Walking down computational chemistry memory lane one acronym at a time

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. To the uninitiated, the computational chemistry literature appears intimating given the pervasive use of acronyms and eponymous method names. While jargons expedite communication of complex ideas between specialists, and add clarity to a discussion (e.g., explaining complicated concepts in plain language may not capture subtle - but important - nuances in meaning), they nevertheless presents a significant barrier to understanding for researchers in other fields. Specifically, an inability to comprehend the meaning of the various terms and jargons used would significantly impede understanding and navigating the literature – and may translate into difficulty in selecting appropriate tools for the task at hand. Scientific progress (incremental and breakthroughs) is built upon prior work. By placing various computational methods and techniques along a chronological thread, a commentary article aims to demystify the tangled web of acronyms and terms that populate the electronic structure calculations literature and highlights the interrelationships between methods – particularly, how one method evolved from another. Additionally, the chronological framework also allows readers to appreciate developments in computational chemistry through the lens 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 charting progress in the field. Finally, the chronological time-line delineated also provides an opportune backdrop for examining the longstanding question of whether computational power (both capacity and speed) or theoretical insights play a more important role in advancing computational chemistry research. 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 article should be useful as a primer for researchers to gain a more holistic understanding of computational chemistry, and students wishing to learn more about the conceptual basis and purpose of various electronic structure calculations methods prior to venturing into the field's expansive literature.

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. But 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 - are not associated with any meaning. With firsthand experience of the difficulty of disentangling and understanding the web of acronyms and terms nestled within dense technical prose, and deciphering the meaning of methods from the corresponding abbreviations, I wrote a commentary article to help readers better understand the main functions and key methodological underpinnings of the methods, and how they are built upon one another. Additionally, viewing the field's development as a sequence of seemingly disparate methods along a time-line revealed three distinct phases in computational chemistry's development. Specifically, (i) theory development for explaining experimental observations of spectroscopic emission lines of elements and prediction of atomic structure, (ii) utilization of simplifying assumptions and experimental 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 of inexpensive 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. Finally, using the chronological thread delineated, and drawing on examples in the field (where simple but elegant insights, such as the Born-Oppenheimer approximation, help open paths to previously inaccessible solutions), an attempt would be made to critically assess the relative importance of theoretical ingenuity and computational power in seeding new developments and breakthroughs in the field.

Scientists want to communicate their research findings to others through simple, clear and effective writing. Nevertheless, there are constrains 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. Additionally, while 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 has an important role to play. What is different in computational chemistry, however, is 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 mentioned, offers no information concerning the purpose or function of the method - and thus, obfuscate understanding by scientists and non-scientists outside of the field.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.552v2 | CC-BY 4.0 Open Access | rec: 18 Sep 2015, publ: 18 Sep 2015

With availability of large amount of 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 peculiar characteristics of its vocabulary (e.g., eponymous method names etc.) presents a formidable challenge. Specifically, anecdotal accounts reveal that many students and practicing researchers are frustrated by the steep learning curve involved in deciphering the complex lexicon necessary for understanding the functions, assumptions, and methodologies of individual methods - details important for selecting an appropriate computational 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 as important enablers for advancing their research in new and previously unanticipated directions. This, depending on the perspective, can be construed as a loss to science. Thus, by presenting various electronic structure calculations methods within a coherent framework, the article should offer some help for students and newcomers in gaining initial understanding of the key functions, assumptions and application areas of important methods.

Electronic structure calculations is a sub-field of computational chemistry that initially 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. Even in such a well-defined area, 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, placement of different computational chemistry methods and tools along a chronological thread lends clarity to their inter-relationships and reveals distinct phases in the field's evolution. Specifically, three distinct phases or "epochs" in the development of electronic structure calculations readily emerges upon closer examination of the historical evolution of the field: (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) the 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 computational 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 fourth 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.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.552v2 | CC-BY 4.0 Open Access | rec: 18 Sep 2015, publ: 18 Sep 2015

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 help significantly reduce the computational load required. More important, such 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.

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 the organization of the atom, 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 nascent field of electronic structure calculations, whose main objective was to explain the emission spectra of various elements obtained by spectroscopy studies. Specifically, peaks present on the emission spectrum of elements (e.g., sodium and hydrogen) result 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 different orbitals, which, in turn, determines the chemical properties of an atom.

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 purpose 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 use of 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 lies in 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 calculating the various 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 to calibrate essential parameters in models of a particular system. These parameters describe key characteristics of atoms and could not be determined 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 are unrealistic.

Increases in computational speed and capacity, and the availability of user-friendly software packages signal the arrival of the current era of computational chemistry and electronic structure calculations. Specifically, greater computing power allows the calculation, from first-principles, of most system properties with minimal reliance on simplifying assumptions – and at system-relevant spatial and time scales, which typically comprise large numbers of polyatomic molecules. In particular, though systems comprising complete proteins or hundreds of atoms remain inaccessible to even the fastest supercomputers available,[1] significantly larger systems of at least few tens of 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, increase in computational capacity also democratizes the practice of 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 super-computers in an earlier era. Although not the sole *ab initio* method available, density functional theory (DFT) utilizing Gaussian type orbitals is the predominant technique for tackling a range of questions concerning reactivity and molecular recognition between molecules, in fields as diverse as material science, biochemistry and physics.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.552v2 | CC-BY 4.0 Open Access | rec: 18 Sep 2015, publ: 18 Sep 2015

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, separate from the current epoch dominated by first-principles calculations in general and density functional theory in particular. This development is driven in part by the computational efficiency and speed of semi-empirical and approximate methods, and the desire of tackling large scale systems at spatiotemporal resolutions more closely resembling those in natural systems. In particular, within the family of model reduction strategies, coarse-graining approaches – which combines the use of first-principles methods with simplifying assumptions – is increasingly used for tackling problems. In essence, coarse-graining seeks to employ the most suitable tool for tackling individual sub-components of a problem. For example, full *ab initio* techniques would be employed for simulating the precise atomic movement during the binding and subsequent 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 via 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. Such a task that would only likely be tackled by large computing clusters, or even a supercomputer. While it is difficult to predict future developments, given current trends where various investigators are employing myriad computational techniques for solving problems at physiologically relevant time and spatial scales, large system size and complexity meant that, in the absence of a significant leap in computational power and reduction in cost, model reduction approaches would remain popular choices for most researchers. Nevertheless, future development of algorithms capable of first-principles simulation of large systems at a fraction of the current computation 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. Overall, the field of electronic structure calculations can be classified into three distinct 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

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.552v2 | CC-BY 4.0 Open Access | rec: 18 Sep 2015, publ: 18 Sep 2015

evolution of electronic structure calculations can be understood chronologically or through the identification of distinct phases each characterized by a dominant trend in method development. Nevertheless, clustering and binning myriad developments into distinct categories inevitably requires the use of a set of arbitrary criteria – which, in this case, comprises the relative importance of simplifying assumptions, experimental data input, theory or computational power in facilitating the solution of defining equations of systems. Choice of criteria for partitioning different methods into distinct categories is closely intertwined with the questions asked and solutions sought. Thus, depending on the criteria used and perspective of examining a question, different eras can be defined.

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, by placing the development of important methods and discoveries pertinent to electronic structure calculations along a time-line, counterfactual analysis may also 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 over the past decades is generally recognized to have propelled computational chemistry forward; however, the article argues that the interplay between computational capacity and theory may be more nuanced. For instance, closer examination of events following the promulgation of the Schrodinger equation reveals that the significant computational challenge posed by the coupled motions of the nucleus and electrons might had presented a stumbling block to research if not for the simplifying assumption offered by Born and Oppenheimer. Specifically, the assumption enabled researchers to solve the simpler case of electron motion in the context of a fixed nucleus – an approximation which progressively approaches the true solution for atoms of increasing mass. Doing so allows partial solutions to be obtained, which although with caveats attached, nevertheless helped inform solution of problems. Thus, theoretical insight's importance in potentiating 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 article is an initial attempt at providing some thoughts on a longstanding debate – and certainly will not provide the last word on the issue. More detailed examination of the question would await the input of 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 from the viewpoint of promoting intellectual debate. Borrowing an illustrative example from the biological

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.552v2 | CC-BY 4.0 Open Access | rec: 18 Sep 2015, publ: 18 Sep 2015

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, various terms and eponymous method names in electronic structure calculations are placed along a time-line to better reveal their inter-relationships – in particular, how different methods are built upon one another – and the context surrounding their development. Analysis of the chronological thread highlights three distinct 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 densities in atoms; (ii) semi-empirical and approximate methods utilizing experimental data and simplifying assumption for calculating electronic structure of single or few atoms; and (iii) first-principles calculations for tackling systems comprising large numbers of polyatomic molecules. Finally, exploration of the relative roles played by computational power and theoretical insights in advancing the field illuminates the importance of theoretical ingenuity in unlocking hitherto intractable problems, while acknowledging the centrality 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).

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.552v2 | CC-BY 4.0 Open Access | rec: 18 Sep 2015, publ: 18 Sep 2015