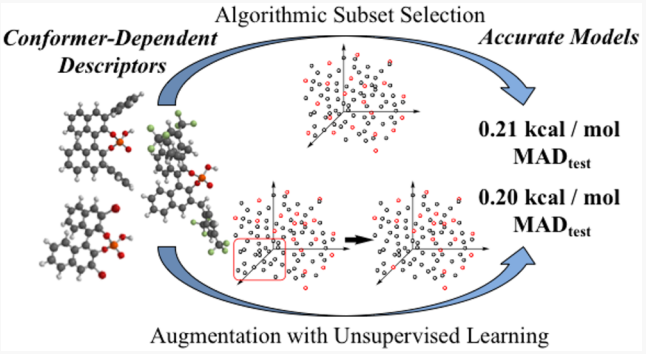
JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY
pubs.acs.org/JACS
Article

Development of a Computer-Guided Workflow for Catalyst Optimization. Descriptor Validation, Subset Selection, and Training Set Analysis

Jeremy J. Henle,[†] Andrew F. Zahrt,[†] Brennan T. Rose, William T. Darrow, Yang Wang, and Scott E. Denmark*

Cite This: https://dx.doi.org/10.1021/jacs.0c04715

ABSTRACT: Modern, enantioselective catalyst development is driven largely by empiricism. Although this approach has fostered the introduction of most of the existing synthetic methods, it is inherently limited by the skill, creativity, and chemical intuition of the practitioner. Herein, we present a complementary approach to catalyst optimization in which statistical methods are used at each stage to streamline development. To construct the optimization informatics workflow, a number of critical components had to be subjected to rigorous validation. First, the critically important molecular descriptors were validated in two case studies to establish the importance of conformation-dependent molecular representations. Next, with a large data set available, it was possible to investigate the amount of data necessary to make predictive models with different modeling methods. Given the commercial availability of many catalyst structures, it was possible to compare models generated with algorithmically selected training sets and commercially available training sets. Finally, the augmentation of limited data sets is demonstrated in a method informed by unsupervised learning to restore the accuracy of the generated models.



INTRODUCTION

Enantioselective catalysis is an enabling technology for the synthesis of chiral, enantiomerically pure organic compounds using substoichiometric quantities of a chiral catalyst.[1,2] Traditionally, catalyst design is guided by empiricism, wherein experimentalists attempt to identify qualitative trends in catalyst structure that dictate catalyst performance. Informed by these trends, new structures are proposed with modifications that are anticipated to afford higher selectivity catalysts. The proposed catalysts must then be synthesized and evaluated, a process which is repeated iteratively until a satisfactory level of performance is achieved. Although frequently successful, this approach is inherently limited because it depends on the ability of the chemist to correctly identify those catalyst features responsible for enantioinduction and for the proposed modifications of the next generation of catalyst structure to improve performance. Even the most experienced practitioner cannot quantitatively discern the many subtle influences each component of the catalyst structure has on dictating the relative energy of competing diastereomeric transition structures. In many cases, the factors influencing selectivity in complex systems are high dimensional; thus, the unaided human mind is incapable of grasping all of the relative contributors related to catalyst performance.

The complexity of this problem has inspired the development of many approaches to aid in catalyst design. Among them, computational methods are particularly appealing owing to the increasing power of computational resources, the decreased demand for materials, and the ability of modern statistical learning methods to recognize patterns in high dimensional data.[3−26] In particular, the capability to evaluate catalysts in silico without preparing and testing them experimentally makes these approaches particularly desirable. The most established method for computational catalyst design is in the application of quantum chemistry to understand the relative energy differentials leading to enantiomers which then enables more informed catalyst design.[4,7,8,10,12−14] Alternatively, surveying catalysts computationally by computing the relative energies of competing transition structures is a viable approach. One such example is Q2MM, in which a transition-state force field is derived for a

Received: April 29, 2020

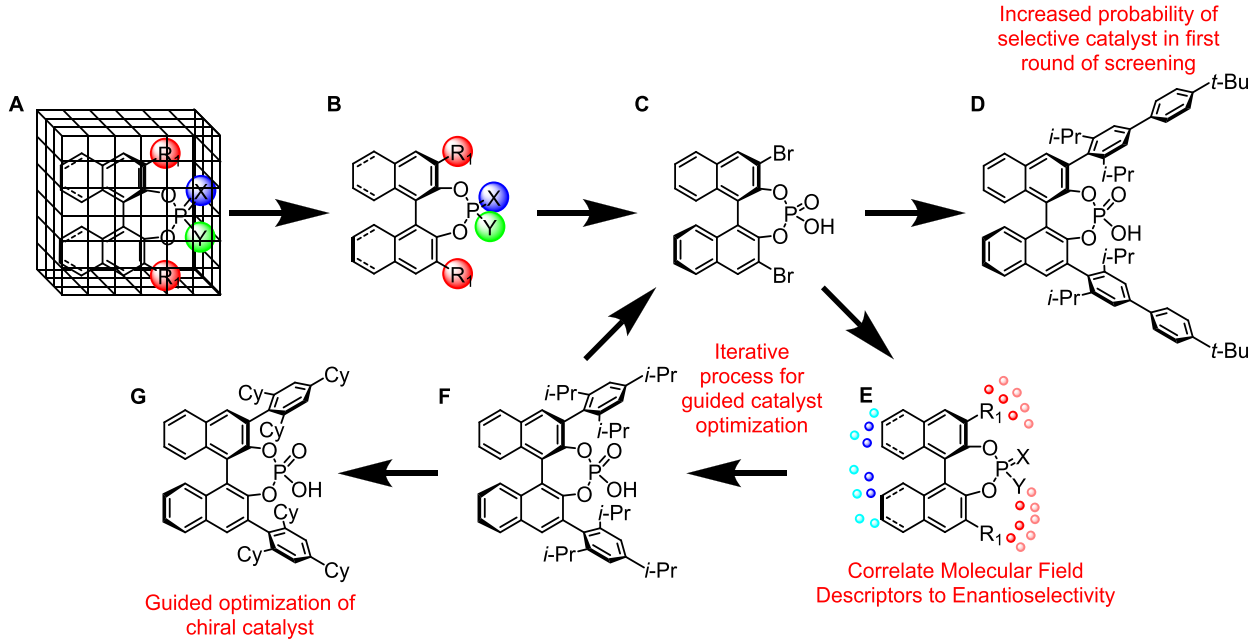


Figure 1. Chemoinformatics-guided process for catalyst discovery and optimization. Adapted with permission from ref 49. Copyright 2019 American Association for the Advancement of Science.

catalytic system, enabling rapid evaluation of many catalyst candidates computationally.[27−29] Similarly, workflows have enabled screening campaigns using quantum chemical methods that are reliable, proceed in a reasonable time frame, and are accessible to the general community.[30] These approaches are attractive in that they do not require any experimental data for evaluation of catalyst candidates. However, they rely on mechanistic knowledge of the enantiodetermining transition structures which is not always available.

An alternative to calculating the relative energies of competing structures is the use of quantitative structure−selectivity relationships (QSSR).[31] In this method, catalyst structures are correlated with experimental data, generating a mathematical model which can be used to evaluate new catalyst structures in silico. Further, QSSR does not require mechanistic information and, in some cases, can be used to extract important mechanistic insights. The seminal example of QSSR in enantioselective catalysis was reported by Norrby and co-workers in which ratios of isomeric products from various nucleophilic substitution reactions on palladium η^3-allyl complexes was predicted.[32] Other early examples of QSSR include the use of molecular interaction fields (MIFs) by Kozlowski,[33−37] Lipkowitz,[38] and Hirst,[39,40] the use of chirality codes by Gasteiger and Aires-de-Sousa,[41−43] and other parameter-based approaches by You[44] and Damen and Hoogenraad.[45,46] More recently, Sigman and co-workers have pioneered a new era of QSSR using linear free energy relationships (LFERs) to identify key structural features of catalysts as well as to predict more selective catalysts.[47]

In our own laboratories, MIF-based approaches have been used in an attempt to elucidate important structural characteristics of phase transfer catalysts.[48,49] More recently, we have used more statistical learning protocols with MIF-type descriptors to evaluate chiral catalysts, culminating in a computer-driven workflow for the optimization of enantioselective catalysts.[50] This workflow is unique in that it contains a combination of critical design features including (1) a systematic method for training set selection that is applicable to any reaction and agnostic to mechanism, (2) a conformer-dependent representation of catalyst properties, (3) accurate predictions for catalyst structures and substrate structures that are novel to the model, and (4) accurate predictions for catalyst structures and substrate structures outside the selectivity range of the training set points. The outline of the general workflow is as follows: (1) a large, in silico library of synthetically accessible catalyst candidates is constructed (Figure 1A); (2) for each member of this library, descriptors are calculated which define the chemical space of the library (Figure 1A); (3) from this library, a representative subset is algorithmically selected, termed the Universal Training Set (UTS) because it is selected only considering catalyst properties and is thus agnostic to reaction and mechanism (Figure 1B); (4) this UTS is synthesized and evaluated in the reaction of interest (Figure 1C); (5) mathematical models are constructed relating the calculated descriptors to experimental outcome (Figure 1E); and (6) the in silico library is virtually screened with the model, and the best catalyst candidates for the particular transformation can be identified for synthesis (Figure 1F,G). This process can be performed interactively, with each subsequent iteration added to the training data, until an ideal catalyst is identified.

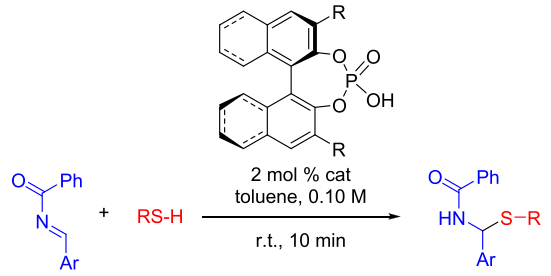
BACKGROUND

In our previous study, the enantioselective formation of N,S-acetals developed by Antilla and co-workers was selected to demonstrate the workflow (Scheme 1).[51]

This system was selected for a number of reasons including short reaction time, clean product profiles, experimental robustness, reproducible reaction outcomes, and sensitivity to perturbations in catalyst structure. In addition to these experimental considerations, other aspects of this reaction made it ideal for benchmarking the viability of the chemoinformatic workflow. Foremost, the reaction had already been

https://dx.doi.org/10.1021/jacs.0c04715
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

Scheme 1. Enantioselective Formation of N,S-Acetals



optimized; thus, if an optimal catalyst could not be identified using our method, it would clearly be the fault of the computational workflow. Further, two other aspects of this system make it an interesting case study for benchmarking: (1) commercially available catalysts give a wide range of selectivity and (2) the catalyst structures are relatively inflexible compared to many other catalyst scaffolds. Because commercially available catalysts cover a wide range of selectivity space, one might expect that readily available catalysts adequately represent the possible chemical diversity in the parent in silico library. Similarly, owing to the relative rigidity of the catalyst scaffold, it is reasonable to hypothesize that for this system the difference between conformer-dependent and single-conformer molecular representations would be negligible. It follows from these points that (1) *if the algorithmically selected subset provides more accurate models than the commercially available subset, one would expect this divergence to increase as the protocol is extended to more complicated systems, and* (2) *if conformer-dependent descriptors are superior in this case, their superiority will only increase as more flexible catalyst scaffolds are investigated.*

To represent the steric properties of molecules, Average Steric Occupancy (ASO) descriptors were calculated as described previously.[50] This work is in essence an application of the 4D-QSAR formalism introduced by Hopfinger and co-workers and first applied to enantioselective catalysis by Hirst and co-workers.[40,52] The reader is referred to the original report for the specific parametrization used for this library, but a general overview is provided here.[50] Descriptors are calculated by first generating a conformer distribution for every member of the in silico library of catalyst candidates. The

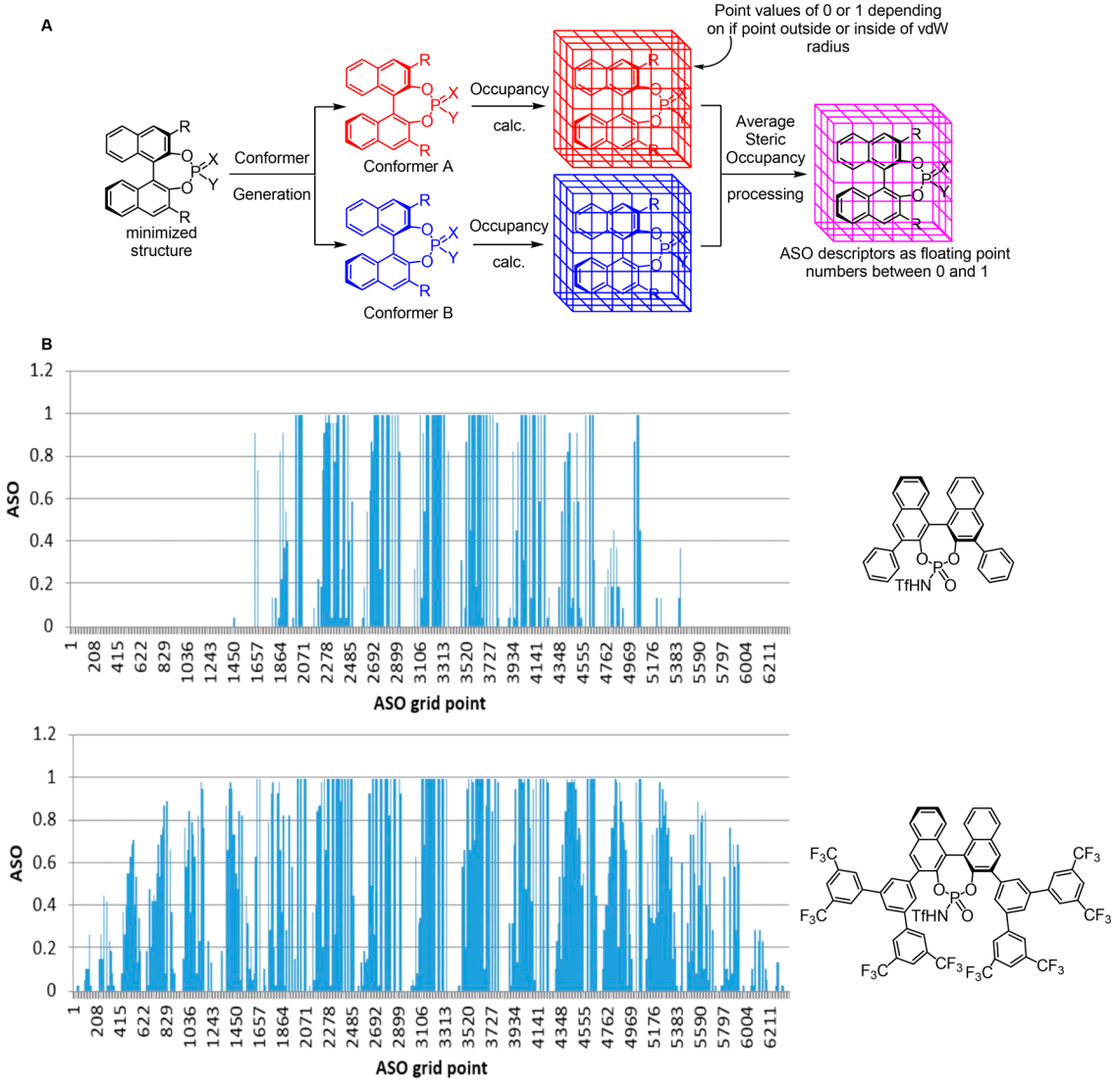


Figure 2. (a) Calculation of ASO descriptors and (b) graphical representation of ASO descriptors for different molecules. Adapted with permission from ref 49. Copyright 2019 American Association for the Advancement of Science.

https://dx.doi.org/10.1021/jacs.0c04715
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

in silico library of BINOL-phosphoric acids used in this study contains 806 members.[53] Every conformer for every catalyst is then superimposed with respect to a common core scaffold and placed in a common grid containing ca. 600 grid points. 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. 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 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 2).[54,55]

To capture the electronic character of the 3,3′-substituents on the phosphoric acid catalysts, a new descriptor was developed to emulate Hammett parameters. Hammett parameters represent the through-bond electronic perturbation a substituent has on a system, whereas other electrostatic potential mapping methods represent through-space effects.[56] However, the 3,3′-substituents in the in silico library are too diverse to be represented with experimentally derived Hammett parameters, owing to the presence of multiple substituents at various positions and other groups that are not simple substituted benzenes. Thus, a new calculable parameter had to be developed that reflects the perturbation of the substituent on a charged particle. First, a tetramethylammonium ion probe is constructed. Then, one of hydrogen atoms of a methyl group is replaced with the 3-substituent of the catalyst. The most positive point on the electrostatic potential map (ESP_{MAX}) still resides around the ammonium residue (specifically, the other methyl groups). Thus, the substituent simply modulates the extent of positive charge around this group. The most positive point on the electrostatic potential energy surface is the ESP_{MAX} descriptor (Figure 3).

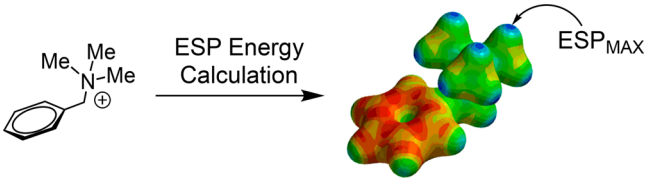


Figure 3. Calculation of the ESP_{MAX} descriptor.

This metric has excellent correlation with Hammett parameters (R^2 = 0.98) and can be rapidly calculated for any desired substituent (Figure 4). Thus, this parameter was selected to represent the electronic character of the 3,3′-substituents. This parameter can be extended to any substituent in any context as an easily calculable electronic descriptor. It is worth noting that this is a substituent-based descriptor, much like Sterimol parameters or Hammett values. Thus, the numerical value for each catalyst structure with the same substituent will be the same.

RESEARCH PLAN

Although our previous report represented the first disclosure of this workflow, the rigorous validation of every step of the workflow was not provided. Accordingly, we present herein the necessary validation underpinning the success of this approach. First, in Case Studies 1 and 2, the importance of conformational flexibility in descriptor calculation is investigated through rigorous comparison with other descriptor classes. Second, a common criticism of more advanced statistical learning methods is that the amount of data necessary to generate accurate models is prohibitively large. In Case Studies 3 and 4, this criticism is investigated by examining the number of data points needed to generate accurate models with different modeling methods. Finally, in our prior work we posed the hypothesis that models generated from a data set containing an algorithmically selected set of molecules will be superior to models generated from a data set containing molecules selected on the basis of their availability. In Case Studies 5 and 6, we more thoroughly investigate this hypothesis by comparing models generated from data sets using algorithmically selected catalysts and commercially available molecules. Then, to extend the utility of our workflow to existing data sets more similar to the latter set, we propose using unsupervised learning as a tool by which to augment such data sets, restoring the accuracy of the models to levels similar to the algorithmically selected set.

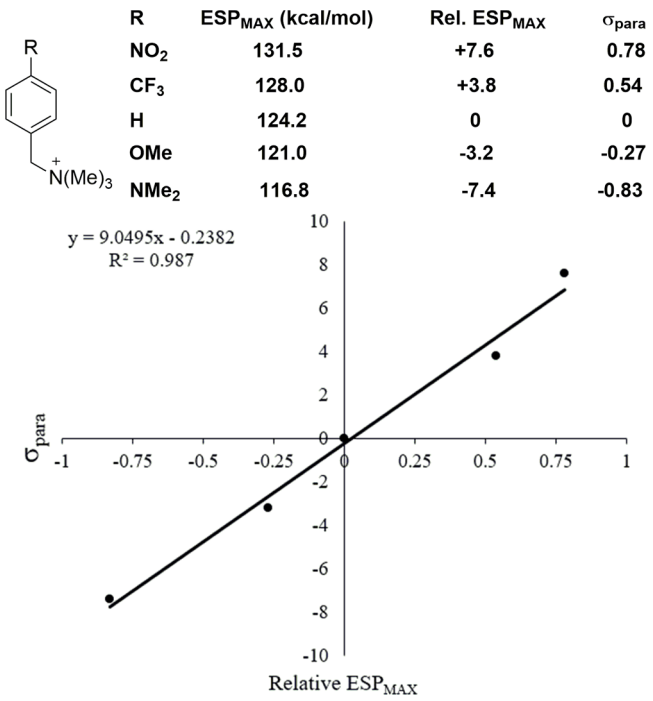


Figure 4. Evaluation of ESP_{MAX} descriptor by correlating relative ESP_{MAX} with Hammett parameters.

RESULTS AND DISCUSSION

2.1. Case Study 1: Increased Accuracy of Conformer-Dependent Descriptors. The suitability of this descriptor set (ASO + ESP_{MAX}) was compared with two different single-conformer representations. The first is a simple steric indicator field (SIF). In this case, the calculation protocol is identical to the ASO, except only one conformer of each catalyst is used in the calculation rather than an ensemble of conformers. Thus, all descriptors are either 1 or 0 in the indicator field. These descriptors were also augmented with the ESP_{MAX} descriptor to represent electronic contributions. The second is the use of electronic and steric molecular interaction fields (MIFs), as employed in comparative molecular field analysis (CoMFA).[57] In this case, a steric MIF was calculated using Lennard-Jones potentials with an sp^3-hybridized carbon atom as a probe atom

https://dx.doi.org/10.1021/jacs.0c04715
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

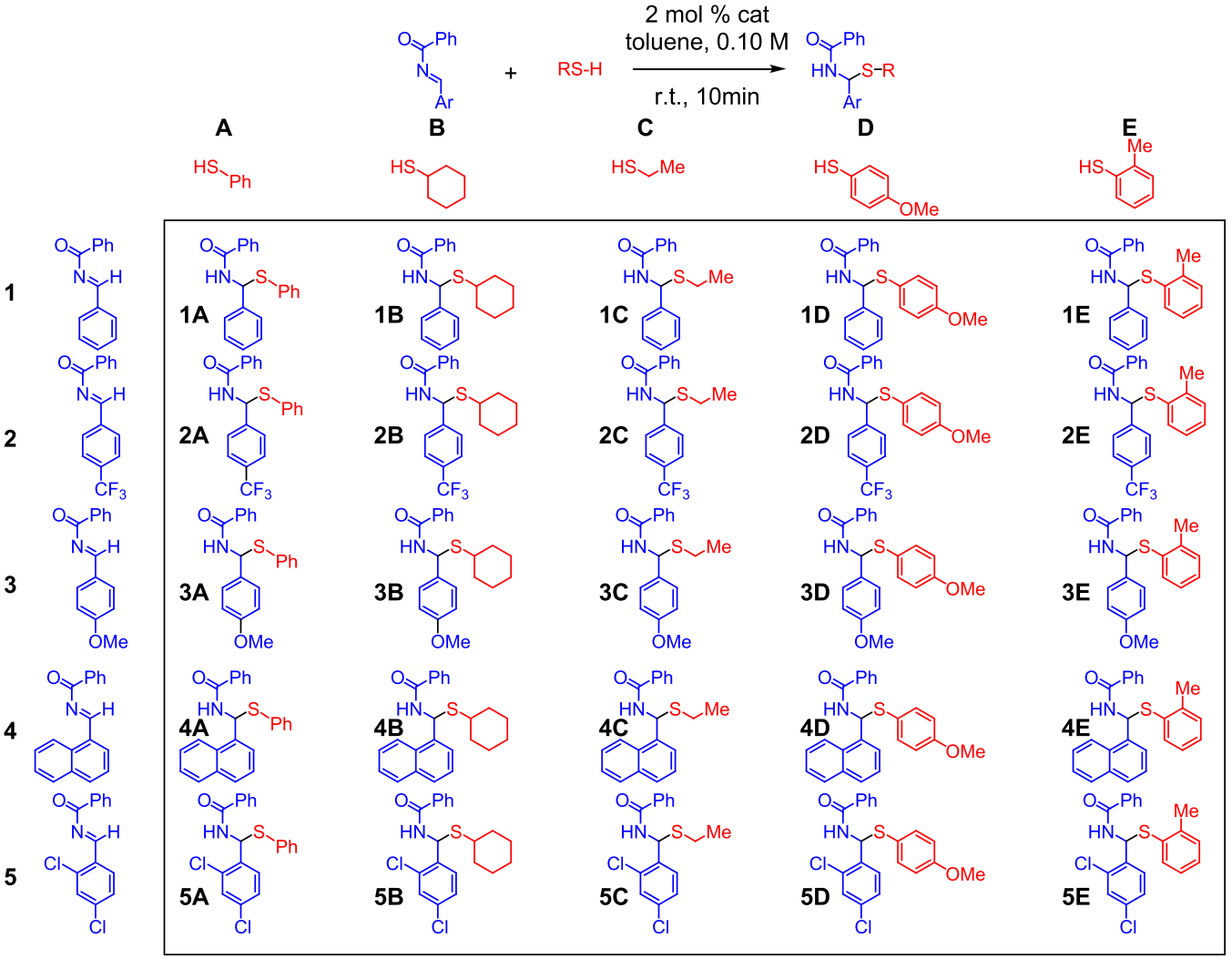


Figure 5. Matrix of 25 different possible substrate combinations derived from imines **1**−**5** and thiols **A**−**E**. Adapted with permission from ref 49. Copyright 2019 American Association for the Advancement of Science.

at each grid point, and an electronic MIF was calculated using the electrostatic potential energy at each grid point using a positron as a probe. For each single-conformer representation, a C_2-symmetric conformer of the molecule was selected for minimization such that each catalyst structure used occupied similar relative conformations. These structures were then minimized at the PM6 level of theory. In this way, differences in descriptor profiles will be attributed to structural differences rather than a difference in the relative conformations of the structures.

As a first experiment, the 1075-reaction data set (generated from 43 catalysts and 25 substrate combinations) published previously was used to compare these different descriptor classes. This data set comprises every substrate/catalyst combination of the 25 substrate combinations in Figure 5, the 24 training set catalysts in Figure 6, and the 19 external set of catalysts in Figure 7. The parametrization of substrates was identical to those found in the original report and was used with all three different catalyst representations to discern only the influence of catalyst representation on accuracy.[50]

In this study, certain catalysts and substrate combinations are intentionally withheld from the training data to ensure some reaction components would be novel to the model. Namely, the 24 UTS catalysts in reactions generating products **1A**−**4D** were used in model training and cross-validation, and reactions with imine **5**, thiol **E**, or any of the external test catalysts in Figure 7 were used as an external test set. Thus, the set used for training and cross-validation was 384 reactions (24 catalysts × 16 substrate combinations = 384 reactions), and the external test set was composed of the remaining 691 reactions. Models were made using support vector machines (with linear, radial basis function, or second-order polynomial kernels, abbreviated herein as SVR_kernel), random forests (RF), gradient boosting regression (GBR), Lasso, Ridge, ElasticNet, Lars, LassoLars, kernel Ridge (with linear, RBF, and second order polynomial kernels), and projection to latent structure (PLS). In each case, the descriptors were first preprocessed with either principle component analysis, mutual information regression, or f-regression to reduce the dimensionality of the input space. Hyperparameter optimization was performed with a Bayesian optimization process when applicable.[58] Models were generated with Sci-Kit learn,[59] and the modeling and evaluation process were automated with in-house Python2 scripts that are available on our GitLab site.[54] The best models were selected of the basis of q^2 from 5-fold cross-validation. A summary of all models generated and their

https://dx.doi.org/10.1021/jacs.0c04715
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

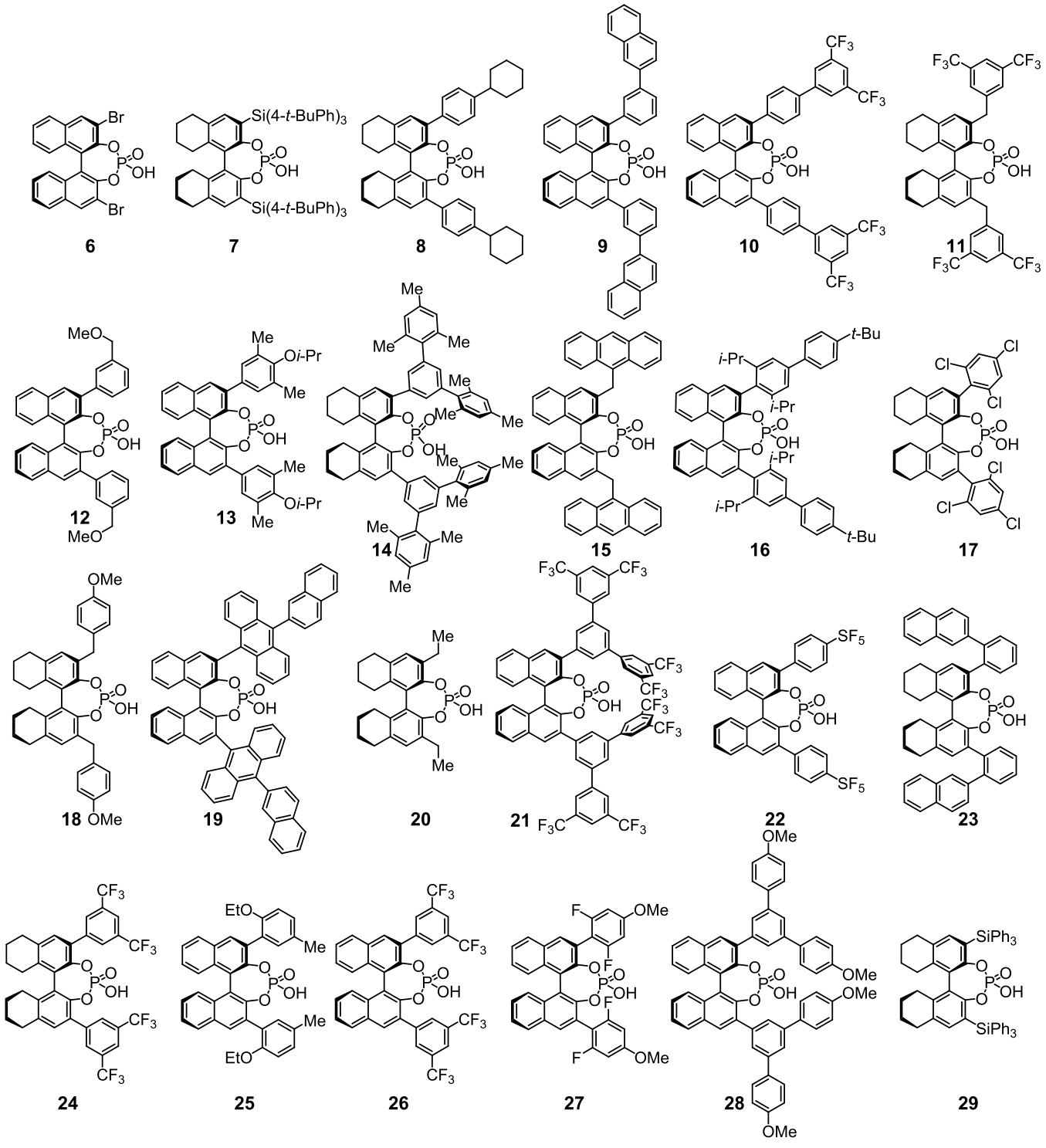


Figure 6. Twenty-four-member UTS of chiral phosphoric acids.

q^2 values are available in the Supporting Information. These best models were then further compared by evaluating their accuracy on the 691 test reactions. As different models were the best performing model for each respective descriptor class, three different models have been evaluated for each descriptor class. Thus, a 3 × 3 matrix was constructed, with the best model for each descriptor class generated with each descriptor class. In this way, bias is avoided by selecting a modeling method suited better for modeling with one of the descriptor type compared to the others (i.e., selecting the best model for one descriptor class to compare all descriptor classes could bias the results in favor of that descriptor class). A summary of best models identified in this study is given in Table 1.

The ASO + ESP_{MAX} descriptors produce models with the lowest MAD for each model type. It is noteworthy that the indicator field-type descriptors dramatically outperformed the interaction field type descriptors, with comparable MADs to conformer-dependent descriptors. To examine the means in more detail, a more rigorous statistical analysis was performed to verify whether the errors are statistically significantly

https://dx.doi.org/10.1021/jacs.0c04715
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

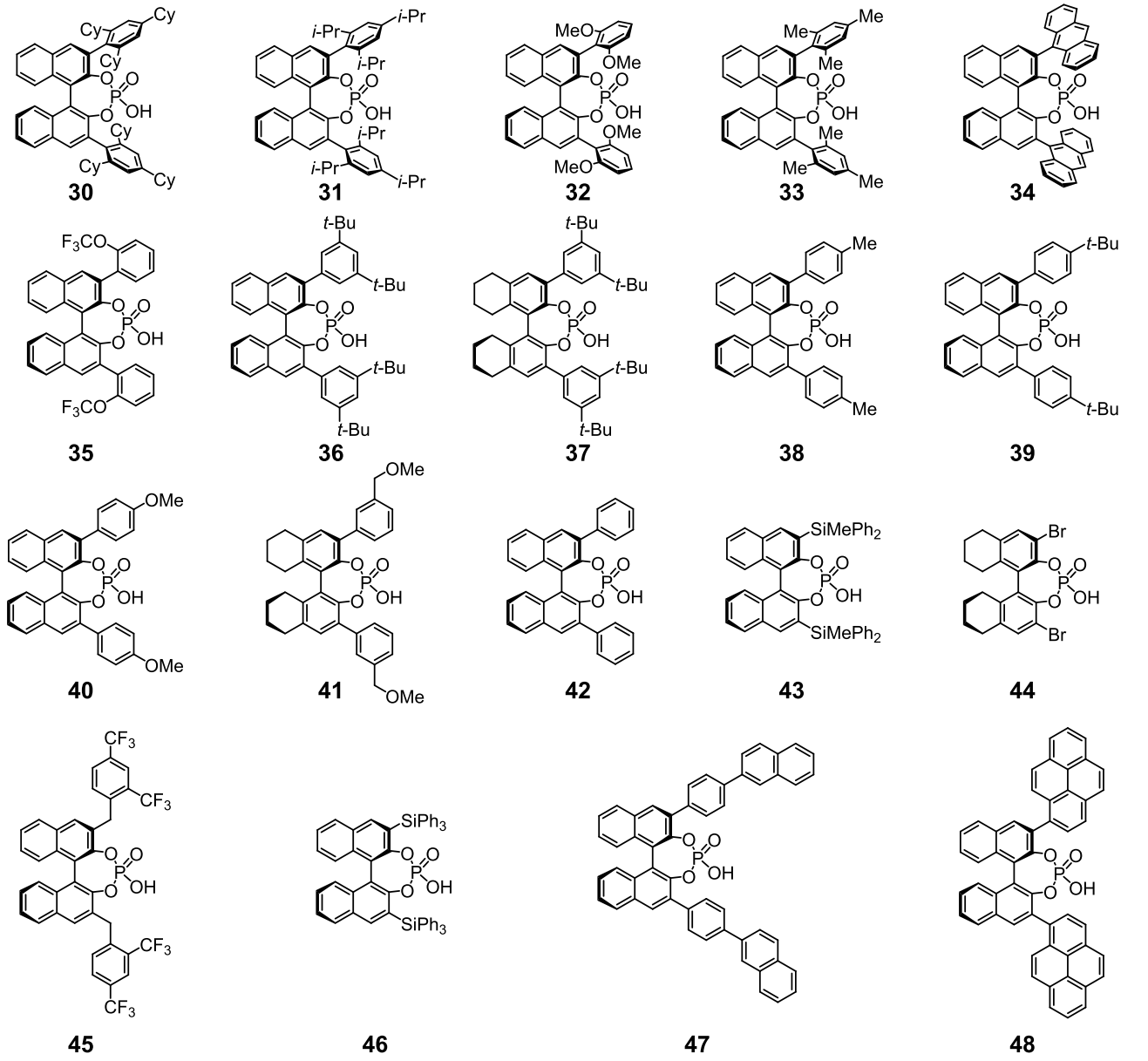


Figure 7. Nineteen-member external catalyst test set of chiral phosphoric acids.

Table 1. Comparison of Descriptor Classes

	descriptor class		
preprocessing and model[a]	ASO + ESPMAX (kcal/mol MAD_{Test})	SIF + ESPMAX (kcal/mol MAD_{Test})	steric and electronic MIFs (kcal/mol MAD_{Test})
FREG500_Kernel_Ridge_poly2	0.24	0.27	0.46
FREG500_SVR_poly2	0.21	0.26	0.44
FREG100_GBR	0.35	0.36	0.36

[a]Preprocessing and model type. FREG500_Kernel_Ridge_poly2 is selection of the top 500 features with f-regression, then modeling with Kernel Ridge using a second order polynomial kernel. FREG500_SVR_poly2 is using f-regression to select the best 500 descriptors, then modeling with a support vector regressor with a second order polynomial kernel. FREG100_GBR is first selecting of the best 100 descriptors with f-regression, then modeling with gradient boosting regression.

different. Because Levene's test for homogeneity indicated variances between groups were statistically significantly different and both the Kolmogorov–Smirnov and Shapiro-Wilk tests for normality indicated the data was not normally distributed, a Kruskal–Wallis analysis of variance was performed with pairwise comparisons between groups and significance values adjusted with the Bonferroni correction for multiple tests. This analysis revealed that indeed the support vector machine trained with ASO descriptors produced statistically significantly lower errors in the test set than every other model except the kernel ridge model trained with ASO descriptors. The kernel ridge models trained with ASO descriptors was not, however, statistically significantly different than the support vector machine or the kernel ridge model

https://dx.doi.org/10.1021/jacs.0c04715
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

trained with steric indicator field descriptors. In general, the molecular interaction field descriptors performed poorly, although the best model for this descriptor class, gradient boosting regression, was not statistically significantly worse than the gradient boosting regression models for ASO and SIF descriptor classes. One plausible explanation for the low performance of this descriptor set is that there are many relatively high-variance data points and extracting relevant stereochemical information from a noisy data set is challenging. However, this hypothesis has not been rigorously evaluated.

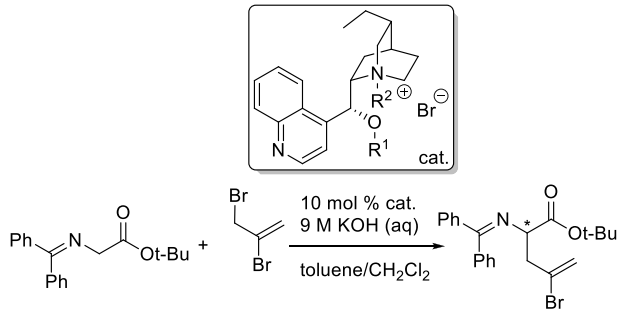
In general, the conformer-dependent descriptors outperform a single conformer descriptors in this case study. However, it is noteworthy that the SIF + ESP_{MAX} descriptors still performed well, generating accurate models. Conceivably, the ASO and SIF descriptors give similar accuracies in this case because the scaffold is very rigid; *we suspect the difference between conformer dependent and single-conformer methods will increase as more flexible conformer scaffolds are investigated.* To probe this hypothesis, a literature data set was selected as an additional case study which uses a catalyst scaffold significantly more flexible than BINOL-phosphoric acids. These descriptors have been compared with fingerprints and 1-hot encoding elsewhere in the literature using this data set.[50,60]

2.2. Case Study 2: APTC Alkylation of a Glycine Imine.

In the selected study, Lygo, Hirst, and co-workers analyzed 88 different enantioselective alkylation reactions performed with different cinchona alkaloid-derived asymmetric phase transfer catalysts (APTC) (Scheme 2).[39] For this study, catalyst structures were represented with both ASO descriptors and our newly implemented, average electronic indicator field (AEIF) descriptors. The AEIF descriptors are conceptually similar to the ASO descriptors, but instead of a binary indicator of occupancy, the atomic charge of the overlapping atom at each point is assigned to that grid point. The sum of these values is then normalized to the number of conformers, furnishing the AEIF value at each grid point. We have developed a Python2 package for calculating AEIF descriptors, which is available on our GitLab page.[54] Notably, these descriptors will change depending on the method of atomic charge calculation used. This conformer dependent representation was compared with a single conformer representation in which a steric indicator field and an electronic indicator field were used. Ten different 70/18 partitions were used to select the 18-member external test set. The 70-member set was then used in training and cross-validation of PLS models, after which the best models were selected to predict the test set (complete modeling details can be found in the Supporting Information). PLS models were chosen because of the large number of descriptors with respect to data points and because of colinearity between descriptor values. The overlaid plots of the external test sets (180 points in total) are depicted in Figure 8.

As is clearly illustrated by the graphs in Figure 8, the conformer dependent representation is much more accurate than the single conformer representation (0.22 kcal/mol MAD_{Test} vs 0.30 kcal/mol MAD_{Test} and q^2 = 0.71 vs 0.53, respectively). Further, the conformer dependent models have a higher R^2 (0.77 vs 0.37), a slope closer to unity (0.68 vs 0.43), and a *y*-intercept closer to zero (0.23 vs 0.40) than the single conformer models. When testing the difference between the test set errors, the models trained with conformer-dependent descriptors are statistically significantly more accurate than the single conformer model (p = 0.0019, full description of test is available in the SI). *This observation supports the hypothesis that the difference in accuracy between conformer dependent and conformer independent models increases as catalyst flexibility increases.*

Scheme 2. Model Reaction System for Enantioselective Alkylation



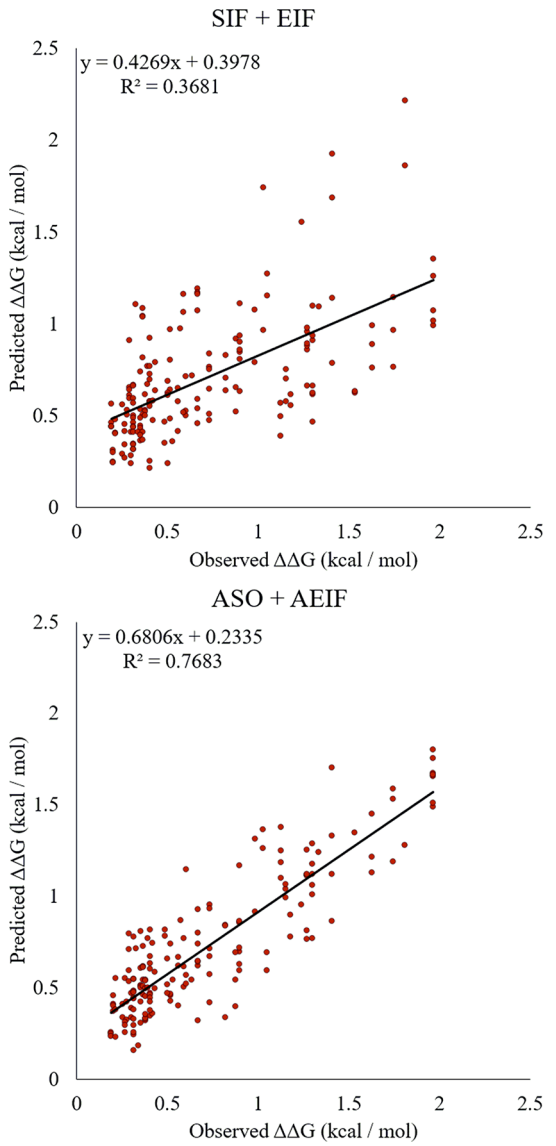


Figure 8. Overlaid test sets for models made with SIF and EIF descriptors (top, MAD = 0.30 kcal/mol) and models made with ASO and AEIF descriptors (bottom, MAD = 0.22 kcal/mol).

https://dx.doi.org/10.1021/jacs.0c04715
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

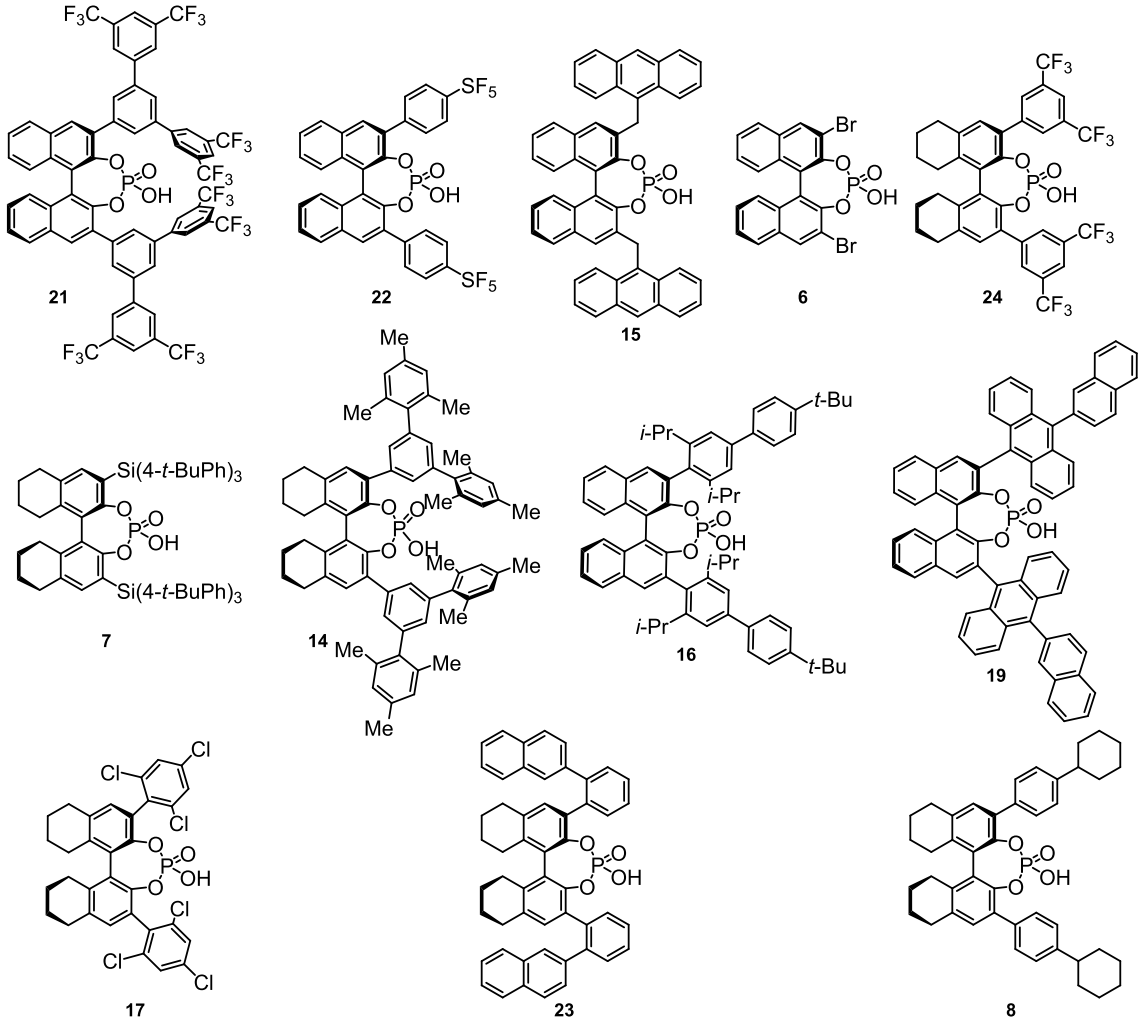


Figure 9. Truncated UTS of chiral phosphoric acids.

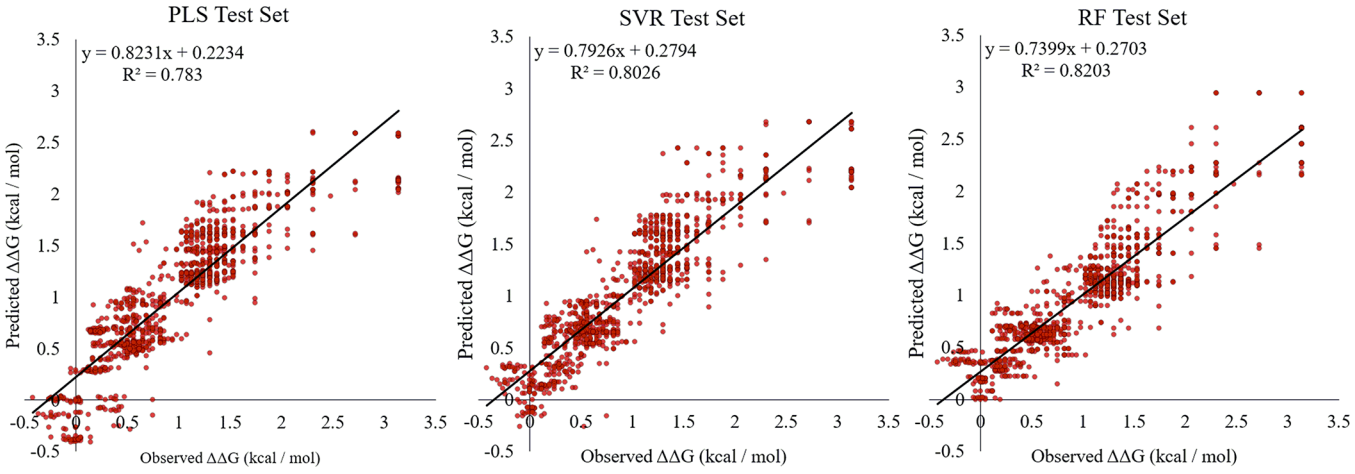


Figure 10. Predicted vs observed for the 1003-member external test sets with PLS models (left, MAD = 0.25 kcal/mol), SVR (middle, MAD = 0.24 kcal/mol), and random forest models (right, MAD = 0.23 kcal/mol).

2.3. Case Study 3: Modeling with Limited Data. As previously mentioned, an often-cited limitation of machine learning methods in enantioselective catalysis is the large number of data points necessary to produce accurate models. Although this assertion is true if broad generality is the goal, *we posit that if the goal of the endeavor is the optimization of a more limited system then accurate predictive models can be made using significantly less experimental overhead.* To illustrate this simplified approach, the modeling of the enantioselective formations of N,S-acetals was revisited using a limited number of data points. As a preliminary exploration of data-limited modeling in this system, only the first 12 catalysts (50%)

https://dx.doi.org/10.1021/jacs.0c04715
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

selected from the *in silico* library of phosphoric acids with the Kennard-Stone algorithm were used in the training set (Figure 9). Further, only reactions containing imines **1** and **2** and thiols **A**−**C** (Figure 5) used in training and cross-validation. Thus, 12 catalysts with 6 substrate combinations gave 72 reactions for model training and validation. Although not an insignificant number of reactions, performing this number of reactions is well within the capability of most synthetic organic chemistry laboratories. The remaining 1003 reactions were used as an external test set.

Three different model types were investigated: PLS, SVR, and random forest (RF). PLS was selected because it is a simple linear model with well precedented use in chemoinformatics using molecular field-type descriptors with relatively limited data sets. SVR and RF were selected because they are popular machine learning methods capable of modeling nonlinear relationships. For the PLS model, the optimal number of latent variables was determined using 3-fold cross-validation. Similarly, hyperparameter optimization for the SVR and RF models were selected by a grid search of hyperparameters, with the best performers identified by q^2. The complete protocol for model optimization and selection is given in the Supporting Information.

When comparing the cross-validation results, the SVR model is the highest performer (q^2 = 0.803), followed by the PLS model (q^2 = 0.785) and finally the RF model (q^2 = 0.693). Using this metric, the more complex model is actually a higher performer in this data-limited case study. When the performance of each model is compared by predicting the external test set, a similar result is observed (Figure 10). In this analysis, the RF most accurately predicts the external test set (MAD_{test} = 0.23 kcal/mol), followed by the SVR model (MAD_{test} = 0.24 kcal/mol), and third the PLS model (MAD_{test} = 0.25 kcal/mol).[61] Thus, even in data limited cases, more complicated machine learning models can be as good or better than simpler modeling techniques. *This illustration indicates that researchers should consider using more complex models even in data-limited scenarios.* It is noteworthy that this phenomenon is likely dependent on the specific system and molecular representation used in any study.

2.4. Case Study 4: Learning Curve Generation. To further examine the amount of data necessary to create accurate models, a learning curve was constructed. To be consistent in analysis of the results, the 384 training + validation/691 test reaction partitioning used in the original study was used at the outset of this experiment. From the 384 possible training reactions, some number of reactions n were randomly selected to use in model training and cross-validation, and those models were further evaluated by the MAD in the 691-member external test set. For each value of n, five training sets were randomly selected. These training sets were used in an ensemble of linear models, and the average MAD of the test set for each value of n was used to evaluate the ensemble. The results are depicted in Figure 11.

As the number of training reactions increases from 24 to 336, a notable increase in q^2 occurs until 96 training reactions is reached. From 144 to 336 training reactions, only incremental improvements in q^2 are observed. However, the MAD_{test} continues to improve with an increasing number of training reactions, reaching 0.21 kcal/mol at 336 reactions. It is worth noting that this model is relating both catalyst and substrate features to enantioselectivity; thus, it is unsurprising that extremely data limited sets (n = 24) have poor

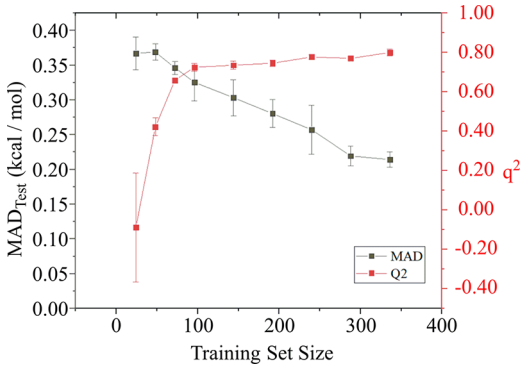


Figure 11. Learning curve, depicting MAD_{test} and q^2(5-fold) plotted against the number of training reactions.

performance. In practice, if one desired to run only 24 reactions, it is likely that a single substrate combination would be used with the Universal Training Set (UTS). To illustrate this, we have constructed a PLS model examining only one reaction, the enantioselective addition of thiol **B** to imine **1**. The 24 UTS catalysts have been used to train the model and the 19 test set catalyst have been used to evaluate the model. This model is indeed very accurate (q^2 = 0.70, MAD_{Test} = 0.156 kcal/mol), and the most selective catalyst for this reaction, which was not included in the training data, is indeed predicted as the most selective catalyst (Figure 12).

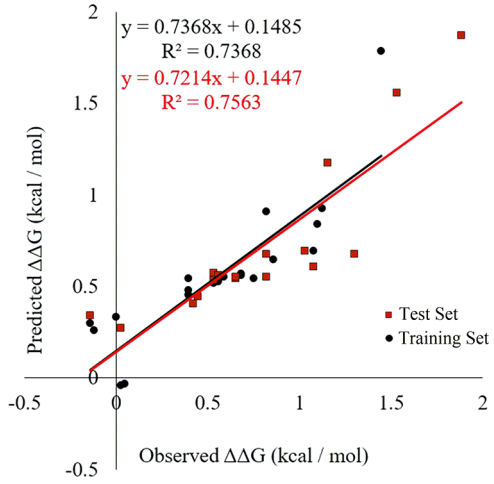


Figure 12. Model for the enantioselective addition of thiol **B** to imine **1**.

2.5. Case Study 5: Improved Predictive Performance with Algorithmic Training Set Selection. In our first publication of this computer-guided workflow,[49] an algorithmically selected set of compounds from the in silico library was identified, termed the UTS. The logic behind this subset selection using the Kennard−Stone algorithm[62] is as follows: (1) the in silico library contains every catalyst candidate that is of interest to be evaluated experimentally on the basis of its synthetic accessibility, (2) the Kennard−Stone algorithm will select boundary cases and sample uniformly over the chemical space of interest, and (3) consequently, all future predictions should be within the convex hull of the training data. Consequently, future predictions will be interpolative; *we hypothesize this process will lead to greater confidence in future predictions*. Alternatively, many researchers might be interested

https://dx.doi.org/10.1021/jacs.0c04715
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

in using readily accessible training sets such as sets of commercially available compounds to use in our workflow rather than synthesizing a training set for the particular catalyst scaffold of interest. To examine the importance of using an algorithmically selected set, two ensembles of linear models were constructed (see the Supporting Information for computational details). The first ensemble was trained and validated using every reaction in our data set which contains one of the first 12 catalysts selected using the Kennard−Stone algorithm (Figure 9). The second ensemble was trained and validated using every reaction in our data set which contains one of the 12 commercially available compounds in the in silico library (Figure 13).[63] Thus, each set for model training and cross-validation was comprised of 300 reactions (12 catalysts × 25 substrate combinations). Reactions which were not included in either set were used as an external test set, and the models generated from training on the truncated UTS and the commercially available set were compared by the performance in predicting the test set (Figure 14).

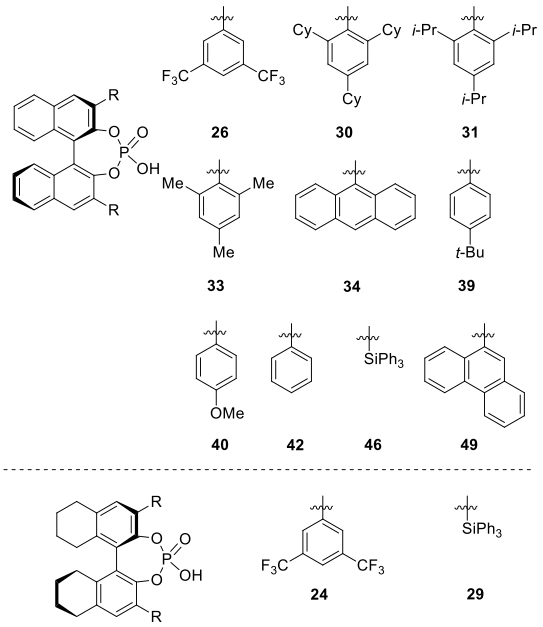


Figure 13. Set of 12 commercially available phosphoric acid catalysts.

As depicted in Figure 14, the models trained using the truncated UTS perform significantly better than those trained with commercially available catalysts (MAD_{Test} = 0.21 and 0.28 kcal/mol, respectively). Further, it is worth considering that the structural diversity in our in silico library of phosphoric acids is relatively limited when compared with other privileged ligand scaffolds in that only the 3,3′-positions of the catalyst structure is varied.[64] Owing to the limited structural variability afforded by the BINOL-phosphoric acid library, the resulting chemical space is quite limited. It follows that the overlap of chemical space coverage between random and systematic subset selection would be highest in this case as compared to more diverse chemical libraries. *Thus, in cases with more diversifiable scaffolds and larger in silico libraries, one would expect that the difference between the readily available set and the algorithmically selected set will be larger, with less accurate models from the readily available set when compared with the algorithmically selected set.*

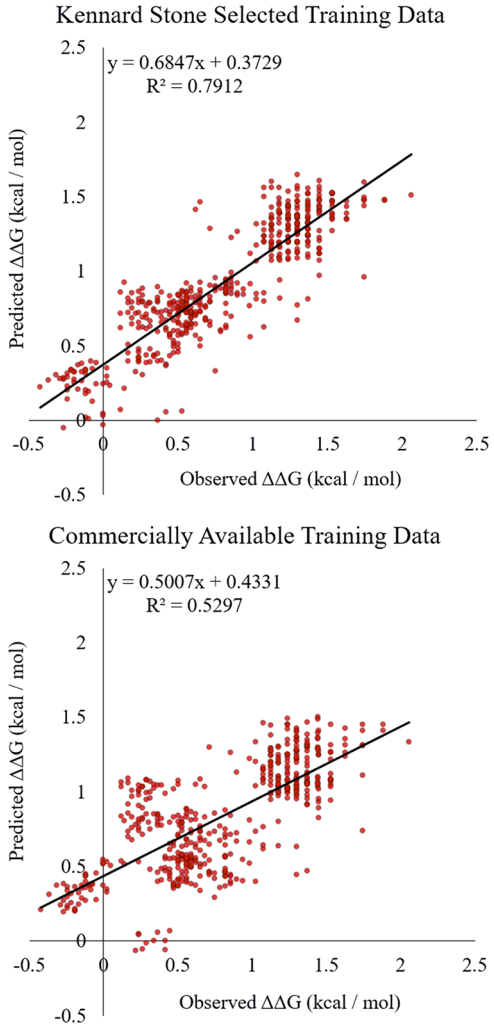


Figure 14. Predicted vs observed plots for models trained with data from 12 UTS catalysts (top, MAD = 0.21 kcal/mol) and 12 commercial catalysts (bottom, MAD = 0.28 kcal/mol). Only the test set is shown.

2.6. Case Study 6: Augmentation of Existing Data Sets Informed by Unsupervised Learning. As demonstrated in Case Study 5, models trained using data sets which are comprised of readily accessible compounds underperform with respect to algorithmically selected sets. However, in many instances, experimentalists may have an existing data set which was not systematically selected and therefore does not span the breadth of chemical space. *Rather than collect data for the entire UTS to use this workflow, it would be desirable to augment the existing data set such that the models generated with that data set are of similar performance to models trained with the UTS. Here, we introduce a tool to achieve this goal by using unsupervised learning to identify underrepresented regions of catalyst space.* Once identified, these catalysts can be introduced to the existing data set by selecting a resident of the unrepresented regions. To illustrate this concept, the commercially available catalyst set in Figure 13 has been augmented with additional structures, informed by performing the K-means clustering algorithm on the entire in silico library of 806 BINOL phosphoric acid catalysts (which contains all of the commercially available catalysts in Figure 14).[65]

Conceptually, the task of augmenting existing data sets was divided into two parts: (1) dividing the space of interest into

https://dx.doi.org/10.1021/jacs.0c04715
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

clusters and (2) selecting representatives from unoccupied clusters. First, the optimal number of clusters, k, is identified using the elbow method (see the Supporting Information for a full description of the elbow method).[66] In this method, some cost function is plotted against the k (Figure 15). In this case, the average distortion (the average distance between catalyst and corresponding cluster centroid) was plotted against k. The location of an elbow, the point at which the change in distortion decreases sharply, is taken at the optimal number of clusters. Loosely, this elbow corresponds to the point at which subdividing the data into additional clusters is unnecessary. In this case, because at least one catalyst must be present per cluster, we equate this with diminishing returns per catalyst synthesized.

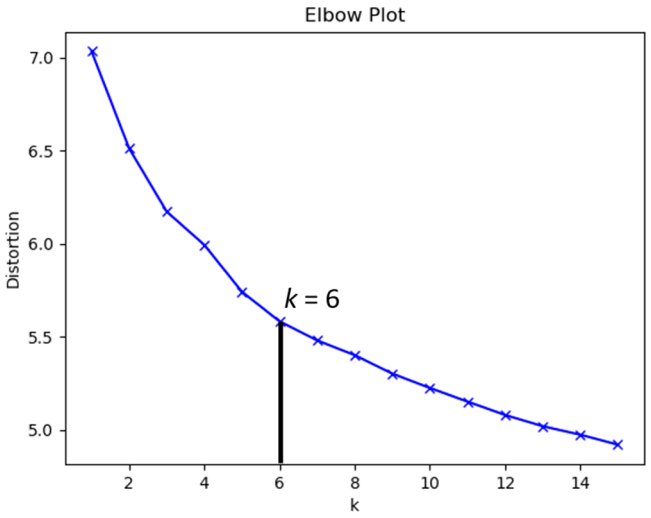


Figure 15. Elbow plot suggesting that the appropriate number of clusters is $k = 6$.

With an optimal number of $k = 6$ identified for the 806-member in silico library, the clusters were inspected for members of the commercially available training set. Five of the six clusters contained members of the commercially available training set, whereas one cluster did not contain any of the commercially available catalysts. Thus, it was postulated that adding a catalyst from this cluster would substantially increase the overall performance of the model. Of the possible catalysts in this cluster, catalyst **18** was selected because it is a representative of this cluster for which experimental data was already available.[67] This 13-member set was then used to create models using the ensemble of linear methods previously employed, and the models were evaluated by examining the predictive performance for the 450-member external test set (identical to the above 475-member external test set minus the 25 reactions of catalyst **18** have been removed). *Remarkably, these new models have similar performance to the models with the truncated UTS (MAD_{Test} = 0.21 kcal/mol for augmented set, 0.20 kcal/mol for Kennard Stone set), supporting the hypothesis that using unsupervised learning to inform augmentation of existing data sets can enable statistically guided optimization endeavors* (Figure 16). This protocol can be used to explore new regions of chemical space and augment data sets for machine learning guided optimization.

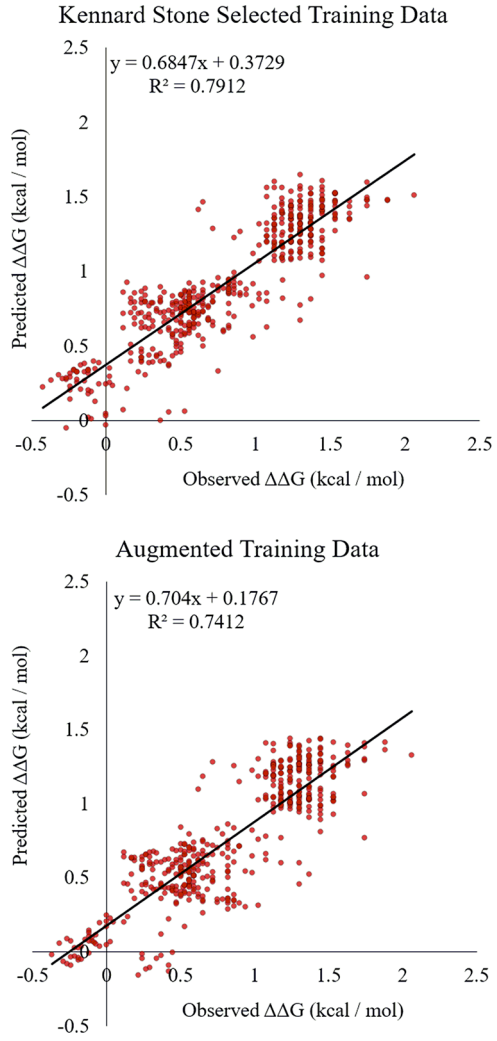


Figure 16. Predicted vs observed plot for models trained on the Kennard Stone selected training set (top, MAD = 0.20 kcal/mol) and the augmented commercial data set (bottom, MAD = 0.21 kcal/mol).

CONCLUSIONS

In summary, validation of every step of our recently disclosed computational workflow, from the conformer-dependent representation of catalyst structures to algorithmically guided subset selection, has been provided. The conformer-dependent catalyst representations outperform their single-conformer analogues even in structurally inflexible systems, with the difference between single-conformer descriptors and conformer-dependent descriptors increasing as the catalyst system becomes more flexible. The superiority of algorithmically selected training sets compared to commercially available training sets has been demonstrated. Finally, the ability to use unsupervised learning to augment existing data sets to achieve predictive performance similar to the algorithmically selected data set has been illustrated. We anticipate that these tools will help experimentalists engaged in challenging optimization problems to leverage a statistically guided solution without prohibitive experimental investment.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/jacs.0c04715.

https://dx.doi.org/10.1021/jacs.0c04715
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

General experimental summary of enantioselective reactions computational methods (PDF)

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

Jeremy J. Henle − *Roger Adams Laboratory, Department of Chemistry, University of Illinois, Urbana, Illinois 61801, United States;* orcid.org/0000-0001-9045-1726

Andrew F. Zahrt − *Roger Adams Laboratory, Department of Chemistry, University of Illinois, Urbana, Illinois 61801, United States;* orcid.org/0000-0002-1835-5163

Brennan T. Rose − *Roger Adams Laboratory, Department of Chemistry, University of Illinois, Urbana, Illinois 61801, United States;* orcid.org/0000-0002-7225-3600

William T. Darrow − *Roger Adams Laboratory, Department of Chemistry, University of Illinois, Urbana, Illinois 61801, United States;* orcid.org/0000-0002-4784-0133

Yang Wang − *Roger Adams Laboratory, Department of Chemistry, University of Illinois, Urbana, Illinois 61801, United States;* orcid.org/0000-0002-7584-6188

Complete contact information is available at: https://pubs.acs.org/10.1021/jacs.0c04715

Author Contributions

†J.J.H. and A.F.Z. contributed equally.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We are grateful to the W. M. Keck Foundation and the National Science Foundation for generous financial support (NSF CHE1900617). A.F.Z. is grateful to the University of Illinois for Graduate Fellowships. B.T.R. thanks the National Science Foundation for Graduate Fellowships. Y.W. thanks Janssen Research Development LLC, San Diego for a postdoctoral fellowship. We thank Dr. Raquel Mendizábal Martell for assistance with statistical analysis. We are also grateful for the support services of the NMR, mass spectrometry X-ray crystallographic, and microanalytical laboratories of the University of Illinois at Urbana−Champaign.

REFERENCES

(1) (a) *Comprehensive Asymmetric Catalysis*; Jacobsen, E. N., Pfaltz, A., Yamamoto, H., Eds.; Springer-Verlag: Heidelberg, 1999; Vols. *I*–*III*. (b) *Comprehensive Chirality*; Carreira, E. M., Yamamoto, H., Eds.; Elsevier: Amsterdam, 2012.

(2) Walsh, P. J.; Kozlowski, M. C. *Fundamentals of Asymmetric Catalysis*; University Science Books: Sausalito, 2009.

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

(4) Ahn, S.; Hong, M.; Sundararajan, M.; Ess, D. H.; Baik, M.-H. Design and Optimization of Catalysts Based on Mechanistic Insights Derived from Quantum Chemical Reaction Modeling. *Chem. Rev.* **2019**, *119*, 6509−6560.

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

(6) 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.

(7) Cheong, P. H.-Y.; Legault, C. Y.; Um, J. M.; Çelebi-Ölçüm, N.; Houk, K. N. Quantum Mechanical Investigations of Organocatalysis: Mechanisms, Reactivities, and Selectivities. *Chem. Rev.* **2011**, *111*, 5042−5137.

(8) Lam, Y.-H.; Grayson, M. N.; Holland, M. C.; Simon, A.; Houk, K. N. Theory and Modeling of Asymmetric Catalytic Reactions. *Acc. Chem. Res.* **2016**, *49*, 750−762.

(9) Peng, Q.; Paton, R. S. Catalytic Control in Cyclizations: From Computational Mechanistic Understanding to Selectivity Prediction. *Acc. Chem. Res.* **2016**, *49*, 1042−1051.

(10) Poree, C.; Schoenebeck, F. A. Holy Grail in Chemistry: Computational Catalyst Design: Feasible or Fiction. *Acc. Chem. Res.* **2017**, *50*, 605−608.

(11) Peng, Q.; Duarte, F.; Paton, R. S. Computing Organic Stereoselectivity - from Concepts to Quantitative Calculations and Predictions. *Chem. Soc. Rev.* **2016**, *45*, 6093−6107.

(12) Wheeler, S. E.; Seguin, T. J.; Guan, Y.; Doney, A. C. Noncovalent Interactions in Organocatalysis and the Prospect of Computational Catalyst Design. *Acc. Chem. Res.* **2016**, *49*, 1061−1069.

(13) Tantillo, D. J. Faster, Catalyst! React! React! Exploiting Computational Chemistry for Catalyst Development and Design. *Acc. Chem. Res.* **2016**, *49*, 1079.

(14) Balcells, D.; Maseras, F. Computational Approaches to Asymmetric Synthesis. *New J. Chem.* **2007**, *31*, 333−343.

(15) Houk, K. N.; Cheong, P. H-Y. Computational Prediction of Small-Molecule Catalysts. *Nature* **2008**, *455*, 309−313.

(16) Fey, N.; Orpen, A. G.; Harvey, J. N. Building Ligand Knowledge Bases for Organometallic Chemistry: Computational Description of Phosphorus(III)-Donor Ligands and the Metal-Phosphorus Bond. *Coord. Chem. Rev.* **2009**, *253*, 704−722.

(17) Corbeil, C. R.; Moitessier, N. Theory and Application of Medium to High Throughput Prediction Method Techniques for Asymmetric Catalyst Design. *J. Mol. Catal. A: Chem.* **2010**, *324*, 146−155.

(18) Fey, N. The Contribution of Computational Studies to Organometallic catalysis: Descriptors, Mechanisms and Models. *Dalton Trans.* **2010**, *39*, 296−310.

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

(20) Neel, A. J.; Hilton, M. J.; Sigman, M. S.; Toste, F. D. Exploiting Non-Covalent π-Interactions for Catalyst Design. *Nature* **2017**, *543*, 637−646.

(21) Baskin, I. I.; Madzhidov, T. I.; Antipin, I. S.; Varnek, A. Artificial Intelligence in Synthetic Chemistry: Achievements and Prospects. *Russ. Chem. Rev.* **2017**, *86*, 1127−1156.

(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) 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.

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

(25) Eksterowicz, J. E.; Houk, K. N. Transition-State Modeling with Empirical Force Fields. *Chem. Rev.* **1993**, *93*, 2439−2461.

(26) 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.

https://dx.doi.org/10.1021/jacs.0c04715
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

(27) Hansen, E.; Rosales, A. R.; Tutkowiski, B.; Norrby, P.-O.; Wiest, O. Prediction of Stereochemistry using Q2MM. *Acc. Chem. Res.* **2016**, *49*, 996−1005.
(28) Rosales, A. R.; Quinn, T. R.; Wahlers, J.; Tomberg, A.; Zhang, X.; Helquist, P.; Wiest, O.; Norrby, P.-O. Application of Q2MM to Predictions in Stereoselective Synthesis. *Chem. Commun.* **2018**, *54*, 8294−8311.
(29) Rosales, A. R.; Wahlers, J.; Limé, E.; Meadows, R. E.; Leslie, K. W.; Savin, R.; Bell, F.; Hansen, E.; Helquist, P.; Munday, R. H.; Wiest, O.; Norrby, P.-O. Rapid Virtual Screening of Enantioselective Catalysts Using CatVS. *Nature Catalysis.* **2019**, *2*, 41−45.
(30) Guan, Y.; Ingman, V. M.; Rooks, B. J.; Wheeler, S. E. AARON: An Automated Reaction Optimizer for New Catalysts. *J. Chem. Theory Comput.* **2018**, *14*, 5249−5261.
(31) 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.
(32) Oslob, J. D.; Åkermark, B.; Helquist, P.; Norrby, P.-O. Steric Influences on the Selectivity in Palladium-Catalyzed Allylation. *Organometallics* **1997**, *16*, 3015−3021.
(33) 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.
(34) Phuan, P.-W.; Ianni, J. C.; Kozlowski, M. C. Is the A-Ring of Sparteine Essential for High Enantioselectivity in the Asymmetric Lithiation-Substitution of N-Boc-pyrrolidine? *J. Am. Chem. Soc.* **2004**, *126*, 15473−15479.
(35) 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.
(36) 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.
(37) Kozlowski, M. C.; Ianni, J. Quantum Molecular Interaction Field Models of Substrate Enantioselection in Asymmetric Processes. *J. Mol. Catal. A: Chem.* **2010**, *324*, 141−145.
(38) Lipkowitz, K.; 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.
(39) Melville, J. L.; Andrews, B. I.; Lygo, B.; Hirst, J. D. Computational Screening of Combinatorial Catalyst Libraries. *Chem. Commun.* **2004**, *0*, 1410−1411.
(40) 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.
(41) Aires-de-Sousa, J.; Gasteiger, J. New Description of Molecular Chirality and Its Application to the Prediction of the Preferred Enantiomer in Stereoselective Reactions. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 369−375.
(42) Aires-de-Sousa, J.; Gasteiger, J. Prediction of Enantiomeric Selectivity in Chromatography: Application of Conformation-Dependent and Conformation-Independed Descriptors of Molecular Chirality. *J. Mol. Graphics Modell.* **2002**, *20*, 373−388.
(43) Aires-de-Sousa, J.; Gasteiger, J. Prediction of Enantiomeric Excess in a Combinatorial Library of Catalytic Enantioselective Reactions. *J. Comb. Chem.* **2005**, *7*, 298−301.
(44) 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.
(45) 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.
(46) van der Linden, J. B.; Ras, I.-J.; Hooijschuur, S. M.; Klaus, G. M.; Luchters, N. T.; Dani, P.; Verspui, G.; Smith, A. A.; Damen, E. W. P.; McKay, B.; Hoogenraad, M. Asymmetric Catalytic Ketone Hydrogenation: Relating Substrate Structure and Product Enantiomeric Excess Using QSPR. *QSAR Comb. Sci.* **2005**, *24*, 94−98.
(47) 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.
(48) Denmark, S. E.; Gould, N. D.; Wolf, L. M. 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.
(49) Denmark, S. E.; Gould, N. D.; Wolf, L. M. 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.
(50) 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*, No. eaau5631.
(51) 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.
(52) 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.
(53) For other notable works investigating structural features of BINOL-phosphoric acids responsible for stereoinduction, see: (a) Reid, J. P.; Ermanis, K.; Goodman, J. M. BINOPtimal: a Web Tool for Optimal Chiral Phosphoric Acid Catalyst Selection. *Chem. Commun.* **2019**, *55*, 1778−1781. (b) Reid, J. P.; Sigman, M. S. Holistic Prediction of Enantioselectivity in Asymmetric Catalysis. *Nature* **2019**, *571*, 343−348.
(54) The code to calculate both representations is available on our Gitlab page: https://gitlab.com/SEDenmarkLab/ccheminfolib/.
(55) Owing to the unreliable accuracy of the molecular mechanics energies, Boltzmann weighting was not reliable. To accurately weight the contribution of each conformer to the ASO, accurate energies computed at a higher level of theory would be required for each conformer. The computational resources required to perform accurate energy calculations on all structures (∼80000) are inaccessible. Thus, given the empirical accuracy of the models generated, no weighting is performed.
(56) Wheeler, S. E.; Houk, K. N. Through-Space Effects of Substituents Dominate Molecular Electrostatic Potentials of Substituted Arenes. *J. Chem. Theory Comput.* **2009**, *5*, 2301−2312.
(57) Kubinyi, H. Comparative Molecular Field Analysis (CoMFA). In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley, 2008.
(58) More information regarding the relevant hyperparameter optimization and preprocessing is available in the Supporting Information.
(59) Pedregosa, F.; Varoquaux, G.; Gramfort, V.; Michel, B.; Thirion, O.; Grisel, M.; Blondel, P.; Prettenhofer, R.; Weiss, V.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(60) Sandfort, F.; Strieth-Kalthoff, F.; Kühnemund, M.; Beecks, C.; Glorius, F. A Structure-Based Platform for Predicting Chemical Reactivity. *Chem.* **2020**, *6*, 1379−1390.
(61) It is worth noting that the test sets, models, and preprocessing used in these studies are different than in case study 1. Because of this, comparisons between different case studies are convoluted and not recommended.
(62) Kennard, R. W.; Stone, L. A. Computer Aided Design of Experiments. *Technometrics* **1969**, *11*, 137−148.
(63) Each catalyst in this set is available from Strem.
(64) It is worth noting that other permutations of this catalyst scaffold are feasible, but the Gen1-in silico library was a limited, proof-

https://dx.doi.org/10.1021/jacs.0c04715
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

of-concept study. Efforts are underway to expand the chemical space of this in silico library.
(65) For more information regarding the K-means clustering algorithm and its specific implementation, we refer the reader to Sci-kit Learn's documentation: https://scikit-learn.org/stable/modules/clustering.html#k-means.
(66) Thorndike, R. L. Who Belongs in the Family? *Psychometrika* **1953**, *18*, 267−276.
(67) Catalyst **18** was chosen because it was present in the cluster with no commercially available representatives and experimental data for that catalyst was already available. We hypothesize that any catalyst from that cluster would improve the model, such that catalysts on the border of the unrepresented cluster would result in a smaller increase in model accuracy.

https://dx.doi.org/10.1021/jacs.0c04715
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX