Usage Report for Fiscal Year 2024

Project Title:

Rigorous DFT across the periodic table

Name:
○Bun Chan (Nagasaki University), Takahito Nakajima (RIKEN)

Laboratory at RIKEN:
Center for Computational Science, Computational Molecular Science Research Team

１. Background and purpose of the project, relationship of the project with other projects

DFT has been the workhorse of computational chemistry. Yet, its full potential has not been reached after decades of development.

As the first principle pursue of the so-called "exact DFT" remains elusive, the continuous progress of DFT relies on statistical and machine-learning techniques. Indeed, this has become the norm in DFT research for some time. As with science and technology in other areas that rely on "training" with big data, the quality of the data plays an important role.

Regarding data set quality, current DFT methods are often formulated with a narrowly focused set of chemical species and molecular properties; major data sets include almost exclusively light p-block species, and they cover just 15% of the periodic table. As a result, the thus formulated DFT often suffer from "overfitting", i.e., they cannot adequately treat systems that do not belong to the types in the training data. Naturally, this would severely hinder the application of DFT.

In this project, we will devise strategies to improve the robustness of new DFT methods by diversifying the range of data used in the formulation of DFT. Specifically, we aim to expand the data sets to cover the periodic table.

２. Specific usage status of the system and calculation method

This project employs the Gaussian and Q-Chem programs on Hokusai BW, as well as a wide range of standard quantum chemistry software packages such as Molpro, MRCC, and ORCA. They enable us to access a diverse range of quantum chemistry methodologies, including highly accurate coupled-cluster and multi-reference methods with which reference data are obtained, and DFT methods with which insights into the fundamentals of a reliable DFT can be revealed.

Regarding the computational determination of reliable reference data, we have previously diversified from light p-block species to cover s-block elements, as well as transition metal species that are more challenging for theoretical methods than the systems that have been considered by us and others.

In addition, within p-block species, we note that most data sets focus on molecular species, with little attention given to nano-sized and bulk materials. A key reason is the excessive computational resources required for their calculation. As we have previously developed substantially lower-cost yet reliable methods, we apply them to larger nano-sized systems to complement existing molecular data.

With the new data sets, the assessment of DFT is trivial in comparison, and we have carried out such benchmarking in each case. Including in our assessments are a small but representative collection of DFTs that are orthogonal to one another in their formulations. Such distinct differences in their natures enables straightforward identification of key ingredients that are important for the different classes of chemical systems; this facilitates the future development of more reliable methods.

３. Result

To acquire reliable data for DFT development, it is of utmost importance that the reliability of the reference data can be verified in the first place. In our work, while we use the most advanced quantum chemistry methods that are computationally viable, this is by no means a guarantee of accuracy.

More generally, this has been a long-standing issue of computational chemistry. Numerous "diagnostics" have been previously proposed by many research groups to gauge the reliability of the quantum chemistry method. However, the diagnostics themselves are not fully reliable. In many cases, they are also being misused.

In one of our works within FY2024, we have illustrated how several most popular diagnostics are misused. Specifically, the A25 and %TAE metrics are designed with and for analyzing the accuracy of the high-level CCSD(T) method for small molecules. However, they are often applied to medium-sized and large molecules, including practically relevant systems such as nano-materials and proteins.

We have conducted several case studies using such systems to demonstrate that how these diagnostics should not be applied to these large systems with heterogeneous regions. Such systems often contain regions that can be accurately computed with CCSD(T), as well as some regions for with CCSD(T) may be inadequate.

The A25 and %TAE diagnostics leads to false negative in these situations, which in turn may lead to acceptance of unreliable reference data in the literature. To remedy this drawback, we have proposed a protocol of combining A25 and/or %TAE with another recently developed diagnostic (N_FOD), which shows the opposite behavior to A25 and %TAE, and thus provides a more balanced analysis.

While it is important to be aware of the quality of the reference data in DFT development, to broaden the scope of DFT also requires a large amount of data for a wide range of chemical systems and properties. The extreme efforts are required in the computation of high-quality data using CCSD(T) or even higher-level methods. Thus, the use of some low-quality data may be unavoidable.

A key question that arises from this reality is how the use of low-quality data affects the quality of the trained DFT. This issue is becoming increasingly important as the development of DFT increasingly uses machine-learning techniques with enormous data sets. While the principle of "garbage-in-garbage-out" is well appreciated, the actual effect has never been precisely quantified.

In another key project of FY2024, we have filled this knowledge gap. We have synthesized a collection of data sets with systematically varying degrees of quality; using these synthetic data, we have formulated a variety of DFT methods and measured their accuracy in the prediction of chemical properties for a diverse set of molecules.

We find that, encouragingly, for DFT methods with a sound physical foundation and a minimal number of trained parameters, the use of a relatively low-quality data set (with an average uncertainty of ~20 kJ mol^{-1}, 5 times larger than the chemical accuracy threshold) does not lead to a significantly worse trained method. In fact, if the use of low-quality data improves the chemical diversity, it benefits the generality of the resulting DFT.

In relation to this last point, in another work of FY2024, we have further established the "transferability principles" to gauge the degree of diversity of a data set. In additional, we have continued to expand our endeavor in producing high-quality data that have yet to be covered. This includes, for example, the new LiCT set of thermochemical data for metal clusters.

４. Conclusion

Our provision of high-quality data expands the scope of DFT development. Our new venture into the discovery of requirements for DFT development in the age of machine-learning provides a clearer direction for further development.

５. Schedule and prospect for the future

The development of a universal DFT has a long road ahead. We will further this by continue to provide independent reliable data and unravel the best development strategy in this machine-learning era.

Usage Report for Fiscal Year 2024

Fiscal Year 2024 List of Publications Resulting from the Use of the supercomputer

[Paper accepted by a journal]

1. Data Quality in the Fitting of Approximate Models: A Computational Chemistry Perspective. Chan, B.; Dawson, W.; Nakajima, T. *J. Chem. Theory Comput.* **2024**, *20*, 10468.

2. The Paradox of Global Multireference Diagnostics. Chan, B. *J. Phys. Chem. A* **2024**, *128*, 9829.

3. The Bond Energy of the Carbon Skeleton in Polyaromatic Halohydrocarbon Molecules. Chan, B.; Karton, A. *ChemPhysChem* **2024**, *25*, e202400234

4. Exploiting the Correlation between the 1s, 2s, and 2p Energies for the Prediction of Core-Level Binding Energies of Si, P, S, and Cl species. Hirao, K.; Nakajima, T.; Chan, B. *J. Phys. Chem. A* **2024**, *128*, 6879.

5. Identifying and Embedding Transferability in Data-Driven Representations of Chemical Space. Gould, T.; Chan, B.; Dale, S. G.; Vuckovic, S. *Chem. Sci.* **2024**, *15*, 11122.

6. Sorting Drug Conformers in Enzyme Active Sites: The XTB Way. Chan, B.; Dawson, W.; Nakajima, T. *Phys. Chem. Chem. Phys.* **2024**, *26*, 12610.

7. Limiting Factors in the Accuracy of DFT Calculation for Redox Potentials. Chan, B. *J. Comput. Chem.* **2024**, *45*, 1177.

8. The Verification of Delta SCF and Slater's Transition State Theory for the Calculation of Core Ionization Energy. Hirao, K.; Nakajima, T.; Chan, B.; Lee, H.-J. *J. Comput. Chem.* **2024**, *45*, 183.