Mathematical challenges in computational chemistry: multiscale, multiconfigurational approaches, machine learning

Sergey Gusarov (National Research Council Canada),
Alexander E. Kobryn (National Research Council Canada),
Stanislav Stoyanov (Natural Resources Canada),
Valera Veryazov (Lund University, Sweden)

07/08/2022–07/10/2022

The meeting took place in July 2022 right after the largest international computational chemistry event WATOC 2022 in Vancouver, BC. The connection to the WATOC allowed us to invite to our symposium leading scientists from around the world (Canada, Japan, Sweden, US). The format of the meeting was in-person with the total number of participants 15 (during the meeting one person participated by video conferencing from the BIRS hotel room because of the positive COVID-19 test on the arriving day).

1 Overview of the Field

In the last two decades theory and modeling turned to become one of the major topics of applied chemistry along with analytic, synthetic, and other chemistry fields. This made possible because of significant improvements in methodology, numerical methods, and computer software and hardware. Much experimental research started to include computational modeling. The role of computer simulation in modern chemistry cannot be overestimated and the use of effective modeling and simulation plays a critical role in practical applications by providing insights into experiments and helping in system optimization. Specifically, simulations are more and more often used to substitute dangerous and expensive experiments with calculations. At the same time, the impressive progress of modern experimental research in material science and biology necessitates further developments and continuous extension of the applicability and accuracy of nowadays computational chemistry methods. The fast but accurate qualitative and quantitative modeling of large biological molecules, nanoparticles, and interfaces becomes the main focus of the research which requires significant computational efforts and is not always achievable at the current technology level. Most of the computational chemistry problems are about solving the Schrödinger equation for electrons in molecules or the Newton equations of motion for a system of classical particles. Consequently, the mathematics should play the central role in the new developments. The primary purpose of this workshop was to analyse the current needs and expectations of computational chemistry based on the experience provided by top leading scientists and discuss them with the methodology and computational software developers. The following sections have their names after the workshop sessions and comprise both the topics suggested in the initial presentations and topics brought to the surface during the round-table discussions and interpersonal talks.

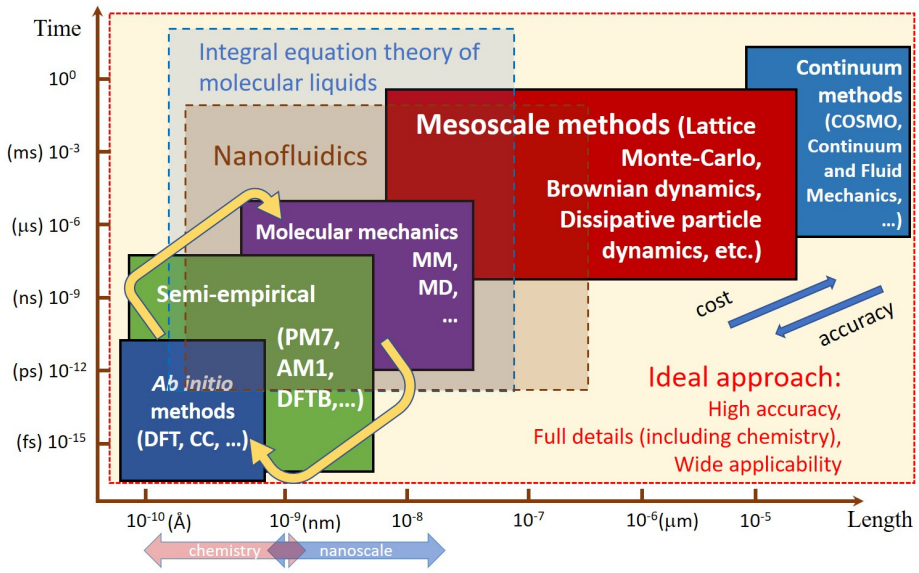

Figure 1: Length-Time scales diagram.

2 Multiscale approach

It is convenient to exhibit the modern computational chemistry methods on length and time scale diagram, Figure 1, where each category of approaches (e.g., ab initio or molecular mechanics) are approximately illustrated by rectangular box positioned according to its applicability (lower left corner) and computational cost (upper right corner). In the case of logarithmic time and length scale the boxes form almost linear hierarchical structure with overlapping regions where the corresponding methods could be applied to model the system of interest. These regions have a very important meaning in the methodology development and practical applications as they allow to verify/estimate the accuracy of coarser methods compare to their more accurate but computationally expensive counterpart. Typically, this could be done by averaging the detailed information from more precise approach and following comparison with the results of higher scale. In the ideal case, we would like to get accurate and detailed information for very large objects which is practically impossible. For example, in this hierarchy the chemical properties are a special interest in modern nano- and bio- sciences but they are only accessible within quantum chemical (ab-initio) and partially semi-empirical approaches which are limited by their polynomial (cubical in the case of LDA and GGA DFT and higher for more accurate approaches) scaling factors in respect to the size of system. Moreover, the application of quantum chemical approaches to the large systems will result in the huge amount of data unavailable to keep with the modern level of hardware. This led to in the principal restriction of modern computational chemistry which might result in the future competitive gap between experimental and computational chemistry.

A very attractive strategy to resolve these restrictions is to use machine learning (ML) methods enabling large-scale exploration of chemical space based on quantum chemical calculations. However, despite being fast and accurate for atomistic chemical properties, the modern ML models do not explicitly capture the electronic degrees of freedom of a molecule, which limits their applicability for reactive chemistry and chemical analysis. So, the new consistent descriptors are needed which are based on deep analysis of the structure of the Schrödinger equation. We will discuss the recent achievements in this area and analyse the pros and cons. Also, all other alternative developments are welcome to discuss during the workshop. For example, the general theory of multiscale techniques is intensively developing in the math community. This could provide a way to optimize the solution to Schrödinger equation, however these general math approaches should be translated to the practical language.

3 New methods in quantum chemistry

The session was moderated by Valera Varyazov, who also delivered a small talk followed by a brief overview by Victor Hugo Malamace da Silva. In conversation it was noticed that many breakthroughs in quantum chemistry have been inspired and driven by mathematical ideas. Use of a wavefunction in the form of determinant to provide a permutation symmetry of the electron wavefunction (the base of the Hartree-Fock theory) is a famous example of such influence. Cholesky decomposition is another example of a mathematical idea, which is used to reduce the amount of computed integrals in many computational codes. Unfortunately, such influence comes to applications with a significant delay (Cholesky decomposition is known from the beginning of 20th century, suggested to be used in 1977, and implemented about 40 years later). Furthermore, from the point of mathematics, quantum chemistry is a huge underdeveloped area: the solved equations can have multiple solutions, or be unstable. Extensive use of numerical approaches and approximations is not always justified. Although we do not have an immediate solution for the better use of mathematical ideas in quantum chemistry, we have to spot the problem with a hope for a change. Advances in the development of new hardware is another game changing factor for quantum chemistry. The "old" paradigm was based on the idea of limited resources (CPUs, memory, storage), and so promotes the reuse of computed data and batching of all calculations. With new hardware architectures the design of computational codes can be significantly revised and simplified at the same time.

4 New computational science ideas

The session, moderated by Stanislav R. Stoyanov, focused mainly on the highly promising and novel applications of ML in the field of computational chemistry. The opening presentation, delivered by Olga Lyubimova, started with an overview of ML, continued with an introduction to molecular featurization and representation, and culminated with a digest of recently proposed chemical descriptors for ML treatment. A heated discussion ensued on the trustworthiness and acceptance of computational chemistry results from ML compared to those from the traditional quantum chemistry-based results. Dr. Lyubimova effectively addressed the concerns, explaining that the mathematics behind ML is not only complex but also robust. Noting her initial cautious attitude towards ML and being trained as a quantum computational chemist, she shared that after taking ML training and carefully and comprehensively reviewing the literature on this novel topic, she started gradually gaining understanding and building confidence in the predictive capability of ML in computational chemistry. The selection of an ML method to employ was noted as a major challenge and an area for improvement towards automation. The consensus was that while quantum chemistry was based on solid physical and mathematical foundation it was very expensive computational and needed transformative performance improvements. In this context, mathematically sound ML approaches could be the necessary tools to speed-up computational chemistry research. It is noteworthy that the representatives with mathematical geology expertise were much more comfortable than computational chemists using ML, e.g., for image analysis, likely because the latter field did not have a solid physics-based predictive and interpretive framework comparable to the quantum theory. Several participants pointed out that while the complete replacement of quantum chemistry with ML would be premature, ML can effectively help address optimization and algorithm selection problems, thus helping accelerate quantum chemistry-based calculations. Moreover, hybrid density functional selection that is typically made based on the users' experience or published recommendations could potentially be determined by using an ML algorithm, as it reflects the percentage of exchange and correlation included in the functional. Another area for improvement by using ML would be the correction for the dispersion interactions that were often done ad hoc, as these important interactions were not accounted by density functional theory, the most widely used quantum chemical method.

New ideas in the areas of mathematical libraries, algorithms, compilers, and hardware were discussed in brief because these were to a large extent covered in the two morning sessions. The importance of unified data formats and their key role in enhancing the communication between computational codes towards automation were also discussed, mainly in the context of the variety of abstract mathematical representations, e.g., using graphs or molecular fingerprinting, that were not always compatible and suffered from reconstruction inaccuracies. These unification and communication challenges were noted to arise due to different conventions used in diverse areas of computational chemistry. Enhancing the communication between computational codes required novel and improved algorithms to transcribe and convert among data formats.

5 Future development

This session was composed of several largely independent topics. All of them were initially suggested by the organizers and later followed by the participants. The session moderator was Alexander E. Kobryn. All the workshop participants actively took place in the debates. In addition, mini presentations on this matter were delivered by A.E. Kobryn and Gabriel Pereira da Costa. The following subsections summarize the most pertinent information and its discussion.

5.1 Acceleration of computations with GPUs

Graphics Processing Units (GPUs) are known as programmable processing units independently working from Central Processor Units (CPUs) and originally responsible for graphics manipulation and output. Because of their high performance in data processing, from the beginning of new millennium parallel GPUs started to be actively used for General Purpose Computing on GPU (GPGPU) and later found its way into fields of material science, computational chemistry, and quantum chemistry. Therefore, contemporary High Performance Computing (HPC) clusters often provide, in addition to the so-called regular compute nodes, the GPU nodes where computations can be run on both CPU and GPU cores. At the same time, big scientific software developers started providing the GPU support in their products. This information is easy to trace and can be found, e.g., at the HandWiki list of quantum chemistry and solid-state physics software [1]. In particular, one can identify that such big developers as ADF, GAMESS, Gaussian, MOLCAS, Quantum Espresso, VASP – to name a few – already support GPUs. At the same time, such popular products as DFTB+/++, DMol3, OpenMX, ORCA – also to name a few – do not yet account for the possibility to accelerate computations with the use of GPUs. The workshop participants have agreed and expressed a hope that the future development of the computational modeling software should include the GPUs support, and the architects of the next generation HPC clusters should continue equipping them with a set of GPU nodes. The workshop participants also agreed to circulate this expectation at any other relevant public events and through the interpersonal communications.

5.2 Quantum computing

In last decade, the most growing expectation with respect to the increase of the computation speed and complexity was about quantum computers – devices that perform quantum computing, a type of computation that harness collective properties of quantum states, such as superposition, interference, and entanglement [2, 3]. The basic unit of quantum information in quantum computing is quantum bit or qbit – a two-state quantum mechanical system often compared for simplicity with an imaginary spin-up/down system and represented as

$$|0\rangle = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad |1\rangle = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

A quantum memory may then be found in any quantum superposition $|\psi\rangle$ of the two states $|0\rangle$ and $|1\rangle$, i.e. $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$, with the coefficients $\alpha, \beta \in \mathbb{C}$, satisfying $|\alpha|^2 + |\beta|^2 = 1$, and called quantum amplitudes. The state of the quantum memory can be manipulated by applying quantum logic gates, in analogy to how classical memory can be manipulated with classical logic gates (AND, OR, XOR, NOT, etc.). With this respect, the most practical type of quantum computers at present seems the quantum circuit model, in which a computation is a sequence of quantum gates and measurements. The great expectations from quantum computing may be explained by the fact that quantum algorithms sometimes offer a polynomial or super-polynomial speed-up over the best known classical algorithms. Figure 2 and Table 1 show a schematic chart of a computing cost and a complexity scaling. For material science, computational chemistry, and quantum chemistry problems this factor is the decisive one and will determine the technological progress in these fields, provided the engineering task of building a powerful quantum computer is solved. The workshop participants noticed in the discussion that technical challenges of this task include not only the problem of physical scalability to increase the number of qbits, but also building quantum gates that are faster than the decoherence time, and lowering the error rates and bringing them to the level of modern classical computers.

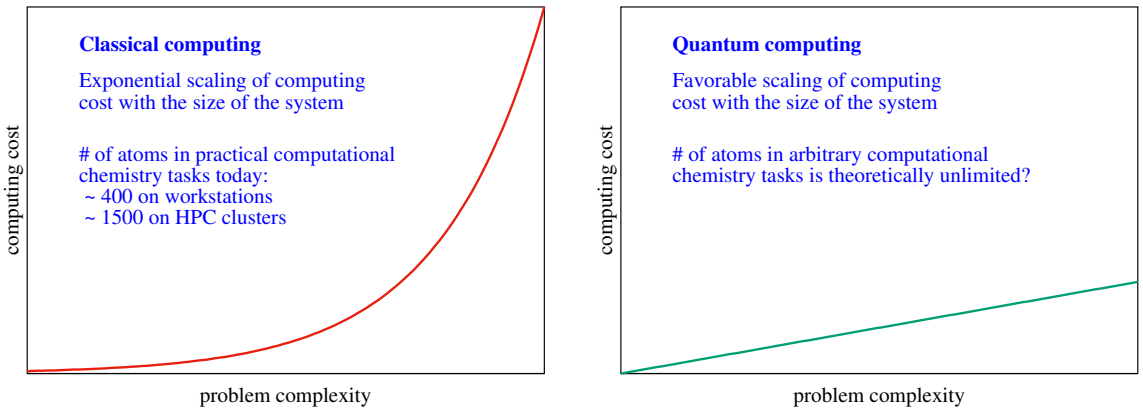

Figure 2: A schematic chart of scaling of computing cost on classical and quantum computers with the increase of the system size or the problem complexity. On charts, the problem complexity increases from left to right and the computing cost increases from bottom to top.

Table 1: The performance of classical vs quantum computers for few selected subroutines that are critical for the execution of the entire algorithm. There is also a comparison of error rates and application areas.

Classical computers		Quantum computers	
Subroutine	Complexity	Subroutine	Complexity
Matrix inversion $AX = B \rightarrow X = A^{-1}B$	$O\left(N \log N\right)$	Matrix inversion $\hat{A}\|X\rangle = \|B\rangle \rightarrow \|X\rangle = \hat{A}^{-1}\|B\rangle$	$O\left((\log N)^2\right)$
Eigenvalues and eigenvectors of sparse/low-rank matrices	$O\left(N^2\right)$	Quantum phase estimation (a.k.a. Q-phase)	$O\left((\log N)^2\right)$
Fast Fourier transform	$O\left(N \log N\right)$	Quantum Fourier transform	$O\left((\log N)^2\right)$
Have low error rates (10^{-15}) and can operate at room temperature		Have high error rates (10^{-3}) and need to be kept at ultralow temperatures	
Are best for everyday numerical processing		Well suited for tasks like optimization problems, data analysis, and simulations	

5.3 Incremental improvements of computer codes vs rewriting from scratch

In computational science each noticeable progress in the hardware development means that the software should be improved and refactored continuously all the time. Then, the principal question of the scale "to be or not to be" is shall the code improvement be incremental, little-by-little, or comprehensive, with rewriting the whole code from the beginning? A general answer to this question does not exist as every situation is worthy a thoughtful consideration. Before making decision it may be not bad idea to start from the time and cost assessments for the following categories: (i) time and cost of improving the existing code; (ii) time and cost of rewriting from scratch; (iii) time and cost of fixing bugs and adding new features; (iv) time and cost of updating and circulating instructions, manuals, tutorials, etc.; (v) time and cost of team management for each of these cases. A separate assessment should be for the level of impact the changes may have on scientific results and therefore appear for the users to be important and valuable or inessential and unappealing. Because of this, the picture we often observe over the years is that both small and big developers prefer improving the existing codes over rewriting them from scratch. The workshop participants agreed that the future development will rather not alter this picture and that the existing balance between

the frequency of the so-called major and minor scientific software updates will be preserved. In both cases, however, one shall be ready to new bugs appearing in the rewritten software. Referring to the Preface of one famous textbook [4]: "In computer science it is generally assumed that any source code over 200 lines contains at least one error". Even so, the story does not end there. Quite oppositely, it marks the beginning of a new cycle in the never-ending line of updates.

5.4 How math can drive and speed up the development of computational chemistry

In recent decades, computational methods became major tools of theoretical studies. Accordingly, the mathematical models and numerical analysis that underlie these methods have an increasingly important and direct role to play in the progress of computational and quantum chemistry [5]. No wonder, the number of mathematical challenges in this area remains high. In our discussion we could mention the need for the following:

- Developing of accurate coarse grained models at moderate computational cost;
- Developing models that exploit the multiscale nature of computational chemistry problems;
- Developing models that properly reflect and describe quantum stochastic processes;
- Development of appropriate treatment for strongly correlated valence electrons.

The mentioned above problems may have a better chance for a quicker and general solutions if they are tackled by teams composed of experts with complementary professional skills: mathematicians, physicists, chemists, programmers, engineers, etc. The workshop participants willingly shared their personal experience of participation in the past in such multi-expert groups. They also expressed the necessity of including additional advanced mathematical courses on non-mathematical university departments, especially if they are related with the material or computational science. In particular, the courses mentioned include complex calculus, functional analysis, mathematical statistics, differential and integral equations, operator calculus.

6 Summary

Based on our discussions we conclude that in order to correspond to the modern level of research on new materials, biology, and medicine the computational chemistry needs significant improvements. The most realistic way of success is to combine different approaches, like more efficient numerical methods and new science (e.g. AI/ML) or increase their efficiency on new hardware (e.g. GPUs). Mathematics plays a central role in this development as the whole field is focused on solving systems of integral and differential equations or their appropriate combination and correlation. Also, we believe that the basic mathematical background of students specializing in computational chemistry should be extended to catalyse the application and development of new methods. In addition to the traditionally studied mathematical fields, such as group theory or differential and integral equations, they need to be more familiarized in operators calculus, projection operator techniques, optimization, data analysis, etc.

To verify and expand the ideas discussed we have decided to apply for a 5-days BIRS workshop. In addition, we have expressed the intention to communicate the most interesting details of our discussions by publishing them in a peer-reviewed scientific journal.

References

[1] List of quantum chemistry and solid-state physics software
`handwiki.org/wiki/List_of_quantum_chemistry_and_solid-state_physics_software`.

[2] J.D. Hidary, *Quantum Computing: An Applied Approach*, 2^{nd} ed. Springer, Cham, 2021.

[3] M.S. Ramkarthik, P.D. Solanki, *Numerical Recipes in Quantum Information Theory and Quantum Computing: An Adventure in FORTRAN 90*, CRC Press, Boca Raton, 2022.

[4] D. Frenkel, B. Smith, *Understanding Molecular Simulation*, 2^{nd} ed. Academic Press, San Diego, 2002.

[5] J. Leszczynski (ed.), *Handbook of Quantum Chemistry*, 2^{nd} ed. Springer, Cham, 2017.