Invited paper

Yuyuan Lu*, Geng Deng and Zhigang Shuai

Future directions of chemical theory and computation

https://doi.org/10.1515/pac-2020-1006

Abstract: Theoretical and computational chemistry aims to develop chemical theory and to apply numerical computation and simulation to reveal the mechanism behind complex chemical phenomena via quantum theory and statistical mechanics. Computation is the third pillar of scientific research together with theory and experiment. Computation enables scientists to test, discover, and build models/theories of the corresponding chemical phenomena. Theoretical and computational chemistry has been advanced to a new era due to the development of high-performance computational facilities and artificial intelligence approaches. The tendency to merge electronic structural theory with quantum chemical dynamics and statistical mechanics is of increasing interest because of the rapid development of on-the-fly dynamic simulations for complex systems plus low-scaling electronic structural theory. Another challenging issue lies in the transition from order to disorder, from thermodynamics to dynamics, and from equilibrium to non-equilibrium. Despite an increasingly rapid emergence of advances in computational power, detailed criteria for databases, effective data sharing strategies, and deep learning workflows have yet to be developed. Here, we outline some challenges and limitations of the current artificial intelligence approaches with an outlook on the potential future directions for chemistry in the big data era.

Keywords: Artificial intelligence; deep learning; density functional theory; emerging technologies; new directions in chemistry research; theoretical and computational chemistry.

Overview of theoretical and computational chemistry

Theoretical and computational chemistry is a branch of chemistry that uses mathematical and physical methods such as thermodynamics, statistical mechanics, and quantum mechanics to explain chemical phenomena and processes via numerical computation and computer simulation [1–5]. It combines theoretical approaches in physics and chemistry and efficient computer programs to calculate the structures and properties of chemical systems. Of course, chemists have been doing computations for centuries, but the term "computational chemistry" is a natural product of the digital age. A significant increase in computing power blurs the boundary between theoretical chemistry and computational chemistry. As Dirac once said [6]: "*The fundamental laws necessary for the mathematical treatment of a large part of physics and the whole of chemistry are thus completely known, and the difficulty lies only in the fact that application of these laws leads to equations that are too complex to be solved*."

Article note: A collection of invited papers on Emerging Technologies and New Directions in Chemistry Research.

***Corresponding author: Yuyuan Lu**, Changchun Institute of Applied Chemistry, Chinese Academy of Sciences, 130022 Changchun, People's Republic of China, e-mail: yylu@ciac.ac.cn
Geng Deng, Institute of Education, Tsinghua University, 100084 Beijing, People's Republic of China
Zhigang Shuai, Department of Chemistry, MOE Key Laboratoy of Organic OptoElectronics and Molecular Engineering, Tsinghua University, 100084 Beijing, People's Republic of China

Great progress has been made in theoretical and computational chemistry because of the improvements and upgrades in computer facilities together with the development and innovation of efficient algorithms. An increasing amount of attention is paid to this research field. Theoretical and computational chemistry deepens our understanding of the chemical phenomenon and enables materials and manufacturing processes to be designed more efficiently.

The basic theory of theoretical and computational chemistry comes from quantum mechanics, classical Newton's mechanics and Maxwell's electrodynamics, and statistical mechanics. All chemical phenomena and processes involving electronic structure and dynamics (such as the breaking and formation of chemical bonds) are solved by quantum mechanics. For their contributions to quantum chemistry computation, Walter Kohn and John Pople won the Nobel Prize in chemistry in 1998. The field of molecular mechanics was developed based on the force field model and the Newtonian mechanics coupled with statistical mechanics. In the same system, the computational complexity of the quantum mechanical model is much higher than that of the molecular mechanics model due to the superposition description of the quantum state. This is because the molecular mechanical model takes the atom as a classical object while the quantum mechanical model also needs to consider the structure and motion of electrons. Martin Karplus, Michael Levitt, and Arieh Warshel won the 2013 Nobel Prize in chemistry for combining quantum mechanics and classical Newtonian mechanics to construct a multi-scale calculation method for complex systems.

More recently, artificial intelligence technologies such as machine learning and big data have been an area of intense study. These emerging artificial intelligence technologies have been introduced into theoretical and computational chemistry with great success. Currently, artificial intelligence in chemistry includes the automatic generation of literature abstracts in chemistry using natural language processing technology [7], intelligent retrieval methods in chemical data [8], automation and robotics in chemical laboratories [9, 10], and chemical applications of neural network methods [11]. Of these, the most active and successful applications identify chemical structures of unknown compounds from spectral data via deep learning [12]. Here, the spectral data include infrared, mass spectrometry, nuclear magnetic resonance, two-dimensional, and high-dimensional nuclear magnetic resonance data [13]. Artificial intelligence also plays an important role in chemical synthesis [14, 15]. There are countless chemical reactions in nature, and strategies to plan a novel and feasible synthetic route are a big problem that puzzles chemists. In the past, researchers were challenged to design a chemical synthesis route because the chemical reaction was infinitely varied under different conditions. Based on big data and artificial intelligence, computers can help researchers design chemical synthesis routes that improve the efficiency of scientific research in developing new drugs, and other chemical compounds [16].

Machine learning and deep learning have continuously produced new applications in the fields of synthetic chemistry, medicinal chemistry, etc. This has led to revolutionary changes in chemistry. For example, Segler et al. [17] collected almost all the chemical reactions published in the past few decades nearly 12.5 million reactions. The team then successfully applied a deep neural network and Monte Carlo tree algorithm to design a new chemical synthesis route. According to the method reported here, it only takes 5 s to design a molecular synthesis route via artificial intelligence. Granda et al. [18] reported an organic synthesis robot that can predict and analyze chemical reactions faster than traditional methods via machine learning algorithms. After training the model with 10 percent of the chemical reactions, the intelligent robot can then predict the chemical reactivity with an accuracy of 86 %.

New chemical reactions can also be discovered using real-time data. Popova et al. [19] developed a novel computational strategy termed ReLeaSE (Reinforcement Learning for Structural Evolution). ReLeaSE can generate targeted chemical libraries of novel compounds with desired properties. These are of great importance in drug discovery where the potential molecules should be optimized according to their properties such as selectivity, solubility, and potency. O'Connor et al. [20] developed a database containing various catalysts with different properties where machine learning and quantum chemistry are combined to search for hidden patterns in the database with the purpose of optimal molecular design of cheaper and more efficient catalysts. There is no doubt that artificial intelligence is a key step in the digitization of chemistry. This can lead to real-

time searching and the efficient design of chemical molecules, a significant reduction in cost and time, and safety improvements.

In this short perspective, we describe the future opportunities and challenges for theoretical and computational chemistry. The discussion includes electronic structure theory, chemical dynamics, molecular design and synthesis planning, complex system, and deep-learning in chemistry.

Electronic structure theory

Density functional theory

Density functional theory (DFT) is a quantum mechanical method that studies the electronic structure of multi-electron systems. It has been widely used in chemistry, physics, and materials science [21–23]. In the framework of the Kohn-Sham formulation of DFT, the most difficult many-body problem due to electron interactions is reduced to a problem of independent electrons moving in an effective potential field [24]. This effective potential field includes the effects of the external potential and Coulombic interactions between electrons including exchange and correlation interactions.

Since 1970, DFT has been widely used in solid-state physics calculations. Density functional theory using local density approximation (LDA) with Slater exchange and VWN correlation from a uniform electron gas usually leads to very satisfactory results with very low computational cost especially when compared with other methods dealing with the many-body problems in quantum mechanics [25]. However, LDA offered less satisfactory results for a molecule than for solids due to the inhomogeneity in electron density. The Parr-Yang book is a milestone in DFT with faster development of better exchange correlation functions suitable for chemical problems [23]. The development of an exchange-correlation function began with LDA and then moved to the general gradient approximation (GGA) such as PBE and PW91. The field progressed with meta-GGA in consideration of orbital kinetic energy and exact exchange such as TPSS and M06-L. Hybrid GGA is now the most widely employed and is represented by B3LYP, hybrid meta-GGA such as TPSSh and M06-2x, and double hybrid-GGA containing unoccupied orbital contributions such as XYG3 and B2PLYP. Savin first proposed to separate long and short range Coulombic interactions [26] that were long ignored by the community but later found to be important in describing the charge transfer excitation and response properties [27]. Range separation for the functional with an optimal ω (the range separation parameter) is now seen in daily practice for modeling the molecular energy materials (often with charge transfer excitations).

One of the main challenges of DFT is to maintain its simplicity while improving its accuracy or expanding its functionals. This is also a problem faced by the entire field of computer science. It is important to keep the theory and calculation concise and to ensure the computational costs are within an acceptable range because this leads to proper density functional approximations. The great success and popularity of DFT lie in that some simple approximations work very well for the calculations of the structural and thermodynamic properties of molecules and solids [28–30]. However, there are still many qualitative failures of DFT which result from the inappropriate approximations of the exchange-correlation functionals. The core issue that how to construct the universally applicable functionals has been elusive, which is full of challenges.

To describe the chemical phenomena and processes more comprehensively, it is necessary to jump out of the equilibrium geometries of molecules and consider the weak interactions between molecules and the transition states in chemical reactions. It remains a great challenge to describe van der Waals forces, covalent bonds, and reaction barriers correctly and efficiently.

It is known that all semilocal density functionals fail to describe the long-range dispersion interactions [31]. Various approaches were proposed to include dispersion corrections into DFT, such as DFT-D-type methods [32–36], vdW-DFs [37, 38], 1ePOT [39–41], and so on. Among them, the DFT-D method proposed by Grimme (DFT-D1 [32], DFT-D2 [33], DFT-D3 [35], and DFT-D4 [36]) is the most widely-used one. The DFT-D method solves the dispersion problem of DFT in a rather general way and the correction can be coupled in a simple form to standard density functionals. The DFT-D3 has been refined with higher accuracy, applies to a

broader range of applicability, and adopts less empiricism [35]. It is just now replacing DFT-D2 as the worldwide *de facto* standard in dispersion corrected DFT calculations. The DFT-D method opens completely new possibilities for the application of DFT in the areas of condensed matter, materials science, and biochemistry where dispersion effects are often of utmost importance [32–36].

The empirical elements are embedded in almost all of the current dispersion corrections in different ways. Hence, it is essential to perform a systemical benchmark on experimental or reliable theoretical data [42]. For the intramolecular and intramolecular interactions, the gold standard is the WF-based singles and doubles coupled-cluster method with perturbative triples (CCSD(T)). The method can provide accurate results, whose error is less than 1 kcal/mol for typical chemical reactions [43, 44]. The domain-based local pair natural orbital coupled-cluster method, abbreviated as DLPNO-CCSD(T), is an approximation to CCSD(T) and it has the following advantages: (1) Accurate. DLPNO-CCSD(T) recovers more than 99.9 % of the CCSD(T) correlation energy. Reaction energies are calculated with a mean deviation of 0.3 kcal/mol for 12 test reactions of medium-sized molecules [45]. (2) The computational cost of DLPNO-CCSD(T) is comparable to DFT but scales linearly with the system size [45, 46]. (3) DLPNO-CCSD(T) operates like a black box without adjusting any complicated parameters.

The majority of the calculation failures of DFT result from its underestimation of the reaction barrier, dissociation energy of ionic molecules, energy gaps, excited states, intermolecular interactions, and charge transfer excitation. Meanwhile, DFT overestimates the binding energies of charge transfer complexes and the response to an electric field in molecules and materials [47]. Both the underestimation and overestimation result from the same origin, *i.e.*, the delocalization error of the approximated functionals. Due to the discrete nature of electrons, the exact energy of the atom as a function of the charge should be a straight-line interpolation between the integers [48]. Yang et al. found that there were delocalization errors when the fractional charge deviates from a linear relationship. They further found that the deviation of fractional spins is responsible for static correlation-related errors, *e.g.*, a strong correlation effect in oxide and magnetic systems, degenerate states, and bond-breaking [49].

Langreth-Perdew's adiabatic connection-fluctuation-dissipation (ACFD) theorem is derived from density fluctuations and has become an important foundation for systematic improvements in functionals from many-body theory, *e.g.*, connecting with the random-phase approximation (RPA-DFT). Parallel to density fluctuation-based ACFD, Yang et al. proposed another adiabatic connection based on a pairing matrix fluctuation that led to the particle-particle RPA formulation of DFT and ppRPA-DFT. The ppRPA-DFT demonstrates the nearly linear fractional charge behavior and thus the delocalization errors are minimized: the ppRPA-DFT correctly describes the van der Waals interaction and has almost no static correlation error for single-bond systems [50]. Both the occupied and unoccupied Kohn-Sham orbitals should be considered. Thus, learning from wavefunction theory in dealing with the correlation problem is critical to developing radically different functionals to advance DFT.

Post-Hartree-Fock and beyond

The wavefunction based methods are increasingly popular because of their systematic accuracy improvements and the development of efficient low-scaling computational techniques. Wavefunctions also have value in fields requiring high accuracy such as chemical reaction dynamics. Wavefunction theory starts with the self-consistent field Hartree-Fock (HF) mean-field theory, which is the origin of the modern *ab initio* quantum chemistry. Later, many post-Hartree-Fock methods have been proposed to treat the correlation effects including configuration interaction (CI), perturbation (MPn), coupled-cluster (CC), multiconfiguration self-consistent field (MCSCF), complete active space self-consistent field (CASSCF), and CAS perturbation. Multi-reference CC remains very time-consuming.

The explicit correlation methods (R12 and F12) can incorporate the interelectron distance r_{12} into the wavefunction ansatz. This idea was originally developed by Hylleraas already in 1929 for the helium atom. It can correctly describe the cusp behavior of the wavefunction [51]. R12 was first introduced in MP2 and then

made popular in coupled cluster theory. Later, a Slater-type correlation F12 form was postulated, which could be fitted through a linear combination of a few Gaussian-type functions. It is still a challenge to formulate an analytical gradient with an explicit correlation form.

The introduction of quantum Monte Carlo (QMC) into traditional wavefunction theory by the Prof. Ali Alavi of Cambridge into perturbation theory was a major step forward. There was later full configuration interaction QMC (FCIQMC) [52]. The imaginary time Schroedinger equation becomes a diffusion equation in QMC. The required antisymmetric property is violated because the lowest energy solution is generally nodeless and symmetric. This has long been a significant problem for QMC. Alavi suggested performing a long-time integration in the space of a Slater determinant; there was a propagation step via population dynamics. The "walker" in the simulation represents an instantaneous wavefunction that carries a sign and a pair of walkers coinciding with the same determinant but with different signs; these are removed from the simulation. Such a new QMC algorithm in the Slater determinant space can efficiently converge to a full CI energy. This approach was quickly applied to coupled cluster even with periodic conditions for solids with great success [53]. The results offered unprecedented accuracy.

The density matrix renormalization group (DMRG) method was initially proposed in condensed matter physics for strongly correlated models [54] and has been quickly applied to quantum chemistry [55, 56]. DMRG can now handle a much larger active space than conventional CASSCF [57]. DMRG iteratively optimizes the eigenvectors of the reduced density matrix, which are used as approximate natural orbitals for constructing a full CI-like but much lower dimension Hilbert space. Recently, DMRG ansatz is understood as a matrix product state (MPS) formalism from applied mathematics and quantum information theory. The optimization procedures are actually consecutive singular value decompositions and maintain a few important singular values in each step. The International Academy of Quantum Molecular Science has twice awarded the annual (single) medal for this field. In 2010, Garnet KL Chan was awarded the IAQMS medal with citation "for his outstanding contribution to the density matrix renormalization group theory of molecular system", and Takeshi Yanai in 2013 "for his development of novel approaches to incorporate dynamical correlation into DMRG using canonical transformation theory".

The most recent remarkable advances are made for the time-dependent MPS formalism for finite-temperature problems, which is quickly emerging as a promising method to deal with quantum dynamics for complex systems [58]. In fact, the MPS ansatz along with the more recent tensor tree network state can grasp the essential quantum entanglements, thus offering proper spaces for time evolution. The applications are envisioned for non-adiabatic dynamics and electronic processes in complex systems such as carrier transport, various optical spectroscopies, singlet fission, and organic light-emitting diodes.

Molecular design and synthesis planning

Molecular design

Molecular design is a hot research topic especially for materials design and drug design. The design and optimization of these two kinds of molecules can bring great benefits and have great potential value. Therefore, related studies have dramatically stimulated the application of deep learning in related fields. However, the process of molecular design is still very challenging because it is expensive and time-consuming considering [49, 59, 60].

Most of today's technologies such as batteries, aerospace, and renewable energy strongly rely on the synthesis and application of advanced materials. Artificial intelligence has only recently begun to affect the field of materials design, and the potential impact of such data-driven materials science is tremendous [61–63]. Computer-aided materials design can significantly reduce the typical cycle for the development and commercialization of new materials [64]. Specifically, the integration of artificial intelligence algorithms could help address the inverse chemical paradigm, automatically discovering the property-biased molecular structures, and thus expediting the design of novel useful compounds [65]. However, the current material

design faces the following challenges: (1) obscure terminology in material informatics makes it difficult for typical material scientists to see how data-driven methods are applied to their own work; (2) limited access to structured data and a lack of data standards; (3) absence of a mature data-sharing mechanism [66].

Synthesis planning

Synthesis planning can be divided into three parts: retrosynthesis, reaction prediction, and reaction optimization. In retrosynthesis, the product is known and the initial reagents from which it can be made are determined. In reaction prediction, the reagents are known, and the main products are predicted. In reaction optimization, both reagents and products are known, and the yield or efficiency of this chemical reaction can be maximized by changing the reaction conditions [67, 68].

The application of machine learning and artificial intelligence technologies in synthesis planning has unprecedented opportunities. Since the machine learning models could discover the hidden relationship between the dataset and the experimental results, efficiently and relatively low-priced, it would promote the development of retrosynthesis and replace the expensive quantum calculations. However, there are still some urgent problems to be solved: (1) Detailed reaction data are limited. (2) The literature prefers to report only successful reactions rather than failed ones, and access to negative reaction data is necessary for more efficient synthesis planning [67].

Theory for complex systems

Challenges in theory and simulations of polymer science

The challenges faced by polymer theory and simulation include the following aspects: The need for design rules for a specific functional polymeric material including hierarchical composites and biomaterials; understanding the transport process of electrons, ions, photons, and energy in polymer systems (solar cells, living cells, batteries, sensors, and membranes); and exploring more efficient and stable processing strategy for polymer materials with complex structures [69, 70].

Polymer processing often involves a strong flow field, temperature gradients, and other external conditions that make polymer systems far away from the equilibrium state. It is a great challenge to predict the properties of polymer materials at the micro, meso, and macro time or length scales. It is impractical to develop a theory considering all variables, and thus it is very important to make proper approximations and reasonable coarse graining.

Electrically charged polymeric systems

Charged polymers involve a series of synthetic materials and consist of the basic unit of most functional biomaterials. The polyelectrolytes in polymer solutions are strongly correlated in terms of both topologies and electrostatic interactions. Moreover, there are lots of factors influencing the polyelectrolyte behaviors, such as the charge distribution on the polyelectrolytes, the electrostatic correlations, size effects, and the polarizability effects around charged polymers [71]. Despite the theoretical efforts devoted to the scaling theories, the analytical theories basing on the field theory, and the liquid-state theories, there are still two major challenges that need to be done. The first challenge associated with electrically charged polymeric systems is that the dielectric properties are spatially heterogeneous, which means that the mean-field theory fails to describe such a system. The second challenge is the theoretical description of molecular conformations or structure factors of charged polymers in saline/salt-free solutions: the correlation length of two charged units is large enough to

even match the size of the molecular conformation, thus coupling with the molecular conformations on larger length scales [69, 71].

Crystallization and glass transition

Although substantial progress has been made in these classic problems in polymer physics, there are still significant challenges in understanding the following situations. (1) The interplay between conformational entropy (favoring disorder) and the short-range attraction between segments (favoring order) results in extremely complex behaviors of polymer crystallization and glass transition. Such processes are far away from the thermodynamic equilibrium and usually take place in supercooled metastable states. The current available theoretical models succeed in describing the polymer crystallization and glass transition processes, but fail to predict these behaviors. Therefore, a universal theoretical framework is urgently needed to describe the kinetics of crystallization and glass transition [72, 73]. (2) It is still far from understood about the origin of the temperature and pressure dependences of the equilibrium dynamics and the corresponding quantitative description of the behaviors remains to be explored. Meanwhile, the origin of the Kauzmann paradox and the relationship between the evolution of the glass structure and physical aging are still unanswered [74]. (3) How the chain connectivity and uncrossability affect the mechanical properties of the entangled polymer system? (4) How to theoretically explain the special nonequilibrium state generated by the thermal history also needs to be further understood, such as the brittle-ductile transition and glass transition [73, 74].

Nonlinear rheological behaviors of entangled polymer fluids

Nonlinear rheology is to study the mechanical behavior of materials under a large strain or rapid deformation conditions. The stress response of materials depends on the deformation amplitude, rate, and flow field type. Owing to the failure of the Boltzmann superposition principle, the nonlinear behavior cannot be predicted by linear viscoelastic behavior. The development of rheological measurement technologies and structural characterization methods has led to the development of nonlinear rheological experimental research. The field moved from focusing on the study of steady-state shear properties to focusing on transient/dynamic shear, transient tension, and even more complex flow field rheological response. The description of nonlinear rheological behavior has also developed from empirical continuum mechanics model to molecular model. However, there is still no universal constitutive model to describe all nonlinear behaviors. Therefore, a key goal of nonlinear rheology is to understand the structural evolution mechanism and dynamic behavior under the condition of complex flow, and to establish a relationship with rheological properties [75–79].

Deep learning in chemistry

As an extremely broad subfield of artificial intelligence, machine learning aims to solve the problems of computer learning from data [13, 14, 80, 81]. Deep learning is a kind of machine learning that uses hierarchical recombination technology to extract relevant information from data and then to learn the specific patterns hidden in the data [2, 3, 68, 82, 83]. The standard workflow of machine learning from data to knowledge is shown in Fig. 1. In the past decades, deep learning has been increasingly applied to a variety of challenging problems in chemistry including drug and materials design, and synthesis planning [15, 80, 83].

Fig. 1: A standard workflow of machine learning.

Accelerated computational models

Utilizing physics-based calculations, the computational models of chemistry are adopted to investigate various properties and behaviors of the modeling system [68]. There are two different ways to apply deep learning in these computational models. The first one is to integrate deep learning methods with physics-based approaches directly, which involves network training to predict the key property. The second method is to predict properties by establishing the relationship between the molecular structure and particular properties. The corresponding examples of the integrated physics-based approach include the prediction of force fields [84], potential energy surfaces [85–87], and corrections of ab initio calculations [88]. Such method is more flexible due to the physical basis, however, it has a slower speed when compared with the second method.

Quantitative structure property/activity relationships

In computational chemistry, there is an alternative approach to realize the deep learning by directly mapping a simple representation of the considered molecules to the target property [68]. Such an approach can be broadly divided into THE quantitative structure property relationship (QSPR) and quantitative structure activity relationship (QSAR). Generally speaking, QSPR predicts the properties of molecular systems. However, QSAR focuses on molecule activity. Both methods are employed to improve the prediction accuracy [89, 90]. The training data available directly determines the properties which can be predicted. Therefore, the choice of the available databases is essential, which are summarized in the excellent review of Butler. It should be noted that some properties can be readily computed with DFT, such as the dipole moments, ionization energies, ground state energies, and so on. The inclusion of computational data sets greatly maximizes the speed and enriches the available data as much as possible.

Challenges in large-scale molecular dynamics

Progress in theoretical and computational chemistry originates from technical research and developments in the field of materials science. Related research can directly solve practical problems leading to the so-called social demand-driven research. Owing to the understanding of fundamental problems, researchers inevitably have to solve these problems when they try to explain the mechanism of experimental phenomena and processes, *i.e.*, problem-driven research. Although the specific problems and systems studied change with time, the current hot research topics will be gradually replaced with new problems. Researchers always have a strong interest in performing larger and larger simulations processing more complex systems, simulating longer dynamics, analyzing more microscopic details, and increasing the diversity of the studied systems [66, 91, 92]. However, large-scale molecular dynamics simulations require access to multicore clusters or supercomputers that are not always available to all researchers and more and more researchers explore the potential applications of GPUs consequently.

Applications in reaction prediction and optimization

The development of technologies for chemical synthesis traditionally required expert knowledge and special laboratory practice. Artificial intelligence might solve these problems to create new knowledge more effectively [15]. By putting a huge amount of experimental and related theoretical data into the computer program, deep learning and other artificial intelligence technologies would help experimental chemists find new important reaction routes [17]. Automation of retrosynthetic analysis has already been put into practice and different learning methods for the prediction of reaction products are also being developed. Using the same logic, one could also optimize reaction conditions and design natural product drugs. Further application of

artificial intelligence technology in those fields must accelerate related research and industry [93]. Although reinforcement learning and machine learning has been used to optimize reaction conditions, access to high-quality, interpretable, and standardized data sets suitable for machine learning is a current bottleneck as the literature on chemical reactions is often unstructured, exists in multiple formats, sometimes behind paywalls, and was collected on different reaction setups [94–97].

Conclusions and outlook

We do not intend to make a comprehensive review of chemical theory and computation, since the scope is vast and changing rapidly. Here we have summarized some major opportunities and challenges from several representative research fields in theoretical and computational chemistry. It should be noted that now many experimentalists can use computational chemistry as a lab tool to supplement analysis and characterization, thanks to the successful developments of package programs operated as a black box. The overall challenges for theoretical chemists lie in developing more efficient and more accurate computational methods for even bigger and more complex systems.

Acknowledgments: This work is a part of "IUPAC organizational structure review" for the "future direction of chemistry survey". The authors are grateful to Dr. Mark Cesa, Prof. Javier Garcia-Martinez, Dr. Michael Droescher, Dr. Lori Ferrins, and Prof. Ito Chao for their encouragement and fruitful discussions.
Research funding: This work was funded by National Key Research and Development Program of China (Grant No. 2020YFA0713601; Funder ID: 10.13039/501100012166), the National Natural Science Foundation of China (Grant Nos. 21790340 and 22073092; Funder ID: 10.13039/501100001809), and the Chinese Chemical Society through the "Young Elite Scientist Lift-Up" (Grant No. YESS20160032).

References

[1] D. C. Young. *Computational Chemistry: A Practical Guide for Applying Techniques to Real World Problems*, John Wiley & Sons, Inc., New York (2001).
[2] J. B. Mitchell. *Future Med. Chem.* **3**, 451 (2011).
[3] G. B. Goh, N. O. Hodas, A. Vishnu. *J. Comput. Chem.* **38**, 1291 (2017).
[4] R. Gómez-Bombarelli, A. Aspuru-Guzik. in *Handbook of Materials Modeling: Methods: Theory and Modeling*, W. Andreoni, S. Yip (Eds.), pp. 1–24, Springer International Publishing, Cham (2018).
[5] K. I. Ramachandran, G. Deepa, K. Namboori. *Computational Chemistry and Molecular Modeling: Principles and Applications*, Springer-Verlag Berlin Heidelberg, Coimbatore (2008).
[6] P. A. M. Dirac. *Math. Proc. Camb. Phil. Soc.* **26**, 376 (1930).
[7] M. D. Yandell, W. H. Majoros. *Nat. Rev. Genet.* **3**, 601 (2002).
[8] L. J. Jensen, J. Saric, P. Bork. *Nat. Rev. Genet.* **7**, 119 (2006).
[9] B. Maryasin, P. Marquetand, N. Maulide. *Angew. Chem. Int. Ed.* **57**, 6978 (2018).
[10] B. Burger, P. M. Maffettone, V. V. Gusev, C. M. Aitchison, Y. Bai, X. Wang, X. Li, B. M. Alston, B. Li, R. Clowes, N. Rankin, B. Harris, R. S. Sprick, A. I. Cooper. *Nature* **583**, 237 (2020).
[11] J. N. Wei, D. Duvenaud, A. Aspuru-Guzik. *ACS Cent. Sci.* **2**, 725 (2016).
[12] K. Ghosh, A. Stuke, M. Todorović, P. B. Jørgensen, M. N. Schmidt, A. Vehtari, P. Rinke. *Adv. Sci.* **6**, 1801367 (2019).
[13] X. Yang, Y. Wang, R. Byrne, G. Schneider, S. Yang. *Chem. Rev.* **119**, 10520 (2019).
[14] F. Peiretti, J. M. Brunel. *ACS Omega* **3**, 13263 (2018).
[15] A. F. de Almeida, R. Moreira, T. Rodrigues. *Nat. Rev. Chem.* **3**, 589 (2019).
[16] P. Carbonell, T. Radivojevic, H. García Martín. *ACS Synth. Biol.* **8**, 1474 (2019).
[17] M. H. S. Segler, M. Preuss, M. P. Waller. *Nature* **555**, 604 (2018).
[18] J. M. Granda, L. Donina, V. Dragone, D.-L. Long, L. Cronin. *Nature* **559**, 377 (2018).
[19] M. Popova, O. Isayev, A. Tropsha. *Sci. Adv.* **4**, eaap7885 (2018).
[20] N. J. O'Connor, A. S. M. Jonayat, M. J. Janik, T. P. Senftle. *Nat. Catal.* **1**, 531 (2018).
[21] W. Kohn. *Rev. Mod. Phys.* **71**, 1253 (1999).

[22] J. A. Pople. *Rev. Mod. Phys.* **71**, 1267 (1999).
[23] W. Koch, M. C. Holthausen. *A Chemist's Guide to Density Functional Theory*, Wiley-VCH Verlag GmbH, New York (2001).
[24] F. M. Bickelhaupt, E. J. Baerends. in *Reviews in Computational Chemistry*, K. B. Lipkowitz, D. B. Boyd (Eds.), pp. 1—86, Wiley-VCH, Inc., New York (2000).
[25] R. G. Parr, W. Yang. *Density Functional Theory of Atoms and Molecules*, Oxford University Press, New York (1989).
[26] A. Savin. *Int. J. Quant. Chem.* **22**, 59 (1988).
[27] H. Iikura, T. Tsuneda, T. Yanai, K. Hirao. *J. Chem. Phys.* **115**, 3540 (2001).
[28] A. D. Becke. *J. Chem. Phys.* **98**, 5648 (1993).
[29] C. Lee, W. Yang, R. G. Parr. *Phys. Rev. B* **37**, 785 (1988).
[30] J. P. Perdew, K. Burke, M. Ernzerhof. *Phys. Rev. Lett.* **77**, 3865 (1996).
[31] S. Grimme. *WIREs Comput. Mol. Sci.* **1**, 211 (2011).
[32] S. Grimme. *J. Comput. Chem.* **25**, 1463 (2004).
[33] S. Grimme. *J. Comput. Chem.* **27**, 1787 (2006).
[34] P. Jurečka, J. Černý, P. Hobza, D. R. Salahub. *J. Comput. Chem.* **28**, 555 (2007).
[35] S. Grimme, J. Antony, S. Ehrlich, H. Krieg. *J. Chem. Phys.* **132**, 154104 (2010).
[36] E. Caldeweyher, S. Ehlert, A. Hansen, H. Neugebauer, S. Spicher, C. Bannwarth, S. Grimme. *J. Chem. Phys.* **150**, 154122 (2019).
[37] M. Dion, H. Rydberg, E. Schröder, D. C. Langreth, B. I. Lundqvist. *Phys. Rev. Lett.* **92**, 246401 (2004).
[38] K. Lee, É. D. Murray, L. Kong, B. I. Lundqvist, D. C. Langreth. *Phys. Rev. B* **82**, 081101 (2010).
[39] E. R. Johnson, I. D. Mackie, G. A. DiLabio. *J. Phys. Org. Chem.* **22**, 1127 (2009).
[40] O. A. von Lilienfeld, I. Tavernelli, U. Rothlisberger, D. Sebastiani. *Phys. Rev. Lett.* **93**, 153004 (2004).
[41] Y. Y. Sun, Y.-H. Kim, K. Lee, S. B. Zhang. *J. Chem. Phys.* **129**, 154102 (2008).
[42] S. Grimme, A. Hansen, J. G. Brandenburg, C. Bannwarth. *Chem. Rev.* **116**, 5105 (2016).
[43] C. Hättig, W. Klopper, A. Köhn, D. P. Tew. *Chem. Rev.* **112**, 4 (2012).
[44] E. G. Hohenstein, C. D. Sherrill. *WIREs Comput. Mol. Sci.* **2**, 304 (2012).
[45] C. Riplinger, P. Pinski, U. Becker, E. F. Valeev, F. Neese. *J. Chem. Phys.* **144**, 024109 (2016).
[46] D. G. Liakos, F. Neese. *J. Chem. Theor. Comput.* **11**, 4054 (2015).
[47] A. J. Cohen, P. Mori-Sánchez, W. Yang. *Science* **321**, 792 (2008).
[48] J. P. Perdew, R. G. Parr, M. Levy, J. L. Balduz. *Phys. Rev. Lett.* **49**, 1691 (1982).
[49] A. J. Cohen, P. Mori-Sánchez, W. Yang. *Chem. Rev.* **112**, 289 (2012).
[50] H. van Aggelen, Y. Yang, W. Yang. *Phys. Rev.* **88**, 030501 (2013).
[51] W. Kutzelnigg, W. Klopper. *J. Chem. Phys.* **94**, 1985 (1991).
[52] G. H. Booth, A. J. W. Thom, A. Alavi. *J. Chem. Phys.* **131**, 054106 (2009).
[53] G. H. Booth, A. Grüneis, G. Kresse, A. Alavi. *Nature* **493**, 365 (2013).
[54] S. R. White. *Phys. Rev. Lett.* **69**, 2863 (1992).
[55] Z. Shuai, J. L. Brédas, S. K. Pati, S. Ramasesha. *Proc. SPIE-Int. Soc. Opt. Eng.* **3145**, 293 (1997).
[56] S. R. White, R. L. Martin. *J. Chem. Phys.* **110**, 4127 (1999).
[57] G. K.-L. Chan, S. Sharma. *Annu. Rev. Phys. Chem.* **62**, 465 (2011).
[58] J. Ren, Z. Shuai, G. Kin-Lic Chan. *J. Chem. Theor. Comput.* **14**, 5027 (2018).
[59] A. Mullard. *Nat. Rev. Drug Discov.* **13**, 877 (2014).
[60] I. Kola, J. Landis. *Nat. Rev. Drug Discov.* **3**, 711 (2004).
[61] V. Vaissier Welborn, T. Head-Gordon. *Chem. Rev.* **119**, 6613 (2019).
[62] Y. S. Meng, M. E. Arroyo-de Dompablo. *Energy Environ. Sci.* **2**, 589 (2009).
[63] A. Jain, Y. Shin, K. A. Persson. *Nat. Rev. Mater.* **1**, 15004 (2016).
[64] A. White. *MRS Bull.* **37**, 715 (2012).
[65] K. Schütt, S. Chmiela, O. von Lilienfeld, A. Tkatchenko, K. Tsuda, K. R. Müller (Eds.), *Machine Learning Meets Quantum Physics*, Springer International Publishing, Cham, Switzerland (2020).
[66] J. Hill, G. Mulholland, K. Persson, R. Seshadri, C. Wolverton, B. Meredig. *MRS Bull.* **41**, 399 (2016).
[67] C. W. Coley, W. H. Green, K. F. Jensen. *Acc. Chem. Res.* **51**, 1281 (2018).
[68] A. C. Mater, M. L. Coote. *J. Chem. Inf. Model.* **59**, 2545 (2019).
[69] C. K. Ober, S. Z. D. Cheng, P. T. Hammond, M. Muthukumar, E. Reichmanis, K. L. Wooley, T. P. Lodge. *Macromolecules* **42**, 465 (2009).
[70] T. E. Gartner, A. Jayaraman. *Macromolecules* **52**, 755 (2019).
[71] M. Muthukumar. *Macromolecules* **50**, 9528 (2017).
[72] M. Muthukumar. *Prog. Polym. Sci.* **100**, 101184 (2019).
[73] B. Lotz, T. Miyoshi, S. Z. D. Cheng. *Macromolecules* **50**, 5995 (2017).
[74] G. B. McKenna, S. L. Simon. *Macromolecules* **50**, 6333 (2017).
[75] Y. Ruan, Z. Wang, Y. Lu, L. An. *Acta Polym. Sin.* **727**, 743 (2017).
[76] Y. Lu, L. Li, W. Yu, L. An. *Acta Polym. Sin.* **12**, 1558 (2018).
[77] Y. Ruan, Y. Lu, L. An. *Acta Polym. Sin.* **12**, 1493 (2018).

[78] Y. Lu, L. An, J. Wang. *Acta Polym. Sin.* **6**, 688 (2016).
[79] Y. Ruan, Y. Lu, L. An, Z.-G. Wang. *Macromolecules* **52**, 4103 (2019).
[80] T. J. Struble, J. C. Alvarez, S. P. Brown, M. Chytil, J. Cisar, R. L. DesJarlais, O. Engkvist, S. A. Frank, D. R. Greve, D. J. Griffin, X. Hou, J. W. Johannes, C. Kreatsoulas, B. Lahue, M. Mathea, G. Mogk, C. A. Nicolaou, A. D. Palmer, D. J. Price, R. I. Robinson, S. Salentin, L. Xing, T. Jaakkola, W. H. Green, R. Barzilay, C. W. Coley, K. F. Jensen. *J. Med. Chem.* **63**, 8667 (2020).
[81] J. G. Freeze, H. R. Kelly, V. S. Batista. *Chem. Rev.* **119**, 6595 (2019).
[82] P. Mamoshina, A. Vieira, E. Putin, A. Zhavoronkov. *Mol. Pharm.* **13**, 1445 (2016).
[83] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, A. Walsh. *Nature* **559**, 547 (2018).
[84] L. Zhang, J. Han, H. Wang, R. Car, W. E. *Phys. Rev. Lett.* **120**, 143001 (2018).
[85] J. S. Smith, O. Isayev, A. E. Roitberg. *Chem. Sci.* **8**, 3192 (2017).
[86] J. Behler. *J. Chem. Phys.* **134**, 074106 (2011).
[87] J. Behler. *Phys. Chem. Chem. Phys.* **13**, 17930 (2011).
[88] R. T. McGibbon, A. G. Taube, A. G. Donchev, K. Siva, F. Hernández, C. Hargus, K.-H. Law, J. L. Klepeis, D. E. Shaw. *J. Chem. Phys.* **147**, 161725 (2017).
[89] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. von Lilienfeld, K.-R. Müller, A. Tkatchenko. *J. Phys. Chem. Lett.* **6**, 2326 (2015).
[90] G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Müller, O. Anatole von Lilienfeld. *New J. Phys.* **15**, 095003 (2013).
[91] A. V. Akimov, O. V. Prezhdo. *Chem. Rev.* **115**, 5797 (2015).
[92] H. Zhu. *Annu. Rev. Pharmacol. Toxicol.* **60**, 573 (2020).
[93] P. Schneider, W. P. Walters, A. T. Plowright, N. Sieroka, J. Listgarten, R. A. Goodnow, J. Fisher, J. M. Jansen, J. S. Duca, T. S. Rush, M. Zentgraf, J. E. Hill, E. Krutoholow, M. Kohler, J. Blaney, K. Funatsu, C. Luebkemann, G. Schneider. *Nat. Rev. Drug Discov.* **19**, 353 (2020).
[94] Z. Zhou, X. Li, R. N. Zare. *ACS Cent. Sci.* **3**, 1337 (2017).
[95] H. Gao, T. J. Struble, C. W. Coley, Y. Wang, W. H. Green, K. F. Jensen. *ACS Cent. Sci.* **4**, 1465 (2018).
[96] C. W. Coley, N. S. Eyke, K. F. Jensen. *Angew. Chem. Int. Ed.* **59**, 22858 (2020).
[97] C. W. Coley, N. S. Eyke, K. F. Jensen. *Angew. Chem. Int. Ed.* **59**, 23414 (2020).