Catalyst Generality Optimisation: A case study on Asymmetric Hydrogenation Reactions

MITACS Globalink Research Internship

Internship supervised by:
Prof. Jolene P. Reid,
Department of Chemistry,
University of British Columbia, Vancouver
jreid@chem.ubc.ca / Tel. (+1) 604-822-3817

Saumya Thakur[a]

[a] *BS+MSc Integrated Dual, Department of Chemistry, Indian Institute of Technology, Bombay*

July, 2022

Abstract

An optimisation workflow introducing the concept of reaction generality, focused on different catalysts, is presented in this report. The project aimed at defining a metric for *Generality* and formulating it as an optimisation problem using a set of 364 asymmetric hydrogenation reaction points. We attempted to quantify this metric for different catalysts present in the dataset based on two factors: diversity of the reactants under that catalyst umbrella and the enantioselectivities obtained by that combination. The final metric was defined as a normalized value over the sum of average ee values obtained from each reactant space cluster. Virtual Augmentation was employed to remove bias from the dataset using predictions from focused Random Forest models. Further, a multivariate linear model was fitted on the obtained generality values with the corresponding ligand descriptors for correlational interpretability. As a validation study, rediscovery studies were performed using the obtained Generality MLR to score for some potential catalysts.

Acknowledgements

I would like to thank Dr. Jolene P. Reid, Department of Chemistry, University of British Columbia for giving me this opportunity to work on this project and Isaiah Betinol for assistance with the ideation, method development and analysis. I would also like to extend my gratitude to the MITACS Globalink for funding my stay and the work for this research internship.

Contents

1. Introduction

Most of the highly reputed and rewarded reactions in the history share a key feature: "They are high yielding and robust and capable of readily accommodating a wide range of substrate functionality and complexity." [3] A very famous example supporting the idea is the Diels-Alder reaction. [1] This concept can potentially be manifested further that might lead to searching for more such high-yielding reactions and catalysts in a faster and more efficient manner. Generality in asymmetric catalysis can be defined on the basis of various aspects of a successful chemical reaction. Spanning from the performance of a catalyst on a diverse set of substrate space in a given reaction to their promising activities on mechanistically distinct reactions, catalyst generality can also be defined in multiple ways. We explored the former idea of generality and defined a *General Catalyst* by its potential to catalyse a diverse space of reactants while providing a high catalytic performance on most of them.

Figure 1 is showing how reactions are discovered and generalised traditionally and it takes decades to achieve a convincing result for a specific reaction class. Years of exhaustive research on the mechanistic and experimental insights leads to a general understanding of a reaction. However, with the recent advances in the machine learning and other statistical modeling tools in various fields including catalytic chemistry, many breakthroughs seem possible [2]. Accelerated discovery of molecules and materials of interest has been a field growing at an exponential pace in the ongoing decade. So, can these tools be used to manifest the information present in already available data to search for more general candidates? Can we model and optimise a metric for generality using datasets scraped from literature? This project was aimed at finding convincing answers to such questions.

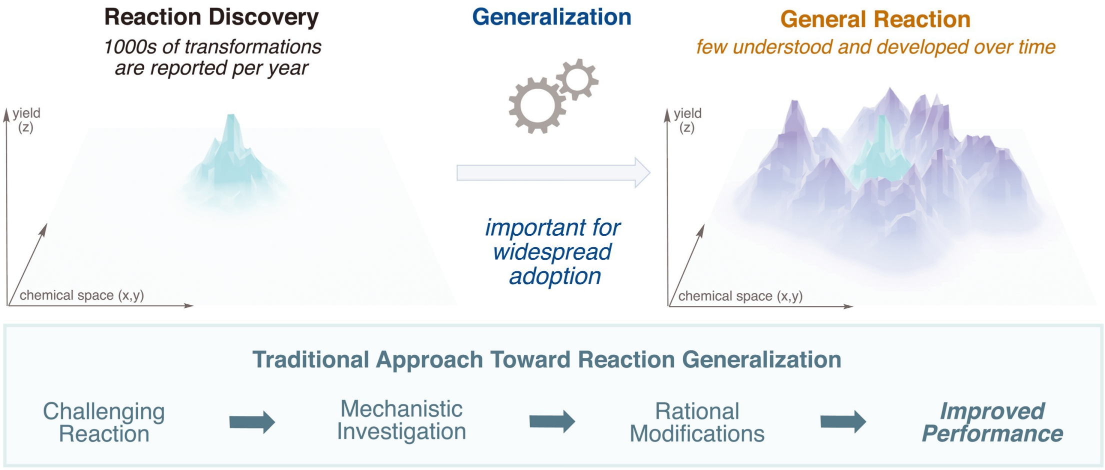

Reaction Discovery
1000s of transformations are reported per year
yield (z)
chemical space (x,y)
Generalization
important for widespread adoption
General Reaction
few understood and developed over time
yield (z)
chemical space (x,y)
Traditional Approach Toward Reaction Generalization
Challenging Reaction
Mechanistic Investigation
Rational Modifications
Improved Performance

Figure 1: Traditional Reaction Generalisation Framework

1.1. Problem Description

Given a class of reactions, can we define and optimise the generality values of a catalyst involving only catalyst's physical organic descriptors and hence rank the catalysts according to their generality scores?

2. Method

The method involved defining a generality metric first followed by the assessment and validation on a reaction dataset. We worked with a literature sourced dataset of 365 asymmetric hydrogenation reactions [4].

2.1. Metric Definition - Generality

The metric definition included two factors affecting the generality of a catalyst: substrate scope diversity and high catalytic performance. The study was performed on a dataset of asymmetric hydrogenation reactions. Reactant space included the electrophile - Iminium ion, nucleophile and the ligand (-Ar group of the catalyst). Hence, the substrate scope diversity was incorporated by clustering the reactant space.

2.1.1. Reactant Space Diversity

The substrate diversity incorporates the element of generality over chemical space. If a catalyst can catalyse a large fraction of the reactant space, it's performance can be analysed next as the second parameter to obtain a generality score. This diversity was assigned using different approaches:

- Distance on UMAP
- Clustering

For the UMAP distance approach, first, a base imine was decided according to their frequencies. The imine which was the highest frequent in the dataset was termed as the base imine. The physical oorganic descriptors of imines were used to obtain the UMAP representations. The UMAP plot for the imines in the transfer hydrogenation dataset is shown in Figure 2.

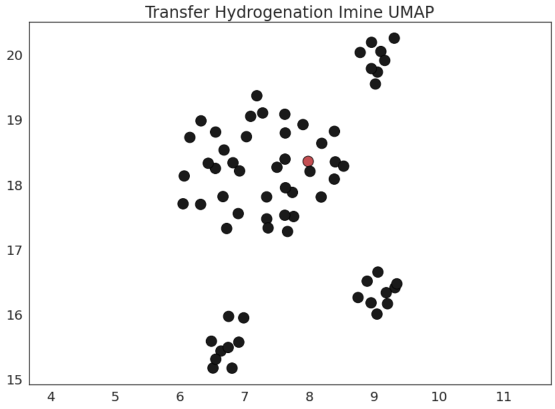


Figure 2: UMAP on Transfer Hydrogenation Imines

The highlighted red circle is for the base imine. Next, the Euclidean distance in this UMAP representation from the base imine was calculated for all the imines and was assigned as the iminium's diversity score. However, this approach saw some potential complications with the assigned values and the concept of Generality. If a catalyst could catalyse a reaction involving an imine with a very high diversity score, and does not perform good for the other areas of the reactant space, it'll still have a competent generality score using this approach.

Hence, the second technique using a clustering algorithm was put into use.

The Clustering approach treated all the imines in a cluster with the same priority and all the clusters as similar. No base imine or cluster was specified in order to eliminate the induced bias due to this. The algorithm used was K-Means Clustering and it was applied first on the imine parameters. The optimal number of clusters was obtained by the Elbow Plots. It later involved clustering the reactant space including both imine and nucleophile.

The generality metric was improvised from here by eliminating the diversity score and instead used the inherent diversity in the space to calculate the generality. The average enantiomeric excess in each cluster was calculated and summed over. It is further normalised by (100 * number of clusters). So, the most general catalyst should have a generality score of 1.

This further involved two different workflows:

- **2-tiered Reaction Generality:** First, the nucleophiles were clustered using nucleophile descriptors and then the imines were clustered using the iminium descriptors
- **Imine-Nucleophile Reaction Generality:** The clustering was performed on pairs of imiines and nucleophiles concatenating both of their descriptors.

The first workflow involved clustering the nucleophiles first and ranking the nucleophiles in a particular cluster according to their distances from the centroid. These nucleophiles were then chosen as the representative nucleophiles because of the huge number of possible imine-nucleophile-ligand combination. Regarding the catalytic performance, since the data used was for asymmetric hydrogenation reactions, the performance metric was enantiomeric excess values. The formula used is:

$$\frac{1}{100*n_c}\sum\frac{\Sigma ee}{n}$$

where, $\frac{\Sigma ee}{n}$ represents the average ee values in each cluster. So, n denotes the number of datapoints corresponding to that cluster, which varies across the clusters.

2.1.2. Biased Dataset

The raw dataset was very skewed in terms of datapoints present for different ligands. The distribution can be visualised in the Figure 3. L14 has 150 reaction points whereas many other ligands have only 1.

In order to obtain an unbiased generality metric and hence rank the ligands, the reactant space should be kept constant. Hence, the Virtual Augmentation of dataset was performed in order to obtain the same substrate space for all the ligands.

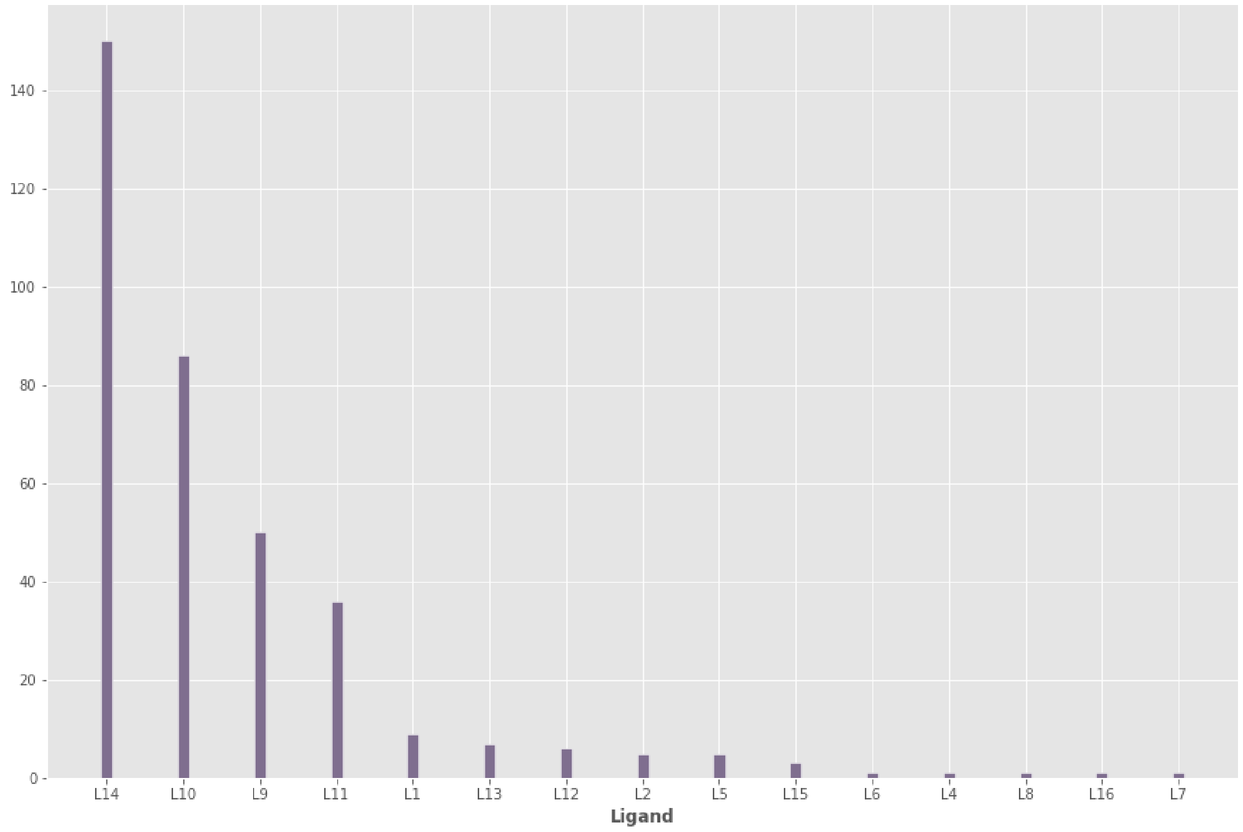


Figure 3: Distribution of reaction points available for a particular ligand

2.2. Optimisation Workflow

The complete optimisation workflow involves the following key steps (preferred in the order listed):

1. Obtaining the Virtual Screening data from the original dataset
2. Clustering the corresponding substrate space from the Virtual Screening data
3. Training a predictor for augmentation of ddG values followed by performing the predictions
4. Converting the predicted ddG values to the corresponding EE values (using T=298K)
5. Obtaining the final generality values from different workflows
6. Repeating the above steps without the most general and least general catalysts to perform the rediscovery studies
7. Training an MLR model on the obtained generality values and using it to predict and validate the scores obtained for the ligands left during rediscovering

2.2.1. Reactant Space Clustering

The reactant space here is referred to the one on which the generality values are being calculated. If it's the augmented ones, the reactant space might be larger compared to the original one. Since the reactant space included imines and nucleophiles, their physical descriptors were combined to obtain the total set of features. Further, the K-Means clustering was performed on this feature set to classify the imine-nucleophile pairs to different cluster labels. If the generality calculation involved augmented data, then the reactant space had all the possible imine-nucleophile combinations. These were obtained from the Virtual Screening

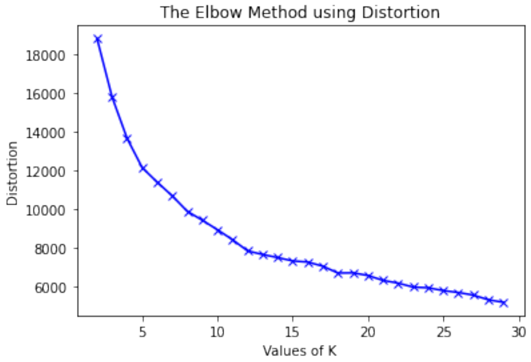


Figure 4: Elbow Plot on imine + nucleophile dataset

setup files. The final elbow plot to obtain the optimal clusters and understand the diversity in the dataset looked something like the Figure 4.

It can be seen that the space is highly diverse as it has so many clusters and those are not well-defined as the elbow is not very sharp. However, this can provide a vague idea about the number of clusters to obtain the generality values.

Once the reactant space is successfully clustered, the reactants in each cluster are extracted for further use in slicing the dataset accordingly. In the 2-tiered approach, both nucleophile clusters and imine clusters were different whereas in the single cluster approach, only the imine clusters are used. Hence, the 2-tiered approach involved a pairing of nucleophiles and imines to search in the virtual dataset.

2.2.2. Virtual Screening

The raw dataset had bias towards a few catalysts; some catalysts had around 150 data points where many had only 1 single reaction reported. Hence, the virtual screening was performed using the predictive models to augment the dataset such that all the catalysts have the same set of reactants. Since the total number of possible combinations in the reactant space shot up to 1.3k, not all the combinations were used in the augmentation. The nucleophiles clustered well into 5 clusters and so the nucleophiles closest to each cluster's centroid were used as the representative nucleophiles.

Steps implemented in setting up Virtual Screening:

1. Extracting the parameters for the cluster representative nucleophiles
2. Obtaining sets of imines, nucleophiles and ligands present in the dataset
3. Assigning solvents according to the imine values; the imine-solvent pairs were also obtained using the original dataset
4. Extracting all imine, ligand and solvent parameters from the original dataset
5. Permuting over all possible combinations of imine, ligand and nucleophiles and assigning the corresponding parameters to each combination added

Once the datapoints for virtual screening are obtained, the predictive models were built to obtain the augmenting predictions. Random Forest models were used as the predictive tool and were trained with many different workflows. Two of them being:

- **Focused Models**: Models trained with different parameters for E and Z imines; features obtained from the Stepwise Selection
- **Top 50 Features**: Models trained with only top 50 features chosen according to their importances obtained by the Random Forest

2.2.3. Obtaining Generality

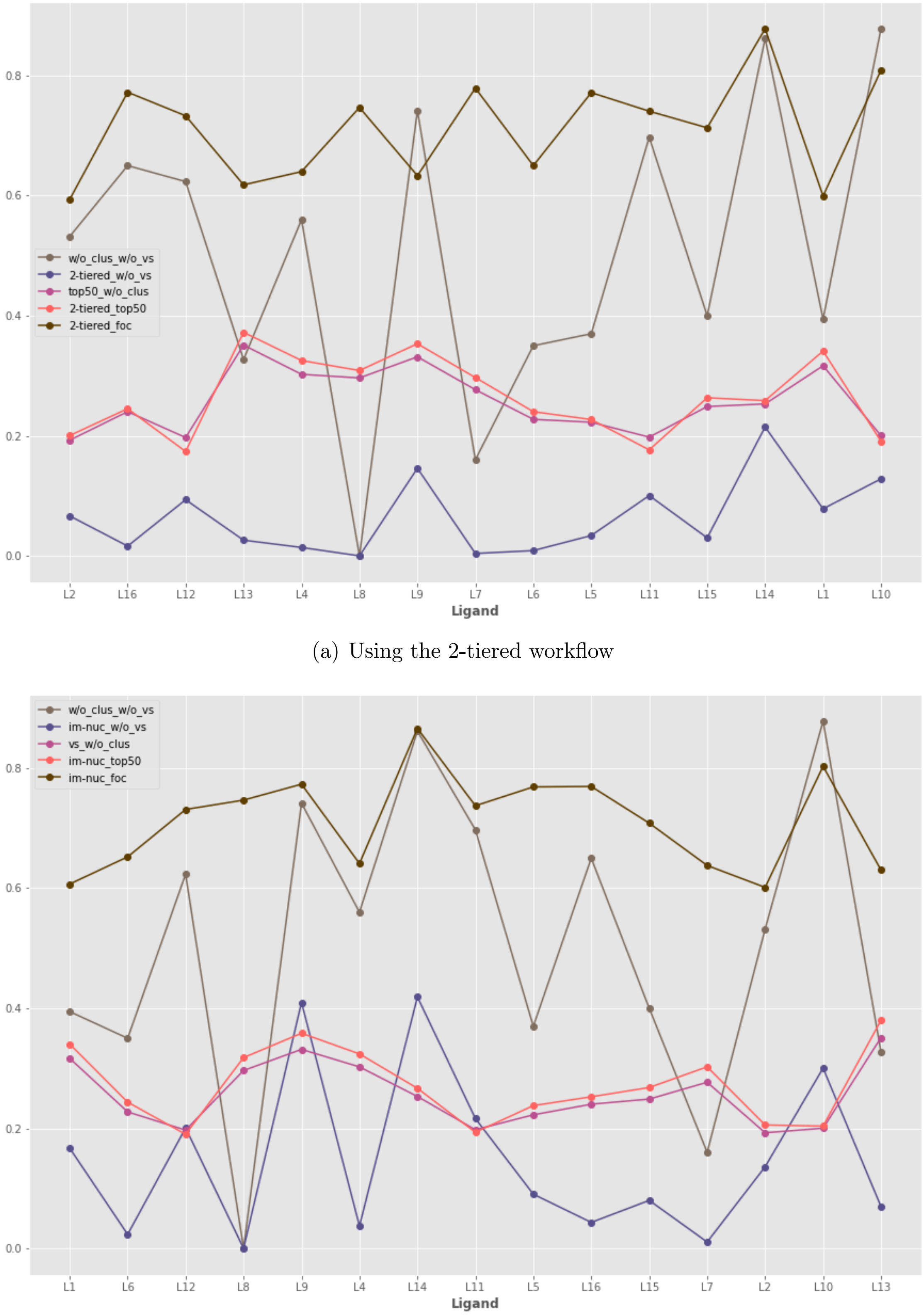
Once the augmentation is successfully performed, generality scores are obtained using the two workflows: **2-tiered** and **Im-nuc** workflows. The major difference between these two are the clusters used to represent the diversity. 2-tiered clusters the nucleophiles first and then clusters the imines; each nucleophile cluster will have the imine clusters under leading to a different formula for generality. Im-nuc clusters on the other hand cluster pairs of imines and nucleophiles. They concatenate the features of imines and nucleophiles followed by the clustering. The predictions were also performed using both top50 and focused RF predictors.

Further, all the different workflows can be classified into **biased** and **unbiased** versions:

- **Biased Versions:**
  - w/o_vs_w/o_clus: the values obtained on raw data without augmentation or without clustering. This involved taking an average of the sum of EE values from all the datapoints present for a particular ligand.
  - 2-tiered_w/o_vs or im-nuc_w/o_vs: the values obtained on raw data without augmentation but with clustering of the reactant space, including the diversity factor.
  - vs_w/o_clus: same as w/o_vs_w/o_clus but with the virtually augmented dataset. The predictions were obtained using the top50 RF predictor
- **Unbiased Versions:**
  - 2-tiered_top50 or im-nuc_top50: the augmentation was performed using the top50 RF predictor and the clustering was done using 2-tiered or im-nuc approach respectively.
  - 2-tiered_foc or im-nuc_foc: the augmentation was performed using the focused models and the clustering was done according to the approach in the names.

As evident, the biased versions either do not have the same reactant space for all the ligands or do not include the concept of diversity. The unbiased versions, on the other hand, use virtually augmented data and clustering of the reactant spaces incorporating substrate diversity in the formula. A summary of results obtained from different workflows can be seen in Figures 5(a) and 5(b).

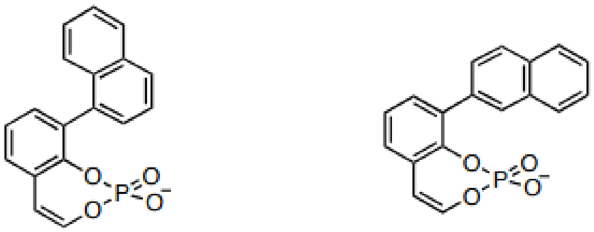
From the graphs, it can be seen that the values using the focused models are very high in comparison to the ones obtained from the top50 RF models. It can be explained by the error metrics of these RF models as the focused models had very high errors compared to the top50 models. So, the further studies were done using the top50 RF predictor and using the imine-nucleophile reactant space.



(a) Using the 2-tiered workflow

(b) Using the im-nuc workflow

Figure 5: Trends and results summarised for 5 different workflows implemented. w/o_clus_w/o_vs stands for the workflow without any augmentation and without any clustering. 2-tiered_w/o_vs or im-nuc_w/o_vs stands for the workflow without augmentation but 2-tiered or im-nuc generality clustering, respectively.



(a) L12: 1-naphthyl (b) L13: 2-naphthyl

Figure 6: Model Catalyst structures that were removed for rediscovery studies

The values obtained using the im-nuc_foc workflow were used for the further optimisation studies. The most general catalyst using this approach was L13 6(b) and the least general was L12 6(a).

2.3. Rediscovery Studies

As a validation, rediscovery studies were performed using the generality values obtained. The ligands having the highest and the lowest generality scores were removed from the whole workflow. This validation study was performed using the im-nuc_top50 workflow since the RF predictor trained with top50 features had better performance on the test data and hence was assumed to be more reliable for the prediction purposes.

3. Results

3.1. Virtual Screening using Random Forest

Since the raw dataset had biased reactant space, virtual screening was performed to obtain a common substrate space for all the ligands. The $\Delta\Delta G$ values for the absent reactant combinations were obtained from the Random Forest models. The dataset had two isomers of imines: (E)-imines and (Z)-imines. So, two different workflows were experimented with to obtain the $\Delta\Delta G$ predictions:

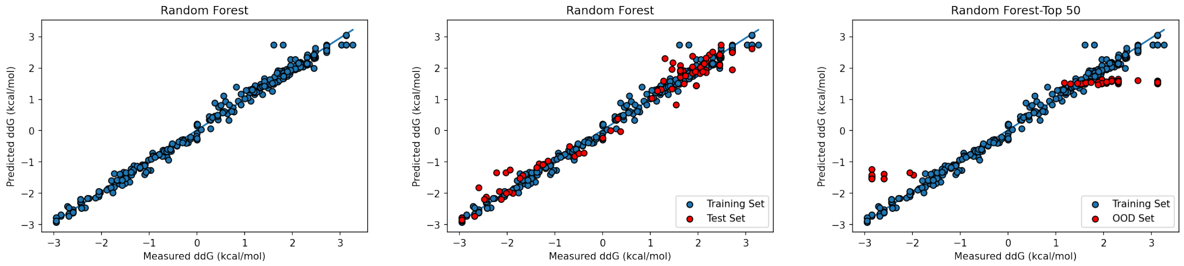
1. **Top 50:** Using the top 50 most important features from Random Forest
2. **Focused:** Using the focused E- and Z-imine RF models using their data and parameters

3.1.1. Top 50 Features

In the ”top 50 features” approach, the Random Forest model was first fitted with all the 230 descriptors available for each reaction. Then, the top 50 features were selected according to the importances obtained by the RF model. Another Random Forest model was fitted using these 50 features and it was used for the final predictions on the augmenting combinations in the Virtual Screening dataset. This single model is used for the predictions of both type-E and type-Z imines.

3.1.2. Focused Models

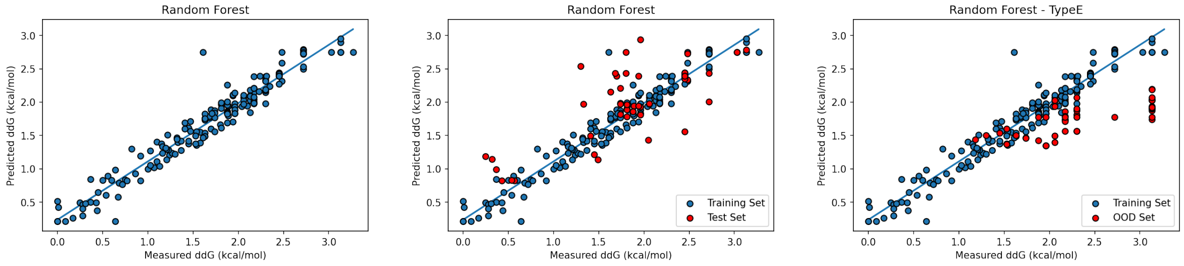
Since the dataset had both E-imines and Z-imines, two different RF models can be fitted on both the data separately. Hence, the whole data can be split further to two smaller datasets: one containing the E-imines and the other containing the Z-imines. According to



(a) Training data (b) Testing data (c) OOD data

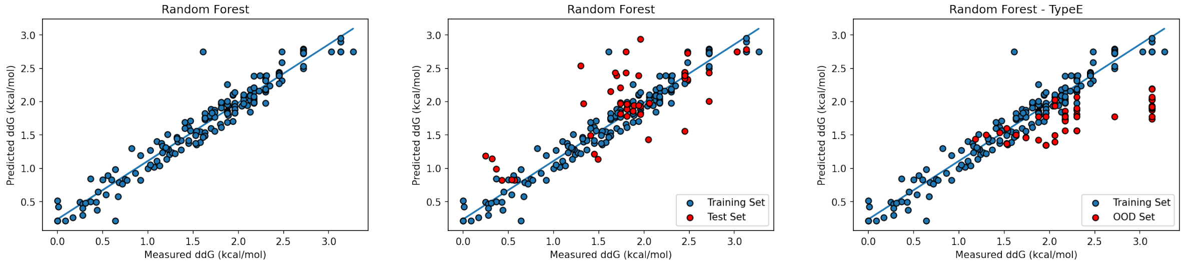
Figure 7: Predictions for training, testing and out-of-distribution datasets obtained using the top 50 features Random Forest Model

[4], the top parameters for E and Z imines were found separately by implementing forward step regression on their separate data respectively. Further, the RF models were fitted using the corresponding data and parameters. For the E-imines, there were 8 parameters and for Z-imines, there were 5 parameters out of a total of 230 parameter set. These RF models were used to perform predictions on the E and Z imine data from the Virtual Screening set, respectively.



(a) Training data (b) Testing data (c) OOD data

Figure 8: Predictions for training, testing and out-of-distribution datasets obtained using the focused RF models trained using E-imine parameters. These parameters were found by implementing the Forward Step Regression only on E-imine dataset.



(a) Training data (b) Testing data (c) OOD data

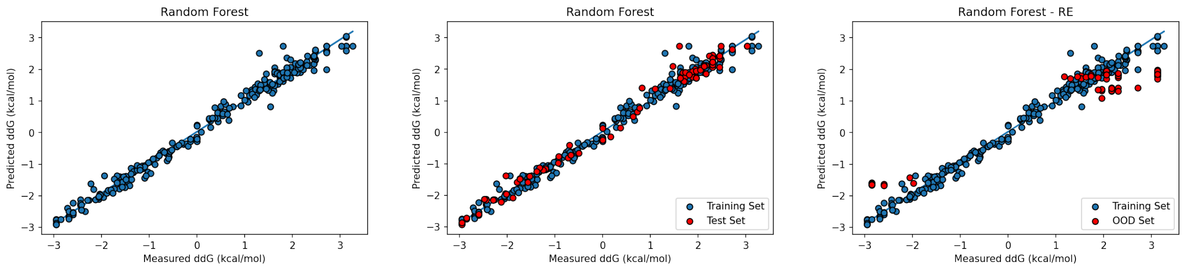
Figure 9: Predictions for training, testing and out-of-distribution datasets obtained using the focused RF models trained using Z-imine parameters. These parameters were found by implementing the Forward Step Regression only on Z-imine dataset.

Table 1: Random Forest metrics for different datasets and features

Workflow	CV R^2	Test R^2	Test MAE	OOD R^2
Top-50	0.934	0.962	0.275	0.77
Type-E	0.63	0.5	0.37	-0.605
Type-Z	0.8	0.844	0.273	-2.574
RE-Top-50	0.94	0.982	0.165	0.823

3.1.3. Rediscovery Results

When L12 and L13 were removed from the original dataset, the whole workflow was repeated to obtain the new generality values without information from L12 and L13 reaction points.



(a) Training data

(b) Testing data

(c) OOD data

Figure 10: Predictions for training, testing and out-of-distribution datasets obtained using the top 50 features from the Random Forest models while doing rediscovery studies (without L12 and L13)

The final geneality values obtained while rediscovering can be seen in the Figure 11.

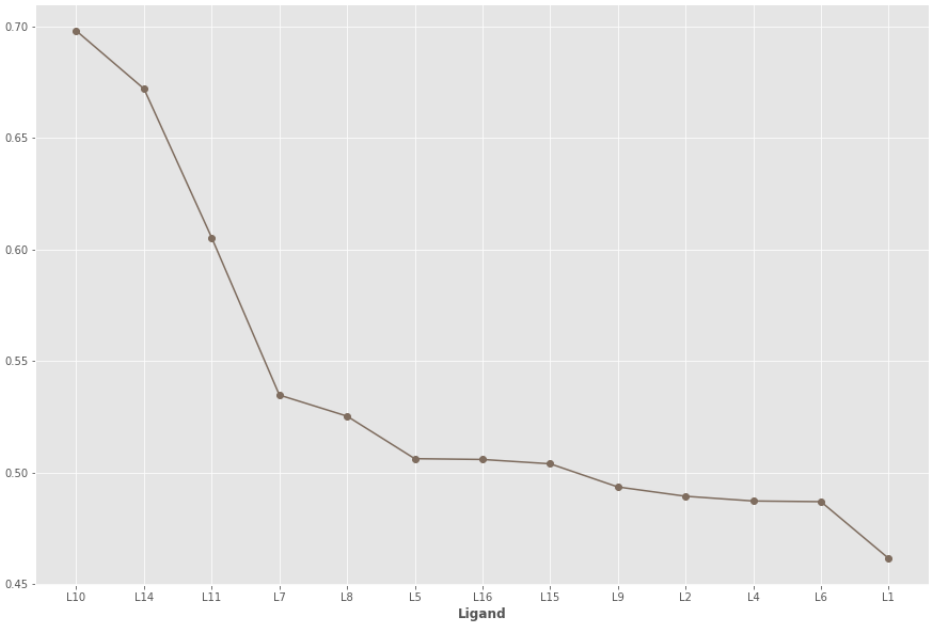


Figure 11: Generality trend obtained while performing rediscovery studies

3.2. Generality Optimisation using Multilinear Regression

Once the final generality scores are obtained from the rediscovery studies, a multi linear regression model was built using catalyst descriptors and the response value as generality. L12 and L13 were kept in the Virtual Screening set to predict the generality values once the model is built. 2-parameter and 3-parameter models were built and the corresponding fit and VS results can be found in Figures 12 and 13. Even though the L12 and L13 do not come out to be the least and most general catalysts, their scores are still close to their real values. The best fit equations were:

- **2-param:** 'y = 0.536 + 0.030Ligand_B1whole + 0.066Ligand_L6'
- **3-param:** 'y = 0.536 + 0.026Ligand_B1whole + 0.063Ligand_L6 + 0.011Ligand_nPOsy'

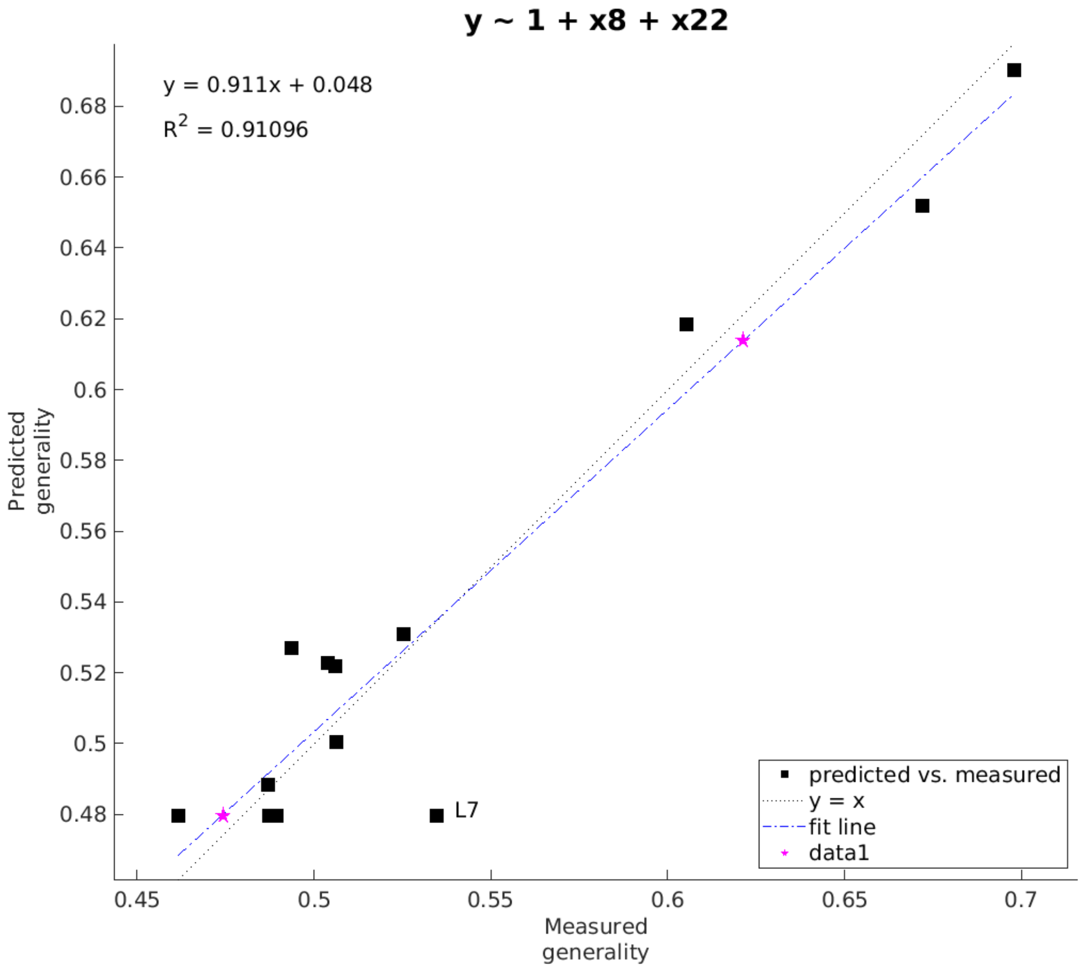


Figure 12: 2-parameter model for Generality

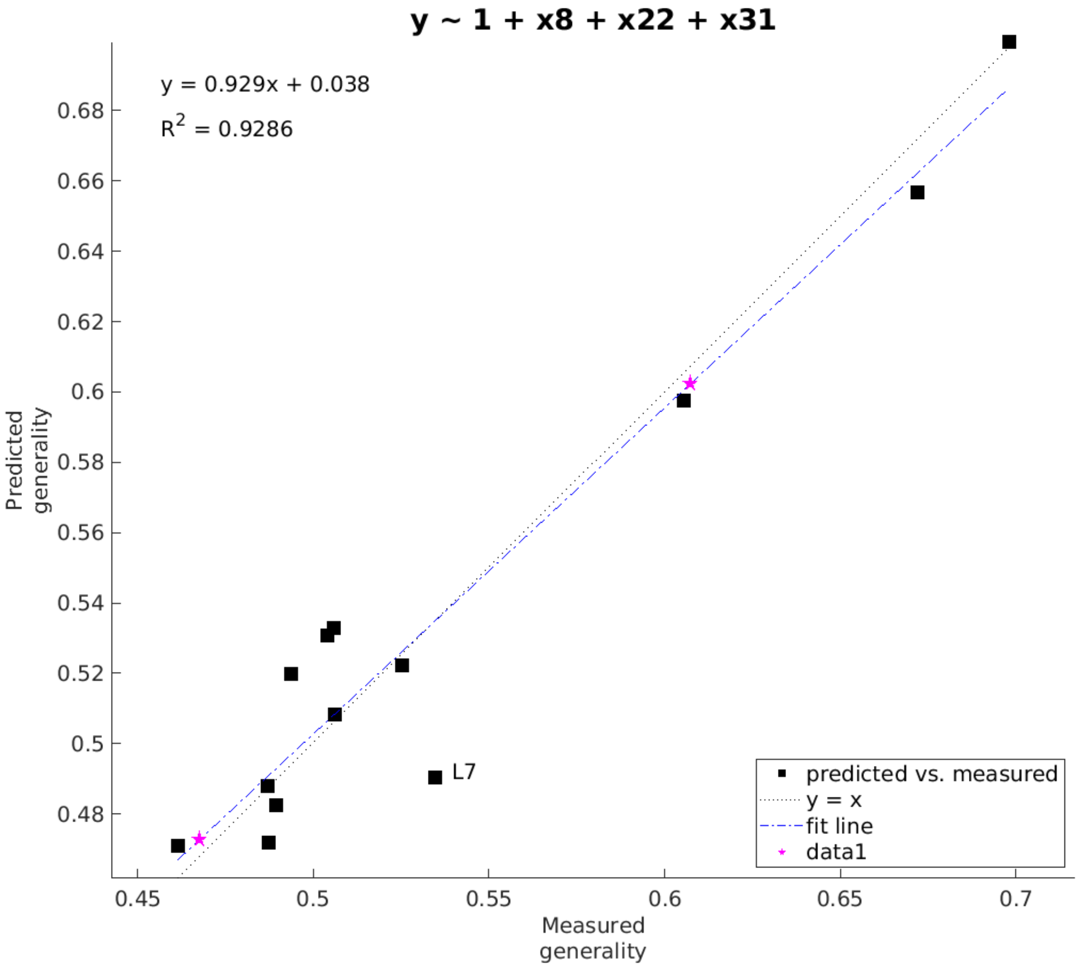


Figure 13: 3-parameter model for Generality

4. Discussion

The final MLR models for generality have either 2 or 3 catalyst descriptors to represent the generality values. The best fit models include B1 and L6 values of the ligands as the most important descriptors which would make sense for a catalyst to be compatible with a huge substrate space and to include its catalytic reactivity. The information extracted from the whole dataset can be represented with these 2 or 3-parameter models. Even though the validation ligand L13 did not come out to be the most general ligand, it had very high score, making it a very general catalyst. Similarly for L12, its predicted generality score was low enough as its true value. The ligand rankings are not very robust and with different predictions and different formula, it experiences slight changes. Hence, the ligands with values on the higher end do not have a robust ranking so the most general catalyst need not be the most general each time. Instead, a combination of values and rankings hold more importance.

5. Conclusion

Even though Catalyst Generality seems a very complex quantity to optimise, it looks possible to obtain a trend of generality scores for a given specific class of reactions. Using the model built from the generalities obtained, rediscovery studies showed that a good model can predict a decent trend of the generality scores. When L12 and L13 were found as the most general and least general catalysts respectively, the model which was built without any information from L12 and L13 could predict an appreciable trend in the reactions. This concept of generality can be further extended with more case studies and different formulae. Extracting relevant information from such big data can help augment research experiments using those new catalysts. It holds the potential to unfold many interesting results that might lead to the community having an entirely new general catalyst to give their experiments a headstart.

References

[1] H. Kim, G. Gerosa, J. Aronow, P. Kasaplar, J. Ouyang, J. B. Lingnau, P. Guerry, C. Farès, and B. List, «A multi-substrate screening approach for the identification of a broadly applicable diels–alder catalyst», Nature communications **10**, 1–6 (2019).

[2] J. R. Kitchin, «Machine learning in catalysis», Nature Catalysis **1**, 230–232 (2018).

[3] C. N. Prieto Kullmer, J. A. Kautzky, S. W. Krska, T. Nowak, S. D. Dreher, and D. W. MacMillan, «Accelerating reaction generality and mechanistic insight through additive mapping», Science **376**, 532–539 (2022).

[4] J. P. Reid and M. S. Sigman, «Holistic prediction of enantioselectivity in asymmetric catalysis», Nature **571**, 343–348 (2019).