Abstract:

Introduction:

Introduce reactions:

Methods:

Results:

Analysis:

Discussion:

Introduction:

The efficacy of a catalytic process is dictated by the possible transition states, which feature core non-covalent interactions that determine their geometries and energies1,2. Such interactions are often difficult to identify and define because they are energetically weak and sensitive to the molecular properties of every reaction component (catalyst, substrates, reagents, solvent and so on)3,4. This overarching issue in reaction optimization is often exacerbated by subtle connections across several reaction variables, wherein modest structural changes to any or a few of these can have a profound effect on the experimental outcome5–7. These factors, combined with the number of dimensions under study in most reactions, are the underlying reasons that optimization is traditionally empirical8,9. This situation is particularly common in the area of asymmetric catalysis, wherein seemingly minor structural variations in any reaction component can have acute and non-intuitive influences on the observed enantioselectivity10. However, it is possible that such mechanistic outliers may be concealed within larger datasets because our pattern recognition skills do not perceive pivotal generalities when reaction situations change.

On this basis, we hypothesized that connecting common mechanistic features through the simultaneous interrogation of all reaction components would provide a holistic view of the key non-covalent interactions responsible for reaction performance. This would enable the transfer of experimental observations to genuinely different substrate combinations with unique catalysts. Here we develop and deploy a workflow that parameterizes all the reaction variables of more than 350 distinct reaction combinations, which allows the development of comprehensive statistical models, resulting in the ability to predict reaction performance for entirely different structural motifs. The workflow includes techniques to probe general mechanistic principles, which provides the basis for transfer learning or generalized identification of the key interactions imparting asymmetric induction.

Asymmetric catalysis is replete with examples of catalysts that can promote disparate reactions through a common mode of activation11–14. However, when ‘similar’ reactions are attempted, many changes to the precise reaction conditions are often required to obtain the desired reaction performance15,16. These changes can be subtle (that is, one aromatic solvent for another) or more profound (one catalyst class for another). This leads us to ask (1) whether mechanistic insight is transferable to a new reaction in the same subclass, given that a standard mechanistic paradigm may exist with a general mode of activation? If so, (2) how could a data-driven workflow that combines data acquisition and a description of the molecules involved mathematically be used to build a statistical model for diverse and multiple reaction profiles? And if such a workflow is achievable, (3) can the observed conditions of one or more reactions be deployed to predict the performance of another? Such analysis could provide a mechanistic understanding of why certain conditions are effective for a general reaction type and the ability to transfer this information quantitatively to out-of-sample predictions streamlining reaction optimization17,18.

To assess a specific workflow that is designed to probe the questions posed above, it would be pragmatic to compare transformations within a reaction class facilitated by a single catalyst chemotype. Although multifarious reports of the same catalyst class for different transformations exist in enantioselective catalysis, comparative studies—even qualitative rather than quantitative—have been sparse. Such an assessment would be challenging because most datasets, often generated under non-uniform conditions, are incomplete and readily comprehensible descriptors for each varying reaction component need to be developed. To address this correlation challenge, we envisioned a strategy for the interrogation of enantioselective catalysis involving the application of modern data-analysis methods and advanced parameter sets. In this approach, integrated descriptor sets—quantitative structure–activity relationships (QSAR), molecular mechanics (MM) and density functional theory (DFT) derived)19—are related to a relatively large library of outputs collected from a general reaction and catalyst type, which are data-mined from multiple literature sources (see the Supplementary Information). By combining appropriate data-organization and trend-analysis techniques, general relationships between reactions can be established. The ability of the statistical models to predict a new reaction type performance is used as a validation of mechanistic transferability (Fig. 1).

Hint:

What is this family of reactions? (CPA chiral phosphoric acid catalysis)

Why this family of reactions is important?

Impact: Traditionally, do these experiment by experiment. Once we collect enough experiments, can we make predictions to find settings

What you want to do, why you want to do It, how we do it and what the results are

Some reactions, don’t need to depend on imine

Some reactions, imine > nucleophile

Catalyst features not important

Possible that catalyst features can be well explained by features of imine, nucleophile

Tell people why we do what we want to do (why we use rf over lasso)

Lasso is simple lr, dt and bt capture variance, rf gi

In stereochemical reactions,

Reactions used:

The reactions used were

Methods:

Data Overview

There were 381 total reactions in the dataset, collected from \_ sources. 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 a DDG value of each reaction (\_\_\_insert explanation\_\_\_). After data processing, we applied various machine learning algorithms to the dataset.

Additionally, 64 out of sample reactions collected from 3 sources were used to test the final pipeline.

Early Assessment of Models

Three sets of relatively simple machine learning models were developed for various purposes.

The first set of models were machine learning regression models in which the properties of the solvent, catalyst, nucleophile, and imine of a reaction were used 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.

The second set of models were machine learning regression models in which the properties of the solvent, catalyst, and nucleophile of a reaction were used to predict the DDG value. Imine properties were excluded. Four separate machine learning models were tested – Lasso, Decision Tree, Boosting Tree, and Random Forest -- and the results of each were compared.

The third set of models were machine learning classification models 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). Six separate machine learning models were tested – K Nearest Neighbor, Decision Tree, Random Forest, Logistic Regression, and Linear Discriminant Analysis -- and the results of each were compared.

Training & Evaluation

We trained the models by randomly splitting the dataset 50:50 into a training set and testing set. The trained model was used to make predictions on the test data, and the predicted values were compared to actual test data values to measure accuracy. For regression models in which we predicted numerical values, r^2 was used as the metric of accuracy. The R^2 value measures how much variance in the dependent variable is explained by the model – the closer the r^2 value is to 1, the more accurate the model. For classification models, accuracy was measured simply by dividing the number of correct predictions by the total number of predictions made.

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 type of model we evaluated, we repeated the model creation/training process 100 times, each time with a different random 50:50 split of train/test data. This process ensured that \_\_\_ The mean accuracy/r^2 value over all 100 iterations was calculated and used to determine the accuracy of the model. Similarly, the mean feature importance of each chemical property across all 100 iterations was calculated.

Figure 1: Results of Random Forest Regression model from second set (excluding iminium properties). The predicted r^2 is 0.933, and the total r^2 is 0.953

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Figure 2: Results 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.

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Models | MSE | Test r2/STD | Train r2/STD | Total r2/STD |
| Lasso | 0.081529 | 0.857335/0.028593 | 0.939504/0.007948 | 0.898993/0.012574 |
| Decision Tree | 0.351993 | 0.883525/0.285952 | 0.997219/0.001423 | 0.940417/0.014141 |
| Boosting Tree | 0.229198 | 0.924112/0.011411 | 0.987517/0.001991 | 0.955898/0.005407 |
| Random Forest | 0.223459 | 0.926373/0.013612 | 0.987074/0.002056 | 0.956674/0.006035 |

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Models | MSE | Test r2/STD | Train r2/STD | Total r2/STD |
| Lasso | 0.637415 | 0.788066/0.040520 | 0.873554/0.016922 | 0.831636/0.014616 |
| Decision Tree | 0.290518 | 0.903708/0.019526 | 0.976915/0.003481 | 0.940484/0.009601 |
| Boosting Tree | 0.233848 | 0.922961/0.011408 | 0.967222/0.004021 | 0.945171/0.005171 |
| Random Forest | 0.203178 | 0.932761/0.011805 | 0.972154/0.003513 | 0.952569/0.005606 |

Table 3: Results of models in the third set

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Test acc | Train acc | Total acc |
| Knn | 0.970366 | 0.977737 | 0.974042 |
| Decision tree | 0.960000 | 0.993053 | 0.976483 |
| Random forest | 0.969581 | 0.993474 | 0.984196 |
| Logistic regression | 0.941571 | 0.973789 | 0.957638 |
| lda | 0.954555 | 0.983947 | 0.969213 |

Early Results:

Out of the first set of models, the best performing model was the Random Forest model with Hyperparameters [insert description].The model had a r^2 value on unseen test data of [insert score], and an overall r^2 value on all data of [insert score]. The performance of the other models developed can be found in [insert table].

The most important features in the Random Forest model were mostly properties of the imine and nucleophile of the reaction [reference to above table]. The Natural Bond Orbital C parameter was by far the most influential on the regressor, while other important features identified by our model include [insert features].

Similarly, the strongest model out of the second set of models was a Random Forest model with an r^2 value on unseen test data of [insert score], and an overall r^2 value of [insert score]. The performance of the other models developed can be found in [insert table].

The most important features in the Random Forest model were mostly properties of the nucleophile of the reaction [reference to above table]. The nucleophilic angles H-X-Nu and H-X-CNu were the most influential on the regressor.

The strongest model out of the third set of models was a K Nearest Neighbors Classifier with [insert accuracy] accuracy on unseen training data and [insert accuracy] accuracy on all data. The performance of the other models developed can be found in [insert table].

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.

Another interesting result was the strong performance of the second set of models (which excluded imine properties). The Random Forest model (test r^2 = [insert score]) from this set was able to slightly outperform the Random Forest model from the first set (test r^2 = [insert score]), despite its disadvantage in not having information about the iminium 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.

Additionally, 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 in its transition state can be predicted using properties of other substances involved in the chemical reaction, especially the nucleophilic reactant involved.

Another interesting finding was that generally, in all three sets of models developed, solvent and catalyst properties did not seem to play a major role in making predictions about the reaction.

that the properties of the iminium and nucleophile are clearly the most important in predicting DDG value of a reaction

Development of Pipeline

One potential flaw we recognized with our random forest models were that 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. (insert explanation as to why this is on a random forest model).

This particular flaw was highlighted when the overall random forest model was tested with 64 out of sample data entries. It especially had trouble with (reaction 19) (with a mean average error of \_\_\_), which had unique, unseen imines, and still struggled with (reaction 18) (with a mean average error of \_\_), which had unique, unseen nucleophiles.

One intriguing solution was to develop multiple models as part of a overall pipeline. When a reaction was predicted, the pipeline first chose the most effective model based on the properties of its imine and nucleophile, and used that model to make the final prediction.

Overall, four robust models were used in this pipeline. One was the aforementioned comprehensive random forest model which utilized all features from the reaction (imine, nucleophile, catalyst, solvent). We also developed two other random forest models, one which was imine focused and one which was nucleophile focused. In the imine focused model, nucleophile properties were excluded, and properties of the solvent, catalyst, and imine of a reaction were used to predict the DDG value. As a result, this model relied heavily on imine properties, and would be able to make predictions on reactions that had outstanding nucleophiles. In the nucleophile focused model, a similar approach was used in that properties of the solvent, catalyst, and nucleophile were used, and imine properties were excluded, leading to a model that mostly relied on nucleophile features. Consequently, the model could be applied to reactions with unseen imines and still make accurate predictions. Finally, as a backup catch all net, we developed a Lasso linear regression model that which could reasonably adapt to unseen imine and nucleophile properties. (explain why lasso is better than rf for extrapolation)

To determine which model to use, we utilized Gaussian Mixture Models, which (insert explanation of how it works). 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 adjusted to the nucleophile properties of (insert properties) of the sample data to develop gaussian models? (probably wrong idk). A new entry/reaction is run through the model, which determines if the overall nucleophile is in a high or low density area in terms of important nucleophile properties. Being in a high density area means that the nucleophile is somewhat similar to the nucleophiles that the model has been trained with, it is familiar with the nucleophile in the reaction and 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 perhaps a model that excluded nucleophile features rather than depended on them would be advantageous. 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, and the pipeline used that determination to decide which model to use (including or excluding imine features).

The overall pipeline functioned as such: once a new reaction was fed in to be predicted, both gaussian mixture models were used to determine its distance from the gaussians?(prob wrong idk) 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, then the nucleophile random forest model was utilized, which excluded imine features. 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, if both gaussian mixture models indicated that the nucleophile and imine were in low density areas, then the Lasso model was used to make the prediction.

Training/Evaluation of Pipeline models:

In the actual models used in the pipeline, we used 100% of our in sample data to train it to get the best trained model.

However, we still used a similar strategy as above to evaluate the models’ performance, splitting the data 50:50 into train and test and finding the average performance over 100 iterations.

(Insert results of each individual model)

INCLUDE HYPERPARAMETER TUNING OF MODELS AND GMM SOMEWHERE

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 (reaction 18) type reactions, the pipeline determined that the imine random forest model (excluding nucleophile) 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 (insert score), which indicates that the prediction was somewhat accurate. For the (reaction 19) tpe reactions, the nucleophile random forest model was used to make predictions, with a low mean average error of (insert score) - once again, a decent score indicating accuracy and good choice of model. Finally, for the (reaction 20) type reaction, the overall random forest model was used to make prediction with a mean average error of (insert score). Across all 64 out of sample predictions, the r^2 value was (insert value).

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 imine and nucleophile specific random forest models in the pipeline show that although both sets of imine and nucleophile properties are important in the overall model, it is still possible to make strong predictions in the absence of one or the other.