Multifidelity methods for design of transition metal complexes

Jon Paul Janet

2.00pm 12/18/2019 in 66-360

Technical Summary

Advisors: Heather Kulik, Department of Chemical Engineering
Youssef Marzouk, Department of Aeronautics and Astronautics

The rational design of materials with tightly controlled properties is crucial to addressing future challenges in energy, electronics and catalysis. While improvements in computing power have made simulation with density functional theory (DFT) an essential tool in screening new materials, it remains too costly to address explore truly high-dimensional design spaces. This problem is especially acute for open-shell transition metal (TM) complexes, which are of central importance in homogeneous catalysis and have applications in molecular electronics, sensors and energy generation and storage. TM complexes consist of metal centers coordinated to a ligand field, and manipulation of this ligand field can precisely tune the electronic properties of the metal. Unfortunately, the space of different metal-ligand combinations is combinatorially-large and poorly characterized (relative to organic chemistry). Further, DFT calculations for these systems are expensive and sensitive to method choice, making it impractical to simulate large numbers of candidates indiscriminately. This makes the search for TM complexes with desired properties a formidable challenge.

This thesis addresses these challenges by formulating algorithmic strategies for materials design that exploit insights from data-driven surrogate models together with first-principles simulations. A framework for data-driven inference of the quantum properties of TM complexes is developed, using artificial neural networks (ANNs) to estimate the quantum mechanical properties of unseen TM complexes at similar accuracy to the baseline uncertainty in DFT calculations, at negligible cost.

A new family of graph-based numerical representations for transition metal complexes is developed that is capable of describing a full range of metal-local and global features while retaining chemical interpretability. In addition to improving predictive inference of trained surrogate models based on purely 2D information, the interpretability of these features allows for extraction of chemical insight from the thousands of DFT evaluations used for model training. Feature selection techniques allow low-dimensional feature sets that show good performance for specific prediction targets (e.g. spin state ordering or redox potential) to be identified, and analysis of the metal-local to metal-distal character of these feature sets provides insights into the relationship between ligand field and properties of the metal center. For example, spin state ordering appears to be strongly controlled by the immediate ligand environment (the first coordination shell) while the redox potential is more sensitive to distal ligand modifications.

The capacity of trained models to extrapolate to new, dissimilar chemistries is thoroughly investigated by assessing prediction accuracy for diverse experimental com-

plexes from the Cambridge Structural Database (CSD). Performance is found to be highly variable, with many well-predicted complexes and a few large errors. This necessitates measures of model confidence, but surrogate model predictive ability is weakly correlated with organic chemical similarity metrics (i.e. Tanimoto distances). We investigate different ways to quantify the extent of extrapolation in chemical space, and determine that model errors can be well-predicted based on extrapolation distance to training data in simple, curated feature spaces.
However, errors are less correlated with extrapolation distances in high-dimensional feature spaces, and therefore chemical extrapolation in the latent space of learned ANN models is proposed as an alternative. This is shown to provide a better qualitative description of out-of-distribution model confidence compared with feature space distances, ensemble averaging and dropout-based standard deviations. To provide an estimate of uncertainty in relevant units, a simple probabilistic error model based on latent distance extrapolation is formulated and calibrated with a small amount of out-of-sample data, giving good quantitative error bounds on both inorganic and organic datasets, as well as good results when used for active learning.
This provides a metric for extrapolative uncertainty, but uncertainty with respect to DFT functional choice is another serious issue for simulation of open-shell transition metal complexes. This is addressed by training surrogate models on data sampled from DFT calculations with different levels of exact exchange. This provides predictions of the unique, system-specific functional sensitivity of transition metal complexes, capturing variations in simulation reliability across chemical space. The developed approach is also adapted to address the difficulty in initializing new simulations of unknown metal-ligand combinations in a spin and oxidation state dependent manner by predicting DFT equilibrium metal-ligand bond lengths. This capacity is incorporated into the open source molSimplify toolkit, which combines these distance predictions with force field calculations on organic bonds to construct initial geometries. This enables future simulations to benefit from high quality initial geometries that were previously only available for organic systems.
The introduced framework is demonstrated in two application areas. Firstly, novel spin crossover (SCO) materials are designed from a space of thousands of candidates, exploiting the surrogate ANN and a newly-introduced genetic algorithm (GA) that balances both model uncertainty (as captured by extrapolation distance from training data) and property optimization. The modified GA generates hundreds of leads in a fraction of the time required for first-principles screening, 60% of which are validated to be SCOs at a DFT level. Finally, multiobjective probabilistic optimization and active learning are used to identity redox couples that balance solubility and redox potential from a design space of approximately three million candidates. A combinatorial strategy is used to construct a diverse and densely-sampled design space and a combination of multitask ANNs and 2D expected improved is used to identify and iteratively refine a Pareto frontier of candidate complexes using a few hundred DFT simulations that are substantially enriched relative to random sampling, providing at least a 500-fold increase in efficiency. The utility of this surrogate-assisted approach is evident from the orders-of-magnitude accelerations obtained over screening purely with DFT, and such strategies open the door for *in silico* design of some of the most challenging molecular systems at a far greater scale than ever before.