Geometric representation learning for chemical property prediction, structure elucidation, and molecular design

Keir Adams

Molecular representation learning has revolutionized computer-aided chemistry by enabling the automatic extraction of arbitrarily complex patterns from datasets of (potentially labeled) molecular structures via deep neural networks. In predictive chemistry, deep learning is increasingly being used to replace expensive physics-based simulations and even experimental measurements of chemical properties. In generative chemistry, deep generative models are powering molecular design and optimization campaigns across chemical industries, particularly drug discovery and functional materials design. Notably, this paradigm shift has been driven by the development of sophisticated representation learning algorithms over the past decade that encode and decode molecular structures with increasing geometric detail – from minimal SMILES strings to elaborate atomistic 3D structures. Yet, many aspects of molecular structure remain neglected by leading geometric representation learning models. Accordingly, this thesis advances the geometric representation learning of molecular structure to create new opportunities in chemical property prediction, structure elucidation, and molecular design.

This thesis begins by highlighting surprising failure modes of graph neural networks when predicting properties dependent on chirality and conformational isomerism. A new stereochemistry-tailored model is then developed to imbue achiral graph neural networks with tetrahedral chiral expressivity while evading the pitfalls plaguing preceding 2D and 3D graph networks. This thesis then examines how the geometric quality of structures encoded by 3D networks impacts their accuracy in property prediction tasks requiring the model to reason about conformational flexibility.

Neglecting certain structural characteristics of molecules that pose difficult or expensive to model is also common in computational chemistry. In nuclear magnetic resonance (NMR) prediction, for instance, quantum chemical calculations typically estimate magnetic shieldings from stationary gas-phase geometries – ignoring vibrational effects and explicit solvent. To advance NMR-based structure elucidation, this thesis next develops neural surrogates for magnetic shielding calculations that, when integrated with molecular dynamics simulations, provide access to unprecedented accuracy in solvent-sensitive NMR spectra prediction.

Finally, this thesis advances *de novo* molecular design by explicitly representing 3D shapes, electrostatics, and non-covalent interactions in deep generative models for small molecules. A shape-conditioned variational autoencoder is first developed to enable the design of chemically diverse molecules that can adopt desired conformational shapes, like ligand binding poses. This strategy is then generalized into a powerful interaction-aware diffusion modeling framework to comprehensively enable bioisosteric replacement in ligand-based drug design.

Overall, this thesis newly designs, applies, and critically analyzes multiple geometric representation learning algorithms to improve the machine learning modeling of molecular chirality, conformational flexibility, vibrational/solvent effects, shape, and intermolecular interactions. Although further advances in our ability to learn powerful and generalizable

representations of all aspects of 3D molecular structure are still sorely needed, I expect that the tools and techniques developed in this thesis will both directly and indirectly contribute to innovative new approaches across computer-aided chemical design and discovery.

Thesis Supervisor: Connor W. Coley
Title: Associate Professor (Without Tenure); Class of 1957 Career Development Professor