Machine Learning for Chemical Reactivity Prediction: Paradigms, Challenges, and Applications

by

Priyanka Raghavan

The discovery of new therapeutic agents in the pharmaceutical industry is a complex, iterative process, often encapsulated by the Design-Make-Test-Analyze (DMTA) cycle, in which chemists ideate, synthesize, and assay compound targets of interest. A significant bottleneck in this cycle is the "Make" phase, where the synthesis of novel compounds can be time-consuming, resource-intensive, and fraught with unpredictable outcomes. Accurate prediction of chemical reactivity, particularly reaction yields and selectivities, is therefore paramount to accelerating drug discovery by enabling more efficient synthesis planning, reducing material waste, and guiding the design of more synthetically accessible molecules. As such, this dissertation explores the application of machine learning (ML) to address critical challenges in chemical reactivity prediction, with a particular focus on low-data regimes and the integration of predictive models into practical drug discovery workflows.

This thesis begins by addressing the pervasive challenge of predicting reaction yields from sparse, literature-derived data. It details the assembly of a large dataset of substrate scopes and evaluates single-task and multi-task ML approaches, highlighting the limitations imposed by data scarcity and noise in real-world chemical literature. Recognizing these challenges, this thesis then provides recommendations for designing experimental datasets that are more conducive to robust machine learning, specifically offering considerations for curating data with the downstream modeling goal in mind.

Building on these insights, this thesis then turns toward specific applications of machine learning in medicinal chemistry, first presenting a direct, impactful implementation of ML to enhance synthetic accessibility in drug design by predicting Suzuki cross-coupling yields from a large, historical pharmaceutical library dataset. ML models are shown to often outperform expert intuition and be successfully integrated into existing workflows for library design and rescue, significantly increasing synthesis efficiency. Finally, this thesis expands from chemical reactions to enzymatic reactions, detailing a computational and ML-based workflow for transaminase enzyme selection, to streamline the enantioselective synthesis of valuable chiral amine building blocks used in medicinal chemistry.

Collectively, this thesis contributes to the growing field of machine learning in chemistry by addressing fundamental challenges in reactivity prediction, particularly in low-data and real-world industrial settings. It provides novel modeling paradigms for existing data and insights into the limitations of current approaches, offers a conceptual framework for improved data generation, and demonstrates the tangible benefits of integrating ML models into the DMTA pipeline. Throughout, the critical interplay between data quality, molecular representation, and model architecture and evaluation is emphasized, paving the way for more reliable and impactful predictive tools that can accelerate the pace of chemical discovery.

Thesis supervisor: Connor W. Coley
Title: Class of 1957 Career Development Professor