Transforming Automated Quantum Chemistry Calculation Workflows *via* Machine Learning: Towards Accelerated Chemical Discovery with Higher Fidelity

Chenru Duan[1,2] and Heather J. Kulik[1]

[1] Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States

[2] Department of Chemistry, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States

Machine learning (ML) has begun to accelerate chemical discovery by providing advances in efficiency needed to overcome the combinatorial challenge of computational materials design. In ML-assisted chemical discovery, an automated quantum chemistry (QC) calculation workflow is used to generate datasets to train surrogate models, which serve as alternatives to rapidly explore a large chemical space to identify candidate materials. However, current automated QC calculation workflows with density functional theory (DFT) as workhorse leads to many attempted calculations that are doomed to fail and brings biases/inaccuracy to the training data that may be out of the domain of applicability of DFT. This includes many compelling functional materials and catalytic processes that are difficult because of their complex electronic structure, such as systems involving strained chemical bonds, open-shell radicals and diradicals, or metal–organic bonds to open-shell transition-metal centers. We address these challenges of computation efficiency and accuracy by integrating ML approaches into conventional DFT-based QC workflows. We build two types of classifiers to predict the likelihood of calculation success: 1. prior to calculations and 2. on-the-fly during calculations. The prior to calculations classifier is a near zero-cost model that rapidly filters out candidate calculations most likely to fail, while the on-the-fly model monitors and terminates an already running calculation if it is predicted to fail with high confidence. Together, these classifiers save half of the computation resources. We also develop multiple types of classifiers to predict the presence of strong correlation, which is usually a sign of a system being out of the domain of applicability of DFT. Our models only require calculations at DFT cost and can classify which systems in a dataset will require more expensive but accurate wavefunction theory calculations, leading to overall high fidelity of the entire dataset. Since electronic structure information is encoded as the inputs, our models are readily transferable to larger systems and systems with unseen elements. All these classifier models represent the first efforts toward autonomous workflows that move past the need for expert determination of the robustness of DFT-based materials discoveries.

Keywords: machine learning, automated computation workflow, chemical discovery