Machine-Learning-Enabled Multiscale Modeling for Catalysis and Engineering: from Data, Algorithms to Software Applications

by Yifan Wang
Advisor: Dr. Dionisios G. Vlachos
Committee Members: Dr. Anthony Beris, Dr. Marianthi Ierapetritou, Dr. Sunita Chandrasekaran

---

ABSTRACT

Catalytic processes produce most of the chemicals today and will continue to provide energy and materials for future generations. Advances in catalyst discovery and efficient industrial processes could mitigate climate change and increase sustainable energy supply. Experimental techniques, such as spectroscopy and microscopy, are often used to characterize catalysts. However, their spatial and temporal resolutions make direct observations under working conditions challenging. Computational tools can complement and potentially extend the experiments. Multiscale modeling simulates the physical and chemical phenomena at multiple time and length scales and provides a first-principles-based understanding of complex catalytic systems. Despite the recent surge in computation power, models are still computationally prohibitive when extensive evaluations are required.

One practical approach is to approximate expensive models with surrogate models. Machine learning can supply efficient surrogates, identify nonlinear correlations, and provide physical insights. In addition, quantitative structure-property relations, which map catalytic structures to performance, can allow further exploration and enable the inverse design. Active learning could accelerate the search for the optimal conditions or materials in the design space. Throughout this thesis, we develop machine-learning-enabled multiscale modeling frameworks for catalysis and engineering systems. The workflow involves high-quality data generation from the first-principles or experiments, efficient algorithm design, and open-source software development. We demonstrate our methodology to subnanometer supported catalysts and biomass utilization.

First, we develop a multiscale modeling framework for supported single atom and subnanometer cluster catalysts. The framework integrates a comprehensive toolset including density functional theory (DFT) calculations performed by collaborators, genetic algorithm-based structure optimization, machine learning, equilibrium-based Metropolis Monte Carlo, and

kinetic Monte Carlo (KMC) simulations. We choose Pd single atoms and subnanometer clusters of a few atoms (size, n = 1-55) on CeO_2(111) in a CO atmosphere as a case study. We first investigate the structures of Pd_n clusters and CO adsorption energies on various sites using DFT. DFT supplies high-quality first-principles data to train machine learning Hamiltonians, which represent efficient structure-to-energy mappings. Combined with the Hamiltonians, Monte-Carlo-based structure optimization algorithms, such as a cluster genetic algorithm, determine low energy structures. Active learning improves the model accuracy by passing the predicted structures to DFT and using the structure-energy DFT data to train the Hamiltonians iteratively. KMC simulations track the structure evolution of the catalysts against the real-time and predict the time scales of elementary events under the working conditions. The framework elucidates the stability, structures, and dynamics of supported metal clusters that are bare or exposed to an adsorbate used for characterization, e.g., CO in infrared spectroscopy. The methodology can be applied to any metal/support system.

Second, we create frameworks and software tools to facilitate experimental design and interpret experimental data for biomass utilization. Active learning algorithms, such as Bayesian optimization, can be used to gain well-informed decisions on what computations or experiments to run to reduce time or materials. We test the framework on various kinetic models and experiments. One example showcases that the surrogate model accurately describes the original microkinetic model. Bayesian optimization locates the maximum 5-hydroxymethyl furfural (HMF) yield in fructose conversion to HMF, a platform chemical to many valuable bio-products. We also develop a multiscale modeling framework to generate feasible lignin structures that match experimental data for various lignin feedstocks. The structures are encoded in both SMILES strings and molecular graphs, allowing fast computation and visualization. The structure libraries generated can enable future kinetics modeling and close the gap between model predictions and experimental observables.