Machine learning aided quantum chemical discovery in the solution phase

Fang Liu (fang.liu@emory.edu)

Modeling the solvent environment in quantum chemical calculations is of great importance to understanding chemistry under real-world experimental conditions. However, rapid computational data set generation of solution-phase molecular properties at the quantum mechanical level of theory was previously hampered by the computational challenges in the implicit and explicit solvent models. We combine quantum chemistry calculations and machine learning (ML) models to improve the efficiency and accuracy of solvent models.

Implicit solvent models are helpful for the quantum chemistry calculations of large molecules, but the computational overhead can still be significant. We exploit graphical processing units (GPUs) to accelerate both the electrostatic interaction integrals and the linear solver in the conductor-like polarizable continuum model (CPCM) and achieved 10X to 140X speedups for density functional theory (DFT) and time-dependent DFT. More recently, we focused on understanding the impacts of CPCM on the self-consistent field (SCF) convergence and DFT delocalization errors in large molecules. Although CPCM has been used as a simple remedy for DFT convergence issues in proteins, the mechanism, applicability, and consequences of using CPCM as an SCF accelerator are not thoroughly investigated. We found that CPCM's SCF acceleration effects are related to the selective stabilization/destabilization of molecular orbitals and are most effective for proteins with charge separations.

Explicit solvent models were rarely used for high throughput quantum chemistry calculations due to the required high degree of configuration sampling and the associated complicated set-up steps. We developed AutoSolvate, an open-source toolkit to streamline the workflow for QC calculation of explicitly solvated molecules, including solvated-structure generation, force field fitting, configuration sampling, and the final extraction of microsolvated cluster structures that QC packages can readily use to predict molecular properties of interest. Another major challenge in solution-phase computational discovery is the discrepancy between computationally predicted molecular properties and experimental measurements. Specifically, prominent errors persist in redox potential calculations compared to experimental measurements. We develop ML models to reduce the errors of redox potential calculations in both implicit and explicit solvent models. We compared and contrasted the performance of models built from the combination of various types of input features and ML methods and found the ML models effective in reducing the gap between computational and experimental redox potential values. The ML models also significantly reduced the sensitivity of the calculated results to DFT functional choice.