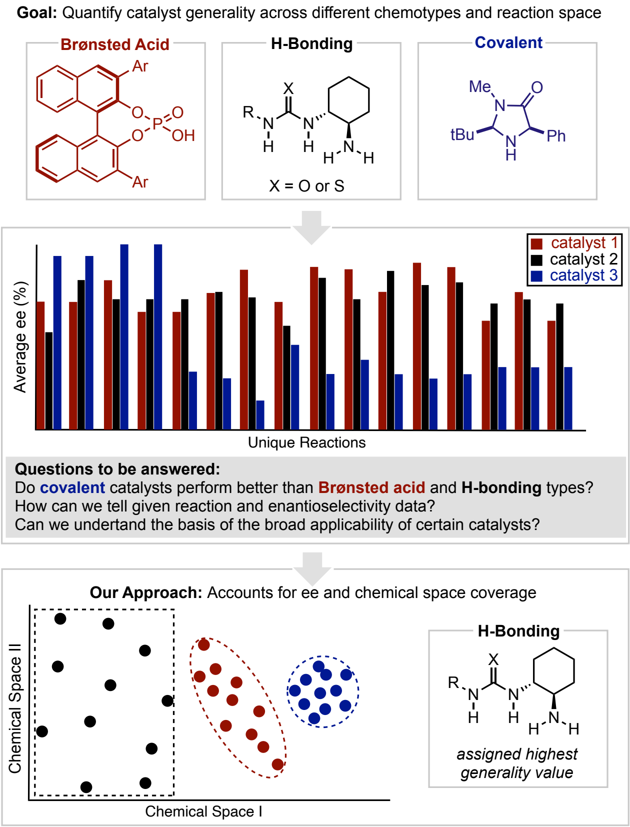
An Unsupervised Machine Learning Workflow for Assigning and Predicting Generality in Asymmetric Catalysis

Isaiah O. Betinol, Saumya Thakur, Jolene P. Reid*

Department of Chemistry, University of British Columbia, Vancouver, British Columbia V6T 1Z1, Canada

ABSTRACT: The development of chiral catalysts that can provide high enantioselectivities across a wide assortment of substrates or reaction range is a priority for many catalyst design efforts. While several approaches are available to aid in the identification of general catalyst systems there is currently no simple procedure for directly measuring how general a given catalyst could be. Herein, we present a catalyst-agnostic workflow centered on unsupervised machine learning that enables the rapid assessment and quantification of catalyst generality. The workflow uses curated literature data sets and reaction descriptors to visualize and cluster chemical space coverage. This reaction network can then be applied to derive a catalyst generality metric through designer equations and interfaced with other regression techniques for general catalyst prediction. As validating case studies, we have successfully applied this method to identify-through-quantification the most general catalyst chemotype for an organocatalytic asymmetric Mannich reaction and predicted the most general chiral phosphoric acid catalyst for the addition of nucleophile to imines. The mechanistic basis for catalyst generality can then be gleaned from the calculated values by deconstructing the contributions of chemical space and enantiomeric excess to the overall result. We conclude that broadly applicable catalysts may be more adaptative to changes in reactant structure because enantioinduction does not rely on a single set of noncovalent interactions. In contrast, some systems work by engaging in robust noncovalent contacts that do not change significantly in nature when the structure of the reaction component is altered. Ultimately, our findings represent a framework for interrogating and predicting catalyst generality, and this strategy should be relevant to other catalytic systems widely applied in asymmetric synthesis.

Chiral catalysts that can be applied to facilitate enantioselective bond constructions between diverse reaction components are prioritized in new synthetic campaigns. Asymmetric organocatalysis exemplifies this where a small subset of catalysts has proven to be remarkably accommodating to changes in reaction component structure.[1–3] To this extent, the privileged status of several catalyst chemotypes has narrowed the focus of asymmetric catalyst discovery, with most modern developments falling within certain boundaries of catalyst space. While this approach of employing broadly applicable catalysts largely drives reaction optimization efforts, it can be especially disadvantageous to underutilized catalytic systems. This issue is very common in enantioselective reaction development where practitioners are generally not willing to explore recently reported or unfamiliar catalysts where considerable synthetic effort is required to generate materials and the results are less certain. Even in cases where extensive reaction surveys have been performed, they rarely include the information necessary to reach reliable conclusions about catalyst generality.[4,5] In other words, while employing different catalyst structures to facilitate a reaction on the same substrate is possible, it is seldom performed, thus diminishing the data available for the direct comparison of catalyst performances. This issue is exasperated when selections are to be made across multiple catalyst chemotypes as conclusions must be drawn from fragmented datasets derived from unique catalyst types. Accordingly, the most general catalyst structure or chemotype can remain largely unknown, hindering catalyst design and application in diverse reaction space. It is for these reasons that the identification and development of general catalyst structures is both necessary and difficult (Figure 1).


Goal: Quantify catalyst generality across different chemotypes and reaction space

Questions to be answered:
Do **covalent** catalysts perform better than **Brønsted acid** and **H-bonding** types?
How can we tell given reaction and enantioselectivity data?
Can we undertand the basis of the broad applicability of certain catalysts?

Our Approach: Accounts for ee and chemical space coverage

Figure 1. Our approach to quantifying catalysts generality accounts for the extent of the chemical space applicable to catalysis in addition to the enantioselectivity values.

Despite this, only recently have research efforts recognized generality as a target property to be optimized for (i.e., yield, selectivity, etc.) with few examples in transition metal-,[6,7] bio-,[8] and photocatalysis.[9] Regarding asymmetric catalysis, recent works have focused on using high-throughput techniques allowing for direct comparative studies of catalyst performance.[10,11] While such protocols assess an important aspect of generality, they do not capture the impact of the catalyst structure in high-dimensional search space. To this end, our group has focused on utilizing comprehensive statistical models that encompass many reaction types and conditions to provide information about the necessary catalyst features for high enantioselectivity across a broad reaction range.[12] Although highly enabling, it is typically limited to one catalyst chemotype which constrains the breadth of structures that can be analyzed in the process.

While these existing approaches probe generality, they typically do not return outcomes that can be interpreted as a value; therefore, an expert chemist is required to analyze the data generated either experimentally or virtually to draw conclusions about the most general catalyst structure. A second limitation in applying these workflows is that the reaction space to be interrogated is not rigorously defined. Therefore, it is difficult to compare catalyst structures that perform well for a large breadth of substrates to those that proceed with high enantioselectivities for a set of similar reactions. Clearly, a new approach that solves the challenging problem of calculating a standalone normalized generality value for a given catalyst chemotype or structure is needed to allow comparisons to made in such difficult cases.

Considering this, we envisioned that deriving a quantitative generality metric would not only provide a physical organic tool for mechanistically assessing general catalyst performance but also allow for the development of a statistical means to optimize general catalyst structures. Herein, we provide a catalyst-agnostic workflow that leverages unsupervised machine learning to capture the breadth of substrates and reactions amenable to a particular catalyst or chemotype (Figure 1). This work represents a new technique in asymmetric catalyst assessment and will prove valuable in applying and developing general catalyst structures for enantioselective synthesis.

General Approach

As a synthetic tool, catalysts that encompass many diverse substrates are more valuable than those systems that facilitate reactions with substrates possessing similar molecular features.[13] Accordingly, our approach to deriving quantitative generality values focuses on accounting for the extent of the chemical space covered by the system in addition to the enantioselectivity values (Figure 2). We realized that a flexible approach that permits end users to define how enantioselectivity and chemical space would be weighted within the workflow would be most useful. To this end, we pursued a method that allows the practitioner to assign an enantioselectivity value that they would consider successful. We postulated that if the catalyst was very general, this enantioselectivity value could be achieved across a wide range of reactions displaying unique structural features. Conversely, if the catalyst was not generalizable the system would provide desirable enantioselectivities only within a narrow portion of reaction space (Figure 2A).

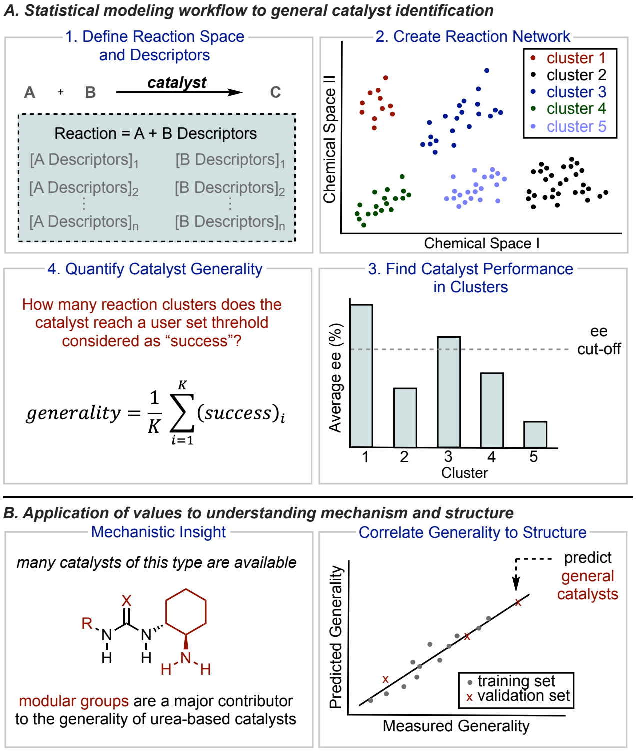

A. Statistical modeling workflow to general catalyst identification
1. Define Reaction Space and Descriptors
A + B $\xrightarrow{catalyst}$ C
Reaction = A + B Descriptors
[A Descriptors]$_1$ [B Descriptors]$_1$
[A Descriptors]$_2$ [B Descriptors]$_2$
[A Descriptors]$_n$ [B Descriptors]$_n$
2. Create Reaction Network
cluster 1
cluster 2
cluster 3
cluster 4
cluster 5
Chemical Space II
Chemical Space I
4. Quantify Catalyst Generality
How many reaction clusters does the catalyst reach a user set threhold considered as "success"?
$generality = \frac{1}{K}\sum_{i=1}^{K}(success)_i$
3. Find Catalyst Performance in Clusters
Average ee (%)
ee cut-off
1
2
3
4
5
Cluster
B. Application of values to understanding mechanism and structure
Mechanistic Insight
many catalysts of this type are available
X
R
N
H
N
H
H
N
H
modular groups are a major contributor to the generality of urea-based catalysts
Correlate Generality to Structure
predict general catalysts
Predicted Generality
Measured Generality
training set
validation set

Figure 2. (A) Overview of the workflow to assign generality scores. (B) The values will then provide mechanistic insight into the features that contribute to catalyst generality and a statistical means to optimize general catalyst structures.

To reveal distinct reaction types within a dataset, reactions with a similar distribution of properties were defined using nonlinear dimensionality reduction and unsupervised clustering.[14] In principle, this step can be implemented with different numerical descriptors and dimensionality reduction algorithms, with the ideal choice likely dependent on the problem at hand. In this study, we utilized either RDKit or quantum mechanical descriptors to efficiently transform the reaction components into numerical descriptors and determined each individual reaction by the linear combination of nucleophile and electrophile properties. The reaction space expressed by these descriptors was reduced by Uniform Manifold Approximation and Projection (UMAP) to provide a way of visualizing the high-dimensional data. UMAP was chosen as the preferred dimensionality reduction algorithm due to its ability to elucidate clusters in complex, non-linear data sets compared to linear algorithms like PCA.[15] The supporting information shows PCA applied to our data sets exhibited increased overlap and ill-defined clusters compared to UMAP. This may suggest an inferior ability to capture important reaction features necessary for a reliable analysis. Importantly, the chemically relevant separation observed from UMAP visualization also serves as a validation for the descriptors chosen. Reducing the dimensionality of the reaction space also functions as a crucial preprocessing step for achieving high performance when clustering (i.e., produce well defined and meaningful partitions) given the noted challenges in clustering high-dimensional data.[16] Specifically, it has been shown that clustering accuracy can drastically increase when first reducing the number of dimensions with UMAP.[17]

While there are unsupervised learning methods that identify the natural clustering present in the data set, we specifically choose k-means as it allows setting the number of clusters via the hyperparameter k. Although the number of clusters is determined by the elbow method, it does provide some degree of flexibility, affording the ability to adapt and update distinct reaction space without compromising the statistical validity of the approach. Specifically with non-ideal data sets where enantioselectivity values and chemical space are not distributed evenly, the choice of number of clusters poses an initial tradeoff decision for the user. With a large number of clusters, one can be more confident in the homogeneity of the cluster; however, a lack of comparative data may lead to a stronger correlation with the popularity of a catalyst. In contrast, a lower number of clusters more adequately adjusts for popularity bias though the clusters may include reactions that are less similar. We show below that an effective approach is to augment experimental data sets with virtual data (i.e., predicted values) that can be obtained from well validated regression models. In this context, it is well known that enantioselectivity often shows complex nonlinear dependencies on the identity of the reaction components and conditions.[18,19] This factor combined with the ability of machine learning models to elucidate non-linear matching effects between input variables are the reasons for why we utilize these methods to account for imbalanced literature data – though having this option is not very common or necessary. Many of the reactions that we consider later can be catalyzed by multiple catalyst chemotypes and the diversity of these structures is immense. Accordingly, it is not straightforward to build well performing regression models that can be used to augment experimental data sets of this type. We were therefore very interested in developing a methodology for generality assignment using unsupervised ML that can also be applied to these situations where the data sets are less than ideal and structurally unique catalyst structures are being interrogated.

The grouping of similar reactions is an inherent feature of dimensionality reduction methods, however, reaction boundaries are often difficult to define.[20] In other words, we are working within the assumption that the reaction space as given by the set dimensions is relatively continuously populated without clusters that are separated by "empty" space. Consequently, the implementation of clustering is a necessary requirement. Taken these steps together, the generality of a catalyst is described as the proportion of clusters with an average performance higher than the user set threshold. This can be formulated as:

(1)

$$generality = \frac{1}{K}\sum_{i=1}^{K}(success)_i$$

where K is the total number of clusters, and successes are defined as clusters wherein the average performance is higher than the set threshold. The implicit supposition taken in this approach is that a reaction point defined within that cluster is representative of all the other reaction points. This is important as it allows generality scores to be derived and compared even in cases where overlapping reaction space between catalyst systems is reduced.

The implementation of our equation does require a suitable success value to be determined - if this value is too high, mildly selective catalysts may not be counted in any cluster and the resulting generality scores will show little variation. If the value is too low, catalysts with very poor enantioselectivity values would not be differentiated from highly enantioselective catalysts. Evidently, this set value is system dependent and users of the method should test a range of values before applying the generality scores.

In considering this value further, it is important to recognize the factors that would impact its reliability. Inclusion of reactions that operate under suboptimal conditions could significantly influence the average enantioselectivity values such that they do not meet or exceed the user set success threshold. Although we determined in our case studies that including this data does not change the final conclusions sufficiently, this may not always be the case. We anticipate that the recorded average enantioselectivity will be particularly altered by the extensiveness of experiments performed during reaction optimization and this will vary widely. Accordingly, we suggest that such bias can be minimized by including only reactions that appear in scope tables. Of note, the inclusion of data pertaining to optimization campaigns is necessary in cases where less general catalysts are also being interrogated in the process. Again, users of the methodology should identify when using or combining certain data types would be problematic.

The steps necessary for calculating catalyst generality closely mirror those involved in building machine learning models for predicting enantioselectivity.[21] Considering this, and the significant research activity in the field, we expect our metric can be simply integrated into these well-established workflows and will find broad applicability in assessing catalyst structures.

Results and Discussion

In this study, we applied our metric to two different case studies which interrogate unique aspects of catalyst generality. In describing these results, we surmised that it would be beneficial to use the first case study as a lesson in assigning the generality values by outlining the necessary steps to be taken. This process is then repeated for the second study, however most of the technical discussion is relegated to the supporting information for this example.

Reaction selection and analysis. The most common assessment of generality in asymmetric catalysis is demonstrated through high enantioselectivity across a diverse set of substrates for a given reaction. Indeed, for many widely explored asymmetric reaction types, different catalyst chemotypes have been applied and in some cases to include the same substrate. However, ranking the different catalyst designs for effective enantiocontrol is difficult to achieve retrospectively as the comparative data required for this direct evaluation is small and the diversity of catalyst structure makes it difficult to trace any differences in superior performance. It is for these reasons that many mechanistic investigations focus on evaluating a single catalyst chemotype. Yet, information on how different structures with unique catalytic modes of activation compare with each other is necessary to determine the catalysts that allow access to the greatest diversity of products.

In this first stage of developing an unsupervised learning platform, we sought to identify a "privileged" reaction type in catalysis wherein many different catalyst chemotypes had been employed. As imine electrophiles and carbonyl nucleophiles are amenable to a variety of catalytic modes of activation, we identified the organocatalytic Mannich reaction as the unifying reaction platform (Figure 3A).[22–25] This reaction type provides a wide range of both substrate and catalyst structures from published sources. We curated a dataset consisting of 3003 reactions from 106 publications wherein diverse chiral H-bonding (1418 reactions), Brønsted acid (256 reactions), covalent catalysts (1182 reactions), and miscellaneous catalysts (147 reactions) had been employed (Figure 3). Considering the large numbers of electrophile and nucleophile structures under evaluation (858 structures), we first implemented RDKit descriptors because these feature sets do not require any calculation. Following the rapid assembly of the nucleophile and electrophile descriptor sets with RDKit, we deployed UMAP to segregate the reaction types for further analysis. Essentially this permitted the reaction space to be visualized by reducing the total number of descriptors from 416 to just two (see SI for more details). It should be noted that because we are interrogating enantioselective catalysts, we only visualize reactions with enantioselectivities measured to be 80% ee or higher. These reaction examples encompassed systems traditionally included in the optimization table (changing solvent, catalyst, loadings, temperature, and time) and those included in the reaction scope (Figure 3B).

To reveal which set of starting materials can effectively undergo different types of catalysis, the points were branded by catalyst chemotype allowing a straightforward analysis of these vast reaction networks. Figure 3B shows that these were visualized as either covalent, H-bond, Brønsted acid (BA), miscellaneous (misc) (a catalyst that doesn't naturally fit into the previous categories), and combinations of catalyst structures to reveal overlapping reaction space. Generally, examples from the literature cover the bottom portion of the reaction space well, while the top depicts more unique reactions and is sparsely sampled. Within this populated space, it is immediately obvious that the various reaction types are reasonably separated by catalytic mode of activation, demonstrating the ability for mechanistic classification with UMAP. This also implies that the RDKit descriptor set contains chemically relevant information that is required to differentiate distinct reaction types responsive to alternative modes of catalysis despite the simplified nature of the parameters. Figure 3B shows that UMAP essentially segregates the reaction network into three important reaction types, those that are amenable to catalysis with H-bond donors (left and upper), Brønsted acid catalyst (middle), and those that facilitate reactivity through covalent bonds (lower right).

Applying k-means to identify and interrogate general catalyst chemotypes. Having demonstrated that UMAP in combination with RDKit descriptors can generate mechanistically relevant reaction networks, we set out to determine the most general catalyst chemotype (i.e., phosphoric acid, cinchona alkaloid, secondary amine, and so on) for the organocatalytic Mannich reaction. The correct identification of the current most general catalyst scaffold would give unique insight in deriving features that lead to generality for this system. Such identification from the literature is currently not possible, with proxy measures like popularity or average selectivity of a catalyst not being reliable metrics.

To simplify the visualization of the reaction space in Figure 3, the H-bond catalysts included cinchona alkaloid, squaramide, and urea-based catalysts. However, such a broad binning of catalyst structures by catalytic mode of activation does not readily allow generality values to be derived from chemotypes of similar structure. Accordingly, with the exception of urea and thioureas which are termed collectively as urea-based catalysts, these have been subdivided into their own distinct category for the generality analysis shown in Figure 4. For the same reasons, covalent catalysts have been further arranged as primary (1°) or secondary (2°) amines.

Prior to assigning a generality score for each catalyst chemotype, the data set was restructured to ensure a comparison of only ideal reaction conditions. As noted above, this is important in removing bias as incorporating results from optimization campaigns could have disproportional influence on the final result. Accordingly, only reactions that applied the optimized conditions (i.e., appeared in the reaction scope tables) were included in this portion of the analysis. Following UMAP reduction to 10 dimensions, the substrate space was clustered with the k means algorithm (k = 12) and the average ee corresponding to each catalyst chemotype was acquired from each cluster. Generality scores were calculated according to equation (1) with a success determined by an 80% ee value. Figure 4 summarizes the results along with extended information like the number of reactions present in the database and average ee.

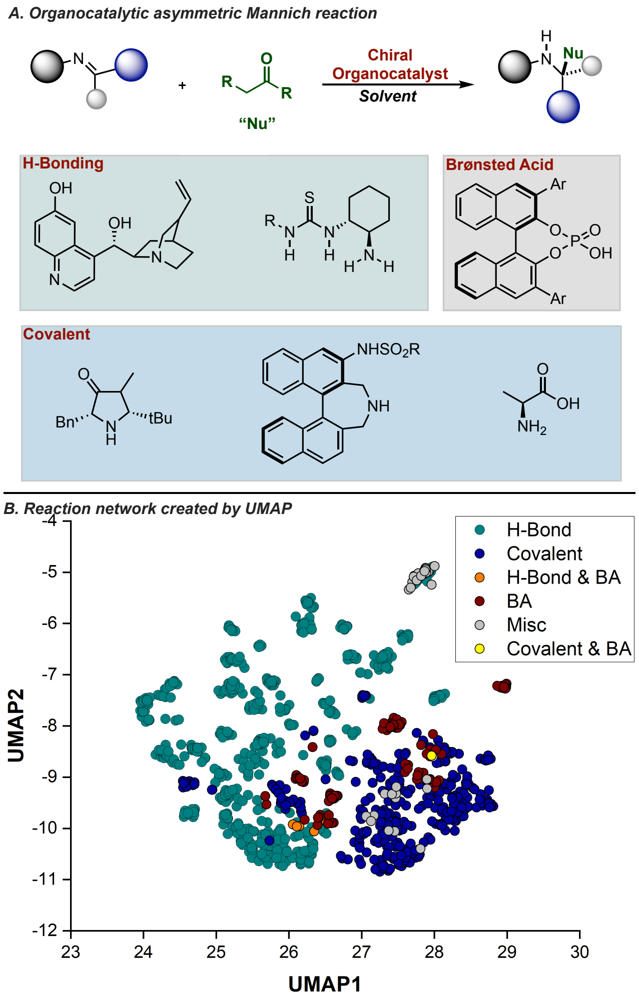


Figure 3. (A) The organocatalytic Mannich reaction. (B) UMAP visualization of the substrate space represented as RDKIT descriptors. Examples of catalysts included in the analysis are shown above. BA refers to Brønsted acid and Misc denotes miscellaneous catalysts.

Although our designer metric is inherently influenced by the reported selectivities, there is minimal correlation between the generality score and the average enantioselectivity value calculated for a catalyst class (Pearson r = -0.26). Because the generality value is derived from chemical space coverage and popular catalysts are typically linked to wider application, naturally, there is a correlation between our generality score and popularity (Pearson r = 0.65). As noted above, we expect this correlation to fluctuate depending on the users set cluster value with higher clusters leading to stronger correlations between the two metrics. There are, however, notable differences between the two values meaning the generality metric reads out a different fundamental feature of catalysis. According to the traditional reaction metrics for gauging generality, secondary amines would be considered the most general for this reaction on the basis of their high enantioselectivity values and substantial popularity. In contrast, urea-based and cinchona alkaloid-based (CA) catalysts measure higher in generality according to our equation. These disparities can be attributed to the breadth of unique reaction space the different catalysts are applied (Figure 5). For example, according to the clustering, many secondary amine catalyzed Mannich reactions are being reported with similar substrates, though high enantioselectivity is observed within the clusters.

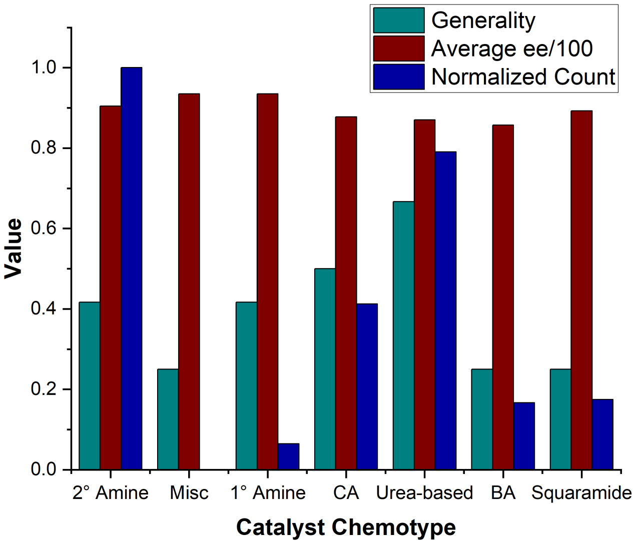


Figure 4. Generality scores obtained for the organocatalytic Mannich reaction with the scaled average ee and min-max normalized popularity of each catalyst chemotype. CA refers to cinchona alkaloid catalysts and urea-based includes both urea and thiourea moieties.

Conversely, urea-based catalysts proceed with generally lower enantioselectivities in overlapping clusters but can catalyze a much larger breadth of reactions. Intriguingly, cinchona alkaloid derived catalysts can also catalyze a greater diversity of substrates than secondary amines despite far less reactions reported with this catalyst. The structural modularity afforded by the catalyst framework is clearly a key feature that accounts for the broad applicability of privileged structures across this reaction type. While a large scope for catalyst modularity can be construed as a possible limitation, it is evident that catalyst optimization requires some level of structural feature tuning. Indeed, H-bonding catalysts like thioureas and cinchona alkaloids have multiple points for introducing a broad set of groups encompassing different steric and electronic properties.[26] In contrast, secondary amines have witnessed significantly less structural diversity at the chiral framework. This can be demonstrated by the number of unique catalysts present in the database that correspond to a particular chemotype. We recorded 27 urea-based catalysts that were good for at least one Mannich reaction which is greater than the number of secondary amines (21) although more reactions have been performed with secondary amines. Similarly, the number of unique cinchona alkaloids (16) is high despite a much lower number of reactions reported.

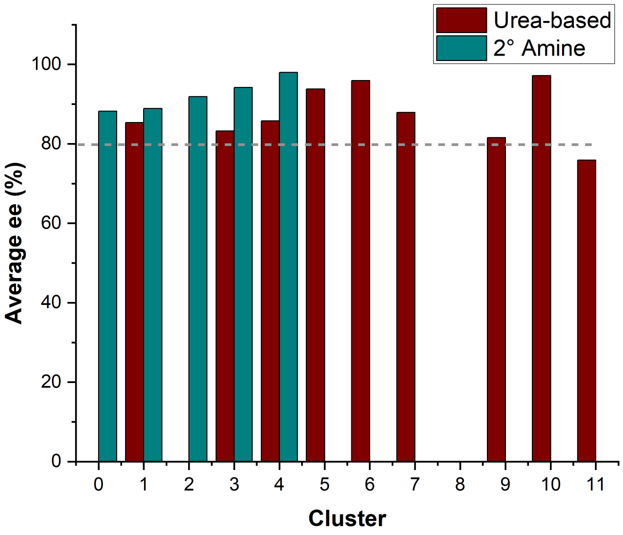
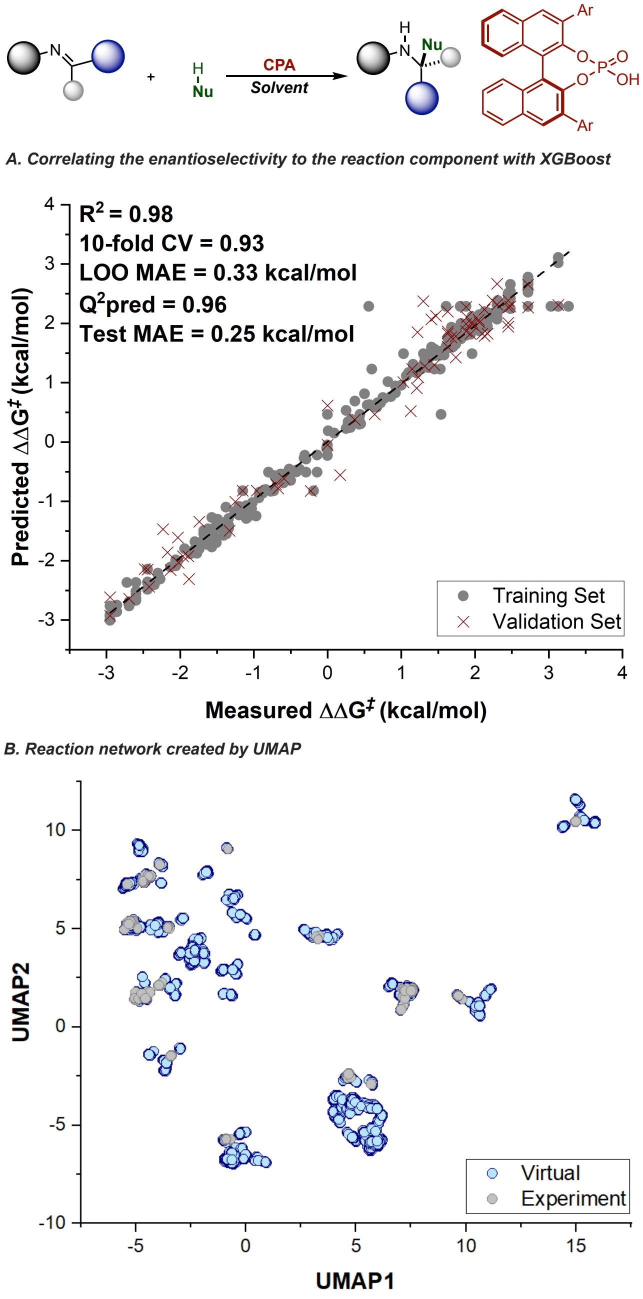


Figure 5. Performance within each cluster for secondary amine (dark cyan) and urea-based (red) catalysts. The success cutoff at 80% ee is shown as a grey dashed line.

Mechanistically, the ability to synthesize many well performing but structurally unique catalysts may provide opportunities to establish different types of noncovalent interactions between the catalyst and various reactants. Essentially, H-bonding catalysts may be more adaptative to changes in reactant structure because enantioinduction does not rely on a single set of noncovalent interactions.[27] Overall, we reason that the privileged nature of such catalysts can be explained by this important factor that relates structure to mechanism. We expect this effect could account for the enhanced generality in scope for other systems, however, insight into the precise structural features contributing to the broad solicitation of certain catalysts remains limited.

Revealing structural features important to generality. To further interrogate the structural effects that permit a chiral catalyst to be impervious to changes in the reaction component structure, multivariate linear regression (MLR) modeling[28–30] was used to complement our generality metric. Specifically, we anticipated that the subtle differences in catalyst generality could be related through the steric and electronic properties of the catalyst. By analyzing the physical organic properties utilized in the mathematical equation, precise structural insight into the molecular features that provide high enantioselectivity for a broad set of reactions could be realized. To probe this idea, we decided to limit our analysis to one catalyst chemotype that has been applied across a reaction range. This would not only provide the necessary structural changes to the reaction component for generality analysis but also incorporate sufficient overlapping catalyst features for modeling.

Considering these constraints, we decided to interrogate the nucleophilic additions to imines catalyzed by chiral phosphoric acids.[31] Earlier studies from our lab established that the enantioselectivity afforded by distinct reaction types can be connected through a mathematical equation that describes the structure of the imine, nucleophile, catalyst and solvent.[32] Although this data set is extensive and achieves substantial coverage of reaction space, the number of occurrences a particular catalyst is used varies widely which may impact the reliability of the generality metrics derived from this literature curated dataset (see above discussion). Consequently, and to explore different substrate-catalyst combinations more comprehensively, we investigated several robust non-linear machine learning (ML) regression techniques for correlating the enantioselectivity outcomes represented as $\Delta\Delta G^{\ddagger}$ to the structure of the reaction components.[33,34] The resulting models could then be deployed to create a virtual dataset[35] by predicting the enantioselectivity for every combination of imine, nucleophile and catalyst contained in the experimental database. Unlike our previous work in correlating and predicting enantioselectivity values, interpolation rather than extrapolation is the overarching goal – this distinction is important as some non-linear models cannot extrapolate outside of the ranges of training data.



A. Correlating the enantioselectivity to the reaction component with XGBoost

B. Reaction network created by UMAP

Figure 6. (A) XGBoost model predicting the $\Delta\Delta G^{\ddagger}$ of the CPA catalyzed nucleophilic addition to imines shown in the reaction scheme above. (B) UMAP visualization of the reactions present in the original dataset (blue) and virtual dataset (red).

To choose the regressor for virtual data set construction, random forest, XGBoost, k-nearest neighbors, and support vector regression models were tested on the experimental dataset.[32] This included a total of 364 reactions that proceed through an E(+ee) or a Z(-ee) imine transition state. Distinguishing between the two imine forms is important in understanding the enantioselectivity outcome as nucleophile addition to the same face will lead to different enantiomers.[36,37] Therefore, the sign of the enantioselectivity value corresponds to a certain imine geometry and this information can be used to predict the absolute product stereochemistry. The training set (80% of the entire dataset) was correlated to the structure of the catalyst, nucleophile, and imine represented by 71 computed parameters. These DFT acquired structural descriptors describe the size and electronic features of the molecules through Sterimol values,[38,39] IR vibrations,[40] NBO charges, energies of molecular orbitals and polarizability. Including parameters to describe the solvent structure here would require significant additional descriptors and this may lead to poor model performance (i.e., overfit). We assume that any subsequent decreases in accuracy in correlating and predicting the $\Delta\Delta G^{\ddagger}$ values will affect both the training and test fits equally, and thus will not change the final conclusions sufficiently to warrant their inclusion. Hyperparameters for each model were tuned using sequential random and grid search algorithms and evaluated by 10-fold cross-validation. Based on its high cross-validation and test set statistics, we chose the XGBoost algorithm shown in Figure 6A to predict the virtual data set (see SI). The high Q2 and low test MAE demonstrate model robustness and considering that every component included in the virtual data set has in some way been represented in model training, the errors in predicting the virtual data can be expected to be similar to the training set.[41,42]

Next, the XGBoost model was applied to predict the enantioselectivity arising from each permutation of imine, nucleophile, and catalyst contained in the experimentally curated data set (125460 reactions consisting of 15 catalysts×8,364 reactants). The differences in structure between tested (experiment) and untested (virtual) reactant combinations is minor (e.g., switching one imine protecting group for another). For example, hemiaminals have been generated from the addition of alcohols to N-Bz protected imines,[43] although mechanistically N-Boc imines should also work well and have been employed as a substrate for reaction with other nucleophile types.[44] Therefore, in this data augmentation tactic we are assuming that the capabilities of the known reactions could be reasonably extended to include other similar electrophile and nucleophile structures that have been successfully applied in at least one other reaction. This straightforward approach may not always be successful (i.e., some reactant combinations may not lead to a reaction), but it does allow us to significantly increase the variety of the reaction partners considered for prediction by the statistical model and ultimately, the reaction space from which the generality value will be derived. Because reaction space will be either active or inactive for all chiral phosphoric acids, this will not affect the final generality score and thus provides a strong incentive for implementing this approach. The UMAP plot in Figure 6B shows the greater coverage of chemical space now covered by the virtual data set. The virtual data points shown in light blue appear in local neighborhoods to the experimental points shown in grey, showing that our augmentation method is simply used to populate the empty space between the known experimental points rather than create genuinely new reaction space where the applicability of our catalysts will be less certain. Importantly, this reaction space now has an enantioselectivity value associated with all tested catalysts and limits the impact of the initially imbalanced dataset.

To ensure the generality scores displayed sufficient variation, a predicted success value of 60% ee was set for this case study. Higher values here led to less selective catalysts not being counted for any cluster. Following UMAP reduction to 10 dimensions, the substrate space was clustered with the k means algorithm (k = 50) and the generality values were determined from the virtual data for the 15 CPA catalysts. The resulting values showed that Ar = 2,4,6-iPr (TRIP) is assigned as the most general catalyst with a value of 0.93 (meaning TRIP will provide on average at least 60% ee in 93% of the clusters), while 9-phenanthryl and 9-anthryl derived chiral phosphoric acids were predicted to be slightly less applicable (Figure 7). This is surprising given the large structural differences between the catalyst systems. Accordingly, we were motivated to understand these results better by deconstructing the contributions of chemical space and enantioselectivity values to the generality score. Figure 8 shows this data simultaneously, where each point represents a cluster of unique reaction space branded by the catalyst.

Inspection of this data shows significant catalyst-substrate matching effects of two superficially structurally similar catalysts: 9-anthryl and 9-phenanthryl (Figure 8A). Interestingly, most substrate clusters provide greater enantioselectivities with one of these catalyst structures. This illustrates that similar catalysts can engage in unique interactions with substrates. The generality values for each catalyst are comparable (9-anthryl = 0.78, 9-phenanthryl = 0.8) which can be explained by the similar number of clusters showing enhanced enantioselectivity with one catalyst. Figure 8B shows the two catalysts with the highest generality scores: TRIP and 9-phenanthryl. While there are some substrates where 9-phenanthryl is the more selective catalyst, it is clear that TRIP is better for a larger range of substrates, explaining the higher generality score.

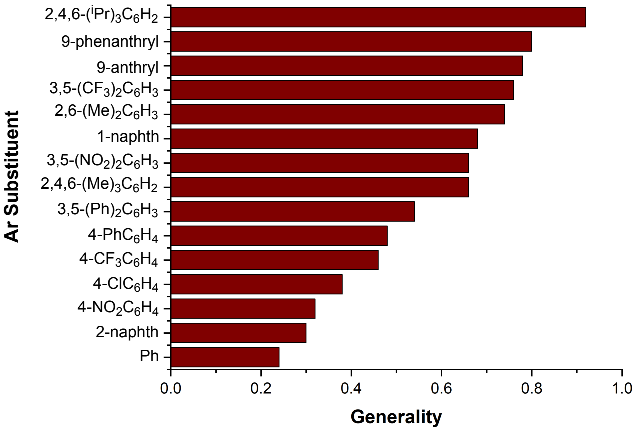
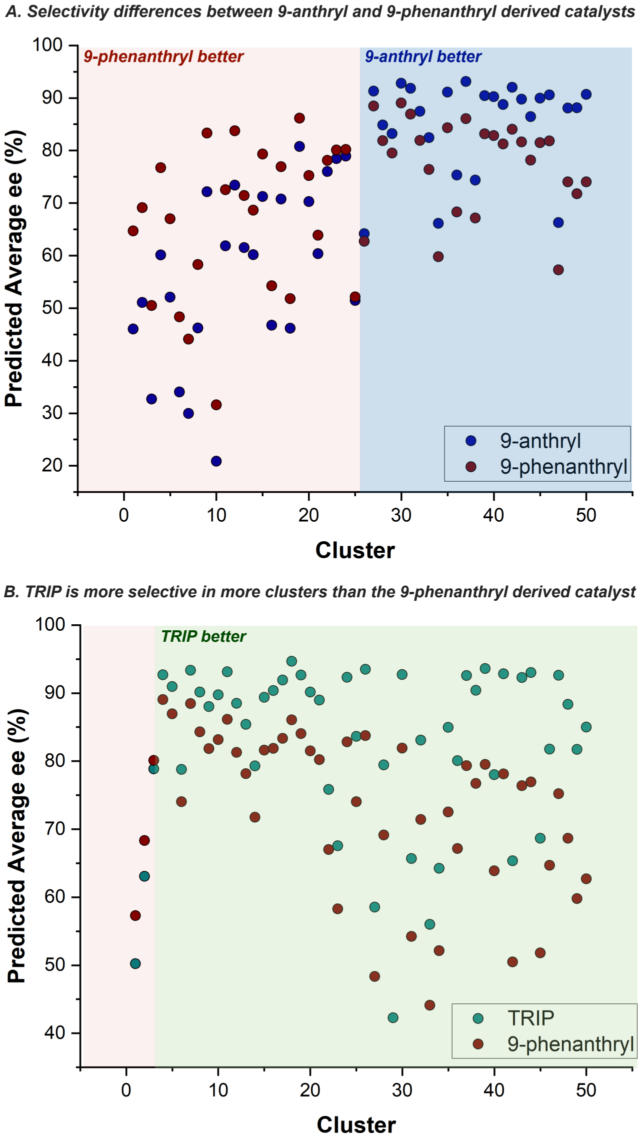


Figure 7. Obtained generality scores for the CPA catalyzed nucleophilic addition to imines.

However, this data set reveals no intuitive global trends regarding the features that impact generality. While higher generality scores were generally associated with larger aromatic groups, increasing the steric bulk did not always produce a more general catalyst (compare 2,4,6-MeC$_6$H$_2$ (0.66) to 2,6-MeC$_6$H$_3$ (0.74). To reveal the precise structural features that contribute to certain catalysts broad applicability, we correlated the molecular features of the catalyst to their associated generality value using MLR. In this approach the catalyst structure represented by a parameter set of steric and electronic descriptors and a forward stepwise linear regression algorithm was applied to the data set. Prior to correlation building, the data set was partitioned 80:20 into training and validation sets. A good relationship was determined using two parameters revealing a simple model consisting of a single steric (Sterimol B5) and electronic (P NMR) term ($R^2 = 0.75$) as shown in Figure 9. Mechanistically, this is consistent with the theory that enantioinduction from CPA catalysts generally stems from repulsive steric interactions between substrates and catalyst and attractive hydrogen-bonding contacts.[31] These features and the analysis shown in Figure 8 could suggest that generally applicable catalysts like TRIP engage in robust non-covalent contacts that do not change significantly in nature when the structure of the reaction component is altered. This mode of generality which relies on transferable non-covalent interactions appears to lead to the highest generality scores.[45]


A. Selectivity differences between 9-anthryl and 9-phenanthryl derived catalysts

B. TRIP is more selective in more clusters than the 9-phenanthryl derived catalyst

Figure 8. Predicted performance within each cluster for the 3 catalysts with the highest generality scores. (A) Comparison of structurally similar catalysts (9-anthryl shown in blue and 9-phenanthryl shown in red). (B) Comparison of structurally dissimilar catalysts (TRIP and 9-phenanthryl, displayed in green and red, respectively).

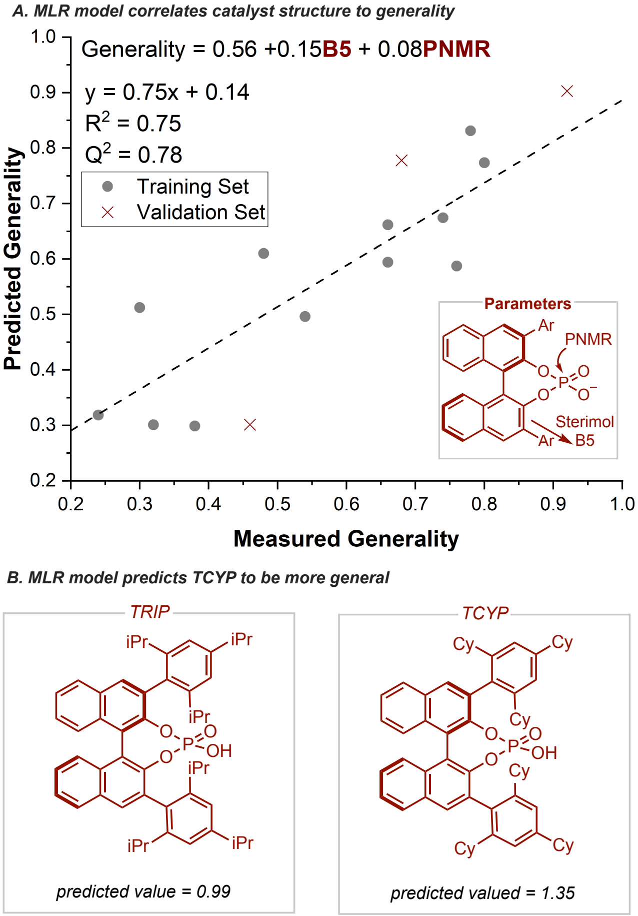


Figure 9. (A) MLR model relating structural parameters derived from the chiral phosphoric acid to the obtained generality values. "Measured Generality" values are generality scores from virtual data generated using a XGBoost model. (B) Utilizing the model to predict a more general catalyst structure.

Generality can also be achieved through adaptable non-covalent contacts where structures like 9-phenanthryl and 9-anthryl may be able to engage in various interactions leading to substrate dependent enantioinduction. These two different modes can be read out as more consistent catalyst performance vs significant substrate-catalyst matching as shown in Figure 8.

A major challenge in the development of our generality metric involved implementing ways to assess the efficacy of the workflow. As generality is not necessarily a physical property that can be measured, it is not straightforward to compare the generality assigned to a catalyst with an experimental value. Accordingly, a final set of studies focused on predicting a more general catalyst for this reaction class. As described above, such prediction tasks are more suited for linear models given the difficulties of extrapolation with decision-tree based models. Our previous work using different statistical tools and experimental methodology had demonstrated that TCYP, a structurally similar but more recently discovered and significantly less utilized catalyst than TRIP, provides higher enantioselectivities across a broader set of reactions than TRIP.[12] Therefore, we were curious to see if our new workflow would lead to the same conclusions. To utilize all available data, all catalysts except for TRIP were added back to the training set and the model retrained. The MLR model was then applied to predict the generality of TRIP and TCYP given the corresponding B5 and NMR terms. Notably, TCYP was not included in any regression model prior to this prediction task. In agreement with our previous study, TCYP is predicted to be a general catalyst with a generality score slightly larger than TRIP (Figure 9B). Although the generality value is predicted to be higher than the maximum of 1 (a consequence of the low cut-off value and boundless linear model), these results demonstrate that new catalysts can be screened for generality using this workflow.

Conclusions

We have developed a new measure centered on unsupervised machine learning to quantify catalyst generality. This approach was evaluated as a method to assess both substrate (broad substrate scope) and reaction generality (applied to construct different bonds) of catalysts. Our metric accounts for the diversity of substrates amenable to catalysis in addition to the recorded enantioselectivity values rather than traditionally subjective measures like popularity or performance. We show that our statistical approach can be applied to identify the most general catalyst chemotype for the organocatalytic Mannich reaction by evaluating the impact of diverse organocatalyst structures. Continued expansion of our approach to diverse bond forming reactions catalyzed by chiral phosphoric acids demonstrates MLR as a statistical means to optimize general catalyst structures. In each example, comparing and deconstructing the generality values reveals several interesting features about the mechanistic basis for generality. Most importantly, robust and adaptable non-covalent interactions are proven to be particularly critical for broad spectrum success with highly diverse substrates and bond forming reactions. We envision that this workflow should facilitate the assessment and mechanistic studies of other enantioselective catalytic reactions and enable the optimization of new general catalyst structure through prediction.

AUTHOR INFORMATION

Corresponding Author

*Jolene P. Reid – Department of Chemistry, University of British Columbia, Vancouver, British Columbia V6T 1Z1, Canada. Email: jreid@chem.ubc.ca

ACKNOWLEDGMENT

Financial support to J.P.R. was provided by the University of British Columbia and the Natural Sciences and Engineering Research Council of Canada (NSERC). I.O.B. acknowledges NSERC for a PGSD Research Fellowship. S.T. is grateful to MITACS for a Globalink Research Internship.

REFERENCES

(1) Bertelsen, S.; Jørgensen, K. A. Organocatalysis—after the Gold Rush. *Chem. Soc. Rev.* **2009**, *38* (8), 2178. https://doi.org/10.1039/b903816g.

(2) Serdyuk, O. V.; Heckel, C. M.; Tsogoeva, S. B. Bifunctional Primary Amine-Thioureas in Asymmetric Organocatalysis. *Org. Biomol. Chem.* **2013**, *11* (41), 7051. https://doi.org/10.1039/c3ob41403e.

(3) Zamfir, A.; Schenker, S.; Freund, M.; Tsogoeva, S. B. Chiral BINOL-Derived Phosphoric Acids: Privileged Brønsted Acid Organocatalysts for C–C Bond Formation Reactions. *Org. Biomol. Chem.* **2010**, 10.1039.c0ob00209g. https://doi.org/10.1039/c0ob00209g.

(4) Kozlowski, M. C. On the Topic of Substrate Scope. *Org. Lett.* **2022**, *24* (40), 7247–7249. https://doi.org/10.1021/acs.orglett.2c03246.

(5) Gensch, T.; Glorius, F. The Straight Dope on the Scope of Chemical Reactions. *Science* **2016**, *352* (6283), 294–295. https://doi.org/10.1126/science.aaf3539.

(6) Prieto Kullmer, C. N.; Kautzky, J. A.; Krska, S. W.; Nowak, T.; Dreher, S. D.; MacMillan, D. W. C. Accelerating Reaction Generality and Mechanistic Insight through Additive Mapping. *Science* **2022**, *376* (6592), 532–539. https://doi.org/10.1126/science.abn1885.

(7) Angello, N. H.; Rathore, V.; Beker, W.; Wołos, A.; Jira, E. R.; Roszak, R.; Wu, T. C.; Schroeder, C. M.; Aspuru-Guzik, A.; Grzybowski, B. A.; Burke, M. D. Closed-Loop Optimization of General Reaction Conditions for Heteroaryl Suzuki-Miyaura Coupling. *Science* **2022**, *378* (6618), 399–405. https://doi.org/10.1126/science.adc8743.

(8) McDonald, A. D.; Higgins, P. M.; Buller, A. R. Substrate Multiplexed Protein Engineering Facilitates Promiscuous Biocatalytic Synthesis. *Nat. Commun.* **2022**, *13* (1), 5242. https://doi.org/10.1038/s41467-022-32789-w.

(9) Speckmeier, E.; Fischer, T. G.; Zeitler, K. A Toolbox Approach To Construct Broadly Applicable Metal-Free Catalysts for Photoredox Chemistry: Deliberate Tuning of Redox Potentials and Importance of Halogens in Donor–Acceptor Cyanoarenes. *J. Am. Chem. Soc.* **2018**, *140* (45), 15353–15365. https://doi.org/10.1021/jacs.8b08933.

(10) Wagen, C. C.; McMinn, S. E.; Kwan, E. E.; Jacobsen, E. N. Screening for Generality in Asymmetric Catalysis. *Nature* **2022**. https://doi.org/10.1038/s41586-022-05263-2.

(11) Kim, H.; Gerosa, G.; Aronow, J.; Kasaplar, P.; Ouyang, J.; Lingnau, J. B.; Guerry, P.; Farès, C.; List, B. A Multi-Substrate Screening Approach for the Identification of a Broadly Applicable Diels–Alder Catalyst. *Nat. Commun.* **2019**, *10* (1), 770. https://doi.org/10.1038/s41467-019-08374-z.

(12) Lai, J.; Li, J.; Betinol, I.; Kuang, Y.; Reid, J. A Statistical Modeling Approach to Catalyst Generality Assessment in Enantioselective Synthesis. *ChemRxiv* **2022**. https://doi.org/10.26434/chemrxiv-2022-80fgz.

(13) Yoon, T. P.; Jacobsen, E. N. Privileged Chiral Catalysts. *Science* **2003**, *299* (5613), 1691–1693. https://doi.org/10.1126/science.1083622.

(14) Hueffel, J. A.; Sperger, T.; Funes-Ardoiz, I.; Ward, J. S.; Rissanen, K.; Schoenebeck, F. Accelerated Dinuclear Palladium Catalyst Identification through Unsupervised Machine Learning. *Science* **2021**, *374* (6571), 1134–1140. https://doi.org/10.1126/science.abj0999.

(15) McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv* **2020**. https://doi.org/10.48550/arXiv.1802.03426.

(16) Steinbach, M.; Ertöz, L.; Kumar, V. The Challenges of Clustering High Dimensional Data. In *New Directions in Statistical Physics*; Wille, L. T., Ed.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2004; pp 273–309. https://doi.org/10.1007/978-3-662-08968-2_16.

(17) Allaoui, M.; Kherfi, M. L.; Cheriet, A. Considerably Improving Clustering Algorithms Using UMAP Dimensionality Reduction Technique: A Comparative Study. In *Image and Signal Processing*; El Moataz, A., Mammass, D., Mansouri, A., Nouboud, F., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, 2020; Vol. 12119, pp 317–325. https://doi.org/10.1007/978-3-030-51935-3_34.

(18) Milo, A.; Neel, A. J.; Toste, F. D.; Sigman, M. S. A Data-Intensive Approach to Mechanistic Elucidation Applied to Chiral Anion Catalysis. *Science* **2015**, *347* (6223), 737–743. https://doi.org/10.1126/science.1261043.

(19) Neel, A. J.; Milo, A.; Sigman, M. S.; Toste, F. D. Enantiodivergent Fluorination of Allylic Alcohols: Data Set Design Reveals Structural Interplay between Achiral Directing Group and Chiral Anion. *J. Am. Chem. Soc.* **2016**, *138* (11), 3863–3875. https://doi.org/10.1021/jacs.6b00356.

(20) van Tilborg, D.; Alenicheva, A.; Grisoni, F. Exposing the Limitations of Molecular Machine Learning with Activity Cliffs. *ChemRxiv* **2022**. https://doi.org/10.26434/chemrxiv-2022-mfq52-v3.

(21) Artrith, N.; Butler, K. T.; Coudert, F.-X.; Han, S.; Isayev, O.; Jain, A.; Walsh, A. Best Practices in Machine Learning for Chemistry. *Nat. Chem.* **2021**, *13* (6), 505–508. https://doi.org/10.1038/s41557-021-00716-z.

(22) Córdova, A. The Direct Catalytic Asymmetric Mannich Reaction. *Acc. Chem. Res.* **2004**, *37* (2), 102–112. https://doi.org/10.1021/ar0302311.

(23) Verkade, J. M. M.; Hemert, L. J. C. van; Quaedflieg, P. J. L. M.; Rutjes, F. P. J. T. Organocatalysed Asymmetric Mannich Reactions. *Chem Soc Rev* **2008**, *37* (1), 29–41. https://doi.org/10.1039/B713885G.

(24) Xiao-Hua, C.; Hui, G.; Bing, X. Recent Progress in the Asymmetric Mannich Reaction. *Eur. J. Chem.* **2012**, *3* (2), 258–266. https://doi.org/10.5155/eurjchem.3.2.258-266.536.

(25) Bagheri, I.; Mohammadi, L.; Zadsirjan, V.; Heravi, M. M. Organocatalyzed Asymmetric Mannich Reaction: An Update. *ChemistrySelect* **2021**, *6* (5), 1008–1066. https://doi.org/10.1002/slct.202003034.

(26) Gallarati, S.; van Gerwen, P.; Laplaza, R.; Vela, S.; Fabrizio, A.; Corminboeuf, C. OSCAR: An Extensive Repository of Chemically and Functionally Diverse Organocatalysts. *Chem. Sci.* **2022**, *13* (46), 13782–13794. https://doi.org/10.1039/D2SC04251G.

(27) Strassfeld, D. A.; Algera, R. F.; Wickens, Z. K.; Jacobsen, E. N. A Case Study in Catalyst Generality: Simultaneous, Highly-Enantioselective Brønsted- and Lewis-Acid Mechanisms in Hydrogen-Bond-Donor Catalyzed Oxetane Openings. *J. Am. Chem. Soc.* **2021**, *143* (25), 9585–9594. https://doi.org/10.1021/jacs.1c03992.

(28) Santiago, C. B.; Guo, J.-Y.; Sigman, M. S. Predictive and Mechanistic Multivariate Linear Regression Models for Reaction Development. *Chem. Sci.* **2018**, *9* (9), 2398–2412. https://doi.org/10.1039/C7SC04679K.

(29) Robinson, S. G.; Sigman, M. S. Integrating Electrochemical and Statistical Analysis Tools for Molecular Design and Mechanistic Understanding. *Acc. Chem. Res.* **2020**, *53* (2), 289–299. https://doi.org/10.1021/acs.accounts.9b00527.

(30) Sigman, M. S.; Harper, K. C.; Bess, E. N.; Milo, A. The Development of Multidimensional Analysis Tools for Asymmetric Catalysis and Beyond. *Acc. Chem. Res.* **2016**, *49* (6), 1292–1301. https://doi.org/10.1021/acs.accounts.6b00194.

(31) Parmar, D.; Sugiono, E.; Raja, S.; Rueping, M. Complete Field Guide to Asymmetric BINOL-Phosphate Derived Brønsted Acid and Metal Catalysis: History and Classification by Mode of Activation; Brønsted Acidity, Hydrogen Bonding, Ion Pairing, and Metal Phosphates. *Chem. Rev.* **2014**, *114* (18), 9047–9153. https://doi.org/10.1021/cr5001496.

(32) Reid, J. P.; Sigman, M. S. Holistic Prediction of Enantioselectivity in Asymmetric Catalysis. *Nature* **2019**, *571* (7765), 343–348. https://doi.org/10.1038/s41586-019-1384-z.

(33) Zahrt, A. F.; Athavale, S. V.; Denmark, S. E. Quantitative Structure–Selectivity Relationships in Enantioselective

Catalysis: Past, Present, and Future. *Chem. Rev.* **2020**, *120* (3), 1620–1689. https://doi.org/10.1021/acs.chemrev.9b00425.

(34) Żurański, A. M.; Martinez Alvarado, J. I.; Shields, B. J.; Doyle, A. G. Predicting Reaction Yields via Supervised Learning. *Acc. Chem. Res.* **2021**, *54* (8), 1856–1865. https://doi.org/10.1021/acs.accounts.0c00770.

(35) Gardner, J. L. A.; Beaulieu, Z. F.; Deringer, V. L. Synthetic Data Enable Experiments in Atomistic Machine Learning. *arXiv* **2022**. https://doi.org/10.48550/ARXIV.2211.16443.

(36) Simón, L.; Goodman, J. M. A Model for the Enantioselectivity of Imine Reactions Catalyzed by BINOL−Phosphoric Acid Catalysts. *J. Org. Chem.* **2011**, *76* (6), 1775–1788. https://doi.org/10.1021/jo102410r.

(37) Reid, J. P.; Simón, L.; Goodman, J. M. A Practical Guide for Predicting the Stereochemistry of Bifunctional Phosphoric Acid Catalyzed Reactions of Imines. *Acc. Chem. Res.* **2016**, *49* (5), 1029–1041. https://doi.org/10.1021/acs.accounts.6b00052.

(38) Brethomé, A. V.; Fletcher, S. P.; Paton, R. S. Conformational Effects on Physical-Organic Descriptors: The Case of Sterimol Steric Parameters. *ACS Catal.* **2019**, *9* (3), 2313–2323. https://doi.org/10.1021/acscatal.8b04043.

(39) Gallegos, L. C.; Luchini, G.; St. John, P. C.; Kim, S.; Paton, R. S. Importance of Engineered and Learned Molecular Representations in Predicting Organic Reactivity, Selectivity, and Chemical Properties. *Acc. Chem. Res.* **2021**, *54* (4), 827–836. https://doi.org/10.1021/acs.accounts.0c00745.

(40) Milo, A.; Bess, E. N.; Sigman, M. S. Interrogating Selectivity in Catalysis Using Molecular Vibrations. *Nature* **2014**, *507* (7491), 210–214. https://doi.org/10.1038/nature13019.

(41) Liu, R.; Wallqvist, A. Molecular Similarity-Based Domain Applicability Metric Efficiently Identifies Out-of-Domain Compounds. *J. Chem. Inf. Model.* **2019**, *59* (1), 181–189. https://doi.org/10.1021/acs.jcim.8b00597.

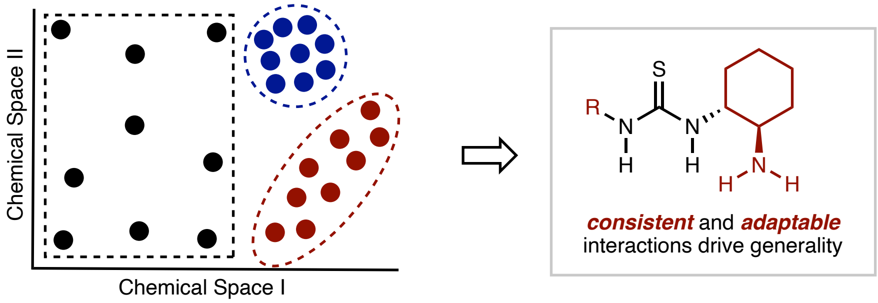
(42) Liu, R.; Glover, K. P.; Feasel, M. G.; Wallqvist, A. General Approach to Estimate Error Bars for Quantitative Structure–Activity Relationship Predictions of Molecular Activity. *J. Chem. Inf. Model.* **2018**, *58* (8), 1561–1575. https://doi.org/10.1021/acs.jcim.8b00114.

(43) Li, G.; Fronczek, F. R.; Antilla, J. C. Catalytic Asymmetric Addition of Alcohols to Imines: Enantioselective Preparation of Chiral *N* , *O* -Aminals. *J. Am. Chem. Soc.* **2008**, *130* (37), 12216–12217. https://doi.org/10.1021/ja8033334.

(44) Rowland, G. B.; Zhang, H.; Rowland, E. B.; Chennamadhavuni, S.; Wang, Y.; Antilla, J. C. Brønsted Acid-Catalyzed Imine Amidation. *J. Am. Chem. Soc.* **2005**, *127* (45), 15696–15697. https://doi.org/10.1021/ja0533085.

(45) Shoja, A.; Reid, J. P. Computational Insights into Privileged Stereocontrolling Interactions Involving Chiral Phosphates and Iminium Intermediates. *J. Am. Chem. Soc.* **2021**, *143* (18), 7209–7215. https://doi.org/10.1021/jacs.1c03829.

TOC Graphic



Quantifying catalyst generality by accounting for enantioselectivity and chemical space coverage with unsupervised machine learning