

What drives computational chemistry forward: theory or computational power?

Wenfa Ng

Novena, Singapore, Email: ngwenfa771@hotmail.com

Abstract

History is often thought to be dull and boring – where large numbers of facts are memorized for passing exams. But the past informs the present and future; particularly in delineating the context surrounding specific events that, in turn, help provide a deeper understanding of their underlying causes and implications. Scientific progress (whether incremental or breakthroughs) is built upon prior work. Chronological examination of the field’s evolution reveals the existence of major “epochs” (e.g., transition from semi-empirical methods to first-principles calculations), and the centrality of key ideas (e.g., Schrodinger equation and Born-Oppenheimer approximation) in potentiating progress in the field. The longstanding question of whether computing power (both capacity and speed) or theoretical insights play a more important role in advancing computational chemistry was examined by taking into account the field’s development holistically. Particularly, availability of large amount of computing power at declining cost, and advent of graphics processing unit (GPU) powered parallel computing are enabling tools for solving hitherto intractable problems. Nevertheless, the article argues, using Born-Oppenheimer approximation as an example, that theoretical insights’ role in unlocking problems through simple (but insightful) assumptions is often overlooked. Collectively, the essay should be useful as a primer for appreciating major development periods in computational chemistry, from which, counterfactual questions illuminate the relative importance of intuition and advances in computing in moving the field forward.

Keywords: simulation; science history; *ab initio*; theory; experiment; electronic calculations; approximations; assumptions; first principles calculations; semi-empirical;

Subject areas: chemistry; physics; biology; computational sciences;

Synopsis

Simulation, together with theory and experiment, comprises the triumvirate of science. Herein, an opinion essay is described which, via examining computational chemistry development path, examines the relative importance of theory and computing power in advancing the field. As an ancillary component, the essay also briefly discusses the peculiar preponderance of eponymous method names in the field, and how jargon affects science communications. Specifically, chronological examination of computational chemistry's development shed light on three major overlapping periods: (i) theory development for explaining experimental observations of spectroscopic emission lines of elements and prediction of atomic structure, (ii) utilization of simplifying assumptions and experiment data for circumventing problems associated with lack of computing power in solving the Schrodinger equation in the pre-computing era, and finally, (iii) dramatic increase in low-cost computing power engendering the rise of first-principles (*ab initio*) methods for solving, with few or no simplifying assumptions, large systems comprising polyatomic and long-chain molecules. Besides the three eras mentioned, the increasing popularity of hybrid methods utilizing fine-grained approaches for simulating critical aspects of a model, and less accurate (but adequate) methods for examining other facets can be construed as a defining feature of the current period. The essay closes by critically assessing the relative importance of theoretical ingenuity and computational power in seeding new developments and breakthroughs, where simple but elegant insights, such as the Born-Oppenheimer approximation, help open paths to previously inaccessible solutions.

Jargon and science communication

Scientists want to communicate their research findings to others through simple, clear and effective writing. Nevertheless, there are constraints on the communication style and use of language, shaped by the requirements of the publishing process and the norms of particular fields. For example, jargon is used in all fields of science, and helps expedite communication of complex concepts between specialists conversant with a field's working language – particularly, during writing of manuscripts where strict page limits are imposed by many journals. Although expressing the same idea in plain language is desirable, the relatively lack of precision and the propensity of introducing subtle changes in meaning through syntax variation meant that technical jargon still play important roles.

Cursory glances at any scientific article on computational chemistry would likely fill the vision field with impenetrable terms and acronyms - which impedes understanding by those unfamiliar with the research area. The situation is not helped by the preponderance of eponymous names for associating particular methods with their developers – which, in contrast to names constituted by abbreviating short phrases describing a method – do not have any meaning. The

practice of naming methods by their eponymous developers (for example, Huckel method or Hartree-Fock Self-Consistent Field), which although a honour for the scientists, offers no information concerning the purpose or function of the method - and thus, obfuscate understanding by scientists and non-specialists outside of the field.

With availability of abundant inexpensive computing power and easy-to-use software packages, and greater appreciation of computational chemistry's utility in validating experimental findings or probing questions inaccessible to experiment, many researchers from other fields are excited by the possibilities afforded. While it is expected that individuals wishing to enter any research area need to invest time in learning a field's lexicon prior to navigating the pertinent literature, the highly abstract nature of computational chemistry, coupled with the peculiarity of its vocabulary (i.e., eponymous method names) presents a formidable challenge. Specifically, anecdotal accounts reveal that many students and practicing researchers are frustrated by the steep learning curve in deciphering the complex lexicon requisite for understanding the functions, assumptions, and methodologies of individual methods - details important for selecting an appropriate modelling tool. In fact, the "lexicon fog" surrounding computational chemistry is so dense - and the time commitment necessary for penetrating it so demanding - that it has dampen time-constrained researchers' enthusiasm in using computational chemistry tools for advancing their research in new and previously unanticipated directions. This, depending on the perspective, can be interpreted as a loss to science.

Distinct eras in the field

Electronic structure calculations, a sub-field of computational chemistry, focuses on explaining and predicting atomic organization and interactions between sub-atomic particles (i.e., neutrons, electrons and protons), and latter, how electron density is distributed between orbitals and their role in mediating bond formation between atoms. As in other research areas, a voluminous body of literature describes myriad methods and tools developed at various junctures in the field's evolution. Though seemingly disparate and not amenable to organization, analyzing different methods and tools chronologically lends clarity to their inter-relationships and reveals different phases in the field's evolution. Specifically, the three overlapping periods are: (i) initial experimental and theoretical studies elucidating the structure of the atom, and the motions of its sub-atomic constituents (a period where theory lagged behind experiment); followed by (ii) use of simplifying assumptions for solving models of single or few atoms and calibration of parameters using data from experimental studies in the era of relatively low computing power (a period where approximate and semi-empirical methods dominate); and finally, (iii) with abundance of low-cost computing power, the emergence of first-principles (*ab initio*) modelling approaches for solving large systems (comprising hundreds to thousands of molecules) with few or no assumptions.

Finally, we may be in the midst of a nascent era where a variety of coarse-graining or model-reduction approaches incorporating simplifying assumptions or experimental data helps researchers tackle problems hitherto only accessible on supercomputers. Specifically, by using fine-grained methods on aspects of a problem that directly informs the answers sought, while allowing some inaccuracies to permeate in other areas, model-reduction approaches significantly reduce the computational load. More important, these approaches allow large systems comprising complex molecules to be tackled using affordable and accessible computing resources such as a small cluster of graphics processing units (GPU) powered computers.

Evolution of the field

The discovery of sub-atomic particles such as the electron, proton and neutron sow the seeds of computational chemistry as an independent field of scientific inquiry. In particular, researchers of the day debated competing theories concerning atom organization, and the mechanistic underpinnings of the forces mediating interactions between sub-atomic particles. Success of the quantum mechanical approach (over classical physics) in explaining the key observation that orbiting negatively-charged electrons do not spiral into the positively-charged nucleus ushered in the emerging field of electronic structure calculations, whose main goal was to explain the emission spectra of various elements. Specifically, peaks present on the emission spectrum of elements (e.g., sodium and hydrogen) arise from the release or absorption of energy during transition of electrons between energy levels. Realization that electrons or, more accurately, electron densities, are arranged in defined energy levels and spatial regions led to the proposal of the atomic and molecular orbital concepts, which, from a quantum mechanical perspective, are regions where electrons of particular energies are located. This era was defined by the promulgation of many of the foundational concepts and tools of computational chemistry, where the theoretical tools of quantum mechanics illuminated spectroscopic observations, highlighting that theory lagged behind experiment in this period. Perhaps the defining contribution in this era was the formulation, by Erwin Schrodinger, of an equation that describes the total energy (or Hamiltonian) of any system. Known simply as the Schrodinger equation, its intractability to solution spawned an entire sub-field seeking to develop methods and strategies for solving it. More specifically, solution of the equation is crucial for understanding the placement of electrons of differing energies in distinct orbitals, which, in turn, determines an atom's chemical properties.

Development of various approximate methods incorporating simplifying assumptions for solving the Schrodinger equation dominated the second era of electronic structure calculations, of which the Born-Oppenheimer approximation is the most iconic. Specifically, the objective was to devise increasingly better and faster techniques for solving the electronic portion of the Hamiltonian with simplifying assumptions such as neglecting the electrostatic repulsions between

electrons (known as electron-electron correlation energy). One example that exemplifies the utility of approximations and assumptions in simplifying previously intractable problems for solution (though at the expense of slight but tolerable inaccuracies) is the Born-Oppenheimer approximation for decoupling electronic and nuclear motions encapsulated within the Schrodinger equation. Specifically, coupled motions of the nucleus and electrons, where electrons' movement influences the atomic nucleus and vice versa, accounts for the mathematical intractability of the Schrodinger equation (a many-body problem). The key to resolving this conundrum hinges on the observation that, for atoms of sufficiently large atomic mass, the nucleus is essentially fixed in space; thus, allowing the entangled motions of the nucleus and electrons to be decoupled. More important, the approximation would increasingly lead to more accurate solution as the atomic mass increases. By applying the approximation, only the electronic component of system energy needs to be solved, thereby, significantly reducing the computation load.

Given the inability of mechanical slide rule and rudimentary calculators in computing properties of atoms with sufficient accuracy and precision, the second era of computational chemistry was also characterized by the emergence of many semi-empirical methods, where experimental data – usually from spectroscopy studies – were used in calibrating essential parameters in models of systems. These parameters describe key characteristics of atoms and could not be calculated from first-principles in the pre-computing era. Additionally, lack of computing power also constrained the types of systems studied to those involving single or few atoms. More important, these systems were also investigated using models incorporating many assumptions - many of which unrealistic.

Increases in computing speed and capacity, and the availability of user-friendly software packages signal the arrival of the third era of electronic structure calculations. Specifically, greater computing power allows the calculation, from first-principles, of most system properties with minimal reliance on simplifying assumptions, in systems comprising large numbers of polyatomic molecules. Although systems of proteins or hundreds of atoms remain inaccessible to even the fastest supercomputers,¹ significantly larger systems of at least few tens of medium-sized molecules – which would allow meaningful answers to questions concerning reactivity, chemical kinetics and evolution of transition states to be obtained - have become increasingly accessible to interrogation. Additionally, advances in computing also democratizes computational chemistry; specifically, by allowing non-specialist researchers to perform routine investigations of simple systems via easy-to-use software packages on desktop computers – compared to command-line programmes on mainframe or supercomputers in an earlier era. Although not the sole *ab initio* method available, density functional theory (DFT) utilizing Gaussian type orbitals is the dominant technique for tackling a wide range of questions concerning reactivity and molecular recognition between molecules, in fields as diverse as material science, biochemistry and physics.

Finally, desire for simulating ever larger systems of long-chain molecules (reminiscent of real-world systems) using less computing time, or on desktop computers and small parallel computing clusters, has driven the development of various model-reduction strategies, in what is emerging as a nascent fourth era. This development taps on the efficiency and speed of semi-empirical and approximate methods, and the need of tackling large scale systems at spatiotemporal resolutions more closely resembling those of natural systems. Specifically, even though multi-fold speedup is afforded by high-performance computers, simulating the dynamics of excited electrons within a single large molecule is the technical limit; thereby, motivating the development of hybrid methods to more effectively simulate a system's most critical aspects in view of a given question. Within the family of model-reduction strategies, coarse-graining approaches, which combines first-principles methods with simplifying assumptions, is increasingly used. Essentially, coarse-graining seeks to employ the most suitable tool for different components of a problem. *Ab initio* techniques, for example, would be used in simulating the precise atomic movement during the binding and cleavage of a molecule at an enzyme's active site, while the important (but less critical) interactions between enzyme and water solvent would be approximated by a mean field that captures, in aggregate, all the electrostatic and *van der Waals* interactions between water molecules, and those between the enzyme and water molecules. Thus, using a mean field - for what would have been more fine-grained calculations - in simulating the solvent effect on enzyme catalysis, significantly reduces the computational requirement that a full explicit treatment of water molecules' interaction with the enzyme would engender. Current trends, where many investigators are employing myriad simulation techniques for solving problems at time and spatial scales relevant to real-world problems, together with large system size and complexity meant that, in the absence of a significant leap in computing power and reduction in cost, model-reduction approaches would likely remain popular. Nevertheless, future development of algorithms capable of first-principles simulation of large systems at a fraction of current cost, would revolutionize the field by making obsolete many of the model-reduction approaches currently in vogue.

History seldom evolves linearly, but rather, is punctuated by sets of related events that arose due to unique circumstances at particular time-points. For example, closer examination of the delineated time-line reveals the clustering of different methods depending on the assumptions used and the extent to which experimental data from spectroscopy and other instruments helps inform model building. One framework for demarcating different periods in the field clusters methods into three overlapping eras: (i) theoretical postulations and experimental elucidation of the structure of atoms and their sub-atomic constituents, (ii) calculations of electron density distribution and understanding the basis of chemical bonding for single or few atoms using simplified models calibrated with experimental data, and (iii) simulations of systems comprising large number of polyatomic molecules with few or no assumptions (i.e., first-principles calculations). Thus, the evolution of electronic structure calculations can be understood chronologically or through the identification of different periods - each characterized by a

dominant methodology trend. Nevertheless, clustering and binning myriad developments into categories inevitably requires using a set of arbitrary criteria – such as the relative importance of simplifying assumptions, experimental data input, theory or computing power. Choice of binning criteria is dependent on the questions asked and solutions sought. Thus, criteria used and perspective in examining a question determines the different eras defined.

Theory or computing power?

Since past events cannot be rewound, counterfactual analysis (i.e., asking “what if” questions) is useful for illuminating the likely consequence or implications of alternative trajectories of events at specific time-points. Similarly, counterfactual analysis could shed new light on the longstanding question concerning the relative importance of computing power and theoretical insights in advancing computational chemistry. Dramatic increases in computing power is generally recognized to have propelled computational chemistry forward; however, the essay argues that the interplay between computing capacity and theory may be more nuanced. For instance, closer examination of events following promulgation of the Schrodinger equation reveals that the significant computational challenge posed by the coupled motions of nucleus and electrons might had presented a stumbling block to research if not for the Born-Oppenheimer approximation. Specifically, the simplification enabled researchers to solve the sub-case of electron motion in the context of a fixed nucleus – an approximation which progressively improves with increasing atomic mass. Doing so allows partial solutions to be obtained, which although with caveats attached, nevertheless helped inform solution of real-world problems. Thus, theoretical insight’s importance in unlocking or initiating developments in computational chemistry is often under-appreciated. Additionally, intuition also serves as a useful check on your thinking – particularly in clarifying the cloud of convoluted equations that might otherwise obfuscate meaning or, act as roadblocks in the smooth flow of logical thought.

The opinion piece is an initial attempt at providing some thoughts on a longstanding debate – and certainly does not provide the last word on the issue. More detailed analysis of the question would come from science historians. Nevertheless, as history is the continuous evaluation, from different perspectives, of existing evidence in light of new developments, and coupled to the fact that successive generations of scholars cast their backward glance on past events from different vintage points, differing interpretations of the same events is expected – and is healthy for promoting intellectual debate. Borrowing an illustrative example from the biological sciences: although Mendel is widely acknowledged to have discovered the laws of genetics, recent research suggests that his research direction – and experiment design – might have been inspired by Imre Festetics.² Similarly, future developments in computational chemistry and re-interpretations of old evidence from fresh perspectives may lead to slightly different conclusions on the above debate.

Collectively, analysis of computational chemistry's chronology highlights four overlapping phases in the field's development, i.e., (i) development of theories for explaining experimental observations of emission lines of elements and prediction of electron density distributions in atoms; (ii) semi-empirical and approximate methods utilizing experimental data and simplifying assumptions for calculating electronic structure of single or few atoms; (iii) first-principles methods for tackling systems comprising large numbers of polyatomic molecules; and (iv) combination of fine-grained and approximate methods in model-reduction approaches for reducing the computational requirement during solution of large systems at meaningful spatiotemporal resolution within reasonable time. Finally, exploration of the relative roles played by computing power and intuition in advancing the field illuminates the importance of theoretical ingenuity in unlocking hitherto intractable problems, while acknowledging the importance of large amount of inexpensive computing power in potentiating the transition from semi-empirical to first-principles methods.

References

1. Van Noorden, R. Modellers react to chemistry award. *Nature* **502**, 280 (2013).
2. Poczai, P., Bell, N. & Hyvönen, J. Imre Festetics and the Sheep Breeders' Society of Moravia: Mendel's Forgotten "Research Network". *PLoS Biology* **12**, e1001772 (2014).