Accelerating Autonomous Molecular Discovery through Automated Quantum Chemistry and Artificial Intelligence

by

Haoyang (Oscar) Wu 吴浩洋

Submitted to the
Department of Chemical Engineering
and
MIT Center for Computational Science and Engineering
on September 8, 2025 in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY IN CHEMICAL ENGINEERING AND COMPUTATION

ABSTRACT

The grand challenges in medicine, energy, and materials science are fundamentally molecular discovery problems. However, the vastness of chemical space renders traditional experimental exploration inefficient and insufficient. Autonomous Molecular Discovery promises to accelerate this process by integrating artificial intelligence (AI), computation, and automation in chemistry, but it faces a critical trilemma: balancing accuracy, speed, and scalability. This thesis documents a systematic effort to alleviate this tension by developing and integrating novel computational frameworks that synergize the first-principles rigor of quantum mechanics (QM) with the predictive efficiency of machine learning (ML) and the scalable automation enabled by AI.

This thesis began by focusing on developing the first *ab initio* kinetic models for the liquid-phase oxidative degradation of Active Pharmaceutical Ingredients. This demonstrates the feasibility and predictive power of automated mechanistic modeling in complex chemical environments, and highlights the acute need for more accurate thermochemical and kinetic data to handle real-world complexity. To address this, we developed a framework for computing systematic thermochemical corrections, and conducted an extensive benchmark of 284 model chemistries, establishing protocols to efficiently achieve chemical accuracy ($\sim$1 kcal/mol) from QM simulations. Recognizing the limitations of speed and data scarcity, we engineered physics-informed ML architectures, notably the QM-GNN, which fuses Graph Neural Networks (GNN) with QM descriptors. This approach significantly improves predictive performance and data efficiency, particularly for reaction regioselectivity in low-data regimes. Finally, to deploy these advances at scale, we designed QuantumPioneer, an automated, high-throughput platform for generating large-scale, high-fidelity QM thermo-kinetic datasets. This platform has produced an extensive database for oxidation reactions, enabling the development of novel ML models for predicting molecular stability and solvation energies. Collectively, this thesis provides a cohesive framework for accelerating molecular discovery, demonstrating that the strategic integration of first-principles simulation and data-driven intelligence can overcome key bottlenecks hindering autonomous chemical design and discovery.

Thesis supervisor: William H. Green, Ph.D.
Title: Hoyt C. Hottel Professor of Chemical Engineering