Machine Learning Strategies for Automatically Identifying and Generating Molecular Structures

Tianfan Jin

Abstract:

Chemistry has been a major beneficiary of machine learning (ML) methods. In chemistry, all ML approaches can be roughly categorized into two types: forward prediction model and inverse prediction model. Forward prediction models take in the information of molecular graph, and use this information to predict structure-related targets such as characterization results or molecular properties. On the other hand, inverse prediction paradigms, take in the information that is relevant to the molecular structure, aiming to reconstruct the molecules based on the input information. My phd work mainly focuses on the latter problem of inverse prediction, and our target is to build ML architecture capable of: (1) automating compound identification given spectral data, and (2) generating satisfied molecular structures given required properties.

Manual chemical structure identification based on spectral data sources remains a time-consuming process in traditional chemical workflows. Although this problem seems susceptible to ML, limited training data and the absence of model architectures suitable for ingesting spectral data from multiple sources has led to limited progress. My phd work tackled this problem by developing transformer-based models that used self and cross-attention mechanisms to compress and integrate the information from 1H-NMR, IR, and EI-MS spectra to predict the chemical structure of unknown analytes. The spectra to structure (StS) models were trained and tested on newly generated spectra for 957,856 distinct organic

CHONSSePFClBrISiB-containing species drawn from the synthetic literature and Pubchem database. Top-1 and top-10 accuracies of 51.2% and 71.1%, respectively were obtained for structure prediction on testing data. The transferability of the StS models were also tested by providing incomplete or contradictory information, testing on structures with experimental spectral references rather than simulated spectra. Near identical performance is achieved in these scenarios illustrating useful domain transferability for this problem.

Though the StS models above displayed satisfied overall accuracy, they are inherently limited by the insufficiency of information even with the combination of three spectral sources. Additional information regarding reaction reactants was introduced when extending StS models to automatically analyze the reaction outcomes, where a new deductive framework was build to predict reaction target using both the information from reactants and characterization results of the products. Compared to the traditional reaction prediction models where the only input is the reactants information, the resulting reaction deduction models could distinguish between intended and unintended reaction outcomes and identify starting material based on a mixture of spectral sources. The deduction models also performed well on tasks that they were not directly trained on, like predicting minor products from named organic chemistry reactions, identifying reagents and isomers as plausible impurities, and handling missing or conflicting information.

Apart from compound identification, inverse molecular design, or automatic molecular generation is also foreseen to be valuable in real scenarios. Generative models for the inverse design of molecules with particular properties have been heavily hyped but have yet to demonstrate significant gains over machine learning augmented expert intuition. A major challenge of such models is their limited

accuracy in predicting molecules with targeted properties in the data scarce regime, which is the regime typical of the prized outliers that inverse models are hoped to discover. For example, activity data for a drug target or stability data for a material may only number in the tens to hundreds of samples, which is insufficient to learn an accurate and reasonably general property-to-structure inverse mapping from scratch. My thesis hypothesized that the property to structure mapping becomes unique when a sufficient number of properties are supplied to the models during training. This hypothesis has several important corollaries if true. It would imply that data scarce properties can be completely determined by a set of more accessible molecular properties. It would also imply that a generative model trained on multiple properties would exhibit an accuracy phase transition after achieving a sufficient size—a process analogous to what has been observed in the context of large language models. To interrogate these behaviors, I have built the first transformers trained on the property to molecular graph task, which this work dub “large property models” (LPMs). A key ingredient is supplementing these models during training with relatively basic but abundant chemical property data. The proof-of-concept study on LPM is based on ~1M molecules sampled from Pubchem database with over 40% of test cases that the generated molecules successfully reproduce all input properties. (within 10% of error range)