Combined QM/MM, Machine Learning Path Integral Approach to Compute Free Energy Profiles and Kinetic Isotope Effects in RNA Cleavage Reactions

Review

Anna Borisova

TCCM
2023

I chose the paper "Combined QM/MM, Machine Learning Path Integral Approach to Compute Free Energy Profiles and Kinetic Isotope Effects in RNA Cleavage Reactions" by Timothy J. Giese, Jinzhe Zeng, Şölen Ekesan, and Darrin M. York for review.

The paper sets out to improve the accuracy and efficiency of QM/MM simulations of chemical reactions in solution. In particular, it aims to correct deficiencies in the commonly used DFTB semiempirical QM method when applied to model phosphoryl transfer reactions relevant to RNA cleavage. The authors develop a QM/MM method that uses the DFTB method as the QM model and adds a machine learning correction potential, termed DPRc. The DPRc potential provides corrections to the DFTB QM-QM and nearby QM-MM interactions.

There are a number of factors that make this task not the most trivial to solve. First and foremost is the size of the molecule involved in the reaction; the structures of biological molecules are too large for pure quantum chemical calculations. Considering that the goal is to study the free energy profile of a reaction in which isotopic and nuclear quantum effects are important, this does not allow the use of conventional molecular dynamics methods. Given the size of the systems studied as well as the computational expense of studying the energy profile of the reaction, the authors' choice to use the QM/MM method is quite obvious and appropriate. QM/MM methods are well-suited for studying chemical reactions in large biological systems, where a small reaction region needs to be modeled quantum mechanically and the surrounding environment classically. Using a semiempirical QM method makes the calculations feasible for extensive sampling and dynamics. However, methods like DFTB may not be sufficiently accurate, motivating the development of the ML correction. PBE0/6-31G* QM/MM umbrella sampling was performed to obtain reference free energy surfaces and to provide structures, energies, and forces to initiate the DFTB2 QM/MM+DPRc network parameter optimization.

It is quite clear that the quantum part is the most difficult for calculations, and the speed of calculations will depend on the degree of approximation of the chosen calculation method. The authors decided to use the second-order density-functional tight-binding method (DFTB2) as a fast, approximate base QM model. Using a semiempirical QM method makes the calculations feasible for extensive sampling and dynamics. However, methods like DFTB may not be sufficiently accurate, motivating the development of the ML correction. In this article, the researchers propose a deep potential range correction that allows, based on DFTB2 calculations, to obtain data comparable in accuracy to *ab initio* DFT (PBE0/6-31G*). The MM region uses the standard AMBER force field parameters with interfacing the DeePMD-kit. The TIP4P/Ew water model is utilized in the initial classical MD simulations. Long-range electrostatic interactions are modeled using the particle-mesh Ewald (PME) method, which enables periodic boundary conditions.

Of course, correction using machine learning methods is not the only method; there are also other methods (perturbative corrections like DFTB3), but the charm of ML methods lies in the low derivatives of computational complexity in applying an already finished model. The bottleneck of such an approach is the accessibility of data for training the model, as well as the problem of the applicability domain, which is the limitation of applying the model beyond the data on which it was trained. During training, an active learning method was used, allowing for the gradual expansion of the training data set. This helps to avoid overfitting on a limited set of structures.

Machine learning is undoubtedly a trendy solution in our time, but a skeptic may ask: Does it make sense to conduct semi-empirical calculations at all if you can simply build a machine learning model that directly predicts energies?

There are several reasons why it still makes sense to use a hybrid approach with semi-empirical and machine learning. Despite the fact that semi-empirical methods are still based on significant approximations, they are still based on the description of the electronic structure and physical principles. This provides interpretability and the possibility of extrapolation beyond the training set. A purely empirical model can make good predictions only within the training data. Another argument against this idea is that a significant set of data for training a purely empirical model still requires calculations, moreover, of a higher order than what is used in this work. In general, if you train a model from scratch, it will require much more costly and accurate data. And finally, the hybrid approach provides a good balance between accuracy, computational efficiency, and interpretability.

Once the network parameters for the DFTB2 QM/MM+DPRc were calibrated using Born-Oppenheimer umbrella sampling data, the authors utilized the optimized set of 4 parameters to evaluate Free Energy Surfaces (FESs) that account for nuclear quantum effects. This was achieved by conducting umbrella sampling integrated with Path Integral Molecular Dynamics (PIMD). The PIMD simulations were executed through a coupling of the i-PI software with a developmental iteration of the SANDER module from the AMBER suite.

The machine learning-enhanced version of the DFTB2 method (DFTB2 combined with DPRc) produces free energy profiles that are remarkably consistent with the high-precision ab initio PBE0/6-31G* calculations, with an average discrepancy of just 1 kcal/mol, illustrating the efficacy of the ML adjustments. The kinetic isotope effects (KIEs) predicted by DFTB2 alone show a slight divergence from the ab initio values, by about 1-2%, but when machine learning corrections are applied, the predictions are almost spot-on, with a variance of merely 0.2%. This serves as strong evidence of the QM/MM+ML model's precision. In terms of speed, the DFTB2 plus DPRc approach is 2 orders of magnitude faster than the ab initio QM/MM methodology, highlighting the significant efficiency gains from using semiempirical quantum mechanics. However, simply reweighting the DFTB2-derived sampling leads to inaccurate free energy profiles due to a lack of alignment with the ab initio findings, underscoring the necessity for ML enhancements instead of relying on reweighting methods alone.