**Predict the Stereoselectivity of Chemical Transformation by Machine Learning**

Justin Li, Dakang Zhang, Yifei Wang, Christopher Ye, Hao Xu, and Pengyu Hong

Introduction:

Stereochemistry plays an essential role in biology. Most biochemical processes in living bodies are extremely sensitive to stereochemistry. For example, our bodies can only digest and make carbohydrates and amino acids of a certain stereochemistry, and all of the proteins that make up our bodies are composed of a single stereoisomer of amino acids. Our bodies can create and digest starch (found in potatoes and bread) but not cellulose (found in wood and plant fibers) despite both being polymers of glucose, however, with different stereochemistry. It is well known in medical practice that stereochemistry is important to drug action. For some therapeutics, single-stereoisomer formulations can lead to improved therapeutic indices because they provide greater selectivity for their biological targets and/or better pharmacokinetics than a mixture of stereoisomers. While one stereoisomer can have positive effects on the body, another stereoisomer may be less effective (D-Isoproterenol vs L-Isoproterenol on the blood pressures or heart rate), ineffective (as in the case of the R enantiomer of ibuprofen), or even toxic (as in the case of thalidomide).

Stereoselectivity or enantioselectivity is the most important aspect of organic transformation. Stereoselectivity can vary greatly in degree depending on reactants, catalysts, and reaction conditions. Quantitatively understanding and controlling the stereoselectivity of a chemical transformation – the relative proportions in which a non-stereospecific chemical transformation generates different stereoisomers under varying reaction conditions – is thus hugely important for organic synthesis. Yet we have only the most basic, qualitative understanding of the stereoselectivity of chemical transformations. We know that the stereoselectivity arises from differences in steric effects and electronic effects in the mechanistic pathways, but we have no rule for accurately, quantitatively predicting stereoselectivity. In addition, the optimizations of asymmetric transformations have been mainly by trial-error. A huge volume of data about the stereoselectivity of chemical transformation has been published over the past 100+ years, and volumes more are now generated. Machine learning has emerged as an effective avenue for taking advantage of these data to build computational models for accurately and quantitatively predicting the stereoselectivity of chemical transformation.

In [Reid & Sigman 2019], Reid and Sigman applied machine learning to predict the stereoselectivity of chiral phosphoric acid (CPA) catalysis, one of the most ubiquitous asymmetric transformation family that adds protic nucleophiles to imines catalyzed by chiral 1,1′-bi-2-naphthol(BINOL)-derived phosphoric acids bearing aromatic groups at the 3 and 3′ positions to produce amine products. The CPA catalysis family is pervasively applicable in both synthetic and biosynthetic settings [Nugent 2010 and Silverio, et al. 2013]. They collected 381 published reactions with varied components and generate a set of molecular features (both geometrical and topological) to describe each imine, nucleophile, catalyst and solvent. Linear regression models were trained to predict enantioselectivity using those features as well as other reaction variables (e.g., concentration of reagents or catalysts, inclusion of molecular sieves, etc.). Although linear regression models are straightforward to interpret, it falls short of capturing interactions between features. In addition, it turns out that the chosen CPA reaction family has a complex data distribution, which is beyond the capacity of merely one linear regression model. In this work, we use the same dataset and investigate machine learning techniques (e.g., LASSO [citation], Regression Tree [citation], and Random Forest [citation]) with capabilities for selecting features, exploring interactions between features, and handling more complex data distributions. Eventually, we develop a composite machine learning model that not only achieves better performance but also offer novel insights.

**Methods:**

**Data**

The training set contains 381 CPA reactions collected from [Reid & Sigman 2019]. Each reaction includes a substrate, solvent, catalyst, nucleophile, and imine. Numerical features were derived from DFT calculations and molecular topologies to describe solvent (160 properties), catalyst (85 properties), nucleophile (15 properties), and imine (22 properties). The activation energy (∆∆G‡) and reaction variables of each reaction were also collected. Additionally, 64 out of sample reactions were also collected from 3 sources [Reid & Sigman 2019]. The goal is to build a robust model that predicts the ∆∆G‡ value of a reaction given the properties of catalyst, imine, nucleophile, and solvent in this reaction.

**Stereoselectivity Prediction Model Development**

We first tested four widely used machine learning techniques (i.e., LASSO [citation], regression tree [citation], random forest [citation], and boosting tree [citation]) for predicting ∆∆G‡ values using all features. LASSO uses the L1 regularization to select a subset of features to be used in describing linear relationships between the input features and the output. Regression tree is able to capture more complex relationships between features in predicting ∆∆G‡ values. To reduce overfitting and increase stability, random forest assembles a collection of regression trees using random subsets of features. Boosting tree uses the boosting technique to iteratively assemble a collection of regression trees. We used the training set and performed 2-fold cross-validation to compare these fourtechniques. The results (Table 1) indicate that random forest performed the best with the mean test r2 value of 0.926 and the test MSE value of 0.223. The results of a typical 2-fold cross validation run is illustrated in Figure 1. In [Reid & Sigman 2019], their linear regression model had a test r^2 value of 0.87, and they identified 6 features in their model – the imine properties “C”, “N”, and “L2”, the nucleophile property “H-X-CNu”, the catalyst properties “SubS”, and the solvent property “Balaban-type index from polarizability weighted distance matrix” . The top five most importance features used by the random forest model contribute almost 80% of importance (see Table 2). There are some similarities between the features selected, such as both the Imine “C” property and the nucleophile “H-X-CNu” property, but one major difference is that their model identified a solvent and a catalyst property as important to their predictions. Our model identified mostly imine properties with some nucleophile properties as important, whereas catalyst and solvent properties had small impacts on the overall prediction. This may be explained by (insert explanation).

Table : Compare LASSO, regression tree, random forest and boosting tree for predicting stereoselectivity. The training set was used. Two-fold cross-validation was run 100 times. The mean results and the corresponding standard deviation (STD) values are listed in this table.

|  |  |  |  |
| --- | --- | --- | --- |
| **Models** | **MSE (STD)** | **Test *r*2 (STD)** | **Train *r*2 (STD)** |
| **Lasso** | 0.343 (0.048) | 0.887 (0.018) | 0.942 (0.008) |
| **DT** | 0.361 (0.083) | 0.880 (0.0278) | 0.999 (0.001) |
| **BT** | 0.225 (0.040) | 0.925 (0.014) | 0.987 (0.002) |
| **RF** | 0.223 (0.048) | 0.926 (0.014) | 0.987 (0.002) |

Figure 1: The results of random forest in a typical 2-fold cross validation run using the training data. The blue dots are training samples, and the red dots represent the validation samples. The training and predicted r^2 values are 0.987 and 0.926 respectively

Chart, scatter chart

Description automatically generated

Table 2: The top five most important properties used by the random forest model. All features are used. The **Feature** column lists the feature names. The **Molecule** column lists the molecule category of each feature. The **Importance** column list the importance weight of each feature.

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature** | **Molecule** | **Importance (out of 100)** | **Description** |
| **C** | imine | 54.59 | Natural Bond Orbital |
| **SL** | imine | 17.26 |  |
| **H-X-Nu** | nucleophile | 2.56 | Nucleophilic Angle |
| **PG** | imine | 2.16 | Natural Bond Orbital |
| **H-X-CNu** | nucleophile | 1.85 | Nucleophilic Angle |

Since the imine gives strong indications towards what the products of a CPA reaction are and it requires extra efforts to obtain imine information, we investigate the possibility of predicting the ∆∆G‡ values without the knowledge of imine features. Interestingly, better machine learning models (see Table 3) could be trained without using information about the imine. Again, random forest performed the best with a mean test r2 of 0.933 and a mean train MSE of 0.203. All top 5 most important features used by the random forest model are nucleophile features (Table 4). Once again, catalyst features are only slightly influential on the final prediction. The most important catalyst properties include “iPOsy”, “B1”, and “C1”, which all have an overall importance of around 0.5. A possible explanation is that there are few catalysts and solvents in our training set, and those that are in our training set are very similar to each other. Elaborate

Based on the previous success of a nucleophile focused model which excluded imine properties, we also examined the performance of a imine focused model which excludes nucleophile properties. Interestingly enough, the models didn’t perform quite as well as our other regression models. Random forest performed the best with a mean test r2 of 0.881 and a mean train MSE of 0.360.

Combining with the results in Tables 1 and 2, we hypothesize that imine properties can be explained by the other molecules involved in the same reactions. Hence, we applied random forest to predict the imine transition state (i.e., E or Z) by using the features of catalyst, nucleophile, and solvent. The model identified that properties of the nucleophile were most influential in determining the transition state. The two-fold cross-validation results (training and test accuracies are 0.993 and 0.970, respectively) indicate that there is a link between the transition states and the corresponding nucleophilic reactants.

Table 3: Compare LASSO, regression tree, random forest and boosting tree for predicting stereoselectivity without using imine information. The training set was used. Two-fold cross-validation was run 100 times. The mean results and the corresponding standard deviation (STD) values are reported here.

|  |  |  |  |
| --- | --- | --- | --- |
| **Models** | **MSE (STD)** | **Test r2 (STD)** | **Train r2 (STD)** |
| **Lasso** | 0.625 (0.086) | 0.793 (0.033) | 0.864 (0.017) |
| **DT** | 0.291 (0.057) | 0.904 (0.020) | 0.977 (0.003) |
| **BT** | 0.234 (0.032) | 0.923 (0.011) | 0.967 (0.004) |
| **RF** | 0.203 (0.0345) | 0.933 (0.012) | 0.972 (0.004) |

Table 4: The top 5 most important properties used by random forest when the imine features are excluded. The **Feature** column lists the feature names. The **Molecule** column lists the molecule category of each feature. The **Importance** column list the importance weight of each feature.

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature** | **Molecule** | **Importance (out of 100)** | **Description** |
| **H-X-Nu** | nucleophile | 20.23839 | Nucleophilic Angle |
| **H-X-CNu** | nucleophile | 17.87433 | Nucleophilic Angle |
| **Nu** | nucleophile | 16.31278 |  |
| **Polarizability (nucleophile)** | nucleophile | 13.01146 | Polarizability of the nucleophile |
| **iXH** | nucleophile | 3.591798 |  |

Table 5: Results of Imine Focused Random Forest model excluding nucleophile features

|  |  |  |  |
| --- | --- | --- | --- |
| **Models** | **MSE (STD)** | **Test r2/STD** | **Train r2/STD** |
| **Lasso** | 0.626 (0.120) | 0.794 (0.040) | 0.874 (0.014) |
| **DT** | 0.524 (0.100) | 0.827 (0.036) | 0.975 (0.005) |
| **BT** | 0.376 (0.041) | 0.876 (0.015) | 0.966 (0.006) |
| **RF** | 0.360 (0.045) | 0.881 (0.015) | 0.966 (0.007) |

Table 6: The top 5 most important properties used by random forest classifier to determine the transition state of the imine. The Feature column lists the feature names. The Molecule column lists the molecule category of each feature. The Importance column list the importance weigh

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature** | **Molecule** | **Importance (out of 100)** | **Description** |
| **Polarizability (nucleophile)** | nucleophile | 5.786139 | Polarizability of the nucleophile |
| **Nu** | nucleophile | 5.099893 |  |
| **HOMO (nucleophile)** | nucleophile | 4.728747 | Highest Occupied Molecular Orbital |
| **H-X-Nu** | nucleophile | 4.518921 | Nucleophilic Angle |
| **H-X-CNu** | nucleophile | 2.852582 | Nucleophilic Angle |

Figure 2: The results of random forest in a typical 2-fold cross validation run using the training data. The blue dots are training samples, and the red dots represent the validation samples. The training and predicted r^2 values are 0.953 and 0.933 respectively

Chart, scatter chart

Description automatically generated

We observed that the training results of random forest are significantly better than the test results in the two-fold cross validation experiments, which is a sign of overfitting. This can be problematic when applying a random forest model trained by the above training data to new CPA reactions whose imines and/or nucleophiles are very different from those in the training data (i.e., new samples could fall in the low-density regions of the training data). In such a scenario, a simpler model (e.g., linear regression), which assumes less about data distribution and requires less amount of training data, can deliver better extrapolation than random forest. One intriguing solution is to train multiple prediction models and combine them into a composite model. We aimed to develop a pipeline that would be select an appropriate model based on whether the imine and nucleophile were in high density areas compared to the training data (which would indicate whether the prediction would require extrapolation). Given a new sample, the composite model first estimates the density of the new sample compared to the training distribution and then chooses a prediction model accordingly. The density can be calculated by Gaussian mixture models (GMM) [citation] fitted on the training data and shows whether the sample is similar to the training data or not. Based on our observations described above, we decided to include three prediction models in this composite model: a random forest model trained by all features (overall model), a second random forest model trained without imine features (nucleophile focused model), and a linear regression model trained via LASSO using all features, choosing to exclude the random forest model without nucleophile features (imine focused model) due to its relatively poor performance compared to other models. Our overall random forest model is able to make strong predictions when both the imine and nucleophile of a reaction are similar to those in our training data, and our nucleophile focused random forest model is able to make strong predictions when the imine of a reaction is not similar to those in our training data, but the nucleophile is. Our Lasso model can handle any reactions in which the nucleophile is not similar to those in our training data. We also trained two GMMs, one for approximating distributions of important nucleophile features (H-X-Nu, H-X-CNu, Nu, Polarizability) and the other for important iminium features (C, SL, PG). The features were selected due to their importance in the earlier overall random forest model and nucleophile focused model.

The composite model works as the following. Given a new reaction sample, it first calculates the density of the sample using the nucleophile GMM. If the density is high, it calculates the density of the sample using the imine GMM. If both densities are high, it applies the overall random forest model trained by all features. If the nucleophile density is high but the imine density is low, it applies the nucleophile focused random forest trained without using the imine features to avoid extrapolation. If the density of the nucleophile is low compared to the GMM, then the linear regression model is used, since our earlier results signify that a random forest struggles slightly when making predictions without nucleophile information (Table 5).

The nucleophile GMM grouped the 381 reactions into 14 components based on four important nucleophile properties – H-X-Nu, H-X-CNu, Nu, and Polarizability. The four properties were chosen due to their respective strong influences on the nucleophile focused random forest model, and we decided to select 14 components through evaluation of AIC and BIC scores. The imine GMM grouped the 381 reactions into 15 components based on three important imine properties – C, SL, and PG.

Figure 3: Actual vs Predicted for Pipeline Out of Sample Predictions

Chart, scatter chart

Description automatically generated

Pipeline Results:

The pipeline performed well on the data. For each of the original 381 reactions, the pipeline determined that the overall random forest model would make the best prediction, which was expected since the GMMs were fitted to those reactions.

We were able to see the full capabilities of the pipeline when it made predictions on the out of sample data that had not been included in the original training or testing data. For the 15 reactions in the “Addition of enecarbamates to benzoyl imines” type reactions, the pipeline determined that Lasso was the best model to make the prediction, due to a low nucleophile GMM score indicating low density. This led to a low mean average error of 0.25, which indicates that the predictions are decently accurate. For the 15 “Hydrogenation of fluorinated alkynyl ketimines” type reactions, the nucleophile focused forest model was chosen to make predictions, with a low mean average error of 0.24 - once again, a decent score indicating accuracy and good choice of model. Finally, for the 34 “Addition of thiols to imines (Denmark)” type reaction, the overall random forest model was used to make prediction with a mean average error of 0.52. Across all 64 out of sample predictions, the mean average error is 0.39 and the r^2 value is 0.951.

The low error scores and the high r^2 value indicate that the pipeline did well in extrapolating out to reactions that it hadn’t been exposed to. Additionally, the strong performance of the nucleophile specific random forest models in the pipeline show that although imine properties are important in the overall model, it is still possible to make strong predictions in the absence of them

Table 6: Results of the pipeline predictions vs Results of [Sigman & Reid 2019]. Note: a \* means that [Sigman & Reid 2019] utilized a E-imine specific model to make the prediction rather than a comprehensive model that dealt with both E/Z imine reactions

|  |  |  |  |
| --- | --- | --- | --- |
| **Source** | Addition of enecarbamates to benzoyl imines | Hydrogenation of fluorinated alkynyl ketimines | Addition of thiols to imines (Denmark) |
| **Model used** | Lasso | Nucleophile focused RF | Overall RF |
| **Average Imine GMM log density score** | 8.430 | -931.349 | 8.896 |
| **Average Nucleophile GMM log density score** | -646434.250 | 14.820 | 10.601 |
| **Mean Average Error of Pipeline prediction** | 0.25 | 0.26 | 0.52 |
| **Mean Average Error in [Sigman & Reid]** | 0.37 (0.24\*) | 0.30 | 0.65 (0.67\*) |

Discussion:

Our project used a small sample of reactions – such a model can definitely be improved on with more data

Our results show that such a model can be implemented effectively

Traditionally, chemists do these experiment by experiment to evaluate the impact of certain reactants. Once we collect enough experiments, can we make predictions to find settings without the need for physical experimentation.

Importance of Features

Nugent, T.C. Chiral Amine Synthesis: Methods, Developments and Applications (Wiley, 2010).

Silverio, D. L. et al. Simple organic molecules as catalysts for enantioselective synthesis of amines and alcohols. Nature 494, 216–221 (2013).