**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 collected the features of 350+ conditions in the CPA chiral phosphoric acid catalysis reaction family, and trained linear regression models for predicting reaction performance. The features include quantitative structure–activity relationships, molecular mechanics, and those derived for iminium, catalysts, and solvents using density functional theory. Why this family of reactions is important?

In this work, we use the dataset provided in [Reid & Sigman 2019] and develop a more sophisticated approach that achieve better performance. A few sentences about how we do it and what the results are.

Methods:

Data Overview

There were 381 total entries in the dataset, from 17 reaction families. Each reaction included a substrate, solvent, catalyst, nucleophile, and imine. Numerical properties of the solvent (160 properties), catalyst (85 properties), nucleophile (15 properties), and imine (22 properties) were identified, as well as the activation energy (∆∆G‡) of each reaction. Additionally, 64 out of sample reactions collected from 3 different reaction families were used to test the final pipeline. More information about the data can be found in the Supplementary Information.

Early Development of Models and Results

The first set of models we developed were regression models which used all properties of the solvent, catalyst, nucleophile, and imine of a reaction to predict the DDG value. Four separate machine learning models were tested – Lasso, Decision Tree, Boosting Tree, and Random Forest -- and the results of each were compared. Lasso is well suited to capturing linear relationships between features and the Y variable, but may struggle when faced with nonlinear relationships. Decision Tree and Boosting Tree are able to capture more complex relationships between features with decisions at each node but may be prone to overfitting. Random Forest improves on decision tree models by creating a collection of decision trees with random subsets of features, reducing overfitting and increasing stability. By implementing three different types of models – linear regression, decision tree, and random forest – we can compare the results and determine the most effective model for the data.

We standardized the data and trained all our models by randomly splitting the dataset 50:50 into a training set and testing set. The predicted values of the test set and train set were compared to actual test and train data values to measure accuracy of the model. We also observed the feature importance of the models to identify chemical properties that were important and influential to the model’s predictions. For every model we evaluated, we repeated the model creation/training process 100 times, each time with a different random split of train/test data, and calculated the average accuracy and feature importance over all 100 iterations.

Out of the regression models, the best performing model was the Random Forest model with an average r^2 value on test data of 0.926 and an average overall r^2 value on all data of 0.956. The performance of the other models developed can be found in Table 1. The most important features in the Random Forest model were mostly properties of the imine of the reaction (Table 4), with nucleophile, catalyst, and solvent properties having slight influence on the reaction. The imine property ‘C’ was by far the most influential on the regressor.

With these results in mind, we developed a second set of regression models in which imine properties were excluded in order to examine whether effective predictions could be made without knowledge of the imine properties – which dominated the earlier models. We tested the same four models – Lasso, Decision Tree, Boosting Tree, and Random Forest – using properties of the solvent, catalyst, and nucleophile. The strongest model out of the second set was a Random Forest model with an average test r^2 value of 0.933 and an average overall r^2 value of 0.953. The performance of the other models developed can be found in Table 2. With the exclusion of imine properties, the most important features in the model were mostly properties of the nucleophile of the reaction (Table 5). The nucleophilic angles H-X-Nu and H-X-CNu and were the most influential on the regressor.

We also hypothesized that imine properties could be explained/predicted by the other compounds involved in the reaction, so we also developed a set of classification model in which the properties of the solvent, catalyst, and nucleophile were used to predict the imine transition state of a reaction (either E or Z). The strongest model out of the third set of models was a K Nearest Neighbors Classifier with an average accuracy of 0.970 on unseen test data and 0.974 on all data. The performance of the other models developed can be found in Table 3.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Figure : Typical random split of Random Forest Regression model from first set (including iminium properties). The predicted r^2 is 0.926, and the total r^2 is 0.957.

Figure : Typical random split of Random Forest Regression model from second set (excluding iminium properties). The predicted r^2 is 0.933, v and the total r^2 is 0.953

Table 1: Results of models in the first set (including imine properties)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Models | MSE | Test r2 (STD) | Train r2 (STD) | Total r2 (STD) |
| Lasso | 0.343 | 0.887 (0.018) | 0.942 (0.008) | 0.914 (0.006) |
| DT | 0.351 | 0.883 (0.285) | 0.997 (0.001) | 0.940 (0.014) |
| BT | 0.229 | 0.924 (0.011) | 0.988 (0.002) | 0.956 (0.005) |
| RF | 0.223 | 0.926 (0.014) | 0.987 (0.002) | 0.956 (0.006) |

Table 2: Results of models in the second set (excluding imine properties)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Models | MSE | Test r2 (STD) | Train r2 (STD) | Total r2 (STD) |
| Lasso | 0.637 | 0.788 (0.041) | 0.874 (0.017) | 0.832 (0.015) |
| DT | 0.291 | 0.904 (0.020) | 0.977 (0.003) | 0.940 (0.010) |
| BT | 0.234 | 0.923 (0.011) | 0.967 (0.004) | 0.945 (0.005) |
| RF | 0.203 | 0.933 (0.012) | 0.972 (0.004) | 0.953 (0.006) |

Table 3: Results of models in the third set

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Test acc | Train acc | Total acc |
| Knn | 0.970 | 0.978 | 0.974 |
| Decision tree | 0.960 | 0.993 | 0.976 |
| Random forest | 0.970 | 0.993 | 0.984 |
| Logistic regression | 0.942 | 0.974 | 0.958 |
| lda | 0.955 | 0.984 | 0.969 |

Table 4: Most important properties in overall random forest

|  |  |  |
| --- | --- | --- |
| **Property** | type | Importance (out of 100) |
| **C** | imine | 54.58841 |
| **SL** | imine | 17.26146 |
| **H-X-Nu** | nucleophile | 2.564404 |
| **PG** | imine | 2.161205 |
| **H-X-CNu** | nucleophile | 1.848946 |

Table 5: Most important properties in nucleophile focused random forest (excluding imine)

|  |  |  |
| --- | --- | --- |
| **Property** | **type** | **Importance (out of 100)** |
| **H-X-Nu** | nucleophile | 20.23839 |
| **H-X-CNu** | nucleophile | 17.87433 |
| **Nu** | nucleophile | 16.31278 |
| **Polarizability (nucleophile)** | nucleophile | 13.01146 |
| **iXH** | nucleophile | 3.591798 |

Analysis:

Firstly, our results indicate a strong correlation between the structural parameters of the various molecules involved in a reaction and the DDG value of the reaction. Our strongest models in both sets of regression models had r^2 values well over 0.9 when making predictions on test data, indicating strong performance, even on data it had not seen yet.

Random Forest models performed the best out of all the models. Its strong ability to capture complex relationships makes it more accurate than linear regression models, and it’s use of multiple trees combats the stability and overfitting issues that Decision Trees often face.

An interesting discovery is the strong performance of the second set of model, which excluded imine properties. The Random Forest model (test r^2 = 0.933) from this set was able to slightly outperform the overall Random Forest model from the first set (test r^2 = 0.926) despite its disadvantage in not having information about the imine involved with the reaction. This was an unexpected finding, as one would expect a decrease in performance after excluding imine properties since properties of the imine were very influential in making predictions in the first set of models. Rather, it was observed that nucleophile properties which were somewhat influential in the first set of models became the most important properties in the second set of models. This suggests that while imine properties are important, they aren’t a necessity for strong predictions.

Generally, solvent and catalyst properties did not seem to play a major role in making predictions about the reaction. A possible conclusion is that catalyst and solvent properties can be well explained and predicted based on the imine and nucleophile involved in the reaction.

The third set of models performed robustly in predicting the transition state of the imine. In addition, most of the important features identified were nucleophile properties that were also important features of the second set of models. This, along with the findings in the earlier paragraph, may signify that imine properties can be predicted and/or explained using properties of other compounds involved in the chemical reaction, especially the nucleophilic reactant involved.

Development of Pipeline:

One potential flaw we recognized with our random forest models was in their extrapolation on data that is very different than the training data. While it was very accurate in predicting most reactions, in a few cases it had trouble when the prominent imine or nucleophile properties were in low density areas that the model was unfamiliar with. This was expected, since Random Forest are very effective dealing with data similar to what it was trained on, but it is tends to have trouble extrapolating on completely new data values that fall outside of the training set because it is difficult for it to discover the trends that would allow it to do so. When faced with extreme outlying data that is far from the rest of the data, a decision tree based model may just generally group it with one side or another of a tree, whereas other models such as linear regression models may be more effective in utilizing captured trends to extrapolate along a spectrum.

This particular flaw was highlighted when the overall random forest model was tested with out of sample data entries. It especially had trouble with entries from the “Hydrogenation of fluorinated alkynyl ketimines” family of reactions, since the reactions have unique and unseen imines, which may cause a problem since the model is dominated by imine properties. The mean absolute error of predictions was 2.967.

One intriguing solution was to develop multiple models as part of an overall pipeline with a series of decision points which determine whether the properties of its imine and nucleophile are similar to our training data. Based on the determinations, we choose the most effective model to make the final prediction. Based on our earlier success with the nucleophile random forest model excluding imine features, we realized that it was possible to exclude parts of the reaction from the model, such as imine and nucleophile, and still produce good results from a random forest model.

Three models from the earlier section are used in this pipeline. One is the overall random forest model which utilized all features from the reaction (imine, nucleophile, catalyst, solvent). This model is imine focused, as the final prediction is heavily influenced by imine properties. We also use the nucleophile focused random forest model, in which imine properties were excluded, leading to a model that could make accurate predictions in the absence of imine properties. Finally, we utilized a Lasso linear regression model (which utilized imine, nucleophile, catalyst, and solvent properties) which could reasonably adapt to both unseen imine and nucleophile properties. While such a linear regression model may not be as accurate as Random Forest models in predicting most data, it would be more effective in extrapolating to new, outlying data than our other Random Forest models would.

To determine which model to use, we utilized Gaussian Mixture Models (GMM), probabilistic models which cluster points in Gaussian distributions. We developed two separate gaussian mixture models: one for important nucleophile features, and one for important iminium features. In the nucleophile mixture model, the model determines if the overall nucleophile of a data entry is in a high or low density area by calculating its log-likelihood score – how well the point fit in the gaussian distributions. A low score indicates that the point does not fit well with the existing gaussian clusters – thus placing it in a low density area far from the other data points. On the other hand, a high score indicates that the point fits well with the existing clusters, and is in a high density area close to the rest of the data. Being in a high density area means that the nucleophile is somewhat similar to the nucleophiles that the model has been trained with, and the model will be able to make accurate predictions based off of the nucleophile properties. Meanwhile, being in a low density area means that the training data had none/few nucleophiles that were similar to the nucleophile in the reaction, indicating that our nucleophile focused random forest model may struggle with said data. A similar imine gaussian mixture model was developed using the imine properties (insert properties) to determine whether the imine was in a high or low density area.

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 depicts a visualization of the imine GMM.

The overall pipeline functioned as such: once a new reaction was fed in to be predicted, the log-likelihood score of the entry in both the nucleophile and imine GMM was determined. If both the nucleophile and imine gaussian mixture models indicated that the nucleophile and imine were in high density areas, then the overall random forest model was used to make predictions. If the nucleophile GMM indicated that the nucleophile was in a high density area but the imine GMM indicated that the imine was in a low density area, the nucleophile focused random forest model is selected since attempts to utilize the imine properties may lead to extrapolation issues. If the nucleophile GMM indicated that the nucleophile was in a low density area but the imine GMM indicated that the imine was in a high density area, then the imine random forest model was utilized, which excluded nucleophile features. Finally, due to the relatively weak performance of a imine-focused random forest model that excluded nucleophile properties (test r^2 = 0.88), if gaussian mixture models indicated that the nucleophile was in a low density area, then the Lasso model was used to make the prediction due to its strong ability to extrapolate.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedFigure 3: Visualization of the clustering mechanisms of a Gaussian Mixture Model using imine properties

Figure 4: Actual vs Predicted for Pipeline Out of Sample Predictions

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 “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.

Analysis of Pipeline:

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

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.