Uncertainty Quantification for Small Molecule Retention Times Using METLIN SMRT Dataset

Priyanka Raghavan, Jenna Fromer, Nathan Stover

10.C51 Group Project

Scientific/Engineering Goal:

Our scientific goal is to predict the retention times of small molecules with valid uncertainty quantification. In order to achieve this goal, we intend to build a quantitative structure-property relationship (QSPR) model to (1) predict the retention time of small molecules in high-performance liquid chromatography (HPLC) with uncertainty and (2) predict the probability that the retention times of two or more molecules will overlap. When performing reactions with little literature precedent, it is often necessary to have a 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. For example, if chemists can predict the retention times of various reagents and their desired products, they may choose to run the reaction using reagents which minimize the chance of overlapping retention times with the desired product. This optimization problem can only be solved with an accurate QSPR model to predict retention times and, importantly, retention time uncertainty. We anticipate that a model that can predict small molecule retention time and uncertainty will accelerate drug discovery in both non-automated small molecule synthesis and in high-throughput autonomous chemical synthesis.

Background:

The METLIN small molecule retention time (SMRT) dataset was first introduced by Domingo-Almenara et al in 2019[1]. This publication also trained a deep neural network on fingerprints and compared its performance to that of a random forest regression model, k nearest-neighbors regression, and NNs trained on various molecular descriptors. Ultimately, they found that the model trained with molecular fingerprints, with relative error slightly above 5%, outperformed other approaches.

Yang et al[2] also attempted to predict retention times using this database and a graph neural network. A mean relative error under 5% was achieved with the GNN, outperforming their own benchmarks on other model structures. While this is similar to our intended architecture, the utilized GNN randomly sampled subgraphs of each molecule instead of using message-passing on the entire molecular graph.

An additional attempt by Kensert et al[3] demonstrated the effectiveness of a graph convolutional network in RT prediction, while benchmarking against other typical models. The best-performing GCN had relative errors below 5%, comparable to that observed by Yang et al.

Some of these papers also test model performance on other datasets. However, we will focus our project on the METLIN SMRT dataset only and have therefore focused on prior work with this data.

Details:

Because the METLIN dataset consists of molecule - retention time pairs, a supervised model is best to take advantage of the structure of the data we are given and make predictions of retention time. Inputs to our model take the form of molecules, and choosing how to represent these molecules is a key challenge for our project. On a high level, we plan on using a message passing neural network, which is consistent with the general trend of the field away from fingerprints and toward learned representations for molecules, especially for datasets of over tens of thousands of molecules. This requires that molecules must be translated from SMILES strings to graph data structures in order to perform the operations of a GNN. This proposition requires many granular decisions about how best to represent molecules in a graph form for our work. Our work will require validating what kinds of node and edge representations work best for our model.

It may appear that this problem has already been solved; impressively low mean relative errors have been achieved, and benchmarks against other models have been evaluated[1-4]. However, our project can be distinguished from previous work in two ways: (1) we intend to perform uncertainty analysis on our estimations. Multiple approaches to uncertainty quantification in machine learning have been reported in the literature[4,5], and we hope to benchmark a few different types of uncertainty quantification in our model. Moreover, we will use estimated uncertainties to predict the probability that two or more molecules will have retention times which significantly overlap. (2) We intend to first use a random split, and then evaluate the effects of a scaffold vs. random split on model performance. Ultimately, the type of split utilized should depend on the application. For example, if a chemist is attempting to predict RTs for a totally new class of molecules, a scaffold split is more appropriate; however, if a chemist synthesizes molecules very similar to others well-represented in literature, a random split may be appropriate. We are curious about the effect the type of split will have on performance.

Data:

The available data[1] consists of 80,038 small molecules and is currently broken down into 3 main files, the data from which can easily be read into Pandas dataframes in Python. The first contains PubChem ID numbers, Inchi keys, and retention times (given in seconds) for each molecule. The second contains several types of descriptors, including molecular properties, functional group counts, and ring/charge descriptors. These can potentially aid in determining any scaffold splitting when training and testing our models. The third file contains 1048-bit extended-connectivity fingerprints (ECFPs) for each molecule.

Necessary pre-processing steps will first include removing any duplicate molecules from the dataset. With regards to the input variable, we will need SMILES strings for all molecules, which we can obtain from the PubChem ID number or Inchi key. Additionally, we will likely evaluate a feedforward neural network baseline using Morgan fingerprints (2048 or 4096-bit), which we can generate from the SMILES strings using RDKit. The graph representations of the molecules can also be built in a similar manner through RDKit. Our desired output variable, retention time, will likely have a large spread and value and will thus likely need to be scaled and normalized. Finally, we have yet to definitively decide which type of data splitting we will use; for our first level of predictions, we would like to evaluate how well the model performs on general small molecules representative of the dataset - thus, we can perform a simple random split. However, we may also be interested in scaffold splits to determine the model's predictive capability on "unseen" molecules.

Thus far, we have determined that there are no duplicates in the provided data, and are able to generate SMILES strings and fingerprints for each molecule. We have also plotted a histogram of the retention times (Fig. 1). From the large spread of this data, it is evident that scaling and normalization will be necessary prior to model training and evaluation.

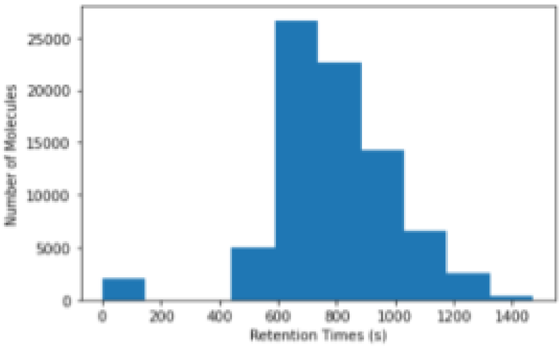

Figure 1: Histogram of retention times from the METLIN database. The majority of molecules fall between 400-1400 seconds, with about 2.5% of molecules being non-retained.

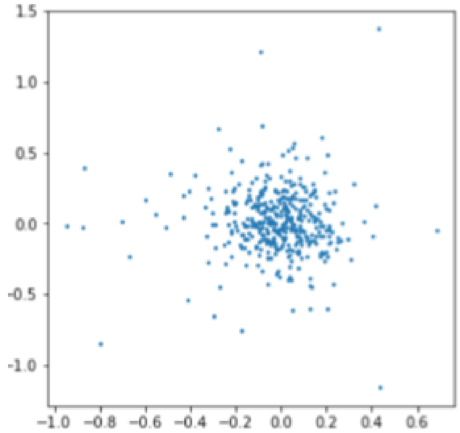

Figure 2: The first two components of 400-component PCA performed on the 2048-bit Morgan fingerprints. A total variance ratio calculation shows that these 400 components account for 80.9% of the overall dataset variance.

In an attempt to further characterize the dataset, we additionally performed PCA using 2048-bit Morgan fingerprints generated from the SMILES. Note that 81 molecules' SMILES strings were not able to be rendered through RDKit, and they have been removed for this and further analyses. This analysis may be helpful for potential dataset splitting based on type of molecule/scaffold.

Additionally, an important note is that this data contains retention time measurements for one specific type of column: a standard, reverse-phase column with water, acetonitrile, and formic acid solvents. Therefore, while this data (and any models trained using this data and their predictions) will not be directly applicable to other columns in practice, this still represents a very standard laboratory column and our work can thus be useful in many experimental applications.

Challenges: The first challenge we expect to face is benchmarking against various models and comparing our performance to other models in the literature. Implementing machine learning models often requires choosing many hyperparameters, such as learning rate, batch size, number of parameters, etc. Some of these hyperparameter choices are arbitrary, as optimizing all of them simultaneously would be extremely computationally expensive. Yet, model performance can certainly be affected by hyperparameters, and some of these values are not included in publications. Therefore, achieving the low relative errors seen by others may prove to be a more challenging task than expected.

Second, we plan to quantify the uncertainty in our RT predictions. We will need to claim how well our model quantifies uncertainty. However, we do not have a test set with given uncertainty, so we cannot simply evaluate our uncertainty quantification against a test set. Instead, we will need to explore some metrics that measure the performance of our uncertainty quantification. Deciding on an appropriate metric for this problem may be a challenge, and we will see what metrics have been proposed in the literature to assess uncertainty quantification performance.

We intend to predict the probability that one or more molecules will have a significantly overlapping retention time. We must ask ourselves: what is considered a "significant overlap" in retention time? This is perhaps a question more suited towards a chemist. Further, how will we train a model to predict the probability of overlap? One option is to simply generate normal distributions for a set of two or more molecules given corresponding estimated uncertainties, and the metric for the probability of retention overlap could be calculated as a function of the overlap of the molecules' distributions. We could take multiple approaches to address this problem, some of which require an additional machine learning model, and some which solely exploit data from the RT predictions and uncertainty predictions. Deciding which approach is most appropriate for this classification problem will certainly be a challenge.

References:

1. 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.

2. 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.

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. Hirschfeld, L.; Swanson, K.; Yang, K.; Barzilay, R.; Coley, C. W. Uncertainty Quantification Using Neural Networks for Molecular Property Prediction. *J. Chem. Inf. Model.* **2020**, *60* (8), 3770–3780. https://doi.org/10.1021/acs.jcim.0c00502.

5. Busk, J.; Jørgensen, P. B.; Bhowmik, A.; Schmidt, M. N.; Winther, O.; Vegge, T. Calibrated Uncertainty for Molecular Property Prediction Using Ensembles of Message Passing Neural Networks. *arXiv:2107.06068 [cs, stat]* **2021**.