Thesis Technical Summary

Advancing Artificial Intelligence for Efficient and Synthesizable In-silico Molecular Design

Author: Wenhao Gao
Supervisor: Connor W. Coley

Small organic molecules possess an astronomical number of structural possibilities and a wide range of functionalities, holding immense potential to provide material-level solutions to critical societal challenges such as health and the environment. However, the discovery of molecules with functionalities tailored to specific applications remains a challenging, time-consuming, and resource-intensive process, often relying on trial-and-error experimentation. Recent advances in computational techniques—particularly in artificial intelligence—offer promising solutions to this inefficiency. These developments are paving the way toward a more systematic and efficient approach to molecular discovery, enabling the design of novel functional molecules tailored to specific needs and accelerating the development of solutions to urgent issues in health, sustainability, and energy.

This thesis presents algorithmic advances in artificial intelligence, particularly deep learning, for de novo molecular discovery, framed as a black-box optimization problem with a focus on small organic molecules. The contributions span three core aspects:

- The first section focuses on improving the **sample efficiency** of molecular optimization. A central capability of any molecular design algorithm is to determine which direction to explore next within chemical space in order to identify molecules with more optimal properties, given a limited set of known examples. Due to the inherent trade-off between computational efficiency and predictive accuracy in modeling methods, it is crucial to evaluate as few candidate molecules as possible to identify the optimal structure. This section introduces the problem formulation and benchmarking efforts for sample-efficient molecular optimization, followed by several approaches aimed at enhancing efficiency.

- The second section addresses the challenge of ensuring **synthetic accessibility** during molecular design. For small organic molecules with non-trivial syntheses, any design that cannot be realized in the lab has limited practical value. This presents a unique constraint in small molecule design that often renders direct adoption of algorithms developed for language or vision tasks ineffective. After framing the problem, this section introduces a generative modeling framework that integrates synthesis and design, ensuring that the search is constrained to synthesizable chemical space. It further introduces the concept of "generative molecular projection" and demonstrates its application in balancing sample efficiency and synthetic feasibility.

- The third section targets the improvement of **oracle accuracy** for molecular discovery. Achieving both accurate and efficient prediction of molecular properties has long been a central goal in computational chemistry. While deep learning has shown promise in breaking the traditional trade-off between accuracy and efficiency by leveraging large-scale historical data, its full potential—especially for directly learning

experimentally measured bioactivities under data-scarce conditions—has yet to be realized. This section presents a benchmarking effort on applying deep learning to therapeutic-related property prediction, and introduces substrate scope contrastive learning as a strategy to learn reactivity-related patterns from published reaction datasets.

Together, these three components present a systematic, data-driven methodology for small organic molecule discovery that minimizes the need for extensive domain expertise. The algorithms developed in this thesis are designed to support autonomous workflows, potentially enabling closed-loop molecular discovery that maximizes efficiency and reduces both cost and reliance on human intuition. While the demonstrations in this thesis primarily target pharmaceutical applications, the methods are task-agnostic and can be readily extended to broader material discovery efforts.