Enantioselectivity, the degree to which one enantiomer of a chiral product (R or S product) is preferentially produced in a chemical reaction, is an important factor in organic chemistry research, because sometimes only one enantiomer is biologically active or synthetically valuable. In enantioselective reactions, the ratio of R and S products is usually determined by the structure of the catalyst, so that the mechanism to form one enantiomer is favored. Being able to predict such selectivity is very helpful for the discovery of new reactions.

In this project, we would like to use GNN, a model naturally suitable for molecules, to predict the enantioselectivity, given a dataset of a specific type of reaction:

形状

中度可信度描述已自动生成

In this reactions, there are three molecules that are variables: the two on the left-hand side of the arrow, and the catalyst. The dataset is collected by Scott Denmark group,1 who used traditional ML like simple MLP, ridge regression, or random forest to solve the prediction problem. The experimental results for enantioselectivity were rigorously collected for ~1000 reactions (5 \* 5 \* 40, by varying the Ar group, R group, and catalyst structures themselves).

Given the suitability of GNN for molecule problem, we would like to formulate the prediction as a graph-level regression task, and try to predict enantioselectivity of the reaction by training a deep GNN.

We would also like to add an unsupervised learning part by trying to generate new catalyst structures, and feed them into the GNN we trained, so that we could search or filter for new molecules that can desirably catalyze the reaction we want.

Reference:

1. Zahrt, Andrew F., et al. "Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning." Science 363.6424 (2019): eaau5631.