried out for a *U-m* set bearing one element only, because the extension to the case of several elements is trivial and similar to the previous Subsect. "A Quadratic Error Restricted First-Order (n+1) Estimation" development, although a brief outline will be given for the sake of completeness. Finally, the remnant plausible restrictions will be only sketched, because the procedure to obtain the application algorithms follows the same trends as in the explicit examples.

a) The $(n + 1)$ framework
For the purpose of finding an alternative restriction to the one described in the previous Subsect. "A Quadratic Error Restricted First-Order (n+1) Estimation", it is necessary to recover the first matrix equation of the partition (202), leading to Eq. (203). Then, upon left multiplying both sides by the row vector $\langle\mathbf{z}|$, one can obtain:

$$\langle\mathbf{z}|\mathbf{x_0}\rangle = \langle\mathbf{z}|\mathbf{Z_0^{-1}}|\mathbf{p_0}\rangle - x\langle\mathbf{z}|\mathbf{Z_0^{-1}}|\mathbf{z}\rangle = \alpha_0 - \alpha_1 x . \quad (225)$$

Then, one can use the resulting Eq. (225) to minimize the scalar product $\langle\mathbf{z}|\mathbf{x_0}\rangle$, appearing on the left-hand side. As in the previous treatment, the right-hand side of Eq. (225) can be considered a difference, which can generate a quadratic error function to be minimized with respect to the unknown parameter, x, which can be evaluated in this manner afterwards. After a trivial manipulation one finds:

$$\varepsilon^{(2)} = |\langle\mathbf{z}|\mathbf{x_0}\rangle|^2 = |\alpha_0 - \alpha_1 x|^2 = \alpha_0^2 - 2\alpha_0\alpha_1 x + \alpha_1^2 x^2 , \quad (226)$$

and in this way the extremum condition imposed upon Eq. (226), becomes a manner to obtain the optimal value of the unknown coefficient:

$$2\alpha_0\alpha_1 - 2\alpha_1^2 x = 0 \rightarrow x_{\perp}^{\mathrm{opt}} = \frac{\alpha_0}{\alpha_1} = \frac{\langle\mathbf{z}|\mathbf{Z_0^{-1}}|\mathbf{p_0}\rangle}{\langle\mathbf{z}|\mathbf{Z_0^{-1}}|\mathbf{z}\rangle} . \quad (227)$$

The result depicted in the quotient of expression (227), turns out to be equivalent to agreeing that the imposed restriction considers the vectors orthogonal in the scalar product (225), or: $\langle\mathbf{z}|\mathbf{x_0}\rangle = 0$. Admitting that, such a restriction produces a second equation in the system (202), which simplifies to:

$$\pi_{\perp}^{\mathrm{opt}} = \theta x_{\perp}^{\mathrm{opt}} = \theta\frac{\langle\mathbf{z}|\mathbf{Z_0^{-1}}|\mathbf{p_0}\rangle}{\langle\mathbf{z}|\mathbf{Z_0^{-1}}|\mathbf{z}\rangle} = \theta\frac{\langle\mathbf{z}|\mathbf{q}\rangle}{\langle\mathbf{z}|\mathbf{a}\rangle} . \quad (228)$$

Expression (228) resembles the limiting case value of the unknown property: $\pi_{\perp}^{\mathrm{opt}}$, for the quadratic error restriction studied in Subsect. "A Quadratic Error Restricted First-Order (n+1) Estimation", as shown into Eq. (217), when the vector of the known properties and the *U-m* discrete representation with respect to the *core set* become linearly dependent. Similar scalar products appear in both expressions. However, in the present orthogonal restriction, they are evaluated using as a metric the inverse of a matrix, which is nothing else than the QSM, associated to the *core set* space. Thus, the scalar products entering Eq. (228) are computed in the *reciprocal* space of the vector space possessing a metric $\mathbf{Z}_0$, as previously commented in Sect. "Linear Quantum QSPR Fundamental Equation". Moreover, the approximate coefficient value in Eq. (227) resembles the exact expression of the *U-m* coefficient, as described in Eq. (208). Rewriting Eq. (228) as:

$$\pi_{\perp}^{\mathrm{opt}} = \left(\frac{\theta}{\langle\mathbf{z}|\mathbf{a}\rangle}\right)\pi^{(0)} = \omega\pi^{(0)} .$$

It can be easily seen how the ratio between the *U-m* self-similarity and the norm of the *U-m* representation with respect to the *core set* in reciprocal space, corrects in this case the rough estimate of the *U-m* property.

b) An alternative restriction case
A variant of the restriction discussed up to now can be easily described. Instead to optimize the scalar product: $\langle\mathbf{z}|\mathbf{x_0}\rangle$, the alternative scalar product: $\langle\mathbf{p_0}|\mathbf{x_0}\rangle$ can

be minimized, so the optimal coefficient will be given by:

$$x^{\text{opt}} = \frac{\langle \mathbf{p}_0 | \mathbf{Z}_0^{-1} | \mathbf{p}_0 \rangle}{\langle \mathbf{p}_0 | \mathbf{Z}_0^{-1} | \mathbf{z} \rangle} \,. \tag{229}$$

c) The simplified $(2+1)$ framework as a naïve application example
The simplified situation $(2+1)$, concerning three molecules, as in the former case of Subsect. "Extended Wave and Density Functions", is simple to solve, for the previous resultant equation in Section 6.1.1, as the involved elements have already been described, when the quadratic error restriction (214) was studied. Under the present orthogonal restriction and using the same notation as the one appearing in the former discussion in Subsect. "Extended Wave and Density Functions", now one can express the estimated property of the U-m as the quotient:

$$\pi^{\text{opt}}_{\perp,AB;U} = z_{UU} \frac{\begin{bmatrix} p_A(z_{BB}z_{AU} - z_{AB}z_{BU}) \\ +p_B(z_{AA}z_{BU} - z_{AB}z_{AU}) \end{bmatrix}}{\begin{bmatrix} z_{AU}(z_{BB}z_{AU} - z_{AB}z_{BU}) \\ +z_{BU}(z_{AA}z_{BU} - z_{AB}z_{AU}) \end{bmatrix}} \,.$$

An equivalent statistical procedure involving the *core set* pairs, as the one described at the end of Subsect. "Extended Wave and Density Functions", can be obviously followed in this case too.

d) The $(n+m)$ framework
The $(n+m)$ algorithm can be also easily set, employing the partition (219) and also Eqs. (220), thus generalizing the algorithm described in Subsect. "Evaluation of Unknown Molecular Properties as Expectation Values". The first equation multiplied by the matrix $\mathbf{Z}_{01}^{\text{T}}$ on the left in both sides of the equality sign, provides:

$$\mathbf{Z}_{01}^{\text{T}} |\mathbf{x}_0\rangle = \mathbf{Z}_{01}^{\text{T}} \mathbf{Z}_{00}^{-1} [|\mathbf{p}_0\rangle - \mathbf{Z}_{01} |\mathbf{x}_1\rangle] \,,$$

and the restriction:

$$\mathbf{Z}_{01}^{\text{T}} |\mathbf{x}_0\rangle = |\mathbf{0}\rangle$$

corresponds to considering the vector $|\mathbf{x}_0\rangle$ as a member of the null space of the matrix $\mathbf{Z}_{01}^{\text{T}}$. Such equality permits, in turn, to write the optimal vector $|\mathbf{x}_1\rangle$ as:

$$|\mathbf{x}^{\text{Opt}}_{\perp;1}\rangle = \left(\mathbf{Z}_{01}^{\text{T}} \mathbf{Z}_{00}^{-1} \mathbf{Z}_{01}\right)^{-1} \left(\mathbf{Z}_{01}^{\text{T}} \mathbf{Z}_{00}^{-1}\right) |\mathbf{p}_0\rangle \,,$$

so, the estimated property vector under the null space restriction simply becomes:

$$|\mathbf{p}^{\text{Opt}}_{\perp;1}\rangle = \mathbf{Z}_{11} |\mathbf{x}^{\text{Opt}}_{\perp;1}\rangle \,,$$

As in the former quadratic error restriction discussed in Subsect. "Expectation Values Within Extended Density Functions Framework", one obtains an expression which shows that this result is nothing else than a linear transformation of the vector of the *core set* properties.

Other Possible Restriction Choices. Besides the previously discussed pair of alternative restriction choices and the one outlined in Section 6.1.2, one can describe other possible procedures to compute the optimal U-m coefficient. They will be briefly explained for a U-m set bearing one element only, as their generalization to $(n+m)$ situations and the $(2+1)$ simplification could be done using the same procedures as before.

a) Quadratic error in reciprocal space vectors: $|\mathbf{x}_0\rangle$ norm restriction
First, one can recall for this purpose Eq. (204), then the *core set* coefficient vector $|\mathbf{x}_0\rangle$ may be expressed with the two basic vectors $\{|\mathbf{p}_0\rangle, |\mathbf{z}_0\rangle\}$ transformed into the reciprocal space: $\{|\mathbf{q}\rangle, |\mathbf{a}\rangle\}$:

$$|\mathbf{x}_0\rangle = |\mathbf{q}\rangle - x|\mathbf{a}\rangle \,. \tag{230}$$

Now, the Euclidean norm of the vector $|\mathbf{x}_0\rangle$ can be optimized, providing the optimal U-m coefficient as:

$$x^{\text{Opt}} = \frac{\langle \mathbf{q} | \mathbf{a} \rangle}{\langle \mathbf{a} | \mathbf{a} \rangle} = \frac{\langle \mathbf{p}_0 | \mathbf{Z}_0^{-2} | \mathbf{z} \rangle}{\langle \mathbf{z} | \mathbf{Z}_0^{-2} | \mathbf{z} \rangle} \,, \tag{231}$$

which is a variant of the form obtained in Eq. (215). The $(n+m)$ case can be easily handled as within the previous discussion on the two described restrictions.

b) Several alternative plausible restrictions within reciprocal space
Finally, another possible restriction set must be described, which can be associated to Eq. (230). Instead of minimizing the norm of the coefficient vector $|\mathbf{x}_0\rangle$ one can minimize either the scalar product $\langle \mathbf{q} | \mathbf{x}_0 \rangle$, or work on the alternative form $\langle \mathbf{a} | \mathbf{x}_0 \rangle$. The first option provides the optimal U-m coefficient:

$$x^{\text{Opt}} = \frac{\langle \mathbf{q} | \mathbf{q} \rangle}{\langle \mathbf{q} | \mathbf{a} \rangle} \,,$$

which constitutes an expression related to Eq. (229), while the form deduced from the second one is equivalent to Eq. (231).

Some Application Remarks. The first-order FQQSPR equation does not possess immediate predictive power. In order to circumvent this limitation though, one can easily show that two alternative approximate algorithms may

be employed, among other possible similar choices. These procedures can be used to estimate the unknown properties of one or various molecules described as QO.

The present algorithms produce similar formal structures, which can be easily connected with classical QSPR points of view. Such a resemblance can be also simply used to manipulate similar, but empirical, equations in the classical QSPR framework, where the computational formalism appears to be of the same characteristics as in linear QQSPR problems. In order to use the algorithms described here in empirical QSPR cases, there is only need to substitute the QSM, which is the basic matrix in QQSPR procedures, by the Gramian matrix of the molecular descriptor set as defined in Eq. (198), which is the comparable molecular space matrix which can be constructed in classical QSPR. In an indirect manner, therefore, the present study provides an alternative to the widespread QSPR algorithms based on the space descriptor path, a new classical QSPR procedure, which appears, from now on, to be accompanied by a quite diverse toolbox set, common to the linear QQSPR framework, in order to obtain predictions of unknown properties in empirical studies.

Finally, the present results, although exhaustive as far as one can see but without discarding the existence of alternative FQQSPR equation restrictions, from the theoretical point of view they lie on the linear QQSPR framework, they have thus to be considered just as a first step in order to generally solve the prediction problem in QQSPR. This is so as, contrary to classical QSPR procedures, the extension of the FQQSPR equation to higher order terms can be easily described, as well as employed within a set of similar ideas and procedures as these herein discussed.

Extensive numerical results and additional study of high order level problems seems therefore to outline the future research in the open QQSPR area of study.

One Molecule at a Time Linear QQSPR

When constructing the linear QQSPR equation one can choose a system of one core molecule and one U-m, which will constitute the simpler case. The similarity matrix is:

$$\mathbf{Z} = \begin{pmatrix} Z_{II} & Z_{IU} \\ Z_{IU} & Z_{UU} \end{pmatrix} .$$

Where the subindex I stands for any molecule in the core set, that is: a well defined molecular structure with a known property p_I and U for any well-defined molecule with an unknown propertyπ, which has to be estimated. It can be written:

$$\begin{pmatrix} Z_{II} & Z_{IU} \\ Z_{IU} & Z_{UU} \end{pmatrix} \begin{pmatrix} c_I \\ c_U \end{pmatrix} = \begin{pmatrix} p_I \\ \pi \end{pmatrix} \rightarrow \begin{cases} Z_{II}c_I + Z_{IU}c_U = p_I \\ Z_{IU}c_I + Z_{UU}c_U = \pi \end{cases} .$$

Taking the first equation and substituting into the second:

$$c_I = \frac{p - Z_{IU}c_U}{Z_{II}} \rightarrow \pi = Z_{IU}\frac{p_I - Z_{IU}c_U}{Z_{II}} + Z_{UU}c_U = \frac{Z_{IU}}{Z_{II}}p_I + \left(Z_{UU} - \frac{(Z_{IU})^2}{Z_{II}}\right)c_U ,$$

an expression which, after rearrangement, provides a way to estimate the unknown property:

$$\pi = \frac{1}{Z_{II}}\left(Z_{IU}p_I + \left(Z_{II}Z_{UU} - (Z_{IU})^2\right)c_U\right) = \frac{1}{Z_{II}}(Z_{IU}p_I + \Delta c_U) ,$$

where $\Delta = \mathrm{Det}(\mathbf{Z})$.

One can see the undetermined coefficient c_U as equivalent to a parameter λ which in turn can be optimized, thus the unknown property could be rewritten as:

$$\pi = \alpha + \beta\lambda_{\mathrm{opt}} \leftarrow \alpha = \frac{Z_{IU}p_I}{Z_{II}} \wedge \beta = \frac{\Delta}{Z_{II}} .$$

There are several ways to obtain the optimal value of the parameter λ, but all of them are equivalent. For example, one can try to make optimal the coefficient c_I in the first equation:

$$c_I = \frac{1}{Z_{II}}(p_I - Z_{IU}\lambda) \rightarrow \frac{\mathrm{d}}{\mathrm{d}\lambda}\left|\frac{1}{Z_{II}}(p_I - Z_{IU}\lambda)\right|^2 = 0 \rightarrow \lambda_{\mathrm{opt}} = \frac{p_I}{Z_{IU}}$$

so, in this way the optimal U-m property is:

$$\pi_{\mathrm{opt}} = \frac{Z_{IU}p_I}{Z_{II}} + \frac{\Delta p_I}{Z_{II}Z_{IU}} = \frac{(Z_{IU})^2 + \Delta}{Z_{II}Z_{IU}}p_I = \frac{Z_{UU}}{Z_{IU}}p_I .$$

Then the problem consists of obtaining the U-m self-similarity and the similarity between the core molecule and the U-m. So, for every core set C molecular structure one can obtain an estimate of the U-m property, say:

$$\forall I \in C\colon \pi_{U,\mathrm{opt}}[I] = \frac{Z_{UU}}{Z_{IU}}p_I .$$

Then, supposing that the cardinality of the core set is N: $\#C = N$, one can obtain a statistical average of all the core set estimates of the U-m property:

$$\langle \pi_U \rangle \approx N^{-1} \sum_{I=1}^{N} \pi_{\text{opt}}[I] = N^{-1} Z_{UU} \sum_{I=1}^{N} \frac{1}{Z_{IU}} p_I$$
$$= N^{-1} \sum_{I=1}^{N} \omega_{IU} p_I \leftarrow \forall I \in C\colon \omega_{IU} = \frac{Z_{UU}}{Z_{IU}} .$$

An expression which proves that this simple QQSPR formulation arrives at the usual QSPR result, consisting of the fact that the estimated value of the property of the U-m is a weighted sum of the properties of the core set:

$$\langle \pi_U \rangle \approx N^{-1} \sum_{I=1}^{N} \omega_{IU} p_I ,$$

the weights being simply the ratios between the U-m quantum self-similarity and the quantum similarity measure of the U-m with every core set molecular structure.

Practical Considerations. However, in this or other more sophisticated cases, the estimation procedure can be achieved in two steps. First, the elements of the core set can be employed as the U-m ones, one by one in front of the remnant $N - 1$ in a kind of Leave One Out procedure. The N optimal estimated values, $\{\langle \pi_I \rangle\}$ say, in this way can be fitted to the experimental property ones, providing in this manner a simple, Hansch-like relationship:

$$p = a \sim \langle \pi \rangle + b ,$$

a relationship which can be further employed to estimate the experimental values of the U-m elements $\{p_U\}$, by using the above-defined equation:

$$p_U = a \langle \pi_U \rangle + b .$$

Moreover, an interesting feature of this procedure is that each estimated value, obtained through solving the fundamental QQSPR equation and irrespective of the fact that the estimation is made over the C or U set elements, can be associated to a mean value, obtained over the set of C-m and also attached to a variance. It is a simple matter of elementary statistical theory application to obtain confidence limits for each estimate, and thus to gather information about, for instance, the outlier nature of some elements and the goodness-of-fit of the whole procedure.

One Molecule at a Time: Quadratic Terms in QQSPR

The operator which can be employed as the source of the fundamental QQSPR equation may be expressed with quadratic and superior terms, within a sequence involving the density elements of the C-m and U-m elements:

$$\Omega(\mathbf{r}) = w_I \rho_I(\mathbf{r}) + w_U \rho_U(\mathbf{r}) + w_I^2 \rho_I^2(\mathbf{r}) + 2 w_I w_U \rho_I(\mathbf{r}) \rho_U(\mathbf{r}) + w_U^2 \rho_U^2(\mathbf{r}) + O(3) .$$

The pair of expectation values of both molecules can be easily written up to third order as:

$$p_I = \langle \Omega \rho_I \rangle = w_I z_{II} + w_U z_{UI} + w_I^2 Z_{III} + 2 w_I w_U Z_{IUI} + w_U^2 Z_{UUI} , \quad (232)$$

and

$$\pi = \langle \Omega \rho_U \rangle = w_I z_{IU} + w_U z_{UU} + w_I^2 Z_{IIU} + 2 w_I w_U Z_{IUU} + w_U^2 Z_{UUU} , \quad (233)$$

where use has been made of the similarity measures like:

$$z_{IU} = \int_D \rho_I(\mathbf{r}) \rho_U(\mathbf{r}) \, \mathrm{d}\mathbf{r} = \int_D \rho_I(\mathbf{r}) \rho_U(\mathbf{r}) \, \mathrm{d}\mathbf{r} = z_{UI} ,$$

and triple similarity measures, for instance:

$$Z_{IUI} = \int_D \rho_I(\mathbf{r}) \rho_U \mathbf{r}) \rho_I(\mathbf{r}) \, \mathrm{d}\mathbf{r} = Z_{IIU} = Z_{UII} = \cdots$$

the properties as expectation values can be rewritten employing the ket-matrix notation:

$$|\mathbf{w}\rangle = \begin{pmatrix} w_I \\ w_U \end{pmatrix} \wedge |\mathbf{z}_I\rangle = \begin{pmatrix} z_{II} \\ z_{UI} \end{pmatrix} \wedge \mathbf{Z}_I = \begin{pmatrix} Z_{III} & Z_{UII} \\ Z_{UII} & Z_{UUI} \end{pmatrix} ,$$

with a similar notation for the ket $|\mathbf{z}_U\rangle$ and the matrix $\mathbf{Z}_U$; the bra notation signifying the corresponding ket transposes. Therefore:

$$\begin{aligned} p_I &= \langle \mathbf{z}_I | \mathbf{w} \rangle + \langle \mathbf{w} | \mathbf{Z}_I | \mathbf{w} \rangle \\ \pi &= \langle \mathbf{z}_U | \mathbf{w} \rangle + \langle \mathbf{w} | \mathbf{Z}_U | \mathbf{w} \rangle . \end{aligned} \quad (234)$$

The QQSPR problem consists of the fact that the coefficient vector is not only unknown $|\mathbf{w}\rangle$ but also the U-m property π. In fact, the quadratic system which corresponds to the quadratic fundamental QQSPR equation in this case, is a set of two different quadratic functions of the same two variables: $|\mathbf{w}\rangle$. The solution may be more complicated than in the linear case, but the procedure can be described in similar terms. That is, first use the C-m equation to express the coefficient w_I in terms of the U-m coefficient and the corresponding similarity measures elements. Then optimize such a coefficient with respect to the U-m one, considered as a parameter. The optimal values of w_U can be employed to evaluate an optimal value of the unknown U-m property π.

A Possible Algorithm. The first fundamental QQSPR equation can be easily transformed into the second-order polynomial root seeking structure.

$$w_I^2 Z_{III} + w_I(z_{II} + 2w_U Z_{IUI}) + (w_U z_{UI} + w_U^2 Z_{UUI}) - p_I = 0\,,$$

which provides:

$$w_I = (2Z_{III})^{-1}\left[-(z_{II} + 2w_U Z_{IUI}) \pm \sqrt{\begin{array}{c}(z_{II} + 2w_U Z_{IUI})^2 - 4Z_{III} \\ \cdot\left(\left(w_U z_{UI} + w_U^2 Z_{UUI}\right) - p_I\right)\end{array}}\right],$$

and after rearranging the square root part:

$$w_I = (Z_{III})^{-1}\left[-\left(\frac{z_{II}}{2} + w_U Z_{IUI}\right) \pm \sqrt{\begin{array}{c}w_U^2(Z_{IUI}^2 - Z_{III}Z_{UUI}) \\ +w_U(z_{II}Z_{IUI} - z_{UI}Z_{III}) \\ +\left(\left(\frac{z_{II}}{2}\right)^2 + Z_{III}p_I\right)\end{array}}\right],$$

the coefficient w_I appears in terms of w_U and the implied similarity integrals. Also, this expression can be employed in the second fundamental QQSPR equation to obtain the U-m property in terms of only one parameter. As in the linear case, the expression of w_I can be optimized with respect to w_U, which can be considered now as a parameter. The expression to be optimized can be written as:

$$\begin{aligned} &w_I = (Z_{III})^{-1}\left[-(\alpha + w_U\beta) \pm \sqrt{w_U^2\gamma_2 + w_U\gamma_1 + \gamma_0}\right], \\ &\alpha = \frac{z_{II}}{2}\,;\beta = Z_{IUI}\,; \\ &\gamma_2 = Z_{IUI}^2 - Z_{III}Z_{UUI}\,;\gamma_1 = z_{II}Z_{IUI} - z_{UI}Z_{III}\,; \\ &\gamma_0 = \alpha^2 + Z_{III}p_I\,. \end{aligned} \tag{235}$$

Thus, the equation yielding the optimal value of w_U can be easily written as:

$$\begin{aligned} 0 = \frac{\mathrm{d}w_I}{\mathrm{d}w_U} &= -\beta \pm \frac{2w_U\gamma_2 + \gamma_1}{2\sqrt{w_U^2\gamma_2 + w_U\gamma_1 + \gamma_0}} \\ &\rightarrow 4\beta^2\left(w_U^2\gamma_2 + w_U\gamma_1 + \gamma_0\right) = (2w_U\gamma_2 + \gamma_1)^2 \\ &\rightarrow \beta^2\left(w_U^2\gamma_2 + w_U\gamma_1 + \gamma_0\right) = \left(w_U\gamma_2 + \frac{\gamma_1}{2}\right)^2 \\ &\qquad = (w_U\gamma_2)^2 + w_U\gamma_2\gamma_1 + \left(\frac{\gamma_1}{2}\right)^2 \\ &\rightarrow w_U^2\left(\beta^2 - \gamma_2\right)\gamma_2 + w_U\left(\beta^2 - \gamma_2\right)\gamma_1 \\ &\qquad + \left(\beta^2\gamma_0 - \left(\frac{\gamma_1}{2}\right)^2\right) = 0\,, \end{aligned}$$

yielding:

$$w_U^{\mathrm{Opt}} = [2\gamma_2]^{-1}\left[-\gamma_1 \pm \sqrt{\gamma_1^2 - 4\gamma_2\left(\beta^2 - \gamma_2\right)^{-1}\left(\beta^2\gamma_0 - \left(\frac{\gamma_1}{2}\right)^2\right)}\right], \tag{236}$$

this value permits us to compute w_I^{Opt} by means of Eq. (235) and therefore π^{Opt} can be obtained with Eqs. (235) using the original form (233).

Alternative Unrestricted Variational Algorithm. Starting again from the quadratic Eq. (234), one can vary both parts of the FQQSPR equation:

$$\begin{aligned} p_I &= \langle \mathbf{z}_I|\mathbf{w}\rangle + \langle \mathbf{w}|\mathbf{Z}_I|\mathbf{w}\rangle \rightarrow |\mathbf{w}[p_I]\rangle = -\frac{1}{2}\mathbf{Z}_I^{-1}|\mathbf{z}_I\rangle\,, \\ \pi &= \langle \mathbf{z}_U|\mathbf{w}\rangle + \langle \mathbf{w}|\mathbf{Z}_U|\mathbf{w}\rangle \rightarrow |\mathbf{w}[\pi]\rangle = -\frac{1}{2}\mathbf{Z}_U^{-1}|\mathbf{z}_U\rangle \end{aligned}$$

so, the optimal estimate values of the C-m and U-m properties will be given by:

$$\begin{aligned} p_I^{\mathrm{est}} &= -\frac{1}{4}\langle \mathbf{z}_I|\mathbf{Z}_I^{-1}|\mathbf{z}_I\rangle\,, \\ \pi^{\mathrm{est}} &= -\frac{1}{4}\langle \mathbf{z}_U|\mathbf{Z}_U^{-1}|\mathbf{z}_U\rangle\,, \end{aligned}$$

which can be associated to minimal values, as the second-order similarity matrices are constructed to be positive definite and thus:

$$\begin{aligned} \mathrm{Det}\,|\mathbf{Z}_I| &= Z_{III}Z_{UUI} - Z_{UII}^2 > 0 \\ &\wedge \mathrm{Det}\,|\mathbf{Z}_U| = Z_{UII}Z_{UUU} - Z_{UUI}^2 > 0\,. \end{aligned}$$

So, an ultimate procedure could be designed, starting to obtain with every one of the core set elements the following linear equation via a least squares procedure:

$$p = ap^{\mathrm{est}} + b\,,$$

in such a way that the linear equation above provides the possibility to obtain the final estimate of the U-m property value:

$$\pi = a\pi^{\mathrm{est}} + b\,.$$

Bibliography

Primary Literature

1. Carbó R, Arnau J, Leyda L (1980) How similar is a molecule to another? An electron density measure of similarity between two molecular structures. Int J Quantum Chem 17:1185–1189

2. Carbó R, Besalú E (1996) Mendeleev conjecture as a consequence of mendeleev postulates. Afinidad 53:77–79
3. Messiah A (1999) Quantum mechanics. Dover Publications, New York
4. Bohm D (1989) Quantum theory. Dover Publications, New York
5. Carbó-Dorca R, Robert D, Amat L, Girones X, Besalu E (2000) Molecular quantum similarity in QSAR and drug design. In: Lecture notes in chemistry, vol 73. Springer, Berlin
6. Carbó-Dorca R (1995) Molecular similarity and reactivity: From quantum chemical to phenomenological approaches. Kluwer Academic, Dordrecht
7. Carbó-Dorca R, Besalu E (1998) A general survey of molecular quantum similarity. J Mol Struct Theochem 451:11–23
8. Carbó-Dorca R (2000) European congress on computational methods in applied sciences and engineering. Barcelona, pp 1–31
9. Carbó-Dorca R, Besalu E (2000) Quantum theory of QSAR. Contributions Sci 1:399–422
10. Bultinck P, Girones X, Carbó-Dorca R (2005) Molecular quantum similarity: Theory and applications. In: Lipkowitz KB, Larter R, Cundari T (eds) Reviews in computational chemistry, vol 21. Wiley, Hoboken, pp 127–207
11. Carbó-Dorca R, Mezey PG (1998) Advances in molecular similarity. JAI Press, London
12. Carbó-Dorca R (2005) Mathematical elements of quantum electronic density functions. In: Advances in quantum chemistry, vol 49. Elsevier Academic, San Diego, pp 121–208
13. Carbó-Dorca R, Besalu E (2001) Extended Sobolev and Hilbert spaces and approximate stationary solutions for electronic systems within non-linear Schrödinger equation. J Math Chem 29:3–20
14. Carbó-Dorca R, Besalu E (2002) Fundamental quantum QSAR (QQSPR) equation: Extensions, nonlinear terms and generalizations within extended Hilbert–Sobolev spaces. Int J Quantum Chem 88:167–182
15. Johnson MA, Maggiora GM (1990) Concepts and applications of molecular similarity. Wiley, New York
16. Carbó R, Calabuig B, Besalú E, Martínez A (1992) Triple density molecular quantum similarity measures: A general connection between theoretical calculations and experimental results. Mol Eng 2:43–64
17. Carbo R, Calabuig B, Vera L, Besalu E (1994) Molecular quantum similarity: Theoretical framework, ordering principles, and visualization techniques. In: Advances in quantum chemistry, vol 25. Academic Press, San Diego, pp 253–313
18. Carbó-Dorca R (2001) Inward matrix products: Extensions and applications to quantum mechanical foundations of QSAR. J Mol Struct Theochem 537:41–54
19. Carbó-Dorca R, Amat L, Besalu E, Girones X, Robert D (2000) Quantum mechanical origin of QSAR: Theory and applications. J Mol Struct Theochem 504:181–228
20. Carbo R, Besalu E, Amat L, Fradera X (1995) Quantum molecular similarity measures (QMSM) as a natural way leading towards a theoretical foundation of quantitative structure-activity relationships (QSPR). J Math Chem 18:237–246
21. Carbó-Dorca R, Besalu E, Amat L, Fradera X (1996) Quantum molecular similarity measures: Concepts, definitions, and applications to quantitative structure-property relationships. In: Carbó-Dorca R, Mezey PG (eds) Advances in molecular similarity. JAI Press, London, pp 1–42
22. Carbó-Dorca R, Amat L, Besalú E, Gironés X, Robert D (2001) Quantum molecular similarity: Theory and applications to the evaluation of molecular properties, biological activities and toxicity. In: Carbó-Dorca R, Gironés X, Mezey PG (eds) Fundamentals of molecular similarity. Proc of 4th girona seminar on molecular similarity. Kluwer Academic/Plenum Publishers, New York, chap 12, pp. 187–320
23. Carbó-Dorca R (1997) Fuzzy sets and boolean tagged sets. J Math Chem 22:143–147
24. Carbó-Dorca R (1998) Tagged sets, convex sets and QS measures. J Math Chem 23:353–364
25. Carbó-Dorca R (1998) Fuzzy sets and boolean tagged sets; vector semispaces and convex sets; quantum similarity measures and ASA density functions; diagonal vector spaces and quantum chemistry. In: Carbó-Dorca R, Mezey PG (eds) Advances in molecular similarity. JAI Press, London, pp 43–72
26. Robert D, Girones X, Carbó-Dorca R (2000) Quantification of the influence of single point mutations on haloalkane dehalogenase activity: A molecular quantum similarity study. J Chem Inf Comput Sci 40:839–846
27. Robert D, Amat L, Carbó-Dorca R (2000) Quantum similarity QSAR: Study of inhibitors binding to trhombin, trypsin and factor Xa, including a comparision with CoMFA and CoMSIA methods. Int J Quantum Chem 80:265–282
28. Girones X, Gallegos A, Carbó-Dorca R (2000) Modelling antimalarial activity: Application of kinetic energy density quantum similarity measures as descriptors in QSAR. J Chem Inf Comput Sci 40:1400–1407
29. Robert D, Girones X, Carbó-Dorca R (2000) Molecular quantum similarity measures as descriptors for quantum QSAR. Polycycl Aromat Compd 19:51–71
30. Renners I, Ludwig LA, Grauel A, Benfenati E, Pellagatti S, Robert D, Carbó-Dorca R, Girones X (2000) Modeling toxicity with molecular descriptors and similarity measures via B-spline networks. IPMU2000 8th international conference on information processing and management of uncertainty in knowledge based systems, Madrid, Spain, pp 1021–1026
31. Renners I, Carbó-Dorca R, Grauel A, Ludwig LA, Robert D, Girones X (2000) Toxicity prediction by using genetically optimized B-spline networks based on molecular quantum similarity. 2nd international ICSC symposium on neural computation, Berlin, Germany
32. Gallegos A, Robert D, Girones X, Carbó-Dorca R (2001) Structure-toxicity relationships of polycyclic aromatic hydrocarbons using molecular quantum similarity. J Comput Aided Mol Design 15:67–80
33. Besalú E, Gallegos A, Carbó-Dorca R (2001) Topological quantum similarity indices and their use in QSAR: Application to several families of antimalarial compounds. Match-Commun Math Comput Chem 44:41–64
34. Gironés X, Carbó-Dorca R (2001) Sobre les relacions lineals d'energia lliure: Mesures de Semblança Molecular Quàntica sobre Funcions de Densitat Electrònica Modificades. Scientia Gerund 25:5–15
35. Ponec R, Gironés X, Carbó-Dorca R (2002) Molecular basis of LFER. On the nature of inductive effects in aliphatic series. J Chem Inf Comput Sci 42:564–570
36. Gironés X, Gallegos A, Carbó-Dorca R (2001) Antimalarial activity of synthetic 1,2,4-trioxabes and cyclid peroxy ke-

tals, a quantum similarity study. J Comput-Aided Mol Design 15:1053–1063
37. Gironés X, Carbó-Dorca R (2002) Molecular quantum similarity-based QSARs for binding affinities of several steroid sets. J Chem Inf Comput Sci 42:1185–1193
38. Gallegos Saliner AG, Amat L, Carbó-Dorca R, Schultz TW, Cronin MTD (2003) Molecular quantum similarity analysis of estrogenic activity. J Chem Inf Comput Sci 43:1166–1176
39. Amat L, Carbó-Dorca R, Cooper DL, Allan NL, Ponec R (2003) Structure-property relationships and momentum-space quantities: Hammett σ-constants. Mol Phys 101:3159–3162
40. Gironés X, Carbó-Dorca R, Ponec R (2003) Molecular basis of LFER. Modeling of the electronic substituent effect using fragment quantum self-similarity measures. J Chem Inf Comput Sci 43:2033–2038
41. Nino A, Munoz-Caro C, Carbó-Dorca R, Gironés X (2003) Rational modelling of the voltage-dependent K+ channel inactivation by aminopyridines. Biophys Chem 104:417–427
42. Gallegos A, Carbó-Dorca R, Ponec R, Waisser K (2004) Similarity approach to QSAR. Application to antimycobacterial benzoxazines. Int J Pharm 269:51–60
43. Gironés X, Carbó-Dorca R (2004) Molecular similarity and quantitative structure-activity relationships. In: Bultinck P, De Winter H, Langenaecker W, Tollenaere JP (eds) Similarity and quantitative structure-activity relationships in computational medicinal chemistry for drug discovery. Marcel Dekker, New York, pp 365–385
44. Giralt F, Espinosa G, Arenas A, Ferre-Gine J, Amat L, Gironés X, Carbó-Dorca R, Cohen Y (2004) Estimation of infinite dilution activity coefficients of organic compounds in Water with neural classifiers. AIChE J 50:1315–1343
45. Carbó-Dorca R, Gironés X (2005) Foundation of quantum similarity measures and their relationship to QSPR: Density function structure, approximations and application examples. Int J Quantum Chem 101:8–20
46. Gironés X, Carbó-Dorca R (2006) Modelling toxicity using molecular quantum similarity measures. QSAR Comb Sci 25:579–589
47. Myers H (1990) Classical and modern regression with applications. Duxbury Press, California
48. Wold S, Sjöström M, Eriksson L (1994) Partial least squares projections to latent structures (PLS) in chemistry. In: von Ragué Schleyer P, Allinger NL, Clark T, Gasteiger J, Kollman PA, Schaefer HF III, Screiner PR (eds) Encyclopedia of computational chemistry. Wiley, Chichester, pp 2006–2021
49. Geladi P, Kowalski BR (1986) Partial least-squares regression: A tutorial. Analytica Chimica Acta 185:1–17
50. Wold S, Johansson E, Cocchi M (1993) PLS-partial least-squares projections to latent structures. In: Kubinyi H (ed) 3D QSAR in drug design. ESCOM, Science Publishers BV, Leiden, chap 5
51. Montgomery DC, Peck EA, Vining GG (1992) Introduction to linear regression analysis. Wiley, New York
52. Whitley DC, Ford MG, Livingstone DJ (2000) J Chem Inf Comput Sci 40:1160–1168
53. Wold S (1978) Cross-validatory estimation of the number of components in factor and principal component models. Technometrics 20:397–405
54. Jolliffe IT (1986) Principal component analysis. Springer, New York
55. Cattell RB (1966) The scree test for the number of factors. Multivar Behav Res 1:245–276
56. Wold S, Eriksson L (1995) Statistical validation of QSAR results. In: Van de Waterbeemd H (ed) Chemometric methods in molecular design. VCH, New York, sect IV, pp 309–318
57. Allen DM (1974) The relationship between variable selection and data augmentation and a method for prediction. Technometrics 16:125–127
58. Shao J (1993) Linear-Model selection by cross-validation. J Am Stat Assoc 88:486–494
59. Carbó-Dorca R (2000) Stochastic transformation of quantum similarity matrices and their use in quantum QSAR models. Int J Quantum Chem 79:163–177
60. Carbó-Dorca R (2007) About the prediction of molecular properties using the fundamental quantum QSPR (QQSPR) equation. SAR QSAR Environ Res 18:265–284
61. Carbó-Dorca R, Van Damme S (2007) Theoretical chemistry accounts: Theory, computation, and modeling. Theoretica Chimica Acta 118:673–679
62. Carbó-Dorca R, Van Damme S (2007) Riemann spaces, molecular density function semispaces, quantum similarity measures and quantum quantitative structure-properties relationships (QQSPR). Afinidad 64:147–153
63. Hansch C, Fujita T (1962) J Am Chem Soc 86:10
64. Oliva JM, Carbó-Dorca R, Mestres J (1996) Conformational analysis from the point of view of quantum molecular similarity. In: Carbó-Dorca R, Mezey PG (eds) Advances in molecular similarity. JAI PRESS, London, pp 135–165
65. Carbó-Dorca R (2000) Quantum quantitative structure-activity relationships (QQSAR): A comprehensive discussion based on inward matrix products, employed as a tool to find approximate solutions of strictly positive linear systems and providings QSAR-quantum similarity measures connection. In: Proc of the european congress on computational methods in applied sciences and engineering (ECCOMAS 2000), Barcelona, chap 12
66. Sen K, Carbó-Dorca R (2000) Inward matrix products, generalised density functions and Rayleigh–Schrödinger perturbation Theory. J Mol Struct Theochem 501:173–176
67. Carbó-Dorca R (2001) Inward matrix product algebra and calculus as tools to construct space-time frames of arbitrary dimensions. J Math Chem 30:227–245
68. Carbó-Dorca R (2003) Applications of inward matrix products and matrix wave functions to Hückel MO theory, Slater extended wave functions, spin extended functions and Hartree method. Int J Quantum Chem 91:607–617
69. Carbó-Dorca R (2003) About some questions relative to the arbitrariness of signs: Their possible consequences in matrix signatures definition and quantum chemical applications. J Math Chem 33:227–244
70. Vinogradov IM (1989) Encyclopaedia of mathematics, vol 4. Kluwer Academic, Dordrecht, p 351
71. Lahey Computer Systems (1998) LF 95 Language Reference. Lahey Computer Systems, Incline Village http://www.lahey.com
72. Carbó-Dorca R (2002) Shell partition and metric semispaces: Minkowski norms, root scalar products, distances and cosines of arbitrary order. J Math Chem 32:201–223
73. Bultinck P, Carbó-Dorca R (2004) A mathematical discussion on density and shape functions, vector semispaces and related questions. J Math Chem 36:191–200

74. Vinogradov IM (1992) Encyclopaedia of mathematics, vol 8. Kluwer Academic, Dordrecht, p 249
75. Constans P, Amat LL, Fradera X, Carbó-Dorca R (1996) Quantum molecular similarity measures (QMSM) and the atomic shell approximation (ASA). In: Carbó-Dorca R, Mezey PG (eds) Advances in molecular similarity, vol 1. JAI Press, London, chap 8, pp 187–211
76. Gironés X, Amat L, Carbó-Dorca R (1998) A comparative study of isodensity surfaces using ab initio and ASA density functions. J Mol Graph Model 16:190–196
77. Constants P, Carbó R (1995) Atomic shell approximation: Electron density fitting algorithm restricting coefficients to positive values. J Chem Inf Comput Sci 35:1046–1053
78. Amat LL, Carbó-Dorca R (1997) Quantum similarity measures under atomic shell approximation: First order density fitting using elementary Jacobi rotations. J Comput Chem 18:2023–2039
79. Amat LL, Carbó-Dorca R (1999) Fitted electronic density functions from H to Rn for use in quantum similarity measures: Cis-diamminedichloroplatinum(II) complex as an application example. J Comput Chem 20:911–920
80. Amat LL, Carbó-Dorca R (2000) Molecular electronic density fitting using elementary Jacoby rotations under atomic shell approximation (ASA). J Chem Inf Comput Sci 40:1188–1198
81. Gironés X, Carbó-Dorca R, Mezey PG (2001) Application of promolecular ASA densities to graphical representation of density functions of macromolecular systems. J Mol Graph Model 19:343–348
82. Mestres J, Solà M, Besalú E, Duran M, Carbó R (1995) Electron density approximations for the fast evaluation of quantum molecular similarity measures. In: Carbó R (ed) Molecular similarity and reactivity: From quantum chemical to phenomenological approaches. Kluwer Academic, Dordrecht, pp 77–85
83. Mestres J, Solà M, Duran M, Carbó R (1995) General suggestions and applications of quantum molecular similarity measures from ab initio fitted electron densities. In: Carbó R (ed) Molecular similarity and reactivity: From quantum chemical to phenomenological approaches. Kluwer Academic, Dordrecht, pp 89–111
84. Carbó R, Calabuig B (1990) Molecular similarity and quantum chemistry. In: Johnson MA, Maggiora GM (eds) Concepts and applications of molecular similarity. Wiley, New York, chapt 6, p 147
85. Carbó R, Besalú E, Amat LL, Fradera X (1996) On Quantum Molecular Similarity Measures (QMSM) and Indices (QMSI). J Math Chem 19:47–56
86. Robert D, Carbó-Dorca R (2000) General trends in atomic and nuclear quantum similarity measures. Int J Quantum Chem 77:685–692
87. Besalú E, Carbó-Dorca R, Karwowski J (2001) Generalized one-electron spin functions and Self-Similarity Measures. J Math Chem 29:41–45
88. Amat L, Besalú E, Carbó-Dorca R (2001) Identification of active molecular sites using quantum-self similarity measures. J Chem Inf Comput Sci 41:978–991
89. Besalú E, Gironés X, Amat L, Carbó-Dorca R (2002) Molecular quantum similarity and the fundaments of QSAR. Acc Chem Res 35:289–295
90. Amat L, Carbó-Dorca R, Cooper DL, Allan NL (2003) Classification of reaction pathways via momentum-space and quantum molecular similarity measures. Chem Phys Lett 367:207–213
91. Bultinck P, Carbó-Dorca R (2003) Molecular quantum similarity matrix based clustering of molecules using dendrograms. J Chem Inf Comput Sci 43:170–177
92. Ponec R, Bultinck P, Van Damme S, Carbó-Dorca R, Tantillo DJ (2005) Geometric and electronic similarities between transition structures for electrocyclizations and sigmatropic hydrogen shifts. Theor Chem Acc 113:205–211
93. Bultinck P, Carbó-Dorca R (2005) Molecular quantum similarity using conceptual DFT descriptors. J Chem Sci 117:425–435
94. Carbó-Dorca R (2008) A quantum similarity matrix (QSM) Aufbau procedure. J Math Chem 44:228–234
95. Berberian SK (1961) Introduction to Hilbert space. Oxford University Press, New York
96. Carbó R, Besalú E (1992) Many center AO integral evaluation using cartesian exponential type orbitals (CETO's). Adv Quantum Chem 24:115–237
97. von Neumann J (1955) Mathematical foundations of quantum mechanics. Princeton University Press, Princeton
98. Born M (1945) Atomic physics. Blackie and Son, London
99. Dirac PAM (1983) The principles of quantum mechanics. Clarendon Press, Oxford
100. Carbó-Dorca R (2000) Quantum QSAR and the eignsystems of stochastic quantum similarity matrices. J Math Chem 27:357–376
101. Gironés X, Amat L, Carbó-Dorca R (2000) Use of electron-electron repulsion energy as a molecular description in QSAR or QSPR studies. J Comput Aided Mol Design 14:477–485
102. Kubinyi H (1993) 3D QSAR in drug design. Theory, methods and applications. ESCOM Science Publishers, Leiden
103. Martin YC (1978) Quantitative drug design. Medicinal research series, vol 8. Marcel Dekker, New York
104. Lien EJ (1987) SAR, side effects and drug design. Medicinal research series, vol 11. Marcel Dekker, New York
105. Martin YC, Kutter E, Austel V (1989) Modern drug research. Medicinal research series, vol 12. Marcel Dekker, New York
106. Wermuth CG (1993) Trends in QSAR molecular modelling 92. Escom, Leiden
107. Kubinyi H (1998) Quantitative structure-activity relationships in drug design. In: Schleyer PVR, Allinger NL, Clark T, Gasteiger J, Kollman PA, Schaefer HF III, Schreiner PR (eds) Encyclopedia of computational chemistry, vol 4. Wiley, Chichester, pp 2309–2319
108. Charton M (1996) Advances in quantitative structure-property relationships, vol 1. JAI Press, London
109. Van de Waterbeemd H (1996) Structure-property correlations in drug research. Academic Press, San Diego
110. Jurs PC (1998) Quantitative structure-property relationships (QSPR). In: Schleyer PVR, Allinger NL, Clark T, Gasteiger J, Kollman PA, Schaefer HF III, Schreiner PR (eds) Encyclopedia of computational chemistry, vol 4. Wiley, Chichester, pp 2320–2330
111. Boethling RS, Mackay D (2000) Handbook of property estimation methods for chemicals. Environmental and health sciences. Lewis Publishers, London
112. Neter J, Wasserman W, Kutner MH (1990) Applied linear statistical models. RD Irwin, Boston
113. Wagner M, Sadowski J, Gasteiger J (1995) Autocorrelation of molecular surface properties for modeling corticosteroid

binding globulin and cytosolic Ah receptor activity by neural networks. J Am Chem Soc 117:7769–7775
114. Cramer RD, Patterson DE, Bunce JD (1988) Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. J Am Chem Soc 110:5959–5967
115. Parretti MF, Kroemer RT, Rothman JH, Richards WG (1997) Alignment of molecules by Monte Carlo optimisation and molecular similarity indices. J Comput Chem 18:1344–1353
116. So SS, Karplus M (1997) Three-dimensional quantitative structure-activity relationships from molecular similarity matrices and genetic neural networks. 1. Method and validations. J Med Chem 40:4347–4359
117. Jain AN, Koile K, Chapman D (1994) Compass: Predicting biological activities from molecular surface properties. Performance comparisons on a steroid benchmark. J Med Chem 37:2315–2327
118. Bravi G, Gancia E, Mascagni P, Pegna M, Todeschini R, Zaliani A (1997) MS-WHIM, new 3D theoretical descriptors derived from molecular surface properties: A comparative 3D QSAR study in a series of steroids. J Comput Aided Mol Design 11:79–92
119. Norinder U (1990) Experimental design based 3D-QSAR analysis of steroid-protein interactions: Application to human CBG complexes. J Comput Aided Mol Design 4:381–389
120. Norinder U (1991) 3D-QSAR analysis of steroid/protein interactions: The use of difference maps. J Comput Aided Mol Design 5:419–426
121. Rum G, Herndon WC (1991) Molecular similarity concepts. 5. Analysis of steroid-protein binding constants. J Am Chem Soc 113:9055–9060
122. Simon Z, Bohl M (1992) Structure-activity relations in gestagenic steroids by the MTD method. The case of hard molecules and soft receptors. Quant Struct Activity Relatsh 11:23–28
123. Waszkowycz B, Clark DE, Frenkel D, Li J, Murray CW, Robson B, Westhead DR (1994) PRO_LIGAND: An approach to de novo molecular design. 2. Design of novel molecules from molecular field analysis (MFA) models and pharmacophores. J Med Chem 37:3994–4002
124. Dunn WJ, Wold S, Edlund U, Hellberg S (1984) Multivariate structure-activity relationships between data from a battery of biological tests and an ensemble of structure descriptors: the PLS method. Quant Struct Activity Relatsh 3: 131–137
125. Good AC, So SS, Richards WG (1993) Structure-activity relationships from molecular similarity matrices. J Med Chem 36:433–438
126. Oprea TI, Ciubotariu D, Sulea TI, Simon Z (1993) Comparison of the minimal steric difference (MTD) and comparative molecular field analysis (CoMFA) methods for analysis of binding of steroids to carrier proteins. Quant Struct Activity Relatsh 12:21–26
127. Klebe G, Abraham U, Mietzner T (1994) Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. J Med Chem 37:4130–4146
128. Hahn M, Rogers D (1995) Receptor Surface Models. 2. Application to quantitative structure-activity relationships studies. J Med Chem 38:2091–2102
129. Silverman BD, Platt DE (1996) Comparative molecular moment analysis (CoMMA): 3D-QSAR without molecular superposition. J Med Chem 39:2129–2140
130. Kellogg GE, Kier LB, Gaillard P, Hall LH (1996) E-state fields: Applications to 3D QSAR. J Comput Aided Mol Design 10:513–520
131. Anzali S, Barnickel G, Krug M, Sadowski J, Wagener M, Gasteiger J, Polanski J (1996) The comparison of geometric and electronic properties of molecular surfaces by neural networks: Application to the analysis of corticosteroid-binding globulin activity of steroids. J Comput Aided Mol Design 10:521–534
132. Norinder U (1996) 3D-QSAR Investigation of the tripos benchmark steroids and some protein-tyrosene kinase inhibitors of styrene type using the TDQ approach. J Chemom 10:533–545
133. Schnitker J, Gopalaswamy R, Crippen GM (1997) Objective models for steroid binding sites of human globulins. J Comput Aided Mol Design 11:93–110
134. Turner DB, Willett P, Ferguson AM, Heritage T (1997) Evaluation of a novel infrared range vibration-based descriptor (EVA) for QSAR studies. 1. General application. J Comput Aided Mol Design 11:409–422
135. Cramer RD III, Depriest SA, Patterson DE, Hecht P (1993) The developing practice of comparative molecular field analysis. In: Kubinyi H (ed) 3D QSAR in drug design. ESCOM, Leiden, part III, pp 443–485
136. Good AC (1995) 3D molecular similarity indices and their application in QSAR studies. In: Dean PM (ed) Molecular similarity in drug design. Blackie Academic & Professional (Capman & Hall), London, chap 2
137. Ortiz AR, Pisabarro MT, Gago F, Wade RC (1995) Prediction of drug binding affinities by comparative binding energy analysis: Application to human synovial fluid phospholipase A2 inhibitiors. In: Sanz F, Giraldo J, Manaut F (eds) QSAR and molecular modelling: Concepts, computational tools and biological applications. Prous Pub, Barcelona, sect VII, pp 439–443
138. Oprea TI, Head RD, Marshall GR (1995) The basis of cross-reactivity for a series of steroids binding to monoclonal antibody (DB3) against progesterone. A molecular modeling and QSAR study. In: Sanz F, Giraldo J, Manaut F (eds) QSAR and molecular modelling: Concepts, computational tools and biological applications. Prous Pub, Barcelona, sect VII, pp 461–462
139. Kimura T, Hasegawa K, Funatsu K (1998) GA strategy for variable selection in QSAR studies: GA-based region selection for CoMFA modeling. Chem Inf Comput Sci 38:276–282
140. Mabilia M, Belvisi L, Bravi G, Catalano G, Scolastico C (1995) A PCA/PLS analysis on nonpeptide angiotensin II receptor antagonists. In: Sanz F, Giraldo J, Manaut F (eds) QSAR and molecular modelling: Concepts, computational tools and biological applications. Prous Pub, Barcelona, sect VII, pp 456–460
141. Sulea T, Oprea T, Muresan S, Ling Chan S (1997) A different method for steric field evaluation in CoMFA improves model robustness. Chem Inf Comput Sci 37:1162–1170
142. Palomer A, Giolitti A, García ML, Fos E, Cabré F, Mauleón D, Carganico G (1995) Molecular modeling and CoMFA investigations on LTD4 receptor antagonists. In: Sanz F, Giraldo J, Manaut F (eds) QSAR and molecular modelling: Concepts, computational tools and biological applications. Prous Pub, Barcelona, sect VII, pp 444–450

143. Tetko IV, Villa AEP, Livingstone DJ (1996) Neural network studies. 2. variable selection. Chem Inf Comput Sci 36: 794–803
144. Klebe G, Abraham U, Mietzner T (1994) Molecular similarity indices in comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. J Med Chem 37:4130–4146
145. Richon AB, Young SS (2008) An introduction to QSAR methodology. A web paper of the network science corporation: http://www.netsci.org/Science/Compchem/feature19.html. Accessed 27 Oct 2008
146. So SS, van Helden SP, van Geerestein VJ, Karplus M (2000) Quantitative structure-activity relationship studies of progesterone receptor binding steroids. Chem Inf Comput Sci 40:762–772
147. Gantmacher F R (1966) Théorie des matrices, vol 2. Dunod, Paris
148. Jacobi CGJ (1846) J Reine Angew Math 30:51–94
149. Carbó-Dorca R, Besalú E, Gironés X (2000) Extended density functions. Adv Quantum Chem 38:3–63
150. Eyring H, Walter J, Kimball GE (1944) Quantum chemistry. Wiley, New York
151. Hoffmann K (1988) Banach spaces and analytic functions. Dover Publications, New York
152. Akhiezer NI, Glazman IM (1993) Theory of linear operators in Hilbert space. Dover Publications, New York
153. Landau LD, Lifshitz EM (1967) Mecánica Cuántica No-Relativista. Editorial Reverté, Barcelona
154. Vinogradov IM (1992) Encyclopaedia of mathematics, vol 8. Reidel-Kluwer Academic, Dordrecht, pp 379
155. Stewart J (1991) Advanced General Relativity. Cambridge University Press, Cambridge
156. Bach A, Amat L, Besalú E, Carbó-Dorca R, Ponec R (2000) Quantum chemistry, Sobolev spaces and SCF. J Math Chem 28:59–70
157. Carbó-Dorca R (2002) Density functions and generating wave functions. In: Sen K (ed) Reviews in modern quantum chemistry, vol I, 15, pp 401–412
158. Edwards CH Jr (1994) Advanced calculus of several variables. Dover Publications, New York
159. Carbó R, Besalú E (1996) Applications of nested summation symbols to quantum chemistry: Formalism and programming techniques. In: Ellinger Y, Defranceschi M (eds) Strategies and applications in quantum chemistry. Kluwer Academic, Dordrecht, pp 229
160. Carbó R, Besalú E (1995) Definition and quantum chemical applications of nested summation symbols and logical functions: Pedagogical artificial intelligence devices for formulae writing, sequential programming and automatic parallel implementation. J Math Chem 18:37–72
161. Breit TG; Phys. Revs. 34 (1929) 553–573; Ibid. 36 (1930) 383–397; Ibid. 39 (1932) 616–624
162. Bethe HA, Salpeter EE (1957) Quantum mechanics of one- and two-electron systems. Springer, Berlin
163. Moss RE (1973) Advanced molecular quantum mechanics. Chapman and Hall, London
164. Carbó-Dorca R, Karwowski J (2001) Theoretical and computational aspects of extended wave functions. Int J Quantum Chem 84:331–337
165. McWeeny R, Sutcliffe BT (1969) Methods of molecular quantum mechanics. Academic Press, London
166. Huzinaga S (1991) Theochem 234:51–73
167. Huzinaga S, Katsuki S, Matsuoka O (1992) J Math Chem 10:25–39
168. Bonifacic V, Huzinaga S (1974) J Chem Phys 60:2779–2786
169. Carbó R, Calabuig B (1992) Molecular quantum similarity measures and n-dimensional representation of quantum objects I. theoretical foundations. Int J Quantum Chem 42:1681–1709
170. Carbó-Dorca R, Besalú E (1996) Extending molecular similarity to electronic energy surfaces: Boltzmann similarity measures and indices. J Math Chem 20:247–261
171. Bultinck P, De Winter H, Langenaecker W, Tollenaere JP (2004) Similarity and quantitative structure-activity relationships in computational medicinal chemistry for drug discovery. Marcel Dekker, New York
172. Vinogradov IM (1988) Encyclopaedia of mathematics, vol 1. Reidel-Kluwer Academic, Dordrecht, p 334
173. Korn GA, Korn TM (1967) Manual of mathematics. Mc Graw-Hill, New York
174. Stewart J (1999) Cálculo. International Thompson Editores, México
175. Carbó R, Besalú E (1993) J Math Chem 13:331–342
176. Carbó R, Besalú E (1994) Comput Chem 18:117–126
177. Carbó-Dorca R (2000) A discussion on an apparent MO theory paradox. J Math Chem 27:35–41
178. Fraga S, García de la Vega JM, Fraga ES (1999) The Schrödinger and Riccati equations. Lecture notes in chemistry, vol 70. Springer, Berlin
179. Meyer CD (2000) Matrix analysis and applied linear algebra. SIAM, Philadelphia
180. Amat LL, Carbó-Dorca R, Ponec R (1999) Simple linear QSAR models based on quantum similarity measures. J Med Chem 42:5169–5170
181. Besalú E, Carbó R, Mestres J, Solà M (1995) In: Sen K (ed) Molecular similarity I. Top Curr Chem 173:31
182. Hansch C, Fujita T (1964) J Am Chem Soc 86:5175
183. Amat LL, Carbó-Dorca R, Ponec R (1998) Molecular quantum similarity studies as an alternative to log P values in QSAR studies. J Comput Chem 14:1575–1583
184. Carbó R, Besalú E (1994) Comput Chem 18:117

Books and Reviews

Al-Fahemi J, Cooper DL, Allan NL (2005) The quantitative use of momentum-space descriptors. Chem Phys Lett 416:376–380
Al-Fahemi J, Cooper DL, Allan NL (2005) The use of momentum-space descriptors for predicting octanol-water partition coefficients. J Mol Struct Theochem 727:57–61
Allan NL, Cooper DL (1995) Topics Curr Chem 173:85–111
Amat LL, Carbó-Dorca R (2002) Use of promolecular ASA density functions as a general algorithm to obtain starting MO in SCF calculations. Int J Quantum Chem 87:59–67
Amat LL, Besalú E, Carbó R, Fradera X (1995) Practical applications of quantum molecular similarity measures (QMSM): Programs and examples. Scientia Gerund 21:17–34
Amat LL, Constans P, Carbó R (1996) Descripció d'un algorisme d'optimització global de les Mesures de Semblança Quàntica Molecular. Scientia Gerund 22:109–121
Amat LL, Pradera X, Carbó R (1996) Sobre els mapes de Semblança Quàntica Molecular. Scientia Gerund 22:97–107

Amat LL, Robert D, Besalú E, Carbó–Dorca R (1998) Molecular quantum similarity measures tuned QSAR: An antitumoral family validation study. J Chem Inf Comput Sci 38:624–631
Amovilli C, McWeeny R (1991) J Mol Struct Theochem 227:1–9
Antolín J, Angulo JC (2008) Quantum similarity indices for atomic ionization processes. Eur Phys J D 46:21–26
Atkins PW, Friedman RS (1997) Molecular quantum mechanics. Oxford University Press, Oxford
Bach A, Carbó-Dorca R (1999) Aplicació de la Semblança Molecular Quàntica en la reducció de l'espai configuracional per a l'estat fonamental i primers exitats de l'àtom d'heli. Scientia Gerund 24:183–196
Bader RFW (1990) Atoms in molecules. Clarendon Press, Oxford
Bell JS (1993) Speakable and unspeakable in quantum mechanics. Cambridge University Press, Cambridge
Benigni R, Cotta-Ramusino M, Giorgi F, Gallo G (1995) J Med Chem 38:629–635
Benigni R, Cotta-Rasmusino M, Gallo G, Giorgi F, Giuliani A, Vari MR (2000) J Med Chem 43:3699–3707
Besalú E (2001) Fast computation of cross-validated properties in full linear leave-many-out procedures. J Math Chem 29:191–204
Besalú E, Carbó R (1995) Quantum similarity topological indices: Definition, analysis and comparison with classical molecular topological indices. Scientia Gerund 21:145–152
Besalú E, Amat L, Fradera X, Carbó R (1995) An application of the molecular quantum similarity: Ordering of some properties of the hexanes. In: QSAR and molecular modelling: Concepts, computational tools and biological applications. Proc of the 10th european symposium on structure-activity relationships. Prous Science, Barcelona, pp 396–399
Besalú E, Carbó R, Mestres J, Solà M (1995) Foundations and recent developments of quantum molecular similarity. In: Topics in current chemistry: Molecular similarity I, vol 173. Springer, Berlin, pp 31–62
Besalú E, Carbó R, Duran M, Mestres J (1995) MESSEM: A density-based quantum molecular similarity system of programs. In: Clementi E (ed) Methods and techniques in computational chemistry METECC-95. STEF, Cagliari, pp 491–508
Bethe HA, Jackiw R (1986) Intermediate quantum mechanics. Benjamin, Menlo Park
Bonaccorsi R, Scrocco E, Tomasi J (1970) J Chem Phys 52:5270–5284
Boon G, Van Alsenoy C, De Proft F, Bultinck P, Geerlings P (2005) Molecular quantum similarity of enantiomers of amino acids: a case study. J Mol Struct Theochem 727:49–56
Borgoo A, Godefroid M, Sen KD, De Proft F, Geerlings P (2004) Quantum similarity of atoms: A numerical Hartree–Fock and information theory approach. Chem Phys Lett 399:363–367
Borgoo A, Torrent-Sucarrat M, De Proft F, Geerlings P (2007) Quantum similarity study of atoms: A bridge between hardness and similarity indices. J Chem Phys 126:234104
Born M (1944) Atomic physics. Blackie & Son, London
Botella V, Pacios LF (1998) Analytic atomic electron densities in molecular se If-similarity measures and electrostatic potentials. J Mol Struct Theochem 426:75–85
Bultinck P, Carbó-Dorca R, Van Alsenoy C (2003) Quality of approximate electron densities and internal consistency of molecular alignment algorithms in molecular quantum similarity. J Chem Inf Comput Sci 43:1208–1217
Bultinck P, Rafat M, Ponec R, Van Gheluwe B, Carbó-Dorca R, Popelier P (2006) Coulomb and overlap self-similarities: A comparative selectivity analysis of structure- electron delocalization and aromaticity in linear polyacenes: Atoms in molecules multicenter delocalization ind. J Phys Chem A 110:7642–7648
Bultinck P, Ponec R, Gallegos A, Fias S, Van Damme S, Carbó-Dorca R (2006) Generalized Polansky index as an aromaticity measure in polycyclic aromatic hydrocarbons. Croatica Chemica Acta 79: 363–371
Bunge M (1979) Causality in modern science. Dover, New York
Burt C, Richards WG, Huxley PH (1990) J Comput Chem 11:1139–1146
Carbó R (1995) Molecular similarity and reactivity: From quantum chemical to phenomenological approaches. Understanding chemical reactivity, vol 14. Kluwer Academic, Dordrecht
Carbó R, Besalú E (1995) Theoretical Foundations of Quantum Molecular Similarity. In: Carbó R (ed) Molecular similarity and reactivity: From quantum chemistry to phenomenological approaches. Kluwer Academic, Dordrecht, pp 3–30
Carbó R, Calabuig B (1989) Molsimil-88: Molecular similarity calculations using a CNDO approximation. Comput Phys Commun 55:117
Carbó R, Calabuig B (1989) Sobre las Medidas de Semejanza Molecular: Una conexión entre Química Quántica e Inteligencia Artificial. Una contribucion a los Anales de la VI Escuela Latinoamericana de Química Teórica, Brasil, vol 1, pp 134
Carbó R, Calabuig B (1992) Molecular quantum similarity measures and n-dimensional representation of quantum objects II. Practical applications. Int J Quantum Chem 42:1695
Carbó R, Calabuig B (1992) Quantum molecular similarity measures and the n-dimensional representation of a molecular set: Phenyldimethylthiazines. J Mol Struct Theochem 254:517–531
Carbó R, Calabuig B (1992) Quantum similarity measures, molecular cloud description and structure-properties relationships. J Chem Inf Comput Sci 32:600–606
Carbó R, Calabuig B (1992) Quantum similarity: Definitions, computational details and applications. In: Fraga S (ed) Computational chemistry: Structure, interactions and reactivity. Elsevier, Amsterdam, pp 300–324
Carbó R, Domingo LL (1987) LCAO MO similarity measures and taxonomy. Int J Quantum Chem 32:517–545
Carbó R, Riera JM (1978) A general SCF theory. Lecture notes in chemistry, vol 5. Springer, Berlin
Carbó R, Martín M, Pons V (1977) Application of quantum mechanical parameters in quantitative structure-activity relationships, Afinidad 34:348–353
Carbó R, Suñé E, Lapeña F, Pérez J (1986) Electrostatic potential comparison and molecular metric spaces. J Biol Phys 14:21
Carbó R, Lapeña F, Suñé E (1986) Similarity measures on electrostatic molecular potentials. Afinidad 43:483
Carbó R, Calabuig B, Martínez A (1991) Semblança molecular: Representació n-dimensional d'un conjunt de molècules. Scientia Gerund 17:133
Carbó R, Molino L, Calabuig B, Besalú E, Martínez A (1992) Medidas de Semejanza Cuántica: Conceptos Clásicos y Nuevas Estrategias. Folia Chimica Theoretica Latina 21–45
Carbó-Dorca R (1998) On the statistical interpretation of density functions: ASA, convex sets, discrete quantum chemical molecular representations, diagonal vector spaces and related problems. J Math Chem 23:365–375
Carbó-Dorca R (2000) Quantum QSAR and the eigensystems

of stochastic quantum similarity matrices. J Math Chem 27:357–376
Carbó-Dorca R (2004) Heisenberg's relations in discrete n-dimensional parameterized metric vector spaces. J Math Chem 36:41–54
Carbó-Dorca R (2004) Infinite-dimensional time vectors as background building blocks of a space-time frame structure. J Math Chem 36:75–81
Carbó-Dorca R (2005) Molecular nuclear fields: A naïve perspective. J Math Chem 38:671–676
Carbó-Dorca R (2004) Non-linear terms & variational approach in quantum QSPR. J Math Chem 36:241–260
Carbó-Dorca R (2006) Descriptors and probability distributions in MO theory: Weighted Mulliken matrices and molecular quantum similarity measures. J Math Chem 39:551–591
Carbó-Dorca R (2008) Mathematical aspects of the LCAO MO first order density function (1): Atomic partition, metric structure and practical applications. J Math Chem 43:1076–1101
Carbó-Dorca R (2008) Mathematical aspects of the LCAO MO first order density function (2): Relationships between density functions. J Math Chem 43:1102–1118
Carbó-Dorca R, Besalú E (2006) Generation of molecular fields, quantum similarity measures and related questions. J Math Chem 39:495–509
Carbó-Dorca R, Bultinck P (2004) A general procedure to obtain quantum mechanical charge and bond order molecular parameters. J Math Chem 36(3):201–210
Carbó-Dorca R, Bultinck P (2004) Quantum mechanical basis for Mulliken population analysis. J Math Chem 36:231–239
Carbó-Dorca R, Bultinck P (2008) Mathematical aspects of the LCAO MO first order density function (3): A general localization procedure. J Math Chem 43:1069–1075
Chaves J, Barroso J, Bultinck P, Carbó-Dorca R (2006) Toward an alternative hardness kernel matrix structure in the electronegativity equalization method (EEM). J Chem Inf Model 46:1657–1665
Cioslowski J, Challacombe M (1991) Int J Quantum Chem S25:81–93
Cioslowski J, Fleishmann ED (1991) J Am Chem Soc 113:64–67
Cioslowski J, Mixon ST (1992) Can J Chem 70:443–449
Cioslowski J, Nanayakkara A (1993) J Am Chem Soc 115:11213–11215
Cioslowski J, Stefanov BB, Constans P (1996) J Comput Chem 17:1352–1358
Constans P, Amat L, Carbó-Dorca R (1997) Towards a global maximization of the molecular similarity function: The superposition of two molecules. J Comput Chem 18:826–846
Cooper DL, Allan NL (1989) J Comput Aided Mol Design 3:253–259
Cooper DL, Allan NL (1992) J Am Chem Soc 114:4773–4776
Cooper DL, Mort KA, Allan NL, Kinchington D, McGuidan CH (1993) J Am Chem Soc 115:12615–12616
Davidson ER (1976) Reduced density matrices in quantum chemistry. Academic Press, New York
Davydov AS (1965) Quantum mechanics. Pergamon Press, New York
Dean PM (1995) Molecular similarity in drug design. Blackie Academic & Professional, London
Dedekind R (1963) Essays on the theory of numbers. Dover, New York
Dunn JF, Nisula BC, Rodbard D (1981) J Clin Endocrinol Metab 53:58–68
Eyring H, Walker J, Kimball GE (1948) Quantum chemistry. Wiley, New York
Ferro N, Gallegos A, Bultinck P, Jacobsen HJ, Carbó-Dorca R, Reinard T (2006) Coulomb and overlap self-similarities: A comparative selectivity analysis of structure-function relationships for Auxin-like molecules. J Chem Inf Model 46:1751–1762
Ferro N, Bultinck P, Gallegos A, Jacobsen HJ, Carbó-Dorca R, Reinard T (2007) Unrevealed structural requrements for auxin-like molecules by theoretical and experimental evidences. Phytochemistry 68:237–250
Fradera X, Amat LL, Besalú E, Carbó-Dorca R (1997) Application of molecular quantum similarity to QSAR. Quant Struct Activity Relatsh 16:25–32
Fratev F, Monev V, Mehlhorn A, Polansky OE (1979) J Mol Struct 56:255–266
Fratev F, Polansky OE, Mehlhorn A, Monev V (1979) J Mol Struct 56:245–253
Gallegos A (2006) Mini-review on chemical similarity and prediction of toxicity. Curr Comput Aided Drug Des 2:105–122
Gallegos A, Girones X (2005) Theoretical Background and QSPR Application. J Chem Inf Model 45:321–326
Gallegos A, Girones X (2005) Topological quantum similarity measures: Applications in QSAR. J Mol Struct Theochem 727:97–106
Gallegos A, Patlewicz G, Worth AP (2005) The use of similarity measures in defining the applicability domain of skin sensitisation SARs. Altex-Alternativen zu Tierexperimenten 22:272
Gallegos A, Netzeva TI, Worth AP (2006) Prediction of estrogenicity: Validation of a classification model. SAR QSAR Environ Res 17:195–223
Gallegos A, Tsakovska I, Pavan M, Patlewicz G, Worth A (2007) Evaluation of SARs for the prediction of skin irritation/corrosion potential-structural inclusion rules in the BfR decision support system. SAR QSAR Environ Res 18:331–342
Gallegos A, Patlewicz G, Worth AP (2008) A review of (Q)SAR models for skin and eye irritation and corrosion. QSAR and Combinatorial Sciences 27:49–59
Geerlings P, De Proft F (2002) Chemical reactivity as described by quantum chemical methods. Int J Mol Sci 3:276–309
Geerlings P, Boon G, Van Alsenoy C, De Proft F (2005) Density functional theory and quantum similarity. Int J Quantum Chem 101:722–732
Gironés X, Amat L, Carbó-Dorca R (1999) Using molecular quantum similarity measures as descriptors in quantitative structure-toxicity relationships. SAR QSAR Environ Res 10:545–556
Gironés X, Robert R, Carbó-Dorca R (2001) TGSA: A molecular superposition program based on topo-geometrical considerations. J Comput Chem 22:255–263
Gironés X, Amat L, Carbó-Dorca R (2002) Modeling large macromolecular structures using promolecular densities. J Chem Inf Comput Sci 42:847–852
Gironés X, Carbó-Dorca R (2004) TGSA-Flex: Extending the capabilities of the topo-geometrical superposition algorithm to handle rotary bonds. J Comput Chem 25:153–159
Goldstein S (1988) Physics Today, March, 42 and April, 38
Good AC (1992) J Mol Graph 10:144–151
Good AC, Richards WG (1993) J Chem Inf Comput Sci 33:112–116
Good AC, Hodgkin EE, Richards WG (1992) J Chem Inf Comput Sci 32:188–191
Greiner W (1997) Relativistic quantum mechanics. Springer, Berlin

Gruber PM, Wills JM (1993) Handbook of convex geometry. North-Holland, Amsterdam
Ho M, Smith VH, Weaver DF, Gatti C (1998) J Chem Phys 108:5469–5475
Hodgkin EE, Richards WG (1987) Int J Quantum Chem 14 :105–110
Huzinaga S, Klobukowski M (1998) J Mol Struct Theochem 167:1–209
Jeffrey A (1995) Handbook of mathematical formulas and integrals. Academic Press, New York
Klein J, Babic D (1997) Chem Inf Comput Sci 37:656
Lee CH, Smithline SH (1994) J Phys Chem 98:1135–1138
Leherte L (2006) Similarity measures based on Gaussian-type promolecular electron density models: Alignment of small rigid molecules. J Comput Chem 27:1800–1816
Lobato M, Besalú E, Carbó R (1996) Relacions Estructura-Propietat per un conjunt d'hidrucarburs a partir de nous descriptors tridimensionals derivats de la Semblança Molecular. Scientia Gerund 22:79–86
Lobato M, Amat L, Besalú E, Carbó-Dorca R (1997) Structure-activity relationships of a steroid family using quantum similarity measures and topological quantum similarity indices. Quant Struct Activity Relatsh 16:465–472
Lobato M, Amat L, Besalú E, Carbó-Dorca R (1998) Estudi QSAR d'una família de quinolones utilitzant indexs de semblança i indexs topològics de semblance. Scientia Gerund 23:17–27
Ludeña EV (1987) In: Fraga S (ed) Química Teórica, Nuevas Tendencias, vol 4. CSIC, Madrid, pp 117–160
Löwdin PO (1955) Quantum theory of many-particle systems. I. Physical interpretations by means of density matrices, natural spin-orbitals, and convergence problems in the method of configuration interaction. Phys Rev 97:1474–1489
Löwdin PO (1955) Quantum theory of many-particle systems. II. Study of the ordinary hartree-fock approximation. Phys Rev 97:1490–1508
Löwdin PO (1955) Quantum theory of many-particle systems. III. Extension of the Hartree–Fock scheme to include degenerate systems and correlation effects. Phys Rev 97:1509–1520
March NH (1992) Electron density theory of atoms and molecules. Academic Press, London
Mc Weeny R (1959) Proc Roy Soc A 253:242–259
Mc Weeny R (1960) Some recent advances in density matrix theory. Rev Mod Phys 32:335–369
Mc Weeny R (1978) Methods of molecular quantum mechanics. Academic Press, London
Mestres J, Solà M, Duran M, Carbó R (1994) On the calculation of ab initio quantum molecular similarities for large systems. J Comput Chem 15:1113–1120
Mestres J, Solà M, Duran M, Carbó R (1994) On the use of ab initio quantum molecular similarities as an interpretative tool for the study of chemical reactions. J Am Chem Soc 116:5909–5915
Mestres J, Solà M, Carbó R (1995) First-order molecular descriptors for molecular steric similarity. Scientia Gerund 21:165–173
Mestres J, Solà M, Carbó R, Luque FJ, Orozco M (1996) Effect of solvation on the charge distribution of a series of anionic, neutral, and cationic species. A quantum molecular similarity study. J Phys Chem 100:606–610
Mezey PG (1993) Shape in chemistry: An introduction to molecular shape and topology. VCH, New York
Mezey PG (1995) Molecular similarity I. In: Sen K (ed) Topics in current chemistry, vol 173. Springer, Berlin, pp 63–83
Mezey PG (1998) Averaged electron densities for averaged conformations. J Comput Chem 19:1337–1344
Mezey PG (2005) Graph representations of molecular similarity measures based on topological resolution. J Comput Methods Sci Eng 5:109–114
Mezey PG, Ponec R, Amat L, Carbó-Dorca R (1999) Quantum similarity approach to the characterization of molecular chirality. Enantiomers 4:371–378
Myers RH (1990) Classical and modern regression with applications. PWS-KENT Publishing company, Boston
Netzeva TI, Gallegos A, Worth AP (2006) Topological quantum similarity indices based on fitted densities: Comparison of the applicability domain of a QSAR for estrogenicity with a large chemical inventory. Environ Toxicol Chem 25:1223–1230
Ortiz JB, Cioslowski J (1991) Chem Phys Lett 185:270–275
Paniagua JC, Alemany P (1999–2000) Química Quàntica. Vols 1 and 2, Barcelona, 1999–2000
Parr RG (1963) The quantum theory of molecular electronic structure. WA Benjamin, New York
Patlewicz G, Jeliazkova N, Gallegos Saliner A, Worth AP (2008) Toxmatch – A new software tool to aid in the development and evaluation of chemically similar groups. SAR QSAR Environ Res 19:397–412
Pauling L, Wilson EB Jr (1985) Introduction to Quantum mechanics. Dover, New York
Petke JD (1993) J Comput Chem 14:928–933
Pierre DA (1969) Optimization theory with applications. Wiley, New York
Pilar FL (1990) Elementary quantum chemistry. McGraw-Hill, Princeton
Pla L (1986) Análisis Multivariado: Método de Componentes Principales Monography no 27. Secretaría General de la OEA, Washington,
Ponec R (1993) J Chem Inf Comput Sci 33:805–811
Ponec R (1995) Overlap determinant method in the theory of pericyclic reactions. Lecture notes in chemistry, vol 65. Springer, Berlin
Ponec R, Strnad M (1990) Collect Czechoslov Chem Commun 55:2583–2589
Ponec R, Strnad M (1990) Collect Czechoslov Chem Commun 55:896–902
Ponec R, Strnad M (1991) J Math Chem 8:103–112
Ponec R, Strnad M (1991) J Phys Organic Chem 4:701–705
Ponec R, Strnad M (1992) Int J Quantum Chem 42:501–508
Ponec R, Amat LL, Carbó-Dorca R (1999) Molecular basis of quantitative structure-properties relationships (QSPR): A quantum similarity approach. J Comput Aided Mol Design 13:259–270
Ponec R, Amat LL, Carbó-Dorca R (1999) Quantum similarity approach to LFER: Substituent and solvent effects on the acidities of carboxylic acids. J Phys Organic Chem 12:447–454
Riera A (1992) J Mol Struct Theochem 259:83–98
Robert D, Carbó-Dorca R (1998) A formal comparison between molecular quantum similarity measures and indices. J Chem Inf Comput Sci 38:469–475
Robert D, Carbó-Dorca R (1998) Analyzing the triple density molecular quantum similarity measures with the INDSCAL model. J Chem Inf Comput Sci 38:620–623
Robert D, Carbó-Dorca R (1998) On the extension of QS to atomic nuclei: Nuclear QS. J Math Chem 23:327–351
Robert D, Carbó-Dorca R (1998) Structure-property relationships in nuclei. Prediction of the binding energy per nucleon

using a quantum similarity approach. Il Nuovo Cimento A111:1311–1321
Robert D, Amat LL, Carbó-Dorca R (1999) Three-dimensional quantitative structure-activity relationships from tuned molecular quantum similarity measures. Prediction of the corticosteroid-binding globulin binding affinity for a steroid family. Chem Inf Comput Sci 39:333–344
Safouhi H (2005) Analytical and numerical development for the two-centre overlap-like quantum similarity integrals over Slater-type functions. J Phys A: Mathematical and General 38:7341–7361
Sen K (1995) Molecular similarity. Topics in current chemistry. vols 173, 174. Springer, Berlin
Shavitt I (1977) In: Schaefer HF III (ed) Methods of electronic structure theory. Modern theoretical chemistry, vol 3. Plenum Press, New York, pp 189–275
Sobolev SL (1938) Math Sb 4:471
Solà M, Mestres J, Duran M, Carbó R (1994) Ab initio quantum molecular similarity measures on metal-substituted carbonic anhydrase (M (II) CA, M=Be, Mg, Mn, Co, Ni, Cu, Zn, and Cd). J Chem Inf Comput Sci 34:1047–1053
Solà M, Mestres J, Carbó R, Duran M (1996) A Comparative analysis by means of Quantum Molecular Similarity Measures of Density Distributions derived from conventional ab initio and Density Functional Methods. J Chem Phys 104:636–647
Solà M, Mestres J, Oliva JM, Duran M, Carbó C (1996) The use of ab initio quantum molecular selfsimilarity measures to analyze electronic charge density distributions. Int J Quantum Chem 58:361–372
Spiegel MR (1968) Mathematical handbook. McGraw-Hill, New York
Stoer J, Witzgall CH (1970) Convexity and optimization in finite dimensions. Die Grundlehren der matematischen Wissenschaften in Einzeldarstellungen, Band 163. Springer, Berlin
Trillas E, Alsina C, Terricabras JM (1995) Introducción a la Lógica Difusa. Ariel Matemática, Barcelona
Tsakovska I, Gallegos A, Netzeva T, Pavan M, Worth AP (2007) Evaluation of SARs for the prediction of eye irritation/corrosion potential-structural inclusion rules in the BfR decision support system. SAR QSAR Environ Res 18:221–235
Vinogradov IM (1989) Encyclopaedia of Mathematics, vol 4. Kluwer Academic, Dordrecht, pp 422–428
Vracko M, Bandelj V, Barbieri P, Benfenati E, Chaudhry Q, Cronin M, Devillers J, Gallegos A, Gini G, Gramatica P, Helma C, Neagu D, Netzeva T, Pavan M, Patlewicz G, Randic M, Tsakovska I, Worth A (2006) Validation of counter propagation neural network models for predictive toxicology according to the OECD principles. A Case Study. SAR QSAR Environ Res 17: 265–284
Wagener M, Sadowski J, Gasteiger J (1995) J Am Chem Soc 117:7769–7775
Whitley DC, Ford MG, Livingstone DJ (2000) Unsupervised forward selection: A method for eliminating redundant variables. Chem Inf Comput Sci 40:1160–1168
Wilkinson JH (1965) The algebraic eigenvalue problem. Clarendon Press, Oxford
Wilkinson JH, Reinsch C (1971) Linear álgebra. Springer, Berlin
Worth AP, Bassan A, de Bruijn J, Gallegos A, Netzeva T, Patlewicz G, Pavan M, Tsakovska I, Eisenreich E (2007) The role of the european chemicals bureau in promoting the regulatory use of (Q)SAR methods. SAR QSAR Environ Res 18:111–125
Zadeh LA (1965) Inform Control 8:338

Quantum Simulations of Ballistic Nanowire Field Effect Transistors

MINCHEOL SHIN
School of Engineering, Information and Communications University, Yuseong, Korea

Article Outline

Glossary

Silicon nanowire field effect transistors A silicon nanowire field effect transistor (SNWFET) has silicon nanowire as the channel, whose cross-sectional area is typically $10 \sim 100\,\mathrm{nm}^2$. The charge transport mainly occurs in one-dimensional subbands which are formed within the channel due to the strong quantum confinement. With enhanced gate control by three-dimensional gates surrounding the channel, SNWFETs can outperform conventional planar metal-oxide-semiconductor field effect transistors (MOSFETs) in their ultimate scaling limit.

Carbon nanotube field effect transistors Carbon nanotube field effect transistors (CNTFETs) resemble SNWFETs except that semiconducting carbon nanotubes are used as the channel material instead of silicon nanowires. Due to intrinsically nano-scale size, good electrical property, and almost ballistic transport nature of carbon nanotubes, CNTFETs exhibit device performance well exceeding conventional MOSFETs. CNTFETs with Schottky-barrier contacts or with doped source/drain can be realized.

Quantum transport in semi-conductor devices As the feature size of semi-conductor devices becomes extremely small, the wave nature of the electrons should prevail in their transport in the devices. The governing equation is the Schrödinger equation and typically single-particle Schrödinger equation with Hartree potential is sufficient for device simulations.

Non-equilibrium green's function The Schrödinger equation with open boundaries can be solved by us-

ing the non-equilibrium Green's Function (NEGF) approach, where the Green's function is defined as the impulse response function of the system Hamiltonian. NEGF approach is formally equivalent to other approaches such as quantum transmitting boundary method, but has advantages in inclusion of individual scattering/interaction terms. A key part of NEGF method is to calculate the self-energies, which contain information about the contacts and scattering or interaction that are considered.

Definition of the Subject

Nanowire FETs such as SNWFETs and CNTFETs are considered as strong candidates for the future nano-electronic devices to replace the today's planar MOSFETs. The devices are intrinsically nano-scale so quantum effects such as the size quantization, barrier tunneling, and interference should prevail. In particular, as the channel length of the nanowire is reduced below 10 nm or so, the source-to-drain tunneling current importantly contributes to the total current. In order to characterize and predict their device performance accurately, quantum device simulations based on the direct solution of the Schrödinger wave equations should be performed. The NEGF method, among others, provides a powerful solution scheme for quantum device simulations. Modeling and simulations are gaining greater importance as computer experiments, particularly in the nano-electronic device area where performance estimation and optimization of newly developed devices are necessary to reduce efforts and costs of real experiments.

Introduction

According to the ITRS 2007 [15], MOSFET devices with sub-10 nm channel length are expected to be fabricated by 2015. In their ultimate scaling regime, performance of conventional planar MOSFETs will be seriously degraded mainly due to the short channel effects [10]. As new device architectures to overcome the problems, nanowire FETs such as SNWFETs and CNTFETs have been recently drawing attention [5,7,16,25,33]. The advantage of SNWFETs mainly lies on the presence of the multiple gates around the channel which can suppress the short channel effects through the enhanced gate control [7,25,33]. In CNTFETs, intrinsically nano-scale carbon nanotubes with excellent electrical properties are used as the channel material so device performance exceeding conventional MOSFETs can be obtained [5,16]. To characterize the new devices and predict their performance accurately in the nano-scale regime, device simulations with solid quantum mechanical treatment are necessary.

For SNWFETs, quantum-mechanical studies have been mostly performed based on the parabolic effective mass theory (PEMT) [6,22,28,30]. PEMT provides a simple Hamiltonian to be dealt with so that efficient simulations employing NEGF method can be performed, especially for SNWFETs with homogeneous cross-sections and ballistic transport. For ultra-thin Si nanowires, however, their bulk property in the transverse direction is not preserved and so their band structure becomes different from that of bulk silicon. In fact, recent studies have revealed that if the cross-sectional area of the silicon nanowire is less than about $3 \times 3\,\mathrm{nm}^2$ non-parabolic and band-edge shift effects becomes so great that PEMT is no longer valid [12,24,31]. In this case, atomistic calculations such as full-band tight binding (TB) or first-principle calculations are necessary to obtain the correct dispersion relationship [12,20,23,24,31]. As practical device simulation approaches, however, they have limitations due to the enormous computational burden. A hybrid approach may give an optimal solution: namely, one continues to use PEMT for ultra-thin Si nanowire channel through appropriate tuning of effective masses and band gap from the atomistic calculations, which has been shown to be valid in recent works [24,31].

For CNTFETs, various approaches ranging from semi-classical to full-band quantum approaches have been implemented for device simulations [2,3,11,13,18,26]. The models yield comparable simulation results in the region above the threshold, but accurate dispersion relationships from atomistic calculations are needed to correctly account for intra-band and inter-band tunneling in the sub-threshold region. For that purpose, the single-band (p_z) TB approach seems to be an optimal one: One may consider more sophisticated Hamiltonians such as full-band or Hückel TB Hamiltonians, but as long as electronic transport is concerned the single-band TB seems to be sufficient because they yield almost the same dispersion relationships [21]. An effective mass approach can be also taken, but the computational complexity is not greatly reduced compared to the TB approach, while loss of accuracy is inevitably incurred due to simplifications involved.

In nanowire FETs of a few tens nanometer in size, diffusive transport due to electron-phonon interaction and surface roughening cannot be entirely ignored and there have recently been considerable efforts to include the scattering effects [14,17,32]. In this article, however, ballistic transport is assumed throughout, since our focus is on the ultimately scaled devices where ballistic transport is expected to prevail and may account for major device characteristics. Simulation based on the ballistic transport alone may also serve as a starting point to build more so-

phisticated simulation tools including scattering and other effects.

In this article, simulation methods employing PEMT and single band TB for SNWFETs and CNTFETs respectively are described in some details as practical approaches for quantum device simulations. Model systems and their Hamiltonians are firstly introduced, followed by NEGF method in the real and mode spaces. Details on the k-space and product-space solutions of cross-sectional Schrödinger equations for SNWFETs are next described. Numerical aspects of three-dimensional Poisson's equations are then addressed. Some simulation results for nanowire FETs are lastly shown, followed by conclusion and future prospects.

Model Systems

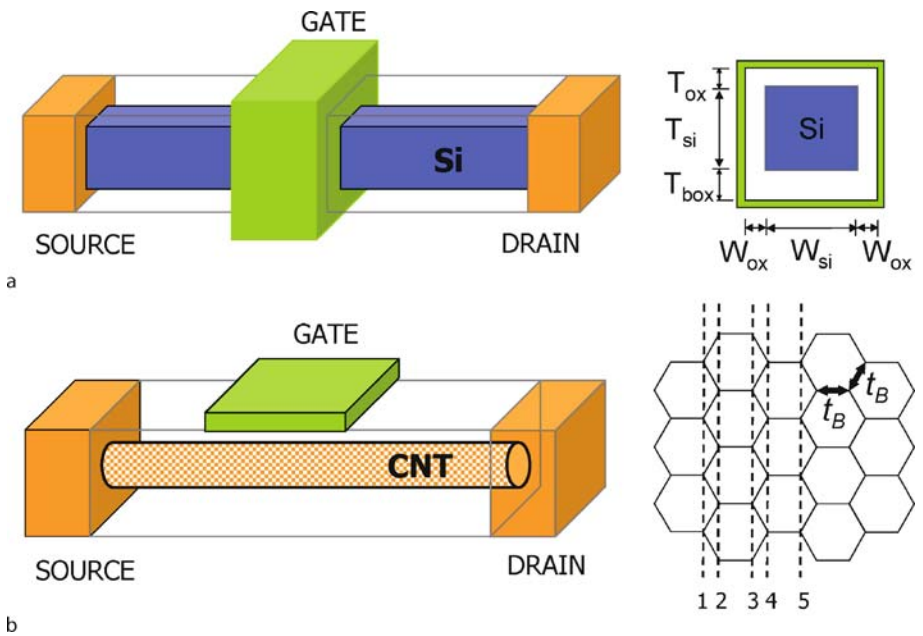
Figure 1a shows a schematic diagram of a SNWFET which is a three-dimensional (3D) structure with multiple gates around the silicon nanowire channel. The source (S) and drain (D) regions are either heavily doped (doped S/D device) or metallic such that Schottky barriers exist at the interfaces with the channel (SB device). The channel region is intrinsic or lightly p-doped and a metallic gate with the mid-gap work function of 4.6 eV is assumed here. Possible types of multiple gates around the nanowire channel are Tri, Pi, Omega, Gate-all-around (GAA) [6,29].

In CNTFETs, single-walled (n,0) zigzag nanotubes are used as the channel material as shown in Fig. 1b. For ohmic-contact devices, source and drain regions are assumed to be infinite carbon nanotubes, which are simple extension of the channel nanotube but heavily doped. If the source and drain regions are both n-doped (or p-doped) and channel is intrinsic, the devices resemble MOSFETs and thus called CNT MOSFETs [4,13]. If source-channel-drain is of $p-i-n$ type (or $n-i-p$), the devices operate like tunnel FETs [4]. For Schottky-contact devices, on the other hand, S/D regions are assumed to be metallic zigzag nanotubes such that Schottky barriers of height ϕ_{Bn}, which is given as an input, is formed at the interfaces with channel. Three gate-types, coaxial, semi-cylindrical, and top gates, are considered. The coaxial gate device is in effect two-dimensional (2D) device, with 2D electrostatics and transport, while the other two are truly 3D devices.

Simulation Methods

Self-consistent Method

In our device simulations of nanowire FETs, the quantum transport of electrons is described by the Schrödinger equation with open boundary conditions and electrostatics are given by the 3D Poisson's equation. To correctly describe highly out-of-equilibrium states due to the finite



Quantum Simulations of Ballistic Nanowire Field Effect Transistors, Figure 1
Device Schematics: a SNWFET with its cross-section on the right and b CNTFET with zigzag carbon nanotube channel

source-drain voltage, the two equations should be solved together to yield the self-consistent solution. Namely, we iteratively solve the Schrödinger equation

$$H\{\phi\}\psi = E\psi \,, \tag{1}$$

to obtain the electron density $n_{3D} = |\psi|^2$, where H is the system Hamiltonian for a given potential profile $\{\phi\}$ and ψ is the wave function, and the Poisson's equation

$$\nabla^2\phi = -\frac{q_0}{\epsilon}\left(N_d - n_{3D}\right) \,, \tag{2}$$

to obtain the potential profile, where N_d is doping density. Once the self-consistency is reached, the drain current is calculated. Among other approaches to solve the Schrödinger equation with open boundaries, the non-equilibrium Green's function approach [9] is employed and described here.

EMT Hamiltonian for SNWFETs

The effective mass Hamiltonian for the SNWFET in Fig. 1a is given by

$$H = H_\perp + H_\parallel \,, \tag{3}$$

where

$$H_\perp = -\frac{\hbar^2}{2}\frac{\partial}{\partial y}\left(\frac{1}{m_y^*}\frac{\partial}{\partial y}\right) - \frac{\hbar^2}{2}\frac{\partial}{\partial z}\left(\frac{1}{m_z^*}\frac{\partial}{\partial z}\right) + V(x,y,z), \tag{4}$$

$$H_\parallel = -\frac{\hbar^2}{2m_x^*}\frac{\partial^2}{\partial x^2}, \tag{5}$$

where m_x^*, m_y^*, and m_z^* are effective masses in the x, y, and z directions, respectively, and $V(x,y,z)$ is the conduction band-edge. Notice that m_y^* and m_z^* are in general functions of y and z if both silicon and oxide regions are taken into account, while m_x^* is the silicon effective mass in the transport direction. For the effective masses, one may simply use the values of bulk Si; for instance, in the [100] transport direction, $(m_x^*, m_y^*, m_z^*) = (m_t, m_t, m_l)$, (m_t, m_l, m_t), and (m_l, m_t, m_t), for the three valleys respectively, where $m_t = 0.18m_0$ and $m_l = 0.98m_0$, and each valley has the degeneracy of 2. For other transport directions, one can refer to Ref. [27]. Effective masses in the oxide region do not importantly affect the vertical confinement energies nor lateral transport, so one may extend the Si effective masses into the oxide region for computational simplicity.

The Si effective masses can be also extracted from the band diagrams produced by full-band TB or first principle calculations. In the original work of Ko et al. [20], a $sp^3d^5s^*$ TB model was considered and the conduction bands of a Si (100) wire were obtained: Two pairs of valleys out of six valleys with k vectors perpendicular to the wire direction were found to fall onto the Γ point in the wire Brillouin zone, and the electron effective masses along the confinement directions are relatively larger for the states than for the two remaining states that become off-Γ states in the wire. Similar works performed by other groups [24,31] and a recent DFT calculation [12] all indicate that the Si effective masses start to deviate from their bulk values if the wire cross-sectional area becomes smaller than about 10 nm^2.

For a nanowire with circular cross-section, $H_\perp$ can be expressed in the circular polar coordinates, in an approximation,

$$H_\perp = -\frac{\hbar^2}{2m_\perp^*}\left(\frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r} + \frac{1}{r^2}\frac{\partial^2}{\partial\theta^2}\right) \,, \tag{6}$$

where

$$\frac{1}{m_\perp^*} = \frac{1}{2}\left(\frac{1}{m_x^*} + \frac{1}{m_y^*}\right). \tag{7}$$

In the real-space, if we use the finite difference scheme and slice the nanowire along the transport direction into N_x cross-sections, $H_\perp$ becomes a block diagonal matrix with diagonal elements $H_{\perp,i}$ of size $N_s \times N_s$ which is the 2D Hamiltonian of the ith cross-section, where N_s is the number of grid points of a cross-section, and $H_\parallel$ becomes a block tridiagonal matrix with diagonal elements

$$\alpha = 2t_x \cdot I_{N_s\times N_s} \,, \tag{8}$$

and off-diagonal elements

$$\beta = -t_x \cdot I_{N_s\times N_s} \,, \tag{9}$$

where $I_{N_s\times N_s}$ is an $N_s \times N_s$ identity matrix and $t_x = \hbar^2/2m_x^*(\Delta x)^2$, where Δx is the grid spacing in the transport direction.

TB Hamiltonian for CNTFETs

A $(n, 0)$ zigzag nanotube has alternating sublattices, as shown in Fig. 1b. In the nearest-neighborhood single-band (p_z) tight binding description, the coupling matrices between the sublattices are given by t and b, where

$$t = t_B \cdot I_{n\times n} \,, \tag{10}$$

$$b = t_B \begin{bmatrix} 1 & 1 & & & 1 \\ & 1 & 1 & & \\ & & & \ddots & \\ 1 & & & & 1 \end{bmatrix}, \tag{11}$$

where $I_{n\times n}$ is an $n \times n$ identity matrix and t_B is the TB coupling parameter. Then the Hamiltonian is written as

$$H = \begin{bmatrix} a_0 & b^\dagger & & & & \\ b & a_1 & t & & & \\ & t & a_2 & b & & \\ & & b^\dagger & a_3 & t & \\ & & & t & a_4 & \\ & & & & & \ddots \end{bmatrix} \tag{12}$$

where a_i is the coupling matrix within the ith sublattice, which is $n \times n$ diagonal matrix with elements

$$[a_i]_{j,j} = -q_0\,\phi(i,j)\,, \tag{13}$$

where j is the index along the circumferential direction of nanotube ($1 \le j \le n$) and $\phi(i,j)$ is the vacuum level potential at the atom site (i,j) as determined by the Poisson's equation.

NEGF Approach

In the NEGF approach [9], the device Green's function $G(E)$ is given by

$$G(E) = (E - H - \Sigma_{\rm L} - \Sigma_{\rm R})^{-1}\,, \tag{14}$$

where H is the device Hamiltonian and $\Sigma_{\rm L,R}$ are contact self-energies of semi-infinite leads to the left (L) and right (R) of the device region, respectively. The ballistic transport is assumed here as only the contact self-energies are considered, which are formally expressed by

$$\Sigma_{\rm L,R} = \chi^\dagger_{\rm L,R}(E - H_{\rm L,R})^{-1}\chi_{\rm L,R}, \tag{15}$$

where $H_{\rm L,R}$ are Hamiltonians of the left and right leads, respectively, and $\chi_{\rm L,R}$ are coupling matrices between the device and the leads.

Given the Green's function $G(E)$, the charge density at position $\vec{r}$ can be calculated as follows. Let us first define

$$\rho_{\rm L,R}(\vec{r},E) = \left[G(E)\Gamma_{\rm L,R}(E)G^\dagger(E)\right]_{\vec{r},\vec{r}} \tag{16}$$

where

$$\Gamma_{\rm L,R}(E) = i\left(\Sigma_{\rm L,R}(E) - \Sigma^\dagger_{\rm L,R}(E)\right) \tag{17}$$

and $[\,\cdot\,]_{\vec{r},\vec{r}}$ denotes the diagonal element of the relevant matrix corresponding to the position $\vec{r}$.

The electron density $n_{\rm 3D}(\vec{r})$ in n-type SNWFETs, where only the electron transport in n-type FETs is considered, is given by

$$n_{\rm 3D}(\vec{r}) = \frac{1}{2\pi\Delta V} \cdot \int_{-\infty}^{\infty} {\rm d}E\left(f^+_{\rm L}\rho_{\rm L}(\vec{r},E) + f^+_{\rm R}\rho_{\rm R}(\vec{r},E)\right)\,, \tag{18}$$

where ΔV is the volume element, whereas the electron density $n_{\rm 2D}(\vec{r})$ and the hole density $p_{\rm 2D}(\vec{r})$ on the CNT surface in CNTFETs are given by

$$n_{\rm 2D}(\vec{r}) = \frac{1}{2\pi\Delta S}\int_{E_{\rm n}(\vec{r})}^{\infty} {\rm d}E\left(f^+_{\rm L}\rho_{\rm L}(\vec{r},E) + f^+_{\rm R}\rho_{\rm R}(\vec{r},E)\right)\,, \tag{19}$$

$$p_{\rm 2D}(\vec{r}) = \frac{1}{2\pi\Delta S}\int_{-\infty}^{E_{\rm n}(\vec{r})} {\rm d}E\left(f^-_{\rm L}\rho_{\rm L}(\vec{r},E) + f^-_{\rm R}\rho_{\rm R}(\vec{r},E)\right)\,, \tag{20}$$

where $E_{\rm n}(\vec{r}) = -q_0\phi(\vec{r})$ is the charge neutrality level and ΔS is the area element. In Eqs. (18)–(20),

$$f^{\pm}_{\rm L,R}(E) = \frac{1}{1 + {\rm e}^{\pm\left(E - E^{L,R}_{\rm F}\right)/k_{\rm B}T}}\,, \tag{21}$$

where $E^L_{\rm F} = E_{\rm F}$ is the Fermi energy at the source and $E^R_{\rm F} = E_{\rm F} - q_0 V_{\rm d}$ is the Fermi energy at the drain, where $V_{\rm d}$ is the drain voltage.

The drain current $I_{\rm d}$ is calculated by using the Landauer–Büttiker formula:

$$I_{\rm d} = \frac{2q}{h}\int_{-\infty}^{\infty} {\rm d}E\; T(E)(f^+_{\rm L}(E) - f^+_{\rm R}(E)) \tag{22}$$

where the transmission probability $T(E)$ is

$$T(E) = {\rm Tr}(\Gamma_{\rm L}G\Gamma_{\rm R}G^\dagger) \tag{23}$$

Mode-Space NEGF

The idea of the mode-space NEGF is basically to perform a unitary transformation from the real space to the "mode" space and to consider only the modes which contribute to the transport [8]. If the unitary matrix is denoted by U, we form

$$\tilde{H} = U^\dagger H U. \tag{24}$$

We can then deal with $\tilde{H}$ instead of the original Hamiltonian and obtain the self-energies, density matrix and current using the transformed Hamiltonian. For the density matrix, however, inverse unitary transformation back to the real-space is needed to obtain the charge density in the real-space.

For SNWFETs described by the effective mass Hamiltonian of Eq. (3), the 2D Hamiltonian $H_{\perp,i}$ of each cross-section is diagonalized by the unitary matrix U_i to yield

$$\tilde{H}_{\perp,i} = U^\dagger_i H_{\perp,i} U_i\,, \tag{25}$$

and the transformation matrix U may be formed as a block diagonal matrix having diagonal elements U_i. On the

other hand, for CNTFETs described by the TB Hamiltonian of Eq. (12), the coupling matrix b is diagonalized as

$$\tilde{b} = U_b^\dagger \, b \, U_b \, , \tag{26}$$

and the block-diagonal unitary matrix U may then be formed to have identical diagonal elements U_b.

The advantage of the mode-space approach is that one may handle with matrices of much smaller size. That is, if the number of modes that effectively contribute to the transport is N_M out of total N_s, the size of H is reduced from $N_x N_s \times N_x N_s$ in the real-space to $N_x N_M \times N_x N_M$ in the mode-space. For instance, for a SNWFET with rectangular cross-section of $5 \times 5\,\mathrm{nm}^2$, N_M is about 10, whereas a moderate real-space meshing would yield $N_s \sim 1000$ so computational efficiency is much greater in the mode-space compared to the real-space.

Uncoupled Mode-space NEGF for SNWFETs

If the coupling between different transport modes in the mode-space is non-existent or can be ignored, the transport problem effectively becomes that of 1D transport modes in parallel [8,13]. For SNWFETs, the so-called uncoupled mode-space NEGF is resulted if we assume that the unitary matrices U_i of Eq. (25) has the following property;

$$U_{i+1}^\dagger U_i \approx I_{N_s \times N_s} \tag{27}$$

i. e., the eigenfunctions of neighboring cross-sectional planes are orthogonal to each other in an approximation. Then

$$\tilde{H} = U^\dagger H U = U^\dagger H_\perp U + U^\dagger H_\parallel U \approx \tilde{H}_\perp + H_\parallel \, , \tag{28}$$

where the block diagonal matrix $\tilde{H}_\perp$ has diagonal elements $\tilde{H}_i$ of Eq. (25). Thus, $\tilde{H}$ becomes a block tridiagonal matrix with elements which are diagonal matrices of size $N_s \times N_s$. By interchanging rows and columns appropriately, $\tilde{H}$ then becomes block-diagonal with elements $\tilde{H}^{(m)}$, where

$$\tilde{H}^{(m)} = \begin{bmatrix} \epsilon_1^{(m)} + 2t_x & -t_x & & & \\ -t_x & \epsilon_2^{(m)} + 2t_x & -t_x & & \\ & & \ddots & & \\ & & & & -t_x \\ & & & -t_x & \epsilon_{N_x}^{(m)} + 2t_x \end{bmatrix}, \tag{29}$$

where $\epsilon_i^{(m)}$ is the mth eigenvalue of the ith cross-sectional plane. $\tilde{H}^{(m)}$ of Eq. (29) corresponds to the 1D Hamiltonian of the mth mode, with "effective" potential $\epsilon^{(m)}(x)$ in the transport direction. The modes are therefore completely uncoupled: With the charge density $n_{1\mathrm{D}}^{(m)}$ of each mode, the 3D charge density is obtained by

$$n_{3\mathrm{D}}(x_i, y_j, z_k) = \sum_m n_{1\mathrm{D}}^{(m)}(x_i) |[U_i]_{jk,m}|^2, \tag{30}$$

and the total drain current is given by

$$I_\mathrm{d} = \sum_m I_{\mathrm{d},m} \, , \tag{31}$$

where $I_{\mathrm{d},m}$ is the current of the mth mode.

In the uncoupled mode-space approach the original 3D problem is therefore split into problems of solving the 2D Schrödinger equations in the cross-sectional planes and the 1D Schrödinger equation in the transport direction. Note that the uncoupled mode-space approach is well suited for the device simulation considered here, because the silicon-on-insulator structure ensures the confinement of the cross-sectional wave functions in the silicon channel and consequently the shape of the 2D wave functions do not change much in the lateral direction.

Uncoupled Mode-Space NEGF for CNTFETs

For CNTFETs with coaxial gates, electrostatic potential $\phi(i)$ at the ith sublattice is same along the circumferential direction. Thus, by the unitary transformation of Eq. (26), a_i's in Eq. (12) remain intact so $\tilde{H}$ becomes block-tridiagonal with elements being $n \times n$ diagonal matrices. As in the case of the SNWFETs, by interchanging rows and columns $\tilde{H}$ becomes block-diagonal with elements $\tilde{H}^{(m)}$, where

$$\tilde{H}^{(m)} = \begin{bmatrix} -q_0\phi(0) & \tilde{b}_m & & & \\ \tilde{b}_m & -q_0\phi(1) & t_\mathrm{B} & & \\ & t_\mathrm{B} & -q_0\phi(2) & \tilde{b}_m & \\ & & \tilde{b}_m & -q_0\phi(3) & t_\mathrm{B} \\ & & & & \ddots \end{bmatrix}. \tag{32}$$

The charge density on the CNT surface and the drain current I_d can be calculated similarly to the case of the SNWFETs. In particular, the electron density

$$n_{2\mathrm{D}}(x_i, y_j) = \sum_m n_{1\mathrm{D}}^{(m)}(x_i) \left|[U_b]_{j,m}\right|^2 = \frac{1}{\Delta y} \sum_m n_{1\mathrm{D}}^{(m)}(x_i) \tag{33}$$

where $\Delta y = 2\pi R_c / n$ where R_c is the radius of the CNT. Notice that the uncoupled mode-space approach for the co-axial CNTFETs is exact [13].

Surface Green's functions

For SNWFETs described by the effective mass Hamiltonian of Eqs. (3)–(5), the contact self-energies $\Sigma_{\rm L}$ and $\Sigma_{\rm R}$ in Eq. (15) have non-zero elements $\beta^\dagger g_{\rm L}\beta$ and $\beta\, g_{\rm R}\, \beta^\dagger$ in the first and last $N_s \times N_s$ blocks, respectively, due to the structure of the coupling matrices $\chi_{\rm L}$ and $\chi_{\rm R}$. The surface Green's functions $g_{\rm L,R}$ can be rather simply obtained as follows. Assuming that the reservoir to the left of the device is simple extension of the left contact, $g_{\rm L}$ can be written as

$$g_{\rm L} = \left(\gamma - \beta^\dagger g_{\rm L}\beta\right)^{-1}, \tag{34}$$

where a $N_s \times N_s$ matrix

$$\gamma = E \cdot I_{N_s\times N_s} - h_0 - \alpha\,, \tag{35}$$

where h_0 is the 2D Hamiltonian at the contact cross-section. If U_0 is the matrix which diagonalizes h_0 and we denote the unitary-transformed matrix $\tilde{A}$ of a matrix A as $\tilde{A} \equiv U_0^\dagger A U_0$, Eq. (34) becomes under the unitary transformation by U_0,

$$\tilde{g}_{\rm L} = \left(\tilde{\gamma} - \beta^\dagger \tilde{g}_{\rm L}\beta\right)^{-1}. \tag{36}$$

The matrices in the equation are all diagonal matrices, thus element kl of the matrix $\tilde{g}_{\rm L}$ is

$$[\tilde{g}_{\rm L}]_{k,l} = \frac{\tilde{\gamma}_k - i\sqrt{4t_x^2 - \tilde{\gamma}_k^2}}{2t_x^2}\delta_{k,l}\,, \tag{37}$$

where $\delta_{k,l}$ is the Kronecker delta. The real-space $g_{\rm L}$ can be finally obtained by inverse unitary transformation. The surface Green's function $g_{\rm R}$ at the drain contact can be calculated likewise.

For CNTFETs described by the TB Hamiltonian of Eq. (12), the mode-space surface Green's function $\tilde{g}_{\rm L}(= U_b^\dagger g_{\rm L} U_b)$ at the source contact can be calculated by solving the following coupled matrix equations:

$$\tilde{g}_{\rm L} = \left(\tilde{c} - \tilde{b}^\dagger \tilde{g}'_{\rm L}\tilde{b}\right)^{-1}, \tag{38}$$

$$\tilde{g}'_{\rm L} = \left(\tilde{c} - t_{\rm B}^2 \tilde{g}_{\rm L}\right)^{-1}, \tag{39}$$

where

$$\tilde{c} = U_b^\dagger \left(E \cdot I_{n\times n} - a_0\right) U_b \tag{40}$$

where a_0 as defined in Eq. (13) is the matrix for the sublattice in contact with the source reservoir. If the potential at the contact is constant along the circumferential direction so that

$$a_0 = -q_0\phi(0) \cdot I_{n\times n} \tag{41}$$

is satisfied, as is true for the case of co-axially-gated CNTFETs, all the matrices in Eqs. (38)–(39) are diagonal. One can then solve for $\tilde{g}_{\rm L}$, similarly to the SNWFET case above, and obtain the real-space surface Green's function via the inverse unitary transformation.

For top-gated CNT MOSFETs, $\tilde{c}$ in Eq. (40) has off-diagonal elements in general due to the fact that the potential at the contact may vary along the circumference of the CNT, and the calculation of the surface Green's function becomes quite involved [11,19]. But if the ideal ohmic contact is assumed, the contact in a simulation can be located arbitrarily far from the gate such that the condition in Eq. (41) is satisfied again. For efficient simulations, one may assume an artificially high dielectric in the immediate vicinity of the contacts, instead of having a very long simulation region, in order to achieve the condition in Eq. (41).

For the Schottky-contact CNTFETs where the source and drain leads are assumed to be metallic zigzag CNTs, $\tilde{g}_{\rm L}$ and $\tilde{g}'_{\rm L}$ in Eqs. (38) and (39) become identical to each other and one has, regardless of the gate types,

$$\tilde{g}_{\rm L} = \left(\tilde{c} - t_{\rm B}^2 \tilde{g}_{\rm L}\right)^{-1}, \tag{42}$$

with $\phi(0)$ in Eq. (41) being a constant built-in potential due to the Schottky contact. In this special case, the real-space surface Green's functions are identical to the mode-space counterparts, because the latter are independent of the modes.

Solution of 2D Schrödinger Equations

In the mode-space approach for SNWFETs, 2D Hamiltonian $H_{\perp,i}$ at each cross-section is diagonalized; i. e., the eigenvalue problem

$$H_{\perp,i}\psi_m(y,z;x_i) = E_m(x)\psi_m(y,z;x_i)\,, \tag{43}$$

where

$$H_{\perp,i} = -\frac{\hbar^2}{2}\frac{\partial}{\partial y}\left(\frac{1}{m_y^*}\frac{\partial}{\partial y}\right) - \frac{\hbar^2}{2}\frac{\partial}{\partial z}\left(\frac{1}{m_z^*}\frac{\partial}{\partial z}\right) + V(y,z;x_i)\,, \tag{44}$$

is solved subject to the boundary condition that the wave functions vanish at the boundaries of the 2D cross-sectional plane. If we write $\psi_m(y,z;x_i)$ in Eq. (43) as

$$\psi_m(y,z;x_i) = \sum_K A_K|K\rangle\,, \tag{45}$$

where $\{|K\rangle\}$ is a basis set and A_K's are expansion coefficients, and insert it in Eq. (43) and multiplying $\langle L|$ to

the both sides of the equation, we obtain an eigenvalue problem

$$\sum_K \mathcal{H}_{LK} A_K = E_m A_L \,, \tag{46}$$

where $\mathcal{H}_{LK} \equiv \langle L|H_{\perp,i}|K\rangle$. In the widely used k-space solution [1], we use

$$|K\rangle = \sqrt{\frac{2}{L_y}}\sqrt{\frac{2}{L_z}}\sin(k_p y)\sin(k_q z)\,, \tag{47}$$

where L_y and L_z are side lengths of the cross-section in the y and z directions, respectively, and

$$k_p = \frac{p\pi}{L_y}\,, \quad (p = 1, \ldots, N_y) \tag{48}$$

$$k_q = \frac{q\pi}{L_z}\,, \quad (q = 1, \ldots, N_z) \tag{49}$$

where N_y and N_z are number of meshes in the y and z directions, respectively. Note that the index K in Eq. (47) is mapped to indices p and q through $K = N_y(p-1)+q$. Specifically, for the Hamiltonian of Eq. (44) with continuous effective masses across the silicon/oxide interfaces, $\mathcal{H}_{LK}$ is given by

$$\mathcal{H}_{LK} = \frac{4}{L_y L_z}\int_0^{L_y}\int_0^{L_z} \mathrm{d}y\,\mathrm{d}z \sin(k_{p'} y)\sin(k_{q'} z) \times \ldots$$
$$\cdots \times \left(\frac{\hbar^2 k_p^2}{2m_y^*} + \frac{\hbar^2 k_q^2}{2m_z^*} + V(y,z;x_i)\right) \times \sin(k_p y)\sin(k_q z), \tag{50}$$

where $k_{p'}$ and $k_{q'}$ are defined similarly to k_p and k_q in Eqs. (48) and (49), respectively, and the index L in the equation is mapped to indices p' and q' through $L = N_y(p'-1)+q'$. Numerically, $\mathcal{H}_{LK}$ in Eq. (50) can be efficiently evaluated using the FFT routines [1].

In the product-space approach [28], on the other hand, we use the basis set

$$|K\rangle \equiv \chi_p(y)\zeta_q(z)\,, \tag{51}$$

where $\chi_p(y) \equiv \chi_p(y;x_i)$ is the pth mode eigenfunction in the y direction with its eigenvalue $\epsilon_p^{(y)} \equiv \epsilon_p^{(y)}(x_i)$, satisfying

$$\left\{-\frac{\hbar^2}{2m_y^*}\frac{d^2}{dy^2} + \bar{V}(y)\right\}\chi_p(y) = \epsilon_p^{(y)}\chi_p(y)\,, \tag{52}$$

where $\bar{V}(y) \equiv \bar{V}(y;x_i)$ is an average potential in the y direction, defined by

$$\bar{V}(y;x_i) = \frac{1}{T_{\mathrm{si}}}\int_{T_{\mathrm{ox}}}^{T_{\mathrm{ox}}+T_{\mathrm{si}}} \mathrm{d}z V(x_i,y,z)\,, \tag{53}$$

where T_{si} is the silicon channel thickness as shown in Fig. 1a. Similarly, we define $\zeta_q(z) \equiv \zeta_q(z;x_i)$ as the qth mode eigenfunction in the z direction with its eigenvalue $\epsilon_q^{(z)} \equiv \epsilon_q^{(z)}(x_i)$, satisfying

$$\left\{-\frac{\hbar^2}{2m_z^*}\frac{d^2}{dz^2} + \bar{V}(z)\right\}\zeta_q(z) = \epsilon_q^{(z)}\zeta_q(z)\,, \tag{54}$$

where $\bar{V}(z) \equiv \bar{V}(z;x_i)$ is an average potential in the z directions, defined by

$$\bar{V}(z;x_i) = \frac{1}{W_{\mathrm{si}}}\int_{W_{\mathrm{ox}}}^{W_{\mathrm{ox}}+W_{\mathrm{si}}} \mathrm{d}y V(x_i,y,z)\,. \tag{55}$$

The 1D eigenvalue problems of Eqs. (52) and (54) can be solved easily in the 1D k-space transformation [1]. If we use a uniform mesh of sizes N_y and N_z in the y and z directions, respectively, K ranges from 1 to $N_y N_z$ and we index it in the order of increasing value of $\epsilon_K \equiv \epsilon_p^{(y)} + \epsilon_q^{(z)}$. Using the relationships of Eqs. (51)–(55), $\mathcal{H}_{LK}$ in the product space becomes

$$\mathcal{H}_{LK} = \epsilon_L \delta_{LK} + \langle L|(V(y,z) - \bar{V}(y) - \bar{V}(z)|K\rangle. \tag{56}$$

An advantage of the product-space approach is that only the first few eigenvalues, which correspond to the subband modes that contribute to the transport, are sufficient to be included in the eigenvalue problem. In other words, if M is the number of subbands that participate in the transport, the size of the matrix $\mathcal{H}$ in Eq. (56) is reduced to M by M. For instance, M is about 10 for the nanowire transistor of cross-sectional area of $5\times5\,\mathrm{nm}^2$. One therefore needs to find the first 10 eigenvalues of a 10 by 10 matrix in the product-space solution. As the area of the cross-sectional plane increase, M increases in proportion to the area. The matrix $\mathcal{H}$ is effectively a banded matrix of band size N_B which ranges from 1 to 10 for the cross sectional areas of $5\times5\,\mathrm{nm}^2$ to $20\times20\,\mathrm{nm}^2$, which also contributes to the efficiency of the numerical calculation.

Solution of 3D Poisson's Equation

In a self-consistent solution of SNWFETs, the following Poisson's equation,

$$\nabla^2\phi^k = -\frac{q}{\epsilon}\left(N_D - n_{3\mathrm{D}}^k\, \mathrm{e}^{(\phi^k-\phi^{k-1})/k_{\mathrm{B}}T}\right)\,, \tag{57}$$

may be used instead of the one in Eq. (2) to achieve faster convergence. In Eq. (57), ϕ^k and $n_{3\mathrm{D}}^k$ are the kth step solutions for the potential and electron density, respectively. For CNTFETs, on the other hand, we solve the Laplace equation in the oxide and interior of CNT,

$$\nabla^2\phi^k = 0\,, \tag{58}$$

together with the interface condition on the CNT surface,

$$\epsilon_{ox}\left.\frac{\partial\phi^k}{\partial\bar{n}}\right|_{ox} - \epsilon_{cnt}\left.\frac{\partial\phi^k}{\partial\bar{n}}\right|_{cnt} = \ldots$$
$$\ldots q_0\left(N_D + p_{2D}^k e^{(\phi^{k-1}-\phi^k)/k_BT} - n_{2D}^k e^{(\phi^k-\phi^{k-1})/k_BT}\right), \quad (59)$$

where $\bar{n}$ denotes the direction normal to the CNT surface, and n_{2D}^k and p_{2D}^k are the kth step solutions for electron and hole densities on the CNT surface given by Eqs. (19) and (20), respectively.

The boundary conditions for the Poisson's equation for ohmic contact devices are such that for the gate contact region,

$$\phi(x,y,z) = V_g + \chi_e - \phi_{mg} \quad (60)$$

where χ_e and ϕ_{mg} are the electron affinity of channel material and the metal gate work function, but for other boundaries including source/drain contacts, the free boundary condition can be used. Especially, the free boundary condition at the source/drain contacts is necessary to achieve the thermal equilibrium conditions in the source/drain [9]. For the Schottky contacts,

$$\phi(x,y,z) = \begin{cases} \phi_{bi} & \text{at the source contact} \\ \phi_{bi} + V_d & \text{at the drain contact}, \end{cases} \quad (61)$$

where the built-in potential $\phi_{bi} \equiv E_c - E_F - \phi_{Bn}$ where E_c is the conduction band edge and ϕ_{Bn} is the Schottky barrier height.

Simulation Results

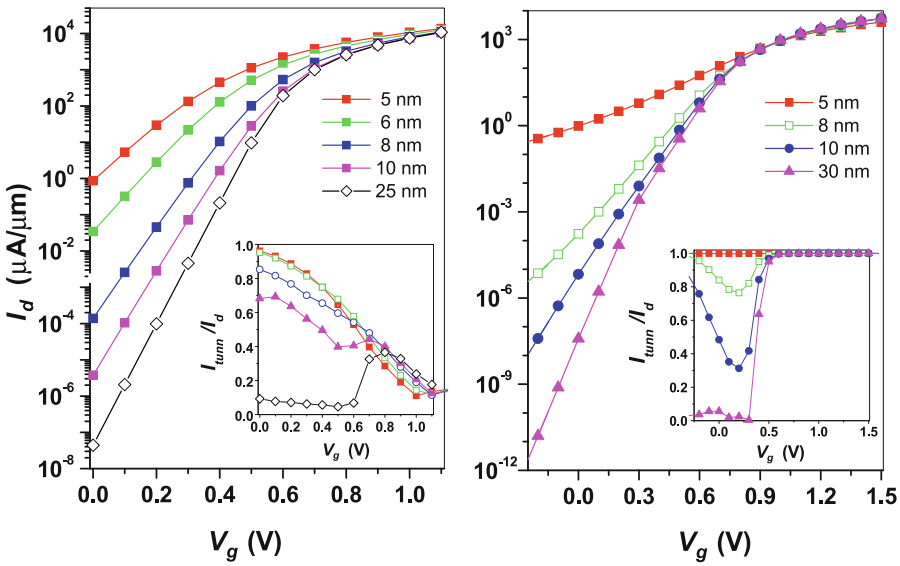
SNWFETs

Simulation results for some device characteristics of SNWFETs such as transfer characteristics and scaling behaviors are presented in the following. In the simulated devices, the channel silicon is lightly p-doped with $N_A = 10^{15}\,\text{cm}^{-3}$ and it is oriented such that its (100) direction is parallel to the transport direction. The silicon nanowire has square cross-section of $5\times5\,\text{nm}^2$, the gate oxide thickness is 1 nm, and the metal gate work function ϕ_{mg} is set at 4.61 eV. In the case of ohmic contact devices, the source/drain are heavily n-doped with the doping concentration of $10^{20}\,\text{cm}^{-3}$ while the Fermi energy of the metal source in the case of Schottky contact devices is assumed to lie 0.5 eV above the (virtual) metal CB edge.

The scaling behavior of GAA SNWFETs with doped S/D with respect to the shrink of the gate length is shown in Fig. 2a. It can be seen from the figure that the device characteristics degrade as the gate length becomes shorter. On-currents do not vary significantly with L, due to the ballistic nature of the transport, but off-current values increase sharply. We note that the SS increases sharply with L_g, especially when L_g is smaller than 10 nm. The behavior of SS with the gate length can be explained in terms of the increasing tunneling-current contribution to the total current, I_{tunn}/I_d, with respect to reduction of gate length, as can be seen in the inset of Fig. 2a. I_{tunn}/I_d becomes increasingly large as L_g becomes shorter, and if L_g = 5 nm, the tunneling contribution at $V_g = 0$ is more than 90% of the total current, giving rise to the high off-current. However, if L_g = 15 nm, for instance, the tunneling contribution becomes much smaller, below 40% in the off-state. In the on-state, the thermionic components dominate for the two cases. Similar trends were also observed for PI-gate and TRI-gate devices, with almost the same I_{tunn}/I_d irrespectively of the gate types (for the device with the same dimensions and at the same biases). Therefore, we can see that the short channel effects in the SNWFETs arise due to the direct source-to-drain tunneling current.

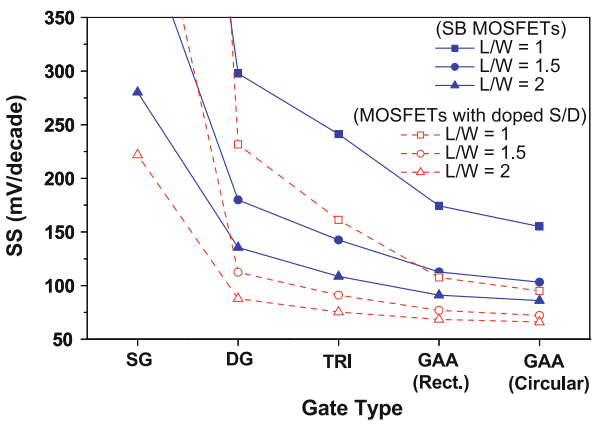
SB-MOSFETs show similar behavior below threshold. Figure 2b shows $I_d - V_g$ characteristics of SB devices with $\phi_{Bn} = 0.2$ eV when the gate length is gradually reduced from 30 nm down to 5 nm. Due to the increase of the direct source-to-drain tunneling current as in the case of the doped S/D devices, drastic deterioration of the off current behavior is noticed. The degree of degradation is severer for SB-MOSFETs, due to the presence of the SB barriers at the contacts as will be discussed shortly. On the other hand, the on-state currents are dominated by the tunneling current as seen in the inset of Fig. 2b contrary to the case of the doped S/D devices.

In a scaling of ballistic SNWFETs, the general trend is that the device performance improves as the channel length L becomes longer (for the same cross-sectional area) or as the channel cross-sectional area $W \equiv W_{si}$ is decreased (for the same channel length). If L and W are varied simultaneously while the aspect ratio L/W is fixed, the characteristics of SNWFETs are largely determined by their aspect ratio [28]. We show in Fig. 3 the device performance, as measured by its subthreshold swing, of SNWFETs with ohmic and Schottky contacts, respectively, for different aspect ratio L/W and different gate types. Compared to their planar counterparts, SNWFETs with multiple gates show enhanced performance, as expected, but it is noteworthy that the multi-gate effect is much greater for devices with smaller aspect ratio. Due to the tunneling current component in the off-state, SS of SB-MOSFETs are seen to be larger than their doped S/D coun-



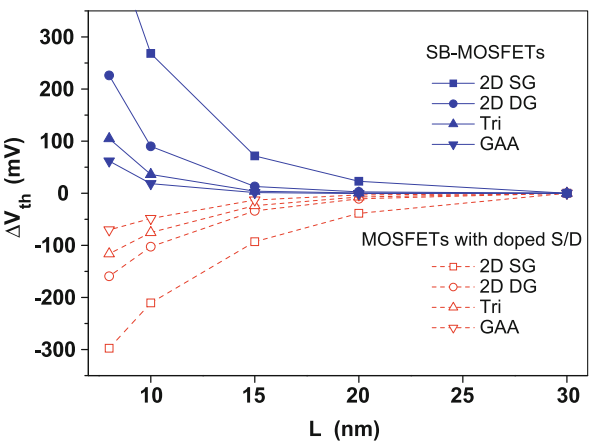
Quantum Simulations of Ballistic Nanowire Field Effect Transistors, Figure 2

$I_d - V_g$ characteristics of GAA SNWFETs with a doped S/D and b Schottky-barrier contacts. In the *insets* are shown the tunneling current contribution I_{tunn}/I_d



Quantum Simulations of Ballistic Nanowire Field Effect Transistors, Figure 3

Subthreshold slopes of SB-MOSFETs and MOSFETs with doped S/D, respectively, for different gate types and aspect ratios



Quantum Simulations of Ballistic Nanowire Field Effect Transistors, Figure 4

Threshold-voltage change ΔV_{th} as a function of channel length L for SB-MOSFETs and MOSFETs with doped S/D. For the SB-MOSFETs, $\phi_{Bn} = 0.2$ eV was assumed

terparts. Notice that SS approaches the theoretical limit of 60 mV/decade for $L/W = 2$ in GAA devices with doped S/D.

Figure 4 shows the change of the threshold voltage ΔV_{th} from 30 nm device for SB and doped S/D devices, respectively, as the channel length is gradually shortened. A sharp difference between the threshold behaviors of SB and doped S/D devices is noticed in the figure: as the channel length is gradually decreased from $L = 30$ nm, ΔV_{th} increases for the SB devices while it decreases for the doped S/D devices. The increasing ΔV_{th} behavior of the SB devices can be explained by the increasing CB bending stiffness: that is, as the channel length becomes shorter, CB bending by the gate voltage becomes less effective, which means that more gate voltage should be applied to enter the on-state and as the consequence, the threshold volt-

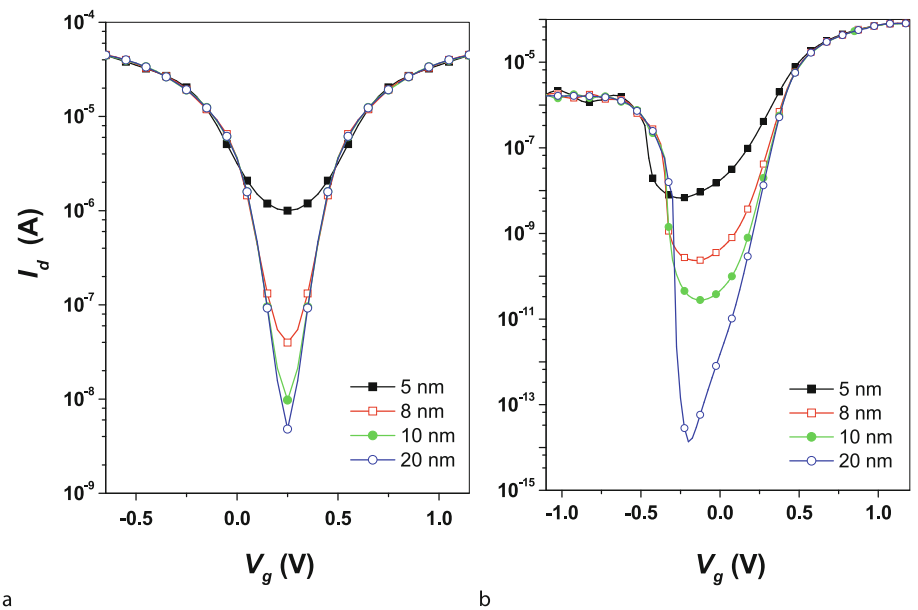
age is increased. For the doped S/D devices, CB also becomes stiffer to bending by the gate voltage as the channel length becomes shorter. This however does not lead to increase of the threshold voltage, because the top of the barrier is gradually lowered as the gate voltage is increased in the case of the doped S/D devices (and the off-state top-of-the-barrier is lower for shorter-channel devices, so less gate voltage is needed to be applied to enter the on-state), while it is clamped by the Schottky barriers in the case of the SB devices. An advantage of multiple-gated devices can be noticed in Fig. 4. Compared to the case of 2DSG devices, the threshold-voltage change to the channel length of the GAA devices is considerably suppressed.

Simulation Results of CNTFETs

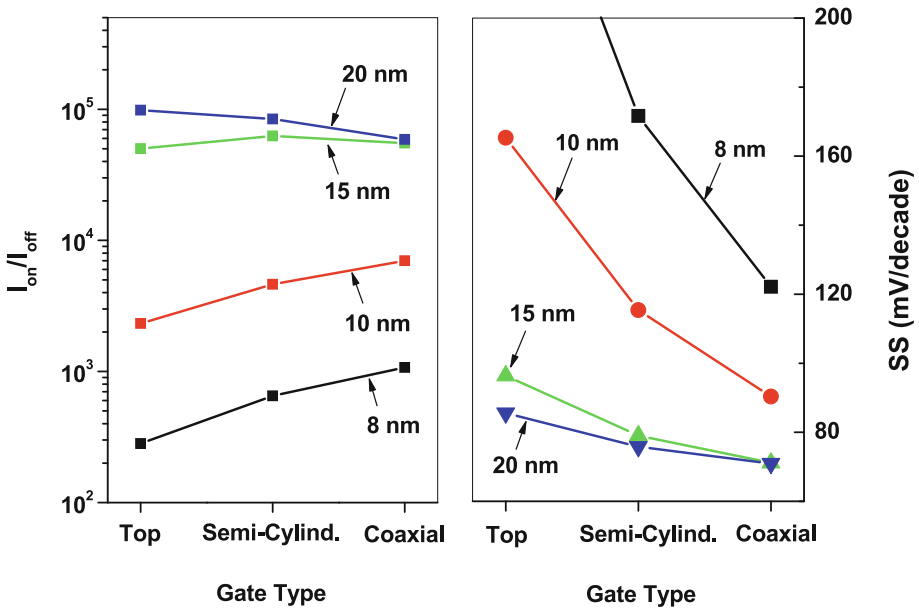
Some simulation results for CNTFETs such as the device-type and gate-type dependence are presented in the following. CNTFETs with (13,0) zigzag nanotube with diameter of ∼1 nm and band gap energy of ∼0.83 eV are considered. Thickness and dielectric constant of the gate oxide are 1 nm and 25, respectively. The channel is assumed to be intrinsic, and the midgap Schottky barrier is assumed for SB CNTFETs while the source/drain regions are n-doped with 10^7 cm^{-1} for CNT MOSFETs.

The transfer characteristics of coaxial SB-CNTFETs as the channel length is scaled down from 20 nm to 5 nm are shown in Fig. 5a. For $L \gtrsim 10$nm, $I_d - V_g$ characteristics are almost same regardless of the channel length because the CNT channel is ballistic and the current is dominated, and limited at the same time, by the tunneling current at the source SB contact. As the device is aggressively scaled down below 10 nm in the channel length, subthreshold currents increase sharply, due to the increased source-to-drain tunneling current as in the case of SNWFETs, and for $L < 5$ nm, the on-off current ratio becomes less than 10^2. Note that ambipolar conduction occurs in the SB-CNTFETs owing to the mirror-symmetric conduction and valence band structures and current is at its minimum at $V_g = V_d/2$ for midgap Schottky barrier height.

Similar scaling behavior can be observed in coaxial CNT MOSFETs, as shown in Fig. 5b. The seemingly ambipolar behavior is however due to band-to-band (BTB) tunneling at negatively high gate voltages and the current minima take place when BTB tunneling begins to be initiated. Compared to SB-CNTFETs, on-current level and on-off current ratio are higher for the CNT MOSFETs, because the thermionic currents prevail in the on-state and the minority carrier injection is suppressed unlike the SB devices. Notice also that SS of the electron branch (the right branch in the figure) approaches the theoretical limit of 60 mV/decade for conventional MOSFETs while that of the BTB branch is less than that because of the BTB tunneling. As the channel length is scaled down, however, the



Quantum Simulations of Ballistic Nanowire Field Effect Transistors, Figure 5
$I_d - V_g$ characteristics of coaxial CNTFETs for various channel lengths: a SB-MOSFETs and b CNT MOSFETs



Quantum Simulations of Ballistic Nanowire Field Effect Transistors, Figure 6
I_{on}/I_{off} ratios and subthreshold slopes for top, semi-cylindrical, and coaxial gated CNT MOSFETs with different channel lengths

device performance becomes degraded as can be seen in Fig. 5b but compared to SB devices, the degree of degradation is less severe.

The gate-type dependence of CNTFETs with Schottky barrier contacts is shown in Fig. 6, where SS and I_{on}/I_{off} of the top, semi-cylindrical, and co-axial gated devices, respectively, are shown for different channel lengths. As expected, device performance is improved as the gate number is increased, especially for devices with sub-10 nm channel length.

Future Directions

A future practical device simulator should possess multi-dimensional, multi-scale features. For simulations of silicon MOSFET-type devices, especially, it should support both planar and wire structures in terms of dimensionality and should encompass a few nanometer to a few micrometer in terms of device size. Efforts should be made to develop a simulator that is most efficient in each regime of interest and seamlessly connects different regimes at the same time. Ballistic nanowire transistor simulation as considered in this article covers only a small portion of the whole area: The ballistic transport regime should be expanded and smoothly transit to diffusive transport regime in one direction and simplified approaches based on the effective-mass theory or single-band tight-binding method as adopted here should be connected with fully atomistic or first-principle calculations in the other direction. For the former, mode-space approach the with one-dimensional subbands where transport can be treated via NEGF, Boltzmann equation or Wigner function approaches seems to be quite attractive as recent works reveal. For the latter, hybrid approach that captures essential physics from rigorous calculations and enables efficient computation at the same time should be further elaborated in such a way to provide an integrated platform for a practical device simulator.

Acknowledgment

This research was supported in part by the university collaboration program of KRISS and in part by grant No. R 01-2005-000-10303-0 from the Basic Research Program of the Korea Science and Engineering Foundation.

Bibliography

1. Abramo A, Cardin A, Selmi L, Sangiorgi E (2000) Two-dimensional quantum mechanical simulation of charge distribution in silicon MOSFETs. IEEE Trans Electron Dev 27:1858–1863
2. Ahn C, Shin M (2007) Quantum simulation of coaxially gated CNTFETs by using an effective mass approach. J Korean Phys Soc 50:1887–1893
3. Alam K, Lake RK (2005) Leakage and performance of zero-Schottky-barrier carbon nanotube transistors. J Appl Phys 98:064037-1–064037-8

4. Appenzeller J, Lin YM, Knoch J, Chen Z, Avouris P (2005) Comparing carbon nanotube transistors – the ideal choice: a novel tunneling device design. IEEE Trans Electron Dev 52: 2568–2576
5. Avouris P, Appenzeller J, Martel R, Wind SJ (2003) Carbon nanotube electronics. Proc of the IEEE 91:1772–1784
6. Bescond M, Nehari K, Autran JL, Cavassilas N, Munteanu D, Lannoo M (2004) 3D quantum modeling and simulation of multiple-gate nanowire MOSFETs. IEDM Tech Dig pp 617–620
7. Cui Y, Zhong Z, Wang D, Wang W, Lieber M (2003) High performance silicon nanowire field effect transistors. Nano Lett 3:149–152
8. Damle PS, Ghosh AW, Datta S (2003) Nanoscale device modeling. In: Reed M et al (eds) Molecular nanoelectronics, American Scientific Publishers, Stevenson Ranch, pp 115–135
9. Datta S (2000) Nanoscale device modeling: the Greens function method. Superlatt Microstruct 28:253–278
10. Doris B, Ieong M, Kanarsky T, Zhang Y, Roy RA, Dokumaci O, Ren Z, Jamin FF, Shi L, Natzle W, Huang HJ, Mezzapelle J, Mocuta A, Womack S, Gribelyuk M, Jones EC, Miller RJ, Wong HSP, Haensch W (2002) Extreme scaling with ultra-thin Si channel MOSFETs. IEDM Tech Dig pp 267–270
11. Fiori G, Iannaccone G, Klimeck G (2006) A three-dimensional simulation study of the performance of carbon nanotube field-effect transistors with doped reservoirs and realistic geometry. IEEE Trans Electron Dev 53:1782–1788
12. Gnani E, Reggiani S, Gnudi A, Parruccini P, Colle R, Rudan M, Baccarani G (2007) Band-structure effects in ultrascaled silicon nanowires. IEEE Trans Electron Devices 54:2243–2254
13. Guo J, Datta S, Lundstrom MS, Anantram MP (2004) Towards multi-scale modeling of carbon nanotube transistors. Int J Multiscale Comput Eng 2:257–276
14. Guo J, Lundstrom M (2005) Role of phonon scattering in carbon nanotube field-effect transistors. Appl Phys Lett 86:193103-1–193103-3
15. (2007) International technology roadmap for semiconductor edition. http://public.itrs.net
16. Javey A, Tu R, Farmer DB, Guo J, Gordon RG, Dai H (2005) High performance n-type carbon nanotube field-effect transistors with chemically doped contacts. Nano Lett 5:345–348
17. Jin S, Park YJ, Min HS (2006) A three-dimensional simulation of quantum transport in silicon nanowire transistor in the presence of electron-phonon interactions. J App Phys 99:123719-1–123719-10
18. John DL, Castro LC, Pereira PJS, Pulfrey PL (2004) A Schrödinger-Poisson solver for modeling carbon nanotube FETs. Proc NSTI Nanotech 3:65–68
19. John DL (2006) Simulation Studies of Carbon Nanotube Field Effect Transistors. University of British Columbia, Vancouver
20. Ko YJ, Shin M, Lee S, Park KW (2000) Effects of atomistic defects on coherent electron transmission in Si nanowires: Full band calculations. J Appl Phys 89:374–379
21. Koswatta SO, Neophytou N, Kienle D, Fiori G, Lundstrom MS (2006) Dependence of DC characteristics of CNT MOSFETs on bandstructure models. IEEE Trans Nanotechnol 5:368–372
22. Luisier M, Schenk A, Fichtner W (2006) Quantum transport in two- and three-dimensional nano-scale transistors: Coupled mode effects in the nonequilibrium Greens function formalism. J App Phys 100:043713-1–043713-12
23. Luisier M, Schenk A, Fichtner W (2006) Atomistic simulation of nanowires in the sp3d5s* tight-binding formalism: From boundary conditions to strain calculations. Phys Rev B 74:205323-1–205323-12
24. Nehari K, Cavassilas N, Autran JL, Bescond M, Munteanu D, Lannoo M (2005) Influence of band-structure on electron ballistic transport in silicon nanowire MOSFET's: an atomistic study. Proc 35th European Solid-State Device Research Conf, pp 229–232
25. Park JT, Colinge JP (2002) Multiple-gate SOI MOSFETs: device design guidelines. IEEE Trans Electron Devices 12:2222–2229
26. Radosavljevic M, Heinze S, Tersoff J, Avouris P (2003) Drain voltage scaling in carbon nantoube transistors. Appl Phys Lett 83:2435–2437
27. Rahman A, Lundstrom MS, Ghosh AW (2005) Generalized effective-mass approach for n-type metal-oxide-semiconductor field-effect transistors on arbitrarily oriented wafers. J Appl Phys 97:053702-1–053702-12
28. Shin M (2007) Efficient simulation of silicon nanowire field effect transistors and their scaling behavior. J Appl Phys 101:024510-1–024510-6
29. Shin M (2007) Quantum simulation of device characteristics of silicon nanowire FETs. IEEE Trans Nanotechnol 6:230–237
30. Wang J, Polizzi E, Lundstrom M (2004) A three-dimensional quantum simulation of silicon nanowire transistors with the effective-mass approximation. J Appl Phys 96:2192–2203
31. Wang J, Rahman A, Ghosh A, Klimeck G, Lundstrom M (2005) On the validity of the parabolic effective-mass approximation for the I-V calculation of silicon nanowire transistors. IEEE Trans Electron Devices 52:1589–1595
32. Wang J, Polizzi E, Ghosh A, Datta S, Lundstrom M (2005) Theoretical investigation of surface roughness scattering in silicon nanowire transistors. Appl Phys Lett 87:043101-1–043101-3
33. Wong HSP (2005) Beyond the conventional transistor. Solid State Electron 49:755–762