Quantum Chemistry Meets Machine Learning: Autonomous Computational Workflow for Chemical Discovery

Chenru Duan

Automation has long been driving the development of modern society since the first industrial revolution and has the potential to provide sufficient productivity forces for Great Harmony. Similar revolution is ongoing in the field of computational sciences, where quantum chemistry software and modern computers have developed to a stage where virtual high throughput screening (VHTS), i.e., running thousands of calculations in parallel, becomes possible. This provides great opportunities for developing automated workflows to utilize the increasing computing power to generate large-scale data sets. Together with machine learning (ML) models trained on these data sets as either surrogate function approximations or generative models, accelerated chemical discovery for functional molecules and materials may be achieved. Current automation workflows, however, are far from perfect. Namely, they produce too many unfruitful results and suffer severely from method selection bias, especially on challenging chemical spaces such as transition metal chemistry. These problems challenge the automated workflows for providing efficiency and accuracy needed for chemical discovery.

In this Thesis, we introduce intelligent ML decision-making models in automation workflows. We build the first set of classifiers to predict the likelihood of calculation success that on-the-fly monitors and terminates an already running calculation if it is predicted to fail with high confidence. These classifiers are extremely transferable and always stays accurate (i.e.,>95%) during the whole process of geometry optimization, saving more than half of the computation resources. We developed the first semi-supervised learning classifier to identify strong static correlation in a system, achieving state of the art for this classification task. Therefore, we can pre-determine which systems require more expensive (yet more accurate) correlated wavefunction theory calculations, thus improving overall data accuracy without adding unnecessary computational cost. We also proposed an approach that utilizes the consensus among multiple density functional approximations (DFAs) to discover robust (i.e., DFA-insensitive) candidate compounds. These compounds discovered based on the DFA consensus are in much better agreement with experimentally observed leads. Lastly, we built a DFA recommender that selects the DFA with the lowest expected error to the reference in a system-dependent manner, which achieves the accuracy needed for inorganic chemical discovery. All these ML-based decision-making models are integrated in workflows for VHTS. We anticipate these “smart” automated computational workflows are keys to autonomous chemical discovery.