ACCOUNTS of chemical research

pubs.acs.org/accounts Article

Dreams, False Starts, Dead Ends, and Redemption: A Chronicle of the Evolution of a Chemoinformatic Workflow for the Optimization of Enantioselective Catalysts

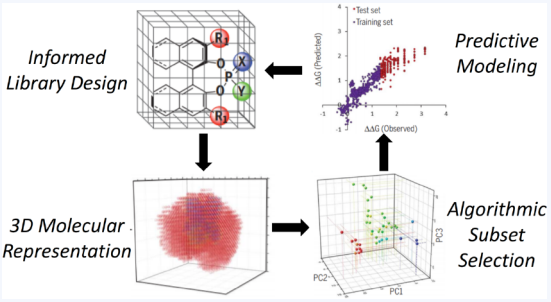
Published as part of the Accounts of Chemical Research special issue "Data Science Meets Chemistry".

N. Ian Rinehart, Andrew F. Zahrt, Jeremy J. Henle, and Scott E. Denmark*

Cite This: *Acc. Chem. Res.* 2021, 54, 2041−2054 Read Online

ACCESS | Metrics & More | Article Recommendations

CONSPECTUS: Catalyst design in enantioselective catalysis has historically been driven by empiricism. In this endeavor, experimentalists attempt to qualitatively identify trends in structure that lead to a desired catalyst function. In this body of work, we lay the groundwork for an improved, alternative workflow that uses quantitative methods to inform decision making at every step of the process. At the outset, we define a library of synthetically accessible permutations of a catalyst scaffold with the philosophy that the library contains every potential catalyst we are willing to make. To represent these chiral molecules, we have developed general 3D representations, which can be calculated for tens of thousands of structures. This defines the total chemical space of a given catalyst scaffold; it is constructed on the basis of catalyst structure only without regard to a specific reaction or mechanism. As such, any algorithmic subset selection method, which is unsupervised (i.e., only considers catalyst structure), should provide an ideal initial screening set for any new reaction that can be catalyzed by that scaffold. Notably, because this design strategy, the same set of catalysts can be used for any reaction that can be catalyzed with that parent catalyst scaffold. These are tested experimentally, and statistical learning tools can be used to create a model relating catalyst structure to catalyst function. Further, this model can be used to predict the performance of each catalyst candidate in the greater database of virtual catalyst candidates. In this way, it is possible estimate the performance of tens of thousands of catalysts by experimentally testing a smaller subset. Using error assessment metrics, it is possible to understand the confidence in new predictions. An experimentalist using this tool can balance the predicted results (reward) with the prediction confidence (risk) when deciding which catalysts to synthesize next in an optimization campaign. These catalysts are synthesized and tested experimentally. At this stage, either the optimization is a success or the predicted values were incorrect and further optimization is required. In the case of the latter, the information can be fed back into the statistical learning model to refine the model, and this iterative process can be used to determine the optimal catalyst. In this body of work, we not only establish this workflow but quantitatively establish how best to execute each step. Herein, we evaluate several 3D molecular representations to determine how best to represent molecules. Several selection protocols are examined to best decide which set of molecules can be used to represent the library of interest. In addition, the number of reactions needed to make accurate, statistical learning models is evaluated. Taken together these components establish a tool ready to progress from the development stage to the utility stage. As such, current research endeavors focus on applying these tools to optimize new reactions.



KEY REFERENCES

- Zahrt, A. F.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. Prediction of Higher Selectivity Catalysts by Computer Driven Workflow and Machine Learning. *Science* **2019**, *363*, eaau5631. DOI: 10.1126/science.aau5631.[1] *First demonstration of fully chemoinformatic workflow for enantioselective catalyst optimization.*
- Henle, J. J.; Zahrt, A. F.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. Development of a Computer-Guided Workflow for Catalyst Optimization. Descriptor Validation, Subset Selection, and Training Set Analysis. *J. Am. Chem. Soc.* **2020**, *142*, 11578−11592.[2] *Comprehensive validation of the various components of the chemoinformatic workflow including descriptor validation,*

Received: December 4, 2020
Published: April 15, 2021

https://doi.org/10.1021/acs.accounts.0c00826
Acc. Chem. Res. 2021, 54, 2041−2054

algorithmic subset selection, and analysis of data requirements for model development.

1. INTRODUCTION

I have not failed, I've just found 10,000 ways that won't work.

Thomas Alva Edison

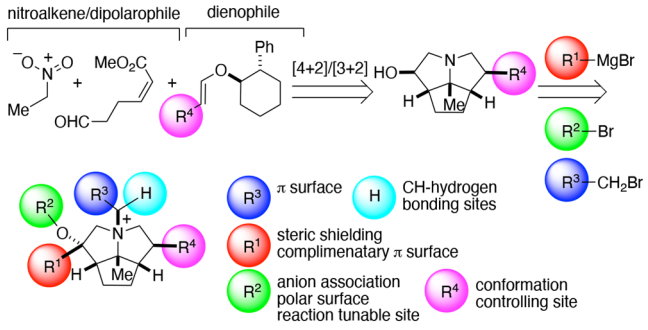
1.1. Disclaimer

This Account is a personal history masquerading as an overview of the development of a new, versatile, fully informatics-guided platform for the identification and optimization of enantioselective catalysts. The corresponding author is taking the opportunity provided by the unique character of *Accounts of Chemical Research* to tell a behind-the-scenes story about how a fundamentally new direction of research evolved. It is a story of unavoidable failures and misguided approaches over the course of over a decade, but it is also a story of the crucial lessons learned from those failures that ultimately led to success. Perhaps most importantly, it is a story of the belief in a vision and the unwavering commitment of many resilient, resourceful, and talented co-workers who never lost faith in that vision.

1.2. How It All Started

In 2007, after 25 years of uninterrupted federal funding for a program on the development and application of tandem [4 + 2]/[3 + 2] cycloadditions of nitroalkenes that lead to over 60 publications, including the enantioselective total synthesis of a dozen alkaloid natural products,[3] it was time to take this program in a new direction. The modular nature of the three-component cycloaddition process particularly in the inter-[4 + 2]/intra-[3 + 2] manifold and availability of multiple sites of diversification especially the quaternization of the pyrrolizidine nitrogen would enable a systematic interrogation of those factors responsible for enantioselectivity in asymmetric phase transfer catalysis, APTC (Scheme 1). Inspired by early studies

Scheme 1. Modular Assembly and Diversification of Chiral APTCs



using Comparative Molecular Field Analysis (CoMFA) from Lipkowitz[4,5] and Kozlowski,[6−9] it was hoped that structural correlates with enantioselectivity could be identified using regression modeling that would enable improvements in catalyst performance. Related studies by Lygo and Hirst[10,11] on a more limited set of APTCs provided additional motivation.

After a heroic effort, over 160 chiral ammonium ions built upon the cyclopentapyrrolizidine (CPP) core were prepared and evaluated[12] in the classic O'Donnell alkylation[13] and a CoMFA model was developed to determine the catalyst design criteria.[14] Although this work generated a 3D-QSAR model with high correlation to enantioselectivity, the model never achieved a high level of predictive accuracy. Owing in part to experimental design, the R^1, R^2, and R^3 positions were varied extensively, but the R^4 position was substituted with only hydrogen and methyl groups. This lack of variation led the model to predict that additional steric bulk at R^4 would result in higher enantioselectivities. Accordingly, lengthy syntheses of ethyl, isopropyl and *tert*-butyl analogs were undertaken and all of these failed to deliver the predicted, superior selectivities. It is clear in hindsight that the model was unable to extrapolate properly the effect of steric bulk at R^4 owing to a lack of variation at this position in the iterations of training set catalysts. Even with a training set consisting of over 160 catalysts, insufficient diversity of the training set from that study resulted in the inability of the model to yield accurate predictions.

The lesson from this years-long enterprise was that to create reliable extrapolative models from 3D-QSAR analyses, better methods are needed to obtain the appropriate training set data. Thus, the next phase required a way to define, for any catalyst scaffold, what constitutes a diverse subset of catalysts and how to quantify meaningful chemical diversity.

This initial failure and many subsequent discussions also stimulated a reevaluation of the project goals and clearer formulation of guiding philosophy for algorithmically guided optimization of enantioselective catalysts.

2. RESEARCH PHILOSOPHY

2.1. Mechanistic and Intuition-Guided Descriptors vs Intuition-Agnostic Descriptors

The use of Linear Free Energy Relationships (LFERs) in modeling selectivity of chiral catalysts has found increasing application in asymmetric transformations, of particular note is the extensive body of work by Sigman and co-workers.[15−19] This workflow identifies catalyst features—typically physically interpretable quantities from physical organic parameters or spectroscopic values—that can be correlated to selectivity for a subset of catalysts. After fitting a multivariate correlation, one can then identify the relative importance (by feature weights) for each descriptor used in these models. The main advantage of using physically interpretable descriptors and modeling with them in this manner is that one can then formulate a mechanistic hypothesis on the basis of the features which are highly correlated with selectivity. This approach requires that (1) features can be found that correlate to selectivity, typically many candidates are found from supervised feature selection, (2) it is possible to select among multiple predictive models that use different features to make predictions, and (3) reliable interpretation of the model can provide meaningful mechanistic insight.

Our approach is complementary to the one described above in that the primary goal is optimization of a reaction by accurate prediction of improved selectivities instead of mechanism elucidation. With a description of catalyst shape and properties, and enough data, models can identify correlations between the catalyst and its selectivity without any knowledge of the mechanism. Since understanding which catalyst/substrate interactions engender selectivity is not required for reaction optimization, we have demonstrated a method wherein a researcher can systematically evaluate a

https://doi.org/10.1021/acs.accounts.0c00826
Acc. Chem. Res. 2021, 54, 2041−2054

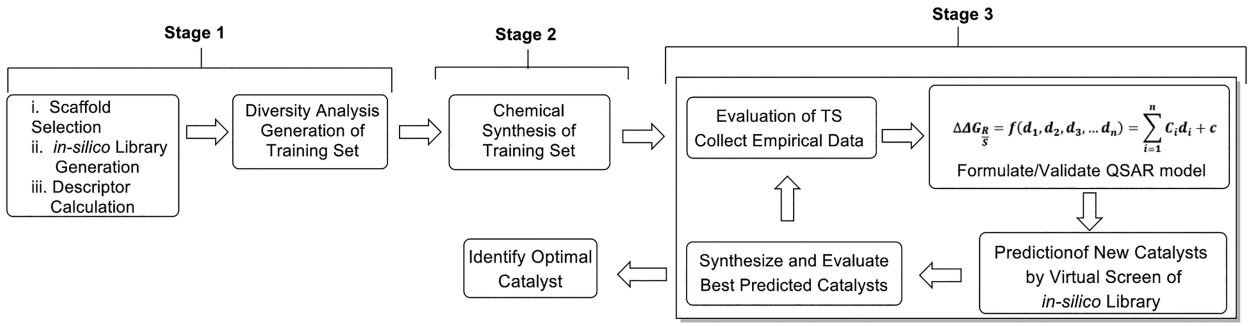


Figure 1. Comprehensive chemoinformatic workflow developed in these laboratories.

library of possible catalyst structures in order to optimize a reaction without simultaneously exploring the reaction mechanism. This approach does not exclude mechanistic inquiry because it is then valid to use the same models to formulate a mechanistic hypothesis as long as it is experimentally tested. This approach has the following requirements: (1) assuring sufficient catalyst and substrate diversity in a data set and (2) using a description of the catalyst which is agnostic to mechanism and captures the steric and electronic properties of the catalyst. If those requirements are met, then the resulting models will be effective at predicting the selectivity of any catalyst with that scaffold, even those which are more selective than any of the training data. Notably, statistical learning approaches have recently gained traction in the chemical sciences for such prediction tasks.[20−42]

3. FULLY CHEMOINFORMATIC-GUIDED WORKFLOW

Despite years of effort to synthesize the CPP library and the conclusion that it was still insufficiently diverse to enable reliable predictive modeling, a conceptually new approach was needed that expanded the scope of chemical space and simultaneously reduced the experimental overhead in generating training data. It became apparent that fulfilling those criteria could be achieved through using algorithmically guided steps from beginning to end. The conceptual framework for this workflow is schematically illustrated in Figure 1.

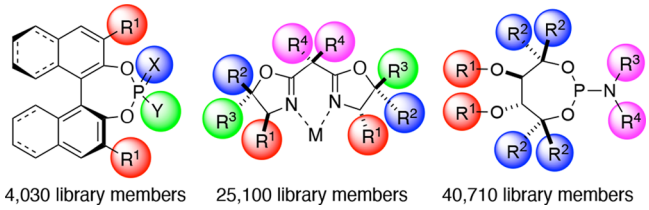


Figure 2. Ligand scaffolds used to generate in silico libraries for cheminformatic development.

The first stage focuses on the generation of an in silico library that contains every synthetically accessible permutation possible for the selected catalyst scaffold. As such, the in silico library is intentionally agnostic to any perceived utility for structural modification. Descriptors are then calculated for each library member, numerically representing the properties in such a way that they can be understood within the context of statistical modeling. The collation of the descriptors of the entire library then defines the chemical space in which the library exists, giving a context to these properties with respect to catalyst identity. This chemical space is then analyzed using different types of selection protocols to identify a subset of these catalysts which comprises a training set whose properties represent the breadth of the library itself within the chemical space. This training set is termed a Universal Training Set (UTS), as it can be used universally with any reaction in which the scaffold can be applied.

The second stage is the synthesis of the UTS. The goal of stage 2 is to acquire enough of each catalyst to be used in reaction screening and optimization. Should synthetic difficulties arise, the nearest-neighbor to the inaccessible one in the chemical space can be selected.

In third stage of the workflow, the UTS is evaluated in a reaction of interest. Experimental data (e.g., enantioselectivity, diastereoselectivity, yield, etc.) is tabulated, from which models are generated using a variety of statistical modeling techniques. Once validated models are acquired, they are used to evaluate every member of the *in silico* library. Those catalysts predicted to be more selective are chosen for synthesis and experimental validation. Once synthesized and tested, two outcomes are possible: (1) the prediction is accurate and the reaction is optimized or, (2) the prediction was inaccurate and the model must be refined. In the case of the latter, this data can be fed back into the training data, iteratively refining the model until optimization is achieved. Of course, there exists the possibility that even with the increased scope of diversity within the in silico library that the optimal catalyst falls short of the desired activity. In this case, a different catalyst scaffold must be considered.

3.1. In Silico Libraries: Chemical Space Diversity Accomplished

In view of the failure of the APTC effort despite synthesizing ∼160 catalyst structures, we recognized the need to generate a much large ensemble of potentially synthesizable compounds to greatly expand the chemical space to be explored. This step was accomplished by constructing large in silico libraries built around several privileged catalyst scaffolds, such as BINOL phosphoric acids (BPA),[43] methylene(bisoxazoline) (BOX),[44] and TADDOL phosphoramidites[45] (Figure 2).

For each catalyst scaffold, a series of substituent databases were created from which the each of the points of diversity illustrated by the colored spheres was populated and then members of the in silico library were constructed in a combinatorial manner. From this process, in silico libraries were generated for three scaffolds of interest: (1) 4030 BPA-derived catalysts, (2) 25 100 BOX ligands, and (3) 40 710 TADDOL-phosphoramidite ligands. Descriptors for each library member were calculated and training subsets were algorithmically selected and synthesized. These training sets were used in optimization campaigns, but the models

https://doi.org/10.1021/acs.accounts.0c00826
Acc. Chem. Res. 2021, 54, 2041−2054

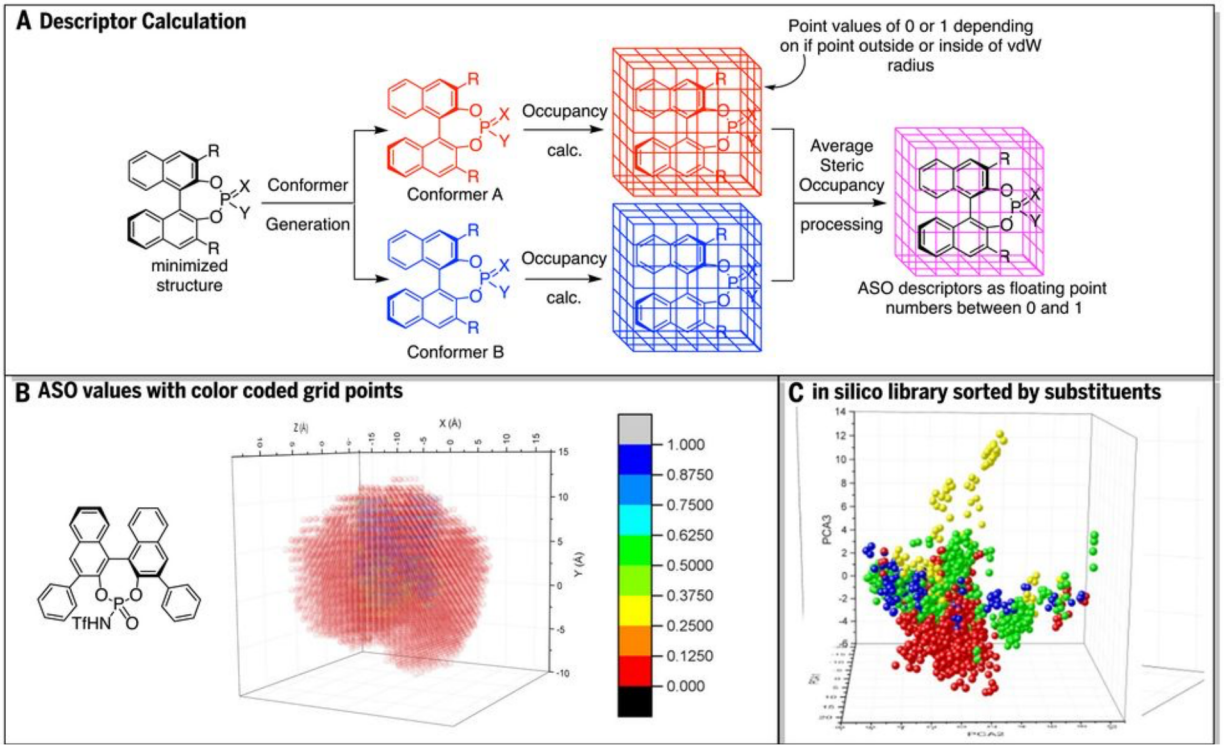


Figure 3. (A) ASO descriptor calculation, (B) ASO feature plot of a BINOL-phosphoryltriflamide, and (C) plot of first three principal components of in silico BPA library. Adapted with permission from ref 1. Copyright 2019 American Association for the Advancement of Science.

generated were not capable of making accurate extrapolative predictions from the data. This failure clearly pointed to the inability of the descriptor sets to capture the necessary chemical features of the library members—new descriptors were needed.

3.2. Chemical Descriptor Development and Evaluation

For algorithmic selection and subsequent modeling to be successful, adequate representation of chemical structure is imperative. In particular, discriminating between similar structures with the same parent scaffold present in these in silico libraries is particularly challenging. In fact, for most common catalyst scaffolds, between 40% and 70% of the atoms and connectivity remains constant between catalysts with vastly different properties. For these reasons, 3D descriptors were employed despite the greater computational cost.

A well-established method of 3D-QSAR modeling employs molecular fields.[4−12,46,47] These methods use molecular interaction or molecular indicator fields to capture the 3D properties of molecules. Molecular interaction fields capture the interaction energy between the molecule of interest and a probe particle at different positions in space around the molecule. Indicator fields, by contrast, denote whether an atom in the molecule occupies a region of space by assigning some value (e.g., binary indicator or atomic property) to gridpoints in that space if they overlap with an atom. Both methods require the alignment of molecules. To circumvent this necessity, grid independent descriptors (GRIND) were developed which use internal distances instead of requiring alignment.[48] Over several years, we have implemented more variations than can be enumerated here; however, every method used to represent a whole molecule failed most likely because only a single, ground-state conformer was used. Obviously, this approach is chemically flawed but lacking the ability to identify and calculate reactive conformers for each library member using quantum chemical methods, no way forward was possible with the current workflow. Clearly, a new molecular representation which captures the dynamic, 3D-nature of molecular structures had to be developed to overcome the limitations encountered in both raw MIF descriptors and GRINDs. Then, those novel descriptors would have to be algorithmically validated, then used to create a UTS to provide experimental data from a new transformation that enables accurate, predictive models to be generated if this program was going to succeed.

3.2.1. Average Steric Occupancy Descriptor: The Solution to the Conformer Issue. Because of the early success of molecular field approaches in asymmetric catalysis by Kozlowski,[6,7] Lipkowitz,[4,5] and Hirst,[10,11] as well as our own efforts,[14] we chose to continue implementing this family of molecular representation. In particular, we were inspired Hirst and co-workers development of 3.5-D QSSR[49,50] and application to asymmetric phase transfer catalysis. In this protocol, multiple conformers of molecules are used in the construction of the molecular field. Accordingly, new descriptors were developed termed Average Steric Occupancy (ASO) descriptors. These descriptors are calculated by first generating a conformer distribution for every member of the in silico library of catalyst candidates. Every conformer of every library member is then superimposed with respect to a common core scaffold and placed in a common grid. The ASO descriptors are then generated from the collection of conformers for a given molecule. For each conformer, every grid point is queried if it falls within the van der Waals radius of an atom in the molecule and assigned a binary value: yes = 1, no = 0 (Figure 3A). This process is repeated for every conformer for a given molecule. Thus, if an individual catalyst candidate has *n* conformers, possible values at each grid point

https://doi.org/10.1021/acs.accounts.0c00826
Acc. Chem. Res. 2021, 54, 2041−2054

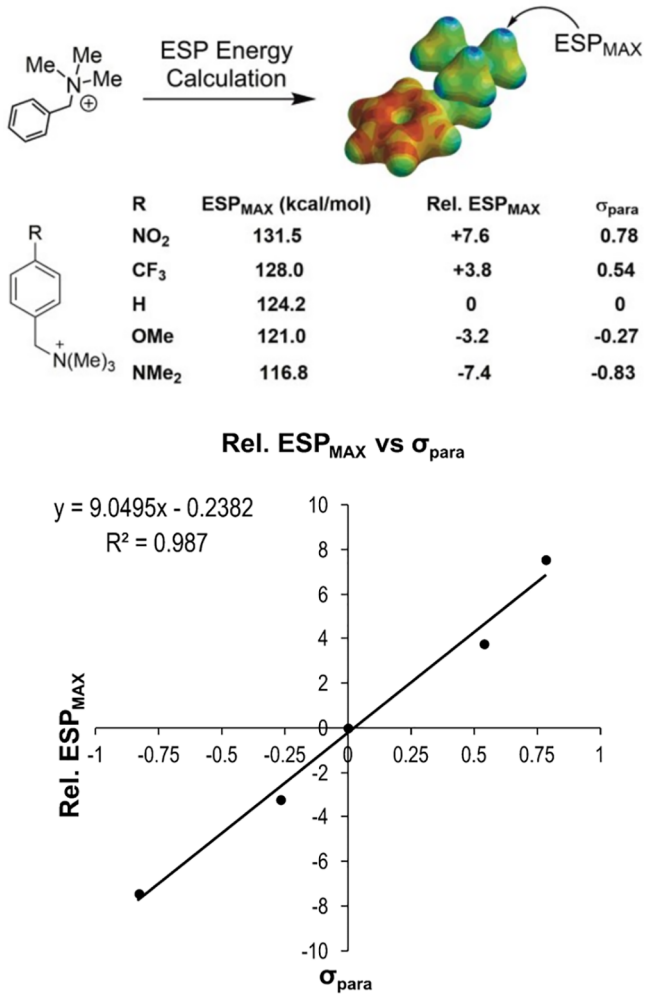


Figure 4. Generation and validation of ESPMax parameter. Adapted with permission from ref 2. Copyright 2020 American Chemical Society.

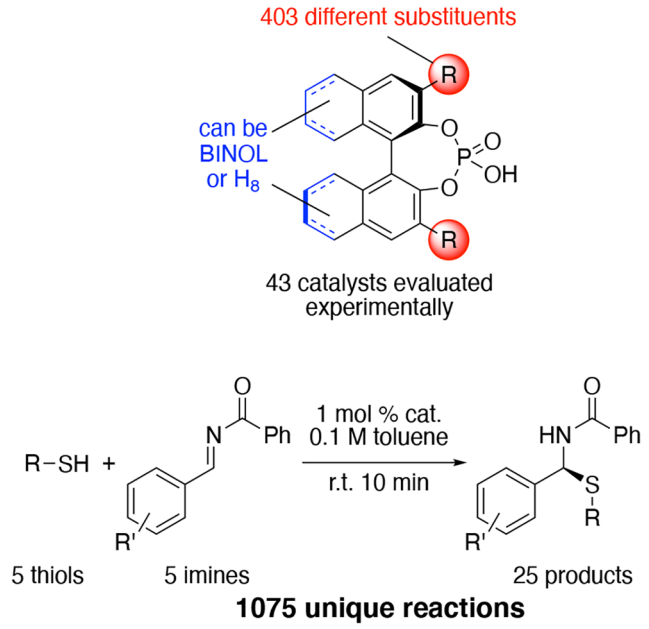


Figure 5. Model reaction system for evaluation of the computational workflow.

range from 0 to n for that molecule. The occupancy values at each grid point are then normalized to the number of conformers, so that every grid point contains a value between 0 and 1. These numbers are the ASO descriptors (Figure 3B). Principal Component Analysis (PCA) was used to visualize the high-dimensional catalyst space for an in silico library of BPAs in three dimensions. These catalysts were color coded on the basis of the designations of 3,3′-substituent classes (Figure 3C). Qualitatively, the different classes of catalyst are separated, indicating that catalyst subclasses are distinguishable using ASO.[51]

3.2.2. Calculable Electronic Descriptor. A calculable electronic descriptor employing electrostatic potential mapping that can capture through-bond effects was invented. This descriptor was an important step because experimental Hammett parameters are not available for the diverse range of substituents within large in silico catalyst libraries. The process for calculating the electrostatic potential maximum (EXPMax) is simple: first, a molecular fragment of interest is appended to a tetramethylammonium ion, then an electrostatic potential (ESP) surface is calculated for that molecule (Figure 4). Then, the maximum charge on that surface is identified. An excellent correlation coefficient (R^2 = 0.98) between ESPMax and Hammett parameters was found.[51] It is not our intention to use this fit to calculate ESPMax for any substituent, but the strong correlation using these five samples means that meaningful electronic information is encoded in this descriptor. This descriptor has recently found use in a QSAR setting for understanding Gram negative cell permeability of positively charged nitrogen compounds.[52]

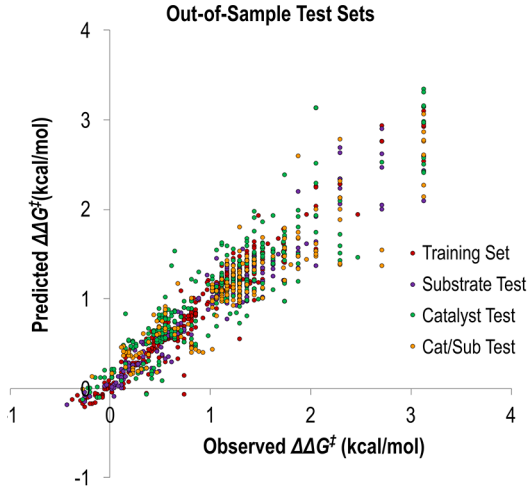
With these molecular representations identified, the workflow was benchmarked on a model system to evaluate if (1) the descriptors were adequate in representing molecular structure and could be used to make models predicting reaction outcomes and (2) if suboptimal reaction conditions could be used to make a model capable of identifying optimal catalyst structure. To this end, this workflow was prototyped with a library of chiral Brønsted acids in a model reaction developed in by Antilla and co-workers (Figure 5).[53]

Investigations were performed with data from 43 different catalysts, which were synthesized and tested experimentally with 25 substrate combinations per catalyst, giving a total of 1075 data points. Although multiple models demonstrated acceptable performance (q^2 > 0.6, R^2 > 0.85 for test set), support vector machines gave the best performance on the basis of mean absolute deviation of predicted values (Figure 6a). To simulate the optimization of reaction using nonoptimal data, the data points below 80% *ee* were used to construct a model that quantitatively reproduced the experimental values of the more selective reactions (Figure 6b). More importantly, the relative catalyst efficacy, on the basis of the average selectivity of the reactions in which that catalyst was employed, matches what is experimentally observed. The success of these studies demonstrates the ability of the novel calculable features (ASO, various electronic parameters) to be used with modern machine learning methods to predict catalyst efficacy with remarkable accuracy. Further, this work has provided a large data set to use in other ML studies.[54−56]

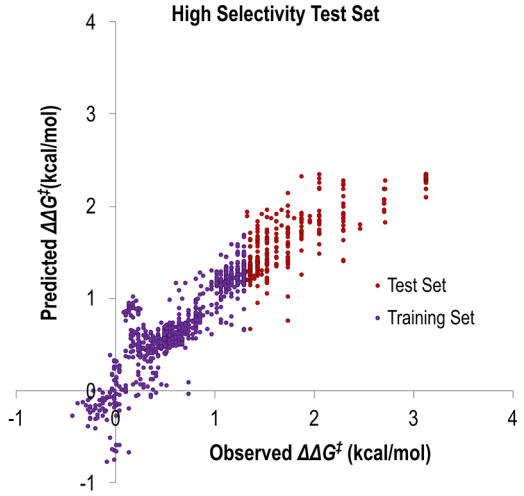
3.3. Benchmarking and Validation of Descriptors

To validate that the model performance using ASO (which uses multiple conformers) is superior to those developed from Steric Indicator Field (SIF) or Molecular Interaction Field

https://doi.org/10.1021/acs.accounts.0c00826
Acc. Chem. Res. 2021, 54, 2041−2054



	MAE kcal/mol	RMSE kcal/mol	R^2
Train	0.09	0.15	0.952
Substrate	0.16	0.22	0.917
Catalyst	0.21	0.29	0.831
Cat/Sub	0.24	0.33	0.830



	MAE (kcal/mol)	RMSE (kcal/mol)	R^2
Train	0.15	0.22	0.784
Test	0.28	0.39	0.652

Figure 6. Predicted versus observed plots for training and test data.

(MIF) descriptors (which use a single conformer), models were trained and cross-validated using 384 examples of the same data set. An external test set of 691 reactions which contained out-of-sample predictions was used for further validation.[2,57] The primary goal of QSSR models is to making predictions for novel examples; therefore, the reliability of predictions for novel products must be assessed when training models. This assessment is best done by using out of sample predictions in test sets. This concept has been adopted in the chemical sciences[58,59] and should be implemented in ML studies in enantioselective catalysis, particularly when dealing with combinatorial data sets. When chemical entities are present in *both* training and test data, the performance of the external test set is often overoptimistic owing to the model identifying patterns in the data rather than fitting to chemically relevant information. Out-of-sample predictions are a more robust test for QSSR performance in ML studies because they prevent these models from arising from pattern recognition.

A modeling experiment was conducted to demonstrate that data set partitioning methods which force out-of-sample predictions can prevent fitting to random patterns in data instead of chemical information. Two data set partition schemes of the BPA data set were constructed in which the models used features, which either contained chemical information or did not. The first partition of 384 training examples was selected randomly from the 1075 unique reactions (the remaining 691 reactions constituted the external test set). The second set of training data was selected by a scheme which defined "train" and "test" entities in the reaction (24 catalysts and 16 products were "train" entities and the remaining 19 catalysts and 9 products were "test" entities), such that any reaction containing a "test" entity was a test set example. This nonrandom partition scheme forces the construction of a test set comprised of out-of-sample predictions. These two data partitioning methods were used to evaluate different descriptor classes, with random features and one-hot encoding used as control experiments. If the control experiments perform as well as the chemical features, it is likely that the models are fitting to the intrinsic structure in the data rather than learning chemically relevant patterns.

The striking results of this study are summarized in Figure 7 (top). Test set performance of all models using random data partitioning were superficially superior (on the basis of MAE) to test sets with out-of-sample predictions. In fact, the performance of the three models using chemical features did not differ significantly from the random control, indicating that test set performance is a result from information shared between train and test sets. As such, conclusions regarding feature performance cannot be made with this data partitioning scheme.

In stark contrast, both the random and 1-hot encoding features *absolutely failed* to make out-of-sample predictions with the nonrandom data partition. The blue squares in Figure 7 (bottom) show examples in which all reaction components are out of sample (imine, thiol, and catalyst). Whereas the control experiments produce no correlation between predicted and observed values, all chemical descriptors depict some correlation with performance ranked ASO > SIF > MIF for this experiment. Clearly, this experimental design is necessary to draw meaningful conclusions, in this case demonstrating the superiority of the ASO descriptors its single-conformer counterparts.

Although ASO performs the best in this study, we hypothesized that the superior performance of ASO compared to SIF would increase as the catalyst structures being modeled become more flexible. To test this hypothesis, and the generality of this molecular representation, a literature data set of enantioselective O'Donnell alkylations using cinchona alkaloids was examined (Figure 8).[11] Using this 88-member data set, 70 data points were used to train and cross-validate Projection to Latent Structure (PLS) models with 18-member tests sets. The result was clear; using SIF and an Electronic Indicator Field (EIF) provided a poor correlation between predicted and observed selectivity of test set predictions (R^2 =

https://doi.org/10.1021/acs.accounts.0c00826
Acc. Chem. Res. 2021, 54, 2041−2054

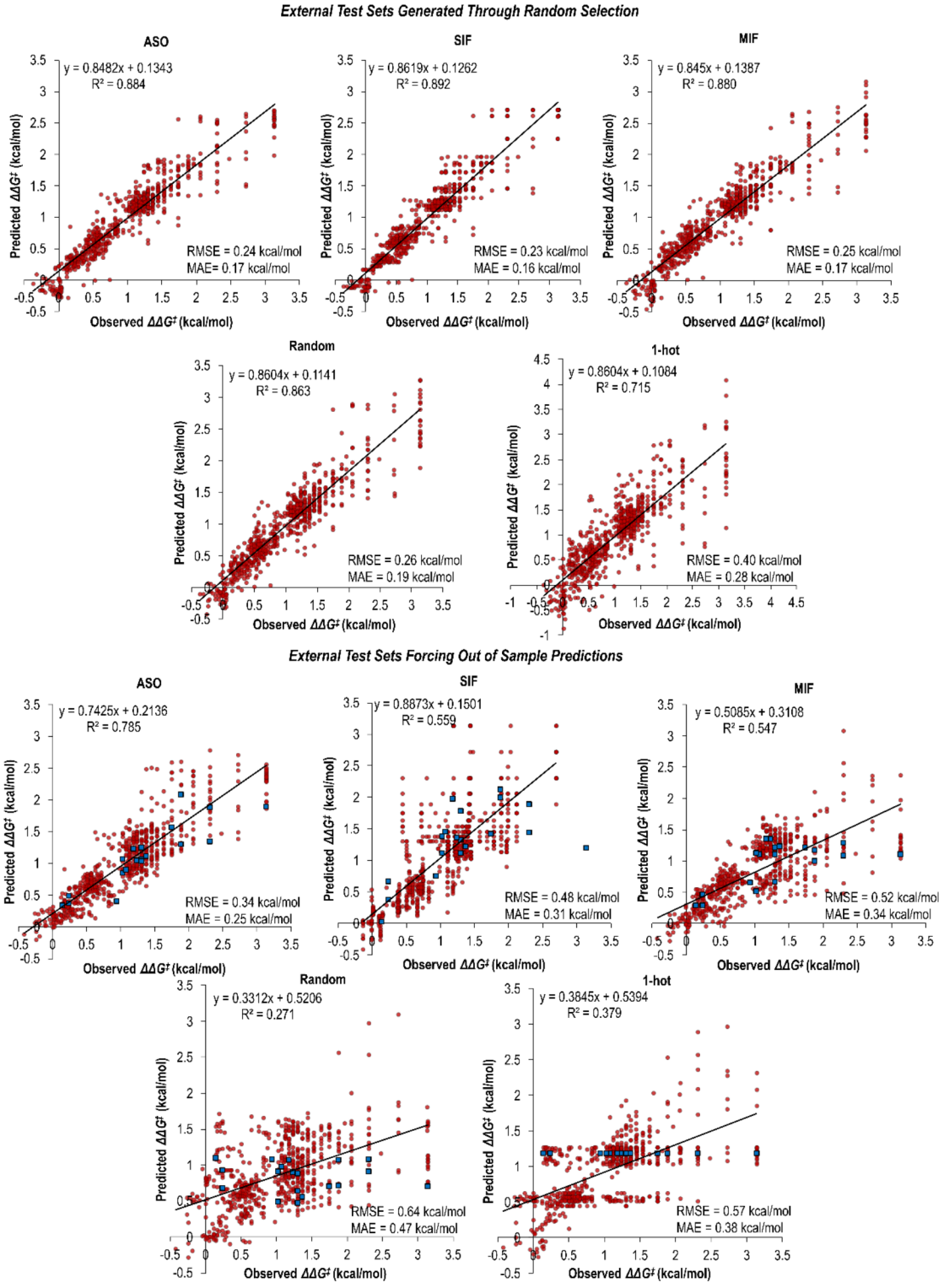


Figure 7. Comparison of random and nonrandom data set partitioning with five different featurization methods. Adapted with permission from ref 56. Copyright 2020 American Chemical Society.

0.368) while models using ASO and Average Electronic Indicator Field (AEIF) descriptors had a higher correlation coefficient for predicted selectivities of test set reactions (R^2 = 0.768). These results suggest that more flexible catalyst systems will show an increased benefit from using multiple conformers (ASO) instead of one (SIF) when modeling asymmetric transformations.

3.4. Algorithmic Subset Selection

When following the established workflow (Figure 1), an experimentalist must first select which catalysts will be synthesized first for data collection and model calculation to commence. One design principle of this workflow is the implementation of a selection protocol that spans the breadth of chemical space contained in the in silico library. As such,

https://doi.org/10.1021/acs.accounts.0c00826
Acc. Chem. Res. 2021, 54, 2041−2054

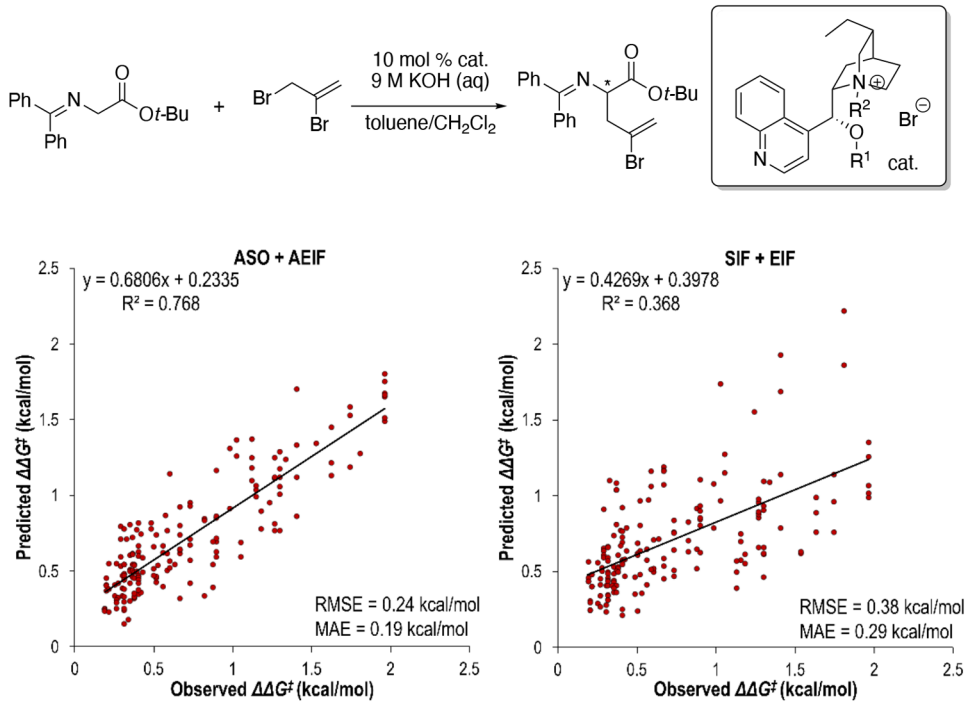


Figure 8. APTC alkylation of glycine imine using SIF and EIF descriptors or ASO and AEIF descriptors. Adapted with permission from ref 2. Copyright 2020 American Chemical Society.

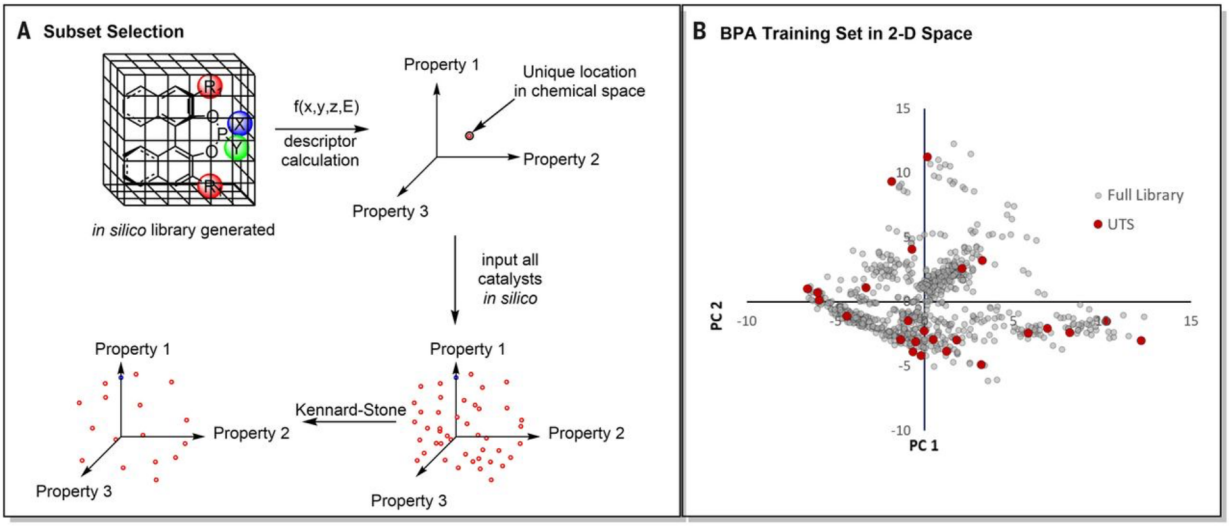


Figure 9. (A) Algorithmic subset selection representing an in silico library and (B) plot (first two principal components) of subset (red) selected from the full in silico library (gray). Adapted with permission from ref 1. Copyright 2019 American Association for the Advancement of Science.

future predictions made within the in silico library will likely be interpolative; as such, these are hypothetically higher confidence predictions (Figure 9).[60,61]

As stated previously, any UTS selected in this manner can be used in any catalytic transformation for which that scaffold is effective. Organic chemists often refer to "privileged" ligand scaffolds;[62] our goal is to apply this approach to such ligand classes and create ideal screening sets for experimentalists to initiate optimization campaigns. In principle, an algorithmically selected subset of in silico catalysts should be the optimum set of catalysts to evaluate in a transformation—this investment represents an excellent return on the resources deployed to synthesize the subset of catalysts.

To evaluate this hypothesis, we tested whether commercially available CPA catalysts could be used to make QSSR models, obviating the need for training set synthesis. Models made with data that used only commercially available catalysts underperformed compared to models using algorithmically selected catalysts, confirming our hypothesis that algorithmic subset selection was important.[56] Clearly, the commercially available catalysts were insufficiently diverse to represent all of the chemical space of possible catalyst structures. However, we also asked if it were possible to augment the set of commercially available catalysts and "rescue" the model performance. To do this, an unsupervised learning technique called *k*-means clustering was used to divide the in silico library into chemically similar clusters. The optimal number of

https://doi.org/10.1021/acs.accounts.0c00826
Acc. Chem. Res. 2021, 54, 2041−2054

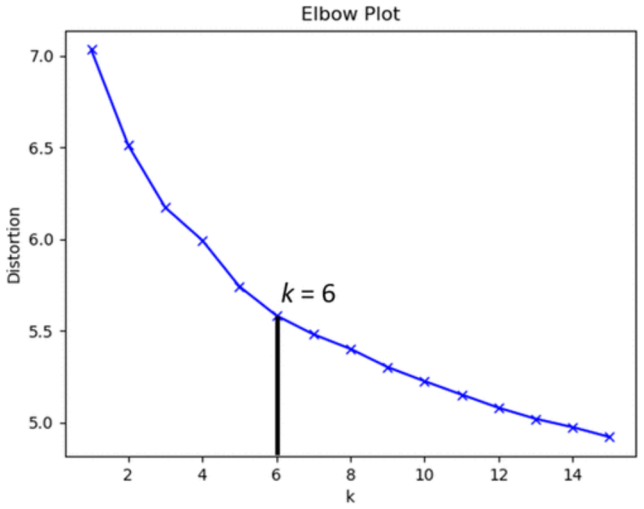


Figure 10. Elbow plot indicating an optimal number of $k = 6$. Adapted with permission from ref 2. Copyright 2020 American Chemical Society.

clusters was identified using the elbow method,[63] in which distortion (defined as the average distance between all catalysts in a cluster the and corresponding cluster centroid) was plotted against the number of clusters, k, for many values of k. The location of an elbow, the point at which the change in distortion decreases sharply, is taken at the optimal number of clusters. In this work, the optimal k was six (Figure 10).

Inspection of the six clusters revealed that commercially available catalysts were completely absent from one cluster. Adding in data from one catalyst from the unrepresented cluster (Figure 11, bottom right) rescued the QSSR model's performance (Figure 11, top right). This method also represents one way to efficiently utilize existing data sets with our workflow.

In principle, randomly selecting catalysts from a diverse library should eventually provide coverage of the chemical space, though it is likely that algorithmic selection would prove more reliable than random selection when selecting a small portion of the library. To test this hypothesis, we quantitatively examined the impact of algorithmic subset selection methods compared to ten randomly selected catalyst training sets in the enantioselective transfer hydrogenation of acetophenone by 331 unique amino acid ligated transition metal catalysts (Figure 12).[64]

Various catalyst subset selection methods were used, and models trained on that subset of reactions were evaluated on the basis of their ability to accurately predict the selectivities of the remaining catalysts in the library to assess out of sample prediction performance. In general, algorithmic subset selection methods (Kennard−Stone, agglomerative clustering,

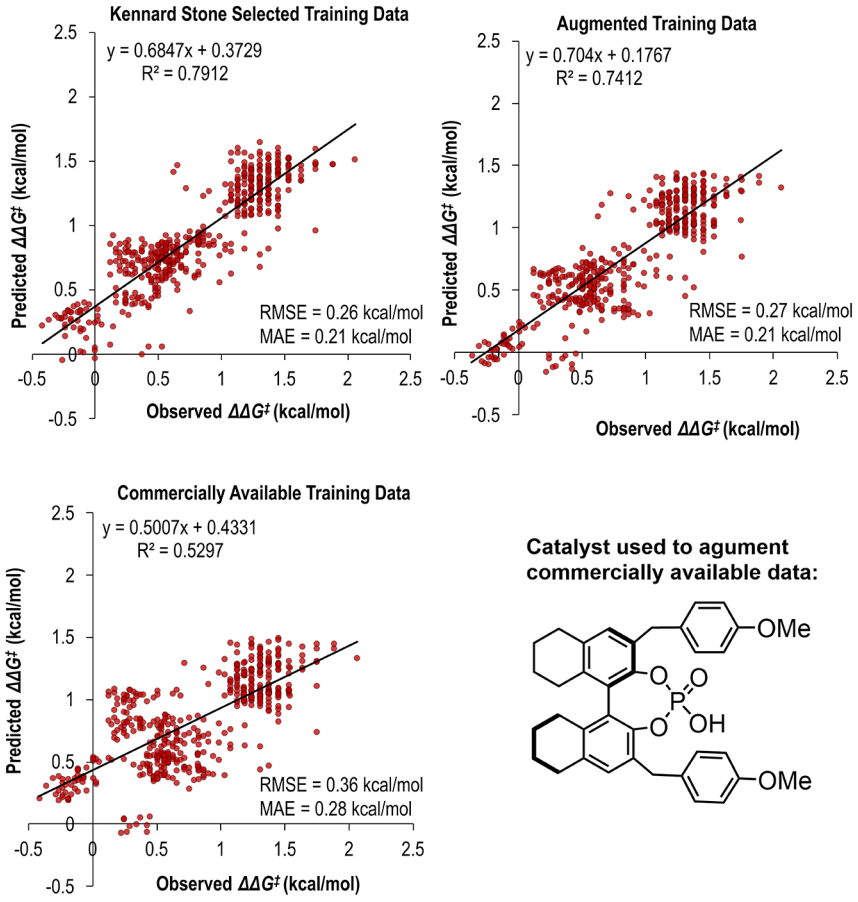


Figure 11. Left: External test set of models trained on diversity algorithmically selected training data vs ETS of models trained on commercially available catalysts. Right: ETS of commercially available catalyst training data augmented with one new catalyst. Adapted with permission from ref 2. Copyright 2020 American Chemical Society.

https://doi.org/10.1021/acs.accounts.0c00826
Acc. Chem. Res. 2021, 54, 2041−2054

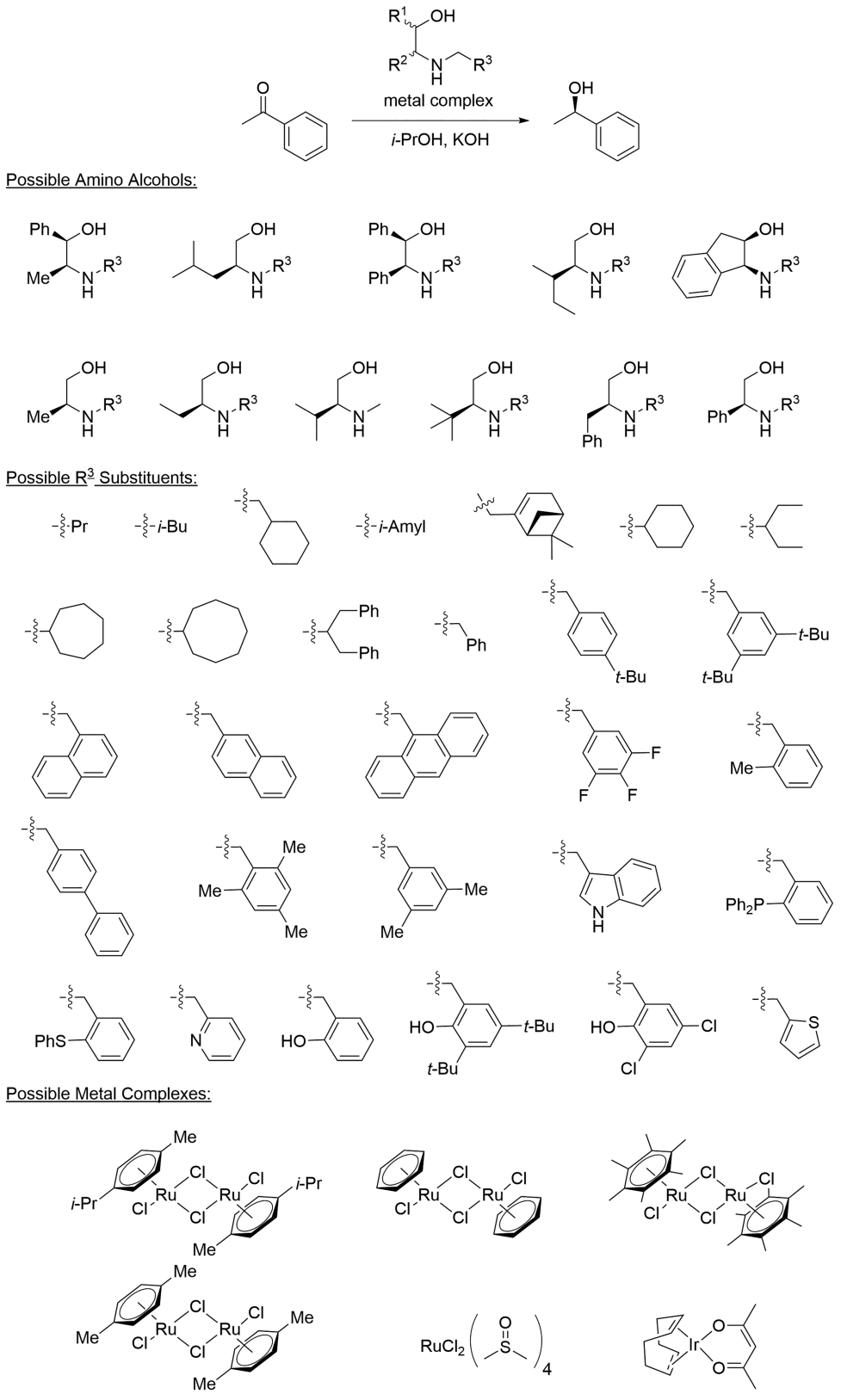


Figure 12. Enantioselective transfer hydrogenation and possible catalyst structures.

affinity propagation, *k*-means, and mean shift) outperformed (average MAE = 0.22 kcal/mol) the average performance (MAE = 0.24 kcal/mol) of ten randomly selected subsets of catalysts (average MAE = 0.22 and 0.24 kcal/mol, respectively). Importantly, the randomly selected subsets had wide variability in their model performances (MAEs from 0.20 to 0.31 kcal/mol). In optimization campaigns in which synthetic overhead is precious, reliably selecting adequate training samples is imperative. Clearly, algorithmic selections

https://doi.org/10.1021/acs.accounts.0c00826
Acc. Chem. Res. 2021, 54, 2041−2054

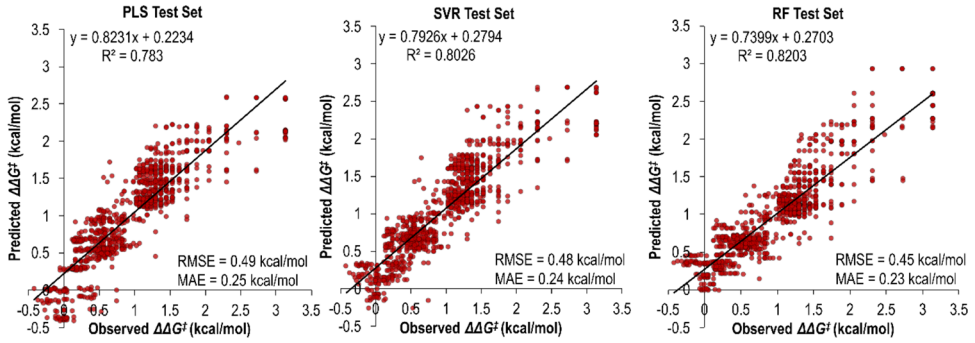


Figure 13. Three different QSSR models of the BPA data set using only 72 training reactions demonstrate effective modeling with limited experimental data. Adapted with permission from ref 2. Copyright 2020 American Chemical Society.

are more reliable than random selection for optimization initialization.

3.5. Examining the Influence of Model Type and Training Set Size on Model Performance

A common critique of machine learning methods in enantioselective catalysis is the requirement of large amounts of data to produce accurate models. We challenged that claim by pushing the limits of modeling using the BPA data set by using 72 reactions for training and cross validation of models and 1003 reactions as an external test set. Three different models were trained with acceptable cross validated correlation coefficients: SVR (q^2 = 0.803), PLS (q^2 = 0.785), RF (q^2 = 0.693). The MAE in test set predictions were 0.24, 0.23, and 0.25 kcal/mol, respectively (Figure 13). Of course, generating predictive models with limited data will be case dependent.

A learning curve is a more quantitative tool for measuring the trade-off between training data and model accuracy; a learning curve generated for the BPA data set is illustrative (Figure 14). Using this plot, an experimentalist can identify when the increase in q^2 and decrease in MAE level off. The MAE of test set predictions is between 0.3 and 0.35 kcal/mol using 96 training examples and continues to decrease to 0.21 kcal/mol after more than tripling the number of training reactions. Any number of training examples in that range would be an acceptable investment to make a useful model to predict selectivities in this system.

Moreover, in the study on APTC alkylation of a glycine imine (vide supra), acceptable QSSR models were generated using 70 data points for training and cross validation and the transfer hydrogenation case study (vide supra) used only 33 or 34 reactions for model training and cross validation, albeit with simpler models.[64] In both cases, a synthetically accessible number of examples were used to train predictive enantioselectivity models. Thus, the complexity of a model and the nature of the system ultimately dictates the minimum amount of training data necessary. The data requirement appears to be highly system dependent. A useful model for selectivity can often be trained on fewer than 100 examples depending on the complexity of the system.

4. CONCLUSIONS AND OUTLOOK

This body of work constitutes the first chapter of a major effort in our laboratory to use data-driven methods to optimize enantioselective catalysts. Taken together, these studies comprise a computational workflow that can be used to inform experimentation. Through the development of numerical representations of chiral molecules, algorithmic

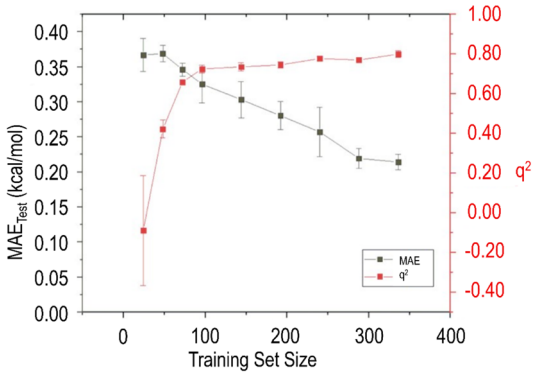


Figure 14. Learning curve relating the quantity of training examples to both the MAE and q^2. Adapted with permission from ref 2. Copyright 2020 American Chemical Society.

subset selection, machine learning, and risk assessment, we have developed a robust tool that can be used reproducibly to optimize many different kinds of reactions. As this first chapter of the program closes, we envision the next in which this tool is used to develop unprecedented catalyst structures in the optimization of new transformations.[65] We look forward to using models generated by this workflow as the basis for formulating mechanistic hypotheses to experimentally test. We believe that the "Golden Age" of this research program lies in the not too distant future.

AUTHOR INFORMATION

Corresponding Author

Scott E. Denmark − *Roger Adams Laboratory, Department of Chemistry, University of Illinois, Urbana, Illinois 61801, United States;* orcid.org/0000-0002-1099-9765; Email: sdenmark@illinois.edu

Authors

N. Ian Rinehart − *Roger Adams Laboratory, Department of Chemistry, University of Illinois, Urbana, Illinois 61801, United States*

Andrew F. Zahrt − *Roger Adams Laboratory, Department of Chemistry, University of Illinois, Urbana, Illinois 61801, United States*

Jeremy J. Henle − *Roger Adams Laboratory, Department of Chemistry, University of Illinois, Urbana, Illinois 61801, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.accounts.0c00826

https://doi.org/10.1021/acs.accounts.0c00826
Acc. Chem. Res. 2021, 54, 2041−2054

Notes

The authors declare no competing financial interest.

Biographies

N. Ian Rinehart graduated with a B.S. in chemistry from the University of Oregon in 2016. He worked in the laboratory of Prof. David R. Tyler as an undergraduate and postbaccalaureate. His current research is focused on the application of machine learning to optimize chemical reactions in the laboratory of Professor Scott E. Denmark at the University of Illinois at Urbana−Champaign.

Andrew F. Zahrt received B.S. degrees in chemistry and biology at Aquinas College in 2014 and completed his Ph.D. studies at the University of Illinois Urbana−Champaign with Prof. Scott E. Denmark in 2020. His research interests include using computational methods to address problems in organic chemistry. He is currently a postdoctoral researcher in the laboratory of Klavs Jensen at MIT.

Jeremy J. Henle obtained a B.S. degree in chemistry from Illinois Wesleyan University in 2011 and a Ph.D. from the University of Illinois at Urbana−Champaign under the supervision of Prof. Scott E. Denmark in 2018. Currently, he is a process chemist at AbbVie. His research interests focus on application of AI/ML to reaction optimization.

Scott E. Denmark is the Reynold C. Fuson Professor of Chemistry at the University of Illinois at Urbana−Champaign (1991-present). He obtained an S.B. degree from MIT in 1975 and a D.Sc.Tech from the ETH Zürich (with Albert Eschenmoser) in 1980. That same year he began his career at the University of Illinois. His research interests include the synthetic, mechanistic and stereochemical aspects of preparatively useful reactions, and the application of AI/machine learning to the optimization of catalysts and reactions.

■ ACKNOWLEDGMENTS

We are grateful to the W. M. Keck Foundation, the National Science Foundation (NSF CHE1900617) and Hoffmann−La Roche, Ltd., for generous financial support. N.I.R. thanks the Robert C. and Carolyn J. Springborn Fund for a graduate fellowship. A.F.Z. thanks the University of Illinois for Graduate Fellowships.

■ DEDICATION

We dedicate this manuscript to the past and present members of the "APTC-Chemoinformatics Subgroup" who spent many years wandering the desert and whose tireless efforts allowed us to finally reach the promised land.

■ REFERENCES

(1) Zahrt, A. F.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. Prediction of Higher Selectivity Catalysts by Computer Driven Workflow and Machine Learning. *Science* **2019**, *363*, eaau5631.

(2) Henle, J. J.; Zahrt, A. F.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. Development of a Computer-Guided Workflow for Catalyst Optimization. Descriptor Validation, Subset Selection, and Training Set Analysis. *J. Am. Chem. Soc.* **2020**, *142*, 11578−11592.

(3) Baiazitov, R. Y.; Denmark, S. E. *Tandem [4 + 2]/[3 + 2] Cycloadditions. In Methods and Applications of Cycloaddition in Organic Syntheses*; Nishiwaki, N., Ed.; Wiley-VCH: Weinheim, 2014; Chapter 16.

(4) Lipkowitz, K. B.; Pradhan, M. Computational Studies of Chiral Catalysts: A Comparative Molecular Field Analysis of an Asymmetric Diels-Alder Reaction with Catalysts Containing Bisoxazoline or Phosphinooxazoline Ligands. *J. Org. Chem.* **2003**, *68*, 4648−4656.

(5) Lipkowitz, K. B.; Kozlowski, M. C. Understanding Stereoinduction in Catalysis via Computer: New Tools for Asymmetric Synthesis. *Synlett* **2003**, *10*, 1547−1565.

(6) Kozlowski, M. C.; Ianni, J. C. Quantum Molecular Interaction Field Models of Substrate Enantioselection in Asymmetric Processes. *J. Mol. Catal. A: Chem.* **2010**, *324*, 141−145.

(7) Kozlowski, M. C.; Dixon, S. L.; Panda, M.; Lauri, G. Quantum Mechanical Models Correlating Structure with Selectivity: Predicting the Enantioselectivity of β-Amino Alcohol Catalysts in Aldehyde Alkylation. *J. Am. Chem. Soc.* **2003**, *125*, 6614−6615.

(8) Ianni, J. C.; Annamalai, V.; Phuan, P.-W.; Panda, M.; Kozlowski, M. C. A Priori Theoretical Prediction of Selectivity in Asymmetric Catalysis: Design of Chiral Catalysts by Using Quantum Molecular Interaction Fields. *Angew. Chem.* **2006**, *118*, 5628−5631.

(9) Huang, J.; Ianni, J. C.; Antoline, J. E.; Hsung, R. P.; Kozlowski, M. C. De Novo Chiral Amino Alcohols in Catalyzing Asymmetric Additions to Aryl Aldehydes. *Org. Lett.* **2006**, *8*, 1565−1568.

(10) Melville, J. L.; Andrews, B. I.; Lygo, B.; Hirst, J. D. Computational Screening of Combinatorial Catalyst Libraries. *Chem. Commun.* **2004**, *4*, 1410−1411.

(11) Melville, J. L.; Lovelock, K. R. J.; Wilson, C.; Allbutt, B.; Burke, E. K.; Lygo, B.; Hirst, J. D. Exploring Phase-Transfer Catalysis with Molecular Dynamics and 3D/4D Quantitative Structure−Selectivity Relationships. *J. Chem. Inf. Model.* **2005**, *45*, 971−981.

(12) Gould, N. D.; Wolf, L. M.; Denmark, S. E. A Systematic Investigation of Quaternary Ammonium Ions as Asymmetric Phase-Transfer Catalysts. Synthesis of Catalyst Libraries and Evaluation of Catalyst Activity. *J. Org. Chem.* **2011**, *76*, 4260−4336.

(13) O'Donnell, M. J. Benzophenone Schiff bases of glycine derivatives: Versatile starting materials for the synthesis of amino acids and their derivatives. *Tetrahedron* **2019**, *75*, 3667−3696.

(14) Gould, N. D.; Wolf, L. M.; Denmark, S. E. A Systematic Investigation of Quaternary Ammonium Ions as Asymmetric Phase-Transfer Catalysts. Application of Quantitative Structure Activity/Selectivity Relationships. *J. Org. Chem.* **2011**, *76*, 4337−4357.

(15) Reid, J. P.; Sigman, M. S. Holistic Prediction of Enantioselectivity in Asymmetric Catalysis. *Nature* **2019**, *571*, 343−348.

(16) Reid, J. P.; Sigman, M. S. Comparing Quantitative Prediction Methods for the Discovery of Small-Molecule Chiral Catalysts. *Nat. Rev. Chem.* **2018**, *2*, 290−305.

(17) Harper, K. C.; Bess, E. N.; Sigman, M. S. Multidimensional Steric Parameters in the Analysis of Asymmetric Catalytic Reactions. *Nat. Chem.* **2012**, *4*, 366−374.

(18) 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*, 1292−1301.

(19) Santiago, C. B.; Guo, J. Y.; Sigman, M. S. Predictive and Mechanistic Multivariate Linear Regression Models for Reaction Development. *Chem. Sci.* **2018**, *9*, 2398−2412.

(20) Burello, E.; Rothenberg, G. *In Silico* Design in Homogeneous Catalysis Using Descriptor Modelling. *Int. J. Mol. Sci.* **2006**, *7*, 375−404.

(21) Maldonado, A. G.; Rothenberg, G. Predictive Modeling in Homogeneous Catalysis: a Tutorial. *Chem. Soc. Rev.* **2010**, *39*, 1891−1902.

(22) Engkvist, O.; Norrby, P.-O.; Selmi, N.; Lam, Y.-H.; Peng, Z.; Sherer, E. C.; Amberg, W.; Erhard, T.; Smyth, L. A. Computational Prediction of Chemical Reactions: Current Status and Outlook. *Drug Discovery Today* **2018**, *23*, 1203−1218.

(23) Durand, D. J.; Fey, N. Computational Ligand Descriptors for Catalyst Design. *Chem. Rev.* **2019**, *119*, 6561−6594.

(24) Strieth-Kalthoff, F.; Sandfort, F.; Segler, M. H. S.; Glorius, F. Machine Learning the Ropes: Principles, Application and Directions in Synthetic Chemistry. *Chem. Soc. Rev.* **2020**, *49*, 6154−6168.

(25) Wodrich, M. D.; Fabrizio, A.; Meyer, B.; Corminboeuf, C. Data-powered augmentation volcano plots for homogeneous catalysis. *Chem. Sci.* **2020**, *11*, 12070−12080.

https://doi.org/10.1021/acs.accounts.0c00826
Acc. Chem. Res. 2021, 54, 2041−2054

(26) Cordova, M.; Wodrich, M. D.; Meyer, B.; Sawatlon, B.; Corminboeuf, C. Data-Driven Advancement of Homogeneous Nickel Catalyst Activity for Aryl Ether Cleavage. *ACS Catal.* **2020**, *10*, 7021−7031.
(27) Oslob, J. D.; Åkermark, B.; Helquist, P.; Norrby, P.-A. Steric Influences on the Selectivity in Palladium-Catalyzed Allylation. *Organometallics* **1997**, *16*, 3015−3021.
(28) Chen, J.; Jiwu, W.; Mingzong, L.; You, T. Calculation on Enantiomeric Excess of a Catalytic Asymmetric Reactions of Diethylzinc Addition to Aldehydes with Topological Indices and Artificial Network. *J. Mol. Catal. A: Chem.* **2006**, *258*, 191−197.
(29) Janet, J. P.; Duan, C.; Yang, T.; Nandy, A.; Kulik, H. J. A Quantitative Uncertainty Metric Controls Error in Neural Network-Driven Chemical Discovery. *Chem. Sci.* **2019**, *10*, 7913−7922.
(30) Hoogenraad, M.; Klaus, G. M; Elders, N.; Hooijschuur, S. M; McKay, B.; Smith, A. A; Damen, E. W.P Oxazaborolidine Mediated Asymmetric Ketone Reduction: Prediction of Enantiomeric Excess Based on Catalyst Structure. *Tetrahedron: Asymmetry* **2004**, *15*, 519−523.
(31) Wang, W.; Yang, T.; Harris, W. H.; Gómez-Bombarelli, R. Active Learning and Neural Network Potentials Accelerate Molecular Screening of Ether-Based Solvate Ionic Liquids. *Chem. Commun.* **2020**, *56*, 8920−8923.
(32) Janet, J. P.; Chan, L.; Kulik, H. J. Accelerating Chemical Discovery with Machine Learning: Simulated Evolution of Spin Crossover Complexes with an Artificial Neural Network. *J. Phys. Chem. Lett.* **2018**, *9*, 1064−1071.
(33) Nandy, A.; Zhu, J.; Janet, J. P.; Duan, C.; Getman, R. B.; Kulik, H. J. Machine Learning Accelerates the Discovery of Design Rules and Exceptions in Stable Metal−Oxo Intermediate Formation. *ACS Catal.* **2019**, *9*, 8243−8255.
(34) Janet, J. P.; Ramesh, S.; Duan, C.; Kulik, H. J. Accurate Multiobjective Design in a Space of Millions of Transition Metal Complexes with Neural-Network-Driven Efficient Global Optimization. *ACS Cent. Sci.* **2020**, *6*, 513−524.
(35) Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. Predicting Reaction Performance in C-N Cross-Coupling using Machine Learning. *Science* **2018**, *360*, 186−190.
(36) Zhou, Z.; Li, X.; Zare, R. N. Optimizing Chemical Reactions with Deep Reinforcement Learning. *ACS Cent. Sci.* **2017**, *3*, 1337−1344.
(37) Nielsen, M.; Ahneman, D. T.; Riera, O.; Doyle, A. G. Deoxyfluorination with Sulfonyl Fluorides: Navigating Reaction Space with Machine Learning. *J. Am. Chem. Soc.* **2018**, *140*, 5004−5008.
(38) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatics Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4*, 268−276.
(39) Friederich, P.; dos Passos Gomes, G.; De Bin, R.; Aspuru-Guzik, A.; Balcells, D. Machine learning dihydrogen activation in the chemical space surrounding Vaska's complex. *Chem. Sci.* **2020**, *11*, 4584−4601.
(40) Foscato, M.; Jensen, V. R. Automated in Silico Design of Homogeneous Catalysts. *ACS Catal.* **2020**, *10*, 2354−2377.
(41) Coley, C. W.; Thomas, D. A.; Lummiss, J. A. M.; Jaworski, J. N.; Breen, C. P.; Schultz, V.; Hart, T.; Fishman, J. S.; Rogers, L.; Gao, H.; Hicklin, R. W.; Plehiers, P. P.; Byington, J.; Piotti, J. S.; Green, W. H.; Hart, A. J.; Jamison, T. F.; Jensen, K. F. A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science* **2019**, *365*, eaax1566.
(42) Shields, B. J.; Stevens, J.; Li, J.; Parasram, M.; Damani, F.; Alvarado, J. I. M.; Janey, J. M.; Adams, R. P.; Doyle, A. G. Baeysian reaction optimization as a tool for chemical synthesis. *Nature* **2021**, *590*, 89−96.
(43) 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*, 9047−9153.
(44) (a) Johnson, J. S.; Evans, D. A. Chiral Bis(oxazoline) Copper(II) Complexes: Versatile Catalysts for Enantioselective Cycloaddition, Aldol, Michael, and Carbonyl Ene Reactions. *Acc. Chem. Res.* **2000**, *33*, 325−335. (b) Desimoni, G.; Faita, G.; Jørgensen, K. A. C_2-Symmetric Chiral Bis(Oxazoline) Ligands in Asymmetric Catalysis. *Chem. Rev.* **2006**, *106*, 3561−3651. (c) Desimoni, G.; Faita, G.; Jørgensen, K. A. Update 1 of: C_2-Symmetric Chiral Bis(oxazoline) Ligands in Asymmetric Catalysis. *Chem. Rev.* **2011**, *111*, PR284−PR437.
(45) Lam, H.-W. TADDOL-Derived Phosphonites, Phosphites, and Phosphoramidites in Asymmetric Catalysis. *Synthesis* **2011**, *2011*, 2011−2043.
(46) Kim, K. H. Comparative Molecular Field Analysis (CoMFA). In *Molecular Similarity in Drug Design*; Dean, P. M., Ed.; Springer: Dordrecht, 1995; 291−331.
(47) Zahrt, A. F.; Athavale, S. V.; Denmark, S. E. Quantitative Structure−Selectivity Relationships in Enantioselective Catalysis: Past, Present, and Future. *Chem. Rev.* **2020**, *120*, 1620−1689.
(48) Pastor, M.; Cruciani, G.; McLay, I.; Pickett, S.; Clementi, S. Grid-INdependent Descriptors (GRIND): A novel Class of Alignment-Independent Three-Dimensional Molecular Descriptors. *J. Med. Chem.* **2000**, *43*, 3233−3243.
(49) Senese, C. L.; Duca, J.; Pan, D.; Hopfinger, A. J.; Tseng, Y. J. 4D-Fingerprints, Universal QSAR and QSPR Descriptors. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1526−1539.
(50) Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B.; Albuquerque, M.; Madhav, P. J.; Duraiswami, C. Construction of 3D-QSAR Models Using the 4D-QSAR Analysis Formalism. *J. Am. Chem. Soc.* **1997**, *119*, 10509−10524.
(51) Zahrt, A. F.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. Prediction of Higher-Selectivity Catalysts by Computer-Driven Workflow and Machine Learning. *Science* **2019**, *363*, eaau5631.
(52) Perlmutter, S. J.; Geddes, E. J.; Drown, B. S.; Motika, S. E.; Lee, M. R.; Hergenrother, P. J. Compound Uptake into *E. Coli* Can Be Facilitated by *N*-Alkyl Guanidiniums and Pyridiniums. *ACS Infect. Dis.* **2020**, DOI: 10.1021/acsinfecdis.0c00715.
(53) Ingle, G. K.; Mormino, M. G.; Wojtas, L.; Antilla, J. C. Chiral Phosphoric Acid-Catalyzed Addition of Thiols to N-Acyl Imines: Access to Chiral N, S-Acetals. *Org. Lett.* **2011**, *13*, 4822−4825.
(54) Sandfort, F.; Strieth-Kalthoff, F.; Kühnemund, M.; Beecks, C.; Glorius, F. A Structure-Based Platform for Predicting Chemical Reactivity. *Chem.* **2020**, *6*, 1379−1390.
(55) Zahrt, A. F.; Denmark, S. E. Evaluating Continuous Chirality Measure as a 3D Descriptor in Chemoinformatics Applied to Asymmetric Catalysis. *Tetrahedron* **2019**, *75*, 1841−1851.
(56) Singh, S.; Pareek, M.; Changotra, A.; Banerjee, S.; Bhaskararao, B.; Balamurugan, P.; Sunoj, R. B. A unified machine-learning protocol for asymmetric catalysis as a proof of concept demonstration using asymmetric hydrogenation. *Proc. Natl. Acad. Sci. U. S. A.* **2020**, *117*, 1339−1345.
(57) Zahrt, A. F.; Henle, J. J.; Denmark, S. E. Cautionary Guidelines for Machine Learning Studies with Combinatorial Datasets. *ACS Comb. Sci.* **2020**, *22*, 586−591.
(58) Chuang, K. v.; Keiser, M. J. Comment on "Predicting Reaction Performance in C−N Cross-Coupling Using Machine Learning. *Science* **2018**, *362*, eaat8603.
(59) Meredig, B.; Antono, E.; Church, C.; Hutchinson, M.; Ling, J.; Paradiso, S.; Blaiszik, B.; Foster, I.; Gibbons, B.; Hattrick-Simpers, J.; Mehta, A.; Ward, L. Can Machine Learning Identify the next High-Temperature Superconductor? Examining Extrapolation Performance for Materials Discovery. *Mol. Syst. Des. Eng.* **2018**, *3*, 819−825.
(60) For related work see: Tropsha, A.; Gramatica, P.; Gombar, V. K. The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. *QSAR Comb. Sci.* **2003**, *22*, 69−77.

(61) For related work see: Golbraikh, A.; Tropsha, A. Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 357−369.
(62) (a) Yoon, T. P.; Jacobsen, E. N. Privileged Chiral Ligands. *Science* **2003**, *299*, 1691−1693. (b) *Privileged Chiral Ligands and Catalysts*; Zhou, Q.-L., Ed.; Wiley-VCH, Weinheim, 2011.
(63) Thorndike, R. L. Who Belongs in the Family? *Psychometrika* **1953**, *18*, 267−276.
(64) Zahrt, A. F.; Rose, B. T.; Darrow, W. T.; Henle, J. J.; Denmark, S. E. Computational Methods for Training Set Selection and Error Assessment Applied to Catalyst Design: Guidelines for Deciding Which Reactions to Run First and Which to Run Next. *React. Chem. Eng.* **2021**, *6*, 694.
(65) We have made our code available on a gitlab repository for anyone to use for handling computational workflow and computing the descriptors used in this work ("ccheminfolib", SHA: b7d907d7e610ccecba1bb1552aeb9cd135fab31c).

https://doi.org/10.1021/acs.accounts.0c00826
Acc. Chem. Res. 2021, 54, 2041−2054