10. C51 Project

Group members: Priyanka Raghavan, Jenna Fromer, Nathan Stover

Modeling Objective: The purpose of this project is to build a quantitative structure-property relationship (QSPR) model to (1) predict the retention time of small molecules in high-performance liquid chromatography (HPLC) and (2) predict the probability that the retention times of two or more molecules will overlap. HPLC is a commonly used purification method to separate a desired product from additional components in a mixture, and thus chromatographic separation is essential in obtaining pure reaction intermediates and products. The retention times of different mixture components ultimately determine the potential separability of the components and the purity of the desired product.

When performing reactions with little literature precedent, it is often necessary to have an *a priori* sense of the retention times of the starting material(s), intermediate(s), and product(s), as overlap of the chemical species in the liquid chromatography column can lead to poor purification. In particular, as autonomous chemical synthesis has been a recent topic of interest, such a model would aid in enabling fully autonomous synthesis, by allowing for prediction of possible peak overlap ahead of purification. Column parameters can then be adjusted as necessary to avoid this overlap. In general, a model to accurately predict HPLC retention times could reduce the cost, time, and resources necessary to synthesize small molecules.

Past work in this area has focused on predicting small molecule retention times using fingerprint-based deep learning methods as well as simple graph neural networks[1,2,3]. Recent benchmarks have shown that graph convolutional neural networks (GCNs) perform best on predicting small molecule retention times[3]. We are proposing to use a message-passing neural network (MPNN), which will use molecular graph representations, to make substrate-specific retention time predictions. MPNNs have been successfully used in QSPR chemistry tasks in the past[4]. We anticipate that an MPNN may take into account the connectivity of molecules, as well as their relation to relevant structurally-based properties (in this case retention times), more accurately than previous approaches.

The prediction of retention times will be our first and primary focus area. If we can develop a well-performing model for this task, we will subsequently attempt to incorporate the prediction of retention time overlap, which would require us to incorporate uncertainty quantification into our model.

Data: We intend to use the METLIN database to extract retention time data for our project[5]. The dataset consists of 80,000 small molecules and their associated retention times on a reverse-phase (RP) ultra-HPLC (UHPLC) column, very similar to the standard column used by many synthetic chemists. Most of the molecules are organic or organic-derived compounds, and mol files/ECFP descriptors are included, which will allow for structural comparison and potential scaffold splitting during testing.

Professor: We think Professor Coley would be the best fit to supervise our project.

References:

1. Yang, Q.; Ji, H.; Lu, H.; Zhang, Z. Prediction of Liquid Chromatographic Retention Time with Graph Neural Networks to Assist in Small Molecule Identification. *Anal. Chem.* **2021**, *93* (4), 2200–2206. https://doi.org/10.1021/acs.analchem.0c04071.

2. Fedorova, E. S.; Matyushin, D. D.; Plyushchenko, I. V.; Stavrianidi, A. N.; Buryak, A. K. Deep Learning for Retention Time Prediction in Reversed-Phase Liquid Chromatography. *Journal of Chromatography A* **2022**, *1664*, 462792. https://doi.org/10.1016/j.chroma.2021.462792.

3. Kensert, A.; Bouwmeester, R.; Efthymiadis, K.; Van Broeck, P.; Desmet, G.; Cabooter, D. Graph Convolutional Networks for Improved Prediction and Interpretability of Chromatographic Retention Data. *Anal. Chem.* **2021**, *93* (47), 15633–15641. https://doi.org/10.1021/acs.analchem.1c02988.

4. Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K.; Barzilay, R. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59* (8), 3370-3388. https://doi.org/10.1021/acs.jcim.9b00237.

5. Domingo-Almenara, X.; Guijas, C.; Billings, E.; Montenegro-Burke, J. R.; Uritboonthai, W.; Aisporna, A. E.; Chen, E.; Benton, H. P.; Siuzdak, G. The METLIN Small Molecule Dataset for Machine Learning-Based Retention Time Prediction. *Nat Commun* **2019**, *10* (1), 5811. https://doi.org/10.1038/s41467-019-13680-7.