Published in final edited form as:

Science. 2019 January 18; 363(6424): . doi:10.1126/science.aau5631.

Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning

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

Roger Adams Laboratory, Department of Chemistry, University of Illinois, Urbana, IL 61801, USA.

Abstract

Catalyst design in asymmetric reaction development has traditionally been driven by empiricism, wherein experimentalists attempt to qualitatively recognize structural patterns to improve selectivity. Machine learning algorithms and chemoinformatics can potentially accelerate this process by recognizing otherwise inscrutable patterns in large datasets. Herein we report a computationally guided workflow for chiral catalyst selection using chemoinformatics at every stage of development. Robust molecular descriptors that are agnostic to the catalyst scaffold allow for selection of a universal training set on the basis of steric and electronic properties. This set can be used to train machine learning methods to make highly accurate predictive models over a broad range of selectivity space. Using support vector machines and deep feed-forward neural networks, we demonstrate accurate predictive modeling in the chiral phosphoric acid-catalyzed thiol addition to *N*-acylimines.

Graphical Abstract

INTRODUCTION: The development of new synthetic methods in organic chemistry is traditionally accomplished through empirical optimization. Catalyst design, wherein experimentalists attempt to qualitatively identify correlations between catalyst structure and catalyst efficiency, is no exception. However, this approach is plagued by numerous deficiencies, including the lack of mechanistic understanding of a new transformation, the inherent limitations of human cognitive abilities to find patterns in large collections of data, and the lack of quantitative guidelines to aid catalyst identification. Chemoinformatics provides an attractive alternative to empiricism for several reasons: Mechanistic information is not a prerequisite, catalyst structures can be characterized by three-dimensional (3D) descriptors (numerical

exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works. The title *Science* is a registered trademark of AAAS.

[†]Corresponding author. sdenmark@illinois.edu.

^{*}These authors contributed equally to this work.

Author contributions: A.F.Z. contributed to catalyst synthesis, acquisition of experimental selectivity data, and computer modeling and composed the manuscript. J.J.H. contributed to creating ccheminfolib, designing and implementing the ASO descriptors, and revising the manuscript. B.T.R., Y.W., and W.T.D. contributed to catalyst synthesis. S.E.D. secured funding, supervised the project, analyzed data, and revised the manuscript.

Competing interests: The authors declare no competing interests.

Data and materials availability: Full experimental procedures, characterization data, and copies of ¹H, ¹³C, ³¹P, and ¹⁹F spectra can be found in the supplementary materials, along with analytical supercritical fluid chromatography traces of all products. The computer code used in these studies is available in GitLab (48).

representations of molecular properties derived from the 3D molecular structure) that quantify the steric and electronic properties of thousands of candidate molecules, and the suitability of a given catalyst candidate can be quantified by comparing its properties with a computationally derived model trained on experimental data. The ability to accurately predict a selective catalyst by using a set of less than optimal data remains a major goal for machine learning with respect to asymmetric catalysis. We report a method to achieve this goal and propose a more efficient alternative to traditional catalyst design.

RATIONALE: The workflow we have created consists of the following components: (i) construction of an in silico library comprising a large collection of conceivable, synthetically accessible catalysts derived from a particular scaffold; (ii) calculation of relevant chemical descriptors for each scaffold; (iii) selection of a representative subset of the catalysts [this subset is termed the universal training set (UTS) because it is agnostic to reaction or mechanism and thus can be used to optimize any reaction catalyzed by that scaffold]; (iv) collection of the training data; and (v) application of machine learning methods to generate models that predict the enantioselectivity of each member of the in silico library. These models are evaluated with an external test set of catalysts (predicting selectivities of catalysts outside of the training data). The validated models can then be used to select the optimal catalyst for a given reaction.

RESULTS: To demonstrate the viability of our method, we predicted reaction outcomes with substrate combinations and catalysts different from the training data and simulated a situation in which highly selective reactions had not been achieved. In the first demonstration, a model was constructed by using support vector machines and validated with three different external test sets. The first test set evaluated the ability of the model to predict the selectivity of only reactions forming new products with catalysts from the training set. The model performed well, with a mean absolute deviation (MAD) of 0.161 kcal/mol. Next, the same model was used to predict the selectivity of an external test set of catalysts with substrate combinations from the training set. The performance of the model was still highly accurate, with a MAD of 0.211 kcal/mol. Lastly, reactions forming new products with the external test catalysts were predicted with a MAD of 0.236 kcal/mol. In the second study, no reactions with selectivity above 80% enantiomeric excess were used as training data. Deep feed-forward neural networks accurately reproduced the experimental selectivity data, successfully predicting the most selective reactions. More notably, the general trends in selectivity, on the basis of average catalyst selectivity, were correctly identified. Despite omitting about half of the experimental free energy range from the training data, we could still make accurate predictions in this region of selectivity space.

CONCLUSION: The capability to predict selective catalysts has the potential to change the way chemists select and optimize chiral catalysts from an empirically guided to a mathematically guided approach. ■

Chemoinformatics-guided optimization protocol. (A) Generation of a large in silico library of catalyst candidates. (B) Calculation of robust chemical descriptors. (C) Selection of a UTS. (D) Acquisition of experimental selectivity data. (E) Application of machine learning to use moderate- to low-selectivity reactions to predict high-selectivity reactions. R, any group; X, O or S; Y, OH, SH, or NHTf; PC, principal component; $\bar{G}$, mean selectivity.

The development of synthetic methods in organic chemistry has historically been driven by Edisonian empiricism. Catalyst design, wherein experimentalists attempt to qualitatively recognize patterns in catalyst structures to improve catalyst selectivity and efficiency, is no exception. However, this approach is hindered by a number of factors, including the lack of mechanistic understanding of a new transformation, the inherent limitations of the human brain to find patterns in large collections of data, and the lack of quantitative guidelines to aid catalyst selection. Chemoinformatics (1–3) provides an attractive alternative for several reasons: No mechanistic information is needed, catalyst structures can be characterized by three-dimensional (3D) descriptors (numerical representations of molecular properties derived from the 3D structure of the molecule) that quantify the steric and electronic properties of thousands of candidate molecules, and the suitability of a given catalyst candidate can be quantified by comparing its properties with a computationally derived model on the basis of experimental data. Although artificial intelligence was applied to problems in chemistry as early as 1965, the use of machine learning methods has yet to affect the daily workflow of organic chemists (4). However, recent developments represent the dawn of a new era in organic chemistry, with the emergence of “big-data” methods to facilitate rapid advances in the field. Computer-assisted synthetic planning (5, 6), the prediction of organic reaction outcomes (7, 8), assisted medicinal chemistry discovery (9, 10), catalyst design (11, 12), the use of continuous molecular representations for automatic generation of new chemical structures (13), materials discovery (14), the enhancement of computer simulation techniques (15), and the optimization of reaction conditions (16) all provide examples in which leveraging machine learning methods facilitates advances in chemistry. The power of these methods is particularly notable for catalyst design; modern machine learning methods have the capacity to find patterns in large sets of data that are incomprehensible to experimental practitioners (17). Discovering these structure-activity relationships may facilitate catalyst identification, thus enabling the rapid optimization of catalytic transformations.

Lipkowitz *et al.* and Kozlowski *et al.* first reported the application of a 3D quantitative structure-activity relationship (QSAR) to asymmetric catalysis, wherein they used different molecular interaction field (MIF) approaches to study copper bis(oxazoline) complexes in enantioselective Diels-Alder reactions and enantioselective alkylations of aryl aldehydes, respectively (18, 19). Although similar MIF-based approaches have since been employed (12, 20, 21), we suspect that such methods have not achieved widespread use because of the reliance on only one conformer in descriptor generation. To address this problem, Sigman and co-workers have employed multivariate regression techniques and catalyst-specific descriptors to glean mechanistic information (22–24). These researchers attribute some of their success to the use of Sterimol values; these substituent-based descriptors have multiple parameters designed to capture the rotation of the group of interest, thus providing a more accurate picture of how the molecule behaves in solution (25). Furthermore, preliminary studies in which predictions are made beyond the bounds of the training data have been described; Sigman and co-workers have demonstrated the ability to predict ~10% enantiomeric excess (ee) beyond the training data (26). However, no examples exist wherein the prediction is far outside the selectivity regime comprising the training data. A very recent example of the utility of machine learning methods in catalysis is the prediction of reaction yields by Doyle and co-workers (27, 28). These authors use many easily calculable descriptors to predict the outcomes of C–N coupling reactions and deoxyfluorination reactions with random forest models (29). Although this method excels in predicting the outcomes of reactions when the predicted value falls within the range of values in the training data, this method has not been used to make predictions beyond the range of observed values in the training set.

The ability to accurately predict a selective catalyst by using a set of nonoptimal data remains a primary objective of machine learning with respect to asymmetric catalysis. This feat is sometimes erroneously referred to as “extrapolation”—an understandable mistake, given that predictions are being made outside the bounds of previously observed selectivities. However, the term “extrapolation” does not refer to the selectivity space of the training data but rather to the descriptor space. Thus, a better statement of this goal is to predict high selectivity values far beyond the bounds of what is encompassed in the training data. Herein, we describe a method to achieve this goal by proposing a more efficient alternative to traditional catalyst design.

This endeavor is challenging for a number of reasons. First, very small energy differences (~1 kcal/mol) can give rise to vastly different selectivities—even modern quantum chemical methods struggle to reproduce these energy differences in diastereomeric transition structures. Subtle changes in catalyst structure can also lead to large changes in catalyst performance, whereas descriptors capable of capturing fundamental chemical properties (30) and the subtle features of catalyst structure responsible for enantioinduction remain imperfect. Moreover, off-cycle or background reactivity can erode enantioselectivity, and selectivity data are rarely uniformly distributed, adding the challenge of modeling on a skewed dataset. Predicting reactions that are more selective than anything in the training data (essential for machine learning to optimize a reaction) requires the model to accurately predict to a fringe case, a formidable challenge in its own right.

Perhaps the greatest impediment to accurate prediction in this manner is that no widely accepted workflow implementing chemoinformatics at all stages of development has been introduced to date. Using training set selection algorithms is essential to guarantee that the maximal breadth of feature space is covered in the training data; thus, by design, there should theoretically be no need for extrapolation. Failure to use training set selection algorithms introduces a greater level of uncertainty for predictions—if the domain of applicability is completely unknown, predictions may be outside the well-described region of feature space, and those predictions may be unfounded. If such methods are to be successful, chemical properties must be represented by robust descriptors. This aspect is especially challenging for asymmetric catalysis, as currently no mathematical representation of organic molecules exists that satisfies the following critical criteria: The descriptors must be rapidly calculable, must contain 3D information about an ensemble of conformers for each molecule, must be general for any given scaffold, and must capture the subtle features of catalyst structure responsible for enantioinduction. We describe the development of a workflow that uses chemoinformatic methods at every stage. Further, we report a molecular representation that facilitates this workflow and that enables the prediction of enantioselective reactions in a manner simulating new reaction optimization.

This new workflow consists of the following components (Fig. 1): (i) construction of an in silico library of a large (31) collection of conceivable, synthetically accessible catalysts of a particular scaffold; (ii) calculation of robust chemical descriptors for each scaffold, thereby creating the chemical space comprising the in silico library; and (iii) selection of a representative subset of the catalysts in this space. This subset is termed the universal training set (UTS), so named because it is agnostic to reaction or mechanism. Thus, the same set of compounds can be used to collect training data for any reaction that can be catalyzed by the common functional group and will cover the maximum breadth of feature space. The continuation of the workflow involves (iv) collection of the training data and (v) application of machine learning methods to generate models that predict the enantioselectivity of each member of the in silico library. These models are evaluated with an external test set of catalysts (predicting selectivities of catalysts outside of the training data). The validated models can then be used to select the optimal catalyst for a given reaction. At this point, either the predicted catalyst obtains the desired level of selectivity (success) or the predicted catalyst data can be recombined with the training data to make more robust models. The process can then be repeated iteratively until optimal selectivity is achieved (Fig. 1).

To develop this workflow, we chose the BINOL (1,1'-bi-2-naphthol)-derived family of chiral phosphoric acids as the catalyst scaffold. This family possesses a number of beneficial features, including synthetic accessibility and ease of diversification by installation of an array of substituents at the 3,3' positions. Additionally, the acidity of the phosphoryl group can be easily modulated, and the backbone can be unsaturated (binaphthyl backbone) or saturated (H₈ backbone). These catalysts can be used for a vast number of synthetically useful reactions; thus, a UTS of this scaffold could be very powerful for method development (32).

Development of average steric occupancy descriptors

The plan began with the formulation of an *in silico* library containing 806 chiral phosphoric acid catalysts. For this class, two scaffolds were selected: catalysts with a fully aromatic binaphthyl backbone and catalysts wherein the second ring of the binaphthyl moiety is saturated (H_8). Then a dataset of 403 synthetically feasible substituents (from a database of readily available commercial sources or fragments that require no more than four well-established synthetic steps) was added to the 3,3' positions of these scaffolds by using Python2 scripts (for full details, see computational methods in the supplementary materials). The substituents were chosen by surveying catalogs of boronic acids, aryl halides, aldehydes, alkyl boranes, and Grignard reagents and adding all members that were compatible with the reaction conditions necessary to install that substituent (such as Suzuki coupling, use of organolithium reagents, and Kumada coupling). Thus, we are confident that our *in silico* library covers a large breadth of chemical space that is synthetically accessible. To construct the chemical space representing this library, chemically meaningful descriptors were calculated. However, using many types of readily available 0D, 1D, 2D, and 3D descriptors (the latter derived mostly from MIFs) led to failure because the calculated features did not adequately represent those catalyst properties responsible for enantioinduction [comparative molecular field analysis (CoMFA), grid-independent descriptors (GRIND), and all descriptors available in RDKit and MOE 2015 are some examples of previous attempts] (33–35). The likely cause of failure was that only a single conformation of each of the catalysts was included. Thus, a new set of descriptors had to be developed that included information about the entire conformer ensemble, could be used for any catalyst scaffold, and would be easily calculable for large libraries of compounds.

To achieve this goal, we invented a new descriptor called average steric occupancy (ASO). The ASO descriptors were inspired by 3.5D and 4D descriptors, simplifying the conformer population information into a location-specific numerical form (36, 37). The protocol for ASO calculation is illustrated in Fig. 2A. First, a conformer distribution for each catalyst in the *in silico* library was obtained. Second, for each molecule, the conformers were aligned and individually placed in identical grids. If a grid point was within the van der Waals radius of an atom, it was assigned a value of 1; otherwise, it was assigned a value of 0. This process was repeated for n conformers, and upon completion each grid point had a cumulative value ranging from 0 to n . The values were then normalized by dividing by n , such that all grid points had a value between 0 and 1. These values constituted the steric descriptors for the structures. These features are represented in Fig. 2B, wherein the ASO values around a phosphoric acid catalyst are depicted. The red grid points mark areas away from the catalyst where ASO values are 0.000 to 0.125, whereas the blue represents grid points where the ASO values are 0.875 to 1.000. Because the catalysts are aligned to the backbone, the corresponding grid points all have a value of nearly 1, and the backbone is visible as the two overlapping blue bands. Below the blue bands are regions of green and yellow; these represent conformers that differ by the rotation of the P–NH–Tf (triflyl) moiety and the phenyl substituents at the 3,3' positions. The capacity of these descriptors to distinguish among catalysts of different classes is illustrated in Fig. 2C. The distribution of the different

catalyst classes in chemical space (from the first three principal components of the ASO chemical space) demonstrates that ASO qualitatively groups like-structured catalysts.

The electronic descriptors were derived from the perturbation that a substituent exerts on the electrostatic potential map of a quaternary ammonium ion (see the computational methods in the supplementary materials for details). These substituent-based electronic descriptors were combined with the ASO descriptors. In total, this process amounted to 16,384 features per catalyst, which was later reduced upon the removal of all features with a variance of zero.

To select a representative subset of the chemical space spanned by the *in silico* library, the dimensionality of the chemical space must be reduced (38). The data were transformed with principal components analysis (PCA) (39), which selects new dimensions such that the variance retained is maximized per dimension kept.

A representative subset (including boundary cases) was selected from this space by using the Kennard-Stone algorithm (40) (Fig. 3). This sampling method is of paramount importance; it guarantees that catalysts from uniform regions of feature space are sampled. Thus, predictions made later in method development should still be in a region of feature space described by the initial training set, giving more confidence in these predictions. The subset of selected catalysts constitutes the UTS, which can then be used to optimize any reaction that can be catalyzed by that catalyst type. The 24 members of the UTS for the chiral phosphoric acid scaffold are given in Fig. 4A. To evaluate the predictions made from the UTS, a separate test set of 19 external catalysts (52 to 70) (Fig. 5B) was selected from the *in silico* library. These external catalysts were selected on the basis of intuitive chemical differences and synthetic accessibility.

Application of the catalyst optimization protocol to asymmetric *N,S*-acetal formation

To validate the ASO and training set selection protocol, the training set was evaluated on a previously optimized model reaction. The enantioselective formation of *N,S*-acetals (Fig. 5A) developed by Antilla and co-workers (41) was selected for several reasons. The reaction is high yielding and highly reproducible; it can be performed under air at room temperature, thus facilitating rapid screening; and a range of selectivities (0 to 99%) has been reported with different catalysts (six reported catalysts). Accordingly, we judged this reaction to be a good candidate for empirically evaluating the selectivity space covered in the UTS. By calculating ASO and electronic descriptors for reactants and products as well and concatenating these descriptors with catalyst descriptors, individual reaction profiles could be constructed that also took into account substrate properties. The inclusion of substrate descriptors also increased the number of data points obtained per catalyst synthesized. As a general note on the use of reactant and product descriptors, we find that it provides the following benefits: More data points can be collected per catalyst synthesized, allowing for stronger models to be produced, and we can use our technology to predict the outcome of reactions with a known catalyst on new substrate combinations, thus creating another powerful tool.

Generation of models by using all selectivity data

To ensure that our descriptors capture the structural information pertaining to enantioinduction and can be used to construct predictive models, 2150 separate experiments were performed, wherein the catalysts shown in Figs. 4A and 5B (43 catalysts in total) were used in the reactions with pairwise combinations of imines and thiols, leading to the 25 different products shown in Fig. 5A. This process creates $43 \times 25 = 1075$ reactions, which were run in duplicate, and the average of duplicate runs was used as the experimental selectivity data. From these 1075 reactions, 475 were randomly selected as an external test set by using a Python random-number generator, and the remaining 600 reactions were used to train the model. To ensure that model efficacy and training set–test set partitioning were unbiased, this process was repeated 10 times. Models were then developed with support vector machines by using a grid-based optimization of hyperparameters with fivefold cross validation (see supplementary materials for details). The average predicted selectivities of the 475 external test set reactions (i.e., those which were not used in the model training process) reveal very good correlation when plotted against the experimental selectivity data (a high coefficient of determination R^2 , a y intercept very close to zero, and a slope approaching unity) (Fig. 6A). The mean absolute deviation (MAD) for each of the 10 randomized trials is listed in Fig. 6B. As is evident from the low MAD of each run, the models make highly accurate predictions of selectivity, confirming that our descriptor set is a valid, numerical representation of molecules capable of capturing the relevant features of catalysts responsible for enantioselectivity.

As experimentalists, we were interested in establishing whether these tools could be used to predict the results of either new substrate combinations or new catalysts that have not previously been tested or to identify new reactions (i.e., substrates and catalysts) that are more selective than any reaction in the training data. We therefore performed two modeling studies to evaluate each hypothesis by partitioning the available data in two different ways. For the first study, the data from reactions of four imines (imines **25a** to **25d**) and four thiols (thiols **26a** to **26d**) (i.e., 16 reactions per catalyst) were evaluated (Fig. 5A). Using the 24-member catalyst training set (Fig. 4A) with each substrate combination then gave rise to $16 \times 24 = 384$ training reactions that could be used for model development. This process also generated $1075 - 384 = 691$ test reactions for external validation (the test reactions were later divided into three different sets, detailed below). For the second study, we investigated whether new, more selective reactions could be predicted. To investigate this possibility, the 1075 experimental selectivity data points were divided such that every catalyst-imine-thiol combination that gave products below 80% ee was included in the training set and no reactions above 80% ee were used at any stage in model development. These remaining, highly selective reactions were instead used as an external test set. Both data division methods violate the iid (independent and identically distributed) assumption (42). Thus, we make no claims as to the generalizability of these studies and simply propose this method as a tool to facilitate the experimental optimization of catalysts and the exploration of substrate scope.

Generation of models derived from the UTS

It was very rewarding to find a highly selective catalyst in the training set (compound **11**) (Fig. 4A), supporting our hypothesis that using the UTS increases the probability of finding an effective catalyst in the first round of screening (catalyst selectivity data are summarized in Fig. 4B). 3,3'-Benzyl-substituted catalysts used in reactions with aliphatic thiols as nucleophiles gave rise to the opposite stereoisomer as the major product compared with the other cases. Thus, the range of selectivities covered by the UTS in the 16 training reactions spans from -43% ee to >99% ee with the same enantiomer of catalyst, further supporting the hypothesis that the UTS covers a broad range of selectivity space, as illustrated in the full compilation of experimental results (table S1). From this dataset, a suite of models was generated and used to predict the selectivity of three families of test sets: a substrate test set of reactions generating new products (i.e., those formed from substrates not included in the training set but using catalysts in the training set), a catalyst test set of reactions generating the same products in the training set but with catalysts not included in the training data, and a substrate-catalyst (sub-cat) test set of reactions creating new products and also using catalysts not included in the training data. For the substrate test set, nine distinct compounds (**31**, **36**, **41**, and **46** to **51**) (Fig. 5A) generated from substrate combinations with unknown results in the model reaction were selected, totaling 216 reactions (24 training catalysts $\times$ 9 test substrates). For the catalyst test set, the 19 external catalysts (**52** to **70**) (Fig. 5B) were evaluated in reactions generating the same products as the training reactions, totaling 304 reactions (19 test catalysts from Fig. 5B $\times$ 16 training substrates from Fig. 5A). For the sub-cat test set, the 19 external test set catalysts were used in reactions producing the nine new products, thus evaluating the capability to predict reaction outcomes with external substrate combinations and external catalysts, totaling 171 reactions [19 test catalysts $\times$ 9 test substrates (Fig. 5B catalysts with Fig. 5A test substrate combinations)].

By using a variety of data preprocessing methods (see supplementary materials for details), we generated a suite of models. Of these, the support vector machines method gave the highest performance on the basis of the MAD from the combined external test sets (Fig. 7A). The first test set evaluated the ability of the models to predict the selectivity only of reactions forming new products. In this role, the model performed well, with an MAD of 0.161 kcal/mol. Next, the same model was used to predict the selectivity of the external test set of catalysts. The performance of the model was still highly accurate, with a MAD of 0.211 kcal/mol. Lastly, reactions forming new products with the external test catalysts were predicted with a MAD of 0.236 kcal/mol. All three test sets were predicted at a level of accuracy equal to or greater than that of most quantum chemistry methods (43). To evaluate catalyst performance, the mean selectivity (G in kilocalories per mole) of each test catalyst across all 25 reactions was calculated (Fig. 5B). The model predicted this efficacy metric with notable accuracy, predicting all catalysts within 0.4 kcal/mol, with only two catalysts (**53** and **54**) predicted outside of 0.3 kcal/mol from the experimentally observed

G . Catalyst **53** gave the best selectivity in the original study (41), and our results were in good agreement with what has been previously reported. Similarly, aliphatic thiols gave diminished selectivity with respect to thiophenol derivatives. The first three principal components of catalyst space also reveal distinct regions of high, medium, and low space

(Fig. 7B). Similarly, the predicted reaction outcomes for the nine test reactions with the best catalyst are illustrated in Fig. 7C. All reaction selectivities except one (49) are predicted within 2% ee of the measured value. Despite not being included at any stage of model development, compound **52** was still predicted to be the most selective catalyst in the in silico library for this transformation. A complete list of predicted selectivity values for the entire in silico library of reactions can be found in data S1.

Reaction prediction beyond the training set

Although modeling with data spanning the entire range (e.g., up to 99% ee) of interest can be valuable for predicting the outcome of new substrate combinations, modeling beyond this range can be leveraged to enhance the rate at which catalytic enantioselective reactions are optimized. To demonstrate this potential in our method, we simulated a situation in which highly selective reactions (i.e., combinations of substrates and catalysts) have not been identified. Accordingly, we partitioned all 1075 reactions as follows: All reactions below 80% ee were used as training data (718 reactions), and all reactions above 80% ee were used as test data (357 reactions). (The identities of the training and test datasets can be found in the supplementary materials.) A variety of modeling methods were tested, and although a number of methods, including support vectors, Lasso, LassoLars, ridge regression, elastic net, and random forest (by no means the state of the art in machine learning; see computational methods for a complete explanation), provided acceptable qualitative results, deep feed-forward neural networks accurately reproduced the experimental selectivities (MAD = 0.33 kcal/mol) (Fig. 8A). More notably, the general trends in selectivity, on the basis of average catalyst selectivity, were correctly identified. As shown in Fig. 8B, the most selective catalyst, **52**, was predicted with the highest accuracy, within 3% ee of the experimental value. Catalysts **53** and **54** were the next two to follow experimentally and computationally (the order is inverted, but they are within experimental error from each other), followed by catalyst **55**. The remaining catalysts shown in Fig. 8B were predicted very accurately, likely because the experimental values are closer to the training set cutoff of 80% ee. Despite omitting about half of the experimental free energy range from the training data, we could still make accurate predictions in this region of selectivity space.

Outlook

The capability to successfully predict the selectivity of higher-performing catalysts has the potential to change the way chemists select and optimize chiral catalysts. This method has not been “pressure tested” in a number of scenarios; reactions that are susceptible to electronic perturbation must be investigated, more flexible catalyst scaffolds need to be explored, and catalyst scaffolds with multiple points of diversity must be examined.

Materials and methods

General information

All reactions were performed in glassware that had been flame-dried under vacuum or oven-dried (140°C) overnight. All reactions were conducted under an atmosphere of dry nitrogen or argon by using a drying tube equipped with phosphorus pentoxide and calcium sulfate.

All reaction temperatures are noted as the oil bath temperature, the internal temperature as monitored by a Teflon-coated thermocouple, or the room temperature (~23°C). Solvents used for extraction were reagent grade, and chromatography solvents were technical grade. Column chromatography was performed using ultrapure silica gel (40 to 69 μm) from Silicycle with a column mixed as a slurry, packed, and eluted at 6 to 8 psi. Retention factors, R_f , are reported for analytical thin-layer chromatography performed on Merck silica gel plates treated with F-254 indicator. Visualizations were accomplished by using ultraviolet (UV) light, aqueous KMnO_4 , ceric ammonium molybdate solution, or iodine. Reaction solvents tetrahydrofuran [Fischer; high-performance liquid chromatography (HPLC) grade], hexanes (Fischer; HPLC grade), diethyl ether [Fischer; butylated hydroxytoluene-stabilized American Chemical Society (ACS) grade], methylene chloride (Fischer; unstabilized HPLC grade), and N,N' -dimethylformamide (Fischer; HPLC grade) were dried by percolation through two columns packed with neutral alumina under positive pressure of argon. Toluene (Fischer; ACS grade) was dried by percolation through a column packed with neutral alumina and a column packed with Q5 reactant, a supported copper catalyst for scavenging oxygen, under a positive pressure of argon. Amines were distilled fresh before use, and pyridine (Fischer; ACS grade) used as a solvent was distilled and stored over 4-Å molecular sieves before use.

Instrumentation

^1H , ^{13}C , ^{19}F , and ^{31}P nuclear magnetic resonance (NMR) spectra were recorded on a Varian Unity Inova 400 spectrometer (^{19}F and ^{31}P), a Varian Unity 500 spectrometer (^1H and ^{13}C), a Bruker Advance 500 spectrometer (^1H , ^{13}C , ^{19}F , ^{29}Si , and ^{31}P), a Varian VXR 500 spectrometer (^1H), or a Unity 500 NB spectrometer (^1H). Spectra are referenced to chloroform [$\delta = 7.26$ parts per million (ppm), ^1H ; 77.0 ppm, ^{13}C], residual benzene ($\delta = 7.15$ ppm, ^1H ; 128.62 ppm, ^{13}C), or dimethyl sulfoxide ($\delta = 2.50$ ppm, ^1H ; 39.52 ppm, ^{13}C). Chemical shifts are reported in parts per million, and multiplicities are indicated by s (singlet), d (doublet), t (triplet), q (quartet), p (pentet), h (hextet), m (multiplet), and br (broad). Coupling constants J are reported in hertz, integration is provided, and assignments are indicated. Mass spectrometry was performed by the University of Illinois Mass Spectrometry Laboratory. Mass spectrometric data were collected on a Waters Q-TOF Ultima (ESI) spectrometer, a Waters Synapt G2-Si spectrometer (ESI), a Waters Quattro Ultima spectrometer (ESI), or a Waters 70-VSE spectrometer (EI). Low-resolution spectral data are reported as (mass, intensity), and high-resolution data are reported as calculated and measured masses to 10^{-4} mass accuracy. Infrared spectra were recorded on a Perkin-Elmer UATR-2 FT-IR spectrophotometer. Peaks are reported as reciprocal centimeters, with relative intensities indicated as s (strong, 67 to 100%), m (medium, 34 to 66%), or w (weak, 0 to 33%). Analytical chiral stationary-phase super-critical fluid chromatography was performed on an Agilent 1100 high-performance liquid chromatograph equipped with an Aurora Systems A-5 supercritical CO_2 adapter for supercritical fluid chromatography and a UV detector (220 or 254 nm) by using Daicel Chiralcel OD, OJ, and OB or Chiralpak AD and AS columns.

Descriptor calculations

To describe the steric environment around a given structure, the strategy taken in these laboratories used grid point–type descriptors. However, instead of using van der Waals potential energy values at grid point locations, this descriptor incorporates steric data from a population of conformers of a given compound. The new calculation process is as follows (see section S3 in the supplementary materials), demonstrated by using a BINOL-based phosphoric acid derivative scaffold. (i) For each base compound within an *in silico* library, a set of conformers within a given energy window (generally 7 to 10 kcal/mol) is generated. (ii) The full set of compounds and associated conformer libraries are aligned to a common core. (iii) A spherical grid of points is then calculated to encompass the entire set of aligned compounds to a depth of 3 Å. (iv) For each conformer, an indicator field is created by determining which grid point locations are within the van der Waals radius of an atom. Locations determined to be within atoms are given a value of 1; those outside are given a value of 0. (v) The ASO of a given catalyst is calculated as the average of the indicator fields for each conformer of that catalyst. This gives a descriptor value of 0 ASO 1 at each grid point. When compiled, the descriptor set acts both to describe the shape of the molecule and to weight that shape with how often the molecule occupies different regions of space. The process of calculating the ASO descriptor set is completed for every catalyst in the *in silico* library (fig. S4). The same protocol is used to calculate starting material and product descriptors, and concatenation of the descriptors generates the reaction profiles.

The ASO descriptor can be used to visually compare the shapes and sizes of different compounds by plotting the descriptor values as bar charts. Shown in figure S5 is a comparison between 3,3'-diphenyl-substituted BINOL-phosphoramidate **1_iv** and the much larger catalyst **182_iv**. As can be seen in the plots, the ASO descriptor values for **182_iv** are much more varied, and nonzero ASO values can be seen for much more of the available descriptor range, indicating that this catalyst is much larger and covers more of the space available to the catalyst. This type of comparative analysis shows that the descriptors are capturing the shape of the molecule as well as distinguishing between catalysts of different sizes and constitutions (Fig. 2D).

To capture the electrostatic effects of substituents on the compounds of interest, a separate set of descriptors was considered. Electrostatic MIF descriptors have underperformed in the applications tested in these laboratories, and these 3D MIF-based electrostatic descriptors do not incorporate conformation-dependent information. Additionally, most descriptor calculation methods based on electrostatic field determinations fail to distinguish between through-bond and through-space effects (44). Although others have used 1D and 2D descriptors, such as Hammett parameters (45), to describe such changes, the substituent libraries used in these laboratories are too diverse to have these parameters derived for them. To that end, a new electrostatic parameter that correlates well with known 1D parameters has been devised.

This electrostatic parameter was calculated for individual substituents represented in the catalyst *in silico* library and is used to estimate the electronic effects of the substituents on the core molecule. The calculation was performed by attaching the substituent group to a

tetramethylammonium cation, generating a benzyltrimethylammonium cation if the substituent is aryl, a homobenzyl-trimethylammonium cation if the substituent is benzyl, or an tetraalkylammonium ion if the substituent is alkyl. An electrostatic potential MIF is then calculated by using NWChem (46) at the B3LYP/6–31G* level of theory, specifying a specific probe and range for the grid to give a single layer of grid points 0.025 Å apart. An example of the grid and calculated electrostatic potential for a 4-nitrobenzyltrimethylammonium cation is shown in figure S6. After the energies are calculated, the maximum and minimum energies calculated in the single-layer MIF are saved, giving the substituent electrostatic potential energy minimum (ESPMIN) and substituent electrostatic potential energy maximum (ESPMAX) descriptors. The ESPMAX descriptors correlated well with known Hammett parameters, suggesting that the descriptor was describing the electron-donating or -withdrawing nature of the given substituents (fig. S7). These electronic parameters were also used for the nucleophiles and electrophiles, wherein the corresponding thioether and aryl moieties, respectively, were appended to the ammonium ion. Further, natural bond orbital (NBO) charges for sulfur and sulfur molecular orbital energies from the NBO calculation were used as electronic descriptors for the thiols.

Model generation

All machine learning methods except deep neural networks were implemented with Python2 scripts by using scikit-learn (47), a Python machine learning package. A collection of models was generated by using a variety of feature selection methods with experimental

G as the observable. Before modeling, all data descriptors were scaled by removing the mean and scaling to unit variance. A variety of feature selection or transform methods were surveyed (variance threshold method, mutual information, f-regression, and PCA). For the feature selection methods, 100, 500, 1000, and 2000 features were selected. Additionally, by using a percentile cutoff, the 10th, 25th, and 50th percentiles were selected. By using PCA, models were generated with 10, 20, 30, 50, and 100 principal components (64, 78, 84, 89, and 94% of variance, respectively). These methods were all performed separately on the scaled descriptor data (PCA and a feature selection method were never used together). The enantioselectivity data (expressed as the free energy differential between the diastereomeric transition structures leading to the different enantiomers) were also highly skewed, so these data were transformed with the Box-Cox transformation by using SciPy before model generation. Each set of preprocessed data (meaning one of the selection methods or PCA on the features with the transformed or untransformed Y-data) were then used to make a collection of models. Models generated include partial least squares PLSn (where $n = 2, 4, 6, 8, 10, 14, 18$ and in which $n < \text{number of principal components}$), random forest, LassoCV, LassoLarsCV, ElasticNetCV, RidgeCV, kernel RidgeCV [kernel = radial basis function (rbf)] [k (number of folds in k -fold cross-validation) = 5], k -nearest neighbors (kNN), and support vector machines with linear, rbf, and polynomial kernels (second-, third-, and fourth-order polynomials). Grid optimization of hyperparameters was performed (example code can be found on the GitLab site) (48). This hyper-parameter optimization was performed with a fivefold train-validation split (e.g., in the case of the UTS data, the 384 “training reactions” were split). Models were evaluated via q^2 (cross-validated R^2), R^2 , and MAD from an external test set of reactions (not used in hyperparameter optimization). Three examples are given in fig. S8. Each model used the transformed (Box-Cox) selectivity data and the top

25% of features selected by mutual information regression. SVR_poly2 (support vector regressor with second-order polynomial kernel) was the only member of the best models to accurately predict the most selective test reactions.

Whereas SVR_poly2 qualitatively selected the best reactions when attempting to predict beyond the range of selectivities in the training data, the models quantitatively underpredicted these reactions. By using Keras (49) with the Theano backend, a Python package that can facilitate deep learning, we generated a deep feed-forward network. Grid-based hyperparameter optimization was used with linear, relu, elu, and selu activation functions; 0.05, 0.1, and 0.2 dropouts on the layers; 4, 40, 400, and 4000 nodes per layer; and 0 to 6 hidden layers. Further, all optimizers available in Keras were tested. This method of hyperparameter optimization was very time intensive, and it is strongly recommended that practitioners instead use a Bayesian optimization of hyperparameters. Attempts to use this kind of optimization and more modern machine learning methods are currently under way.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

We thank K. A. Robb and Z. Wickenhauser for experimental assistance and N. Russell for informative discussions about machine learning. We are also grateful for the support services of the NMR, mass spectrometry, and microanalytical laboratories of the University of Illinois at Urbana-Champaign.

Funding: We are grateful for generous financial support from the W. M. Keck Foundation. A.F.Z. is grateful to the University of Illinois for graduate fellowships. Y.W. thanks Janssen Research Development, San Diego, CA, for a postdoctoral fellowship.

REFERENCES AND NOTES

1. Engel T, Basic overview of chemoinformatics. *J. Chem. Inf. Model* 46, 2267–2277 (2006). doi: 10.1021/ci600234z; pmid: 17125169 [PubMed: 17125169]
2. Willett P, Chemoinformatics: A history. *Wiley Interdiscip. Rev. Comput. Mol. Sci* 1, 46–56 (2011). doi: 10.1002/wcms.1
3. Agrafiotis DK, Bandyopadhyay D, Wegner JK, Vlijmen Hv., Recent advances in chemoinformatics. *J. Chem. Inf. Model* 47, 1279–1293 (2007). doi: 10.1021/ci700059g; pmid: 17511441 [PubMed: 17511441]
4. Feigenbaum EA, Buchanan BG, DENDRAL and Meta-DENDRAL: Roots of knowledge systems and expert system applications. *Artif. Intell* 59, 233–240 (1993). doi: 10.1016/0004-3702(93)90191-D
5. Segler MHS, Preuss M, Waller MP, Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* 555, 604–610 (2018). doi: 10.1038/nature25978; pmid: 29595767 [PubMed: 29595767]
6. Szymku S et al., Computer-assisted synthetic planning: The end of the beginning. *Angew. Chem. Int. Ed. Engl* 55, 5904–5937 (2016). doi: 10.1002/anie.201506101; pmid: 27062365 [PubMed: 27062365]
7. Wei JN, Duvenaud D, Aspuru-Guzik A, Neural networks for the prediction of organic chemistry reactions. *ACS Cent. Sci* 2, 725–732 (2016). doi: 10.1021/acscentsci.6b00219; pmid: 27800555 [PubMed: 27800555]

8. Coley CW, Barzilay R, Jaakkola TS, Green WH, Jensen KF, Prediction of organic reaction outcomes using machine learning. *ACS Cent. Sci* 3, 434–443 (2017).doi: 10.1021/acscentsci.7b00064; pmid: 28573205 [PubMed: 28573205]
9. Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T, The rise of deep learning in drug discovery. *Drug Discov. Today* 23, 1241–1250 (2018). doi: 10.1016/j.drudis.2018.01.039; pmid: 29366762 [PubMed: 29366762]
10. Ma J, Sheridan RP, Liaw A, Dahl GE, Svetnik V, Deep neural nets as a method for quantitative structure-activity relationships. *J. Chem. Inf. Model* 55, 263–274 (2015). doi: 10.1021/ci500747n; pmid: 25635324 [PubMed: 25635324]
11. Denmark SE, Gould ND, Wolf LM, A systematic investigation of quaternary ammonium ions as asymmetric phase-transfer catalysts. Synthesis of catalyst libraries and evaluation of catalyst activity. *J. Org. Chem* 76, 4260–4336 (2011). doi: 10.1021/jo2005445; pmid: 21446721 [PubMed: 21446721]
12. Denmark SE, Gould ND, Wolf LM, A systematic investigation of quaternary ammonium ions as asymmetric phase-transfer catalysts. Application of quantitative structure activity/selectivity relationships. *J. Org. Chem* 76, 4337–4357 (2011). doi: 10.1021/jo2005457; pmid: 21446723 [PubMed: 21446723]
13. Gómez-Bombarelli R et al., Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci* 4, 268–276 (2018). doi: 10.1021/acscentsci.7b00572; pmid: 29532027 [PubMed: 29532027]
14. Raccuglia P et al., Machine-learning-assisted materials discovery using failed experiments. *Nature* 533, 73–76 (2016). doi: 10.1038/nature17439; pmid: 27147027 [PubMed: 27147027]
15. Bartók AP, Gillan MJ, Manby FR, Csányi G, Machine-learning approach for one- and two-body corrections to density functional theory: Applications to molecular and condensed water. *Phys. Rev. B* 88, 054104 (2013). doi: 10.1103/PhysRevB.88.054104
16. Zhou Z, Li X, Zare RN, Optimizing chemical reactions with deep reinforcement learning. *ACS Cent. Sci* 3, 1337–1344 (2017). doi: 10.1021/acscentsci.7b00492; pmid: 29296675 [PubMed: 29296675]
17. Mitchell JBO, Machine learning methods in chemoinformatics. *Wiley Interdiscip. Rev. Comput. Mol. Sci* 4, 468–481 (2014). doi: 10.1002/wcms.1183; pmid: 25285160 [PubMed: 25285160]
18. Lipkowitz KB, Pradhan M, Computational studies of chiral catalysts: A comparative molecular field analysis of an asymmetric Diels-Alder reaction with catalysts containing bisoxazoline or phosphinoxazoline ligands. *J. Org. Chem* 68, 4648–4656 (2003). doi: 10.1021/jo0267697; pmid: 12790567 [PubMed: 12790567]
19. Kozlowski MC, Dixon SL, Panda M, Lauri G, Quantum mechanical models correlating structure with selectivity: Predicting the enantioselectivity of α -amino alcohol catalysts in aldehyde alkylation. *J. Am. Chem. Soc* 125, 6614–6615 (2003). doi: 10.1021/ja0293195; pmid: 12769554 [PubMed: 12769554]
20. Melville JL, Andrews BI, Lygo B, Hirst JD, Computational screening of combinatorial catalyst libraries. *Chem. Commun* 2004, 1410–1411 (2004). doi: 10.1039/b402378a; pmid: 15179489
21. Sciabola S et al., Theoretical prediction of the enantiomeric excess in asymmetric catalysis. An alignment-independent molecular interaction field based approach. *J. Org. Chem* 70, 9025–9027 (2005). doi: 10.1021/jo051496b; pmid: 16238344 [PubMed: 16238344]
22. Harper KC, Sigman MS, Predicting and optimizing asymmetric catalyst performance using the principles of experimental design and steric parameters. *Proc. Natl. Acad. Sci. U.S.A.* 108, 2179–2183 (2011). doi: 10.1073/pnas.1013331108; pmid: 21262844 [PubMed: 21262844]
23. Harper KC, Sigman MS, Three-dimensional correlation of steric and electronic free energy relationships guides asymmetric propargylation. *Science* 333, 1875–1878 (2011). doi: 10.1126/science.1206997; pmid: 21960632 [PubMed: 21960632]
24. Sigman MS, Harper KC, Bess EN, Milo A, The development of multidimensional analysis tools for asymmetric catalysis and beyond. *Acc. Chem. Res* 49, 1292–1301 (2016). doi: 10.1021/acs.accounts.6b00194; pmid: 27220055 [PubMed: 27220055]

25. Harper KC, Bess EN, Sigman MS, Multidimensional steric parameters in the analysis of asymmetric catalytic reactions. *Nat. Chem* 4, 366–374 (2012). doi: 10.1038/nchem.1297; pmid: 22522256 [PubMed: 22522256]
26. Park Y, Niemeyer ZL, Yu J-Q, Sigman MS, Quantifying structural effects of amino acid ligands in Pd(II)-catalyzed enantioselective C–H functionalization reactions. *Organometallics* 37, 203–210 (2018). doi: 10.1021/acs.organomet.7b00751
27. Ahneman DT, Estrada JG, Lin S, Dreher SD, Doyle AG, Predicting reaction performance in C–N cross-coupling using machine learning. *Science* 360, 186–190 (2018).doi: 10.1126/science.aar5169; pmid: 29449509 [PubMed: 29449509]
28. Nielsen MK, Ahneman DT, Riera O, Doyle AG, Deoxyfluorination with sulfonyl fluorides: Navigating reaction space with machine learning. *J. Am. Chem. Soc* 140, 5004–5008 (2018). doi: 10.1021/jacs.8b01523;pmid: 29584953 [PubMed: 29584953]
29. Breiman L, Random forests. *Mach. Learn* 45, 5–32 (2001). doi: 10.1023/A:1010933404324
30. Skoraczyski G et al., Predicting the outcomes of organic reactions via machine learning: Are current descriptors sufficient? *Sci. Rep* 7, 3582 (2017). doi: 10.1038/s41598-017-02303-0; pmid: 28620199 [PubMed: 28620199]
31. In this case, meaning as many as can be accessed by the synthesis of fragments that require no more than four well-established synthetic steps before being combined with a common scaffold.
32. Parmar D, Sugiono E, Raja S, Rueping M, Addition and correction to 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* 117, 10608–10620 (2017). doi: 10.1021/acs.chemrev.7b00197; pmid: 28737901 [PubMed: 28737901]
33. Roy K, Kar S, Das RN, in *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment*, Roy K, Kar S, Das RN, Eds. (Academic Press, 2015), pp. 291–317.
34. Cruz VL, Martinez S, Ramos J, Martinez-Salazar J, 3D-QSAR as a tool for understanding and improving single-site polymerization catalysts. A review. *Organometallics* 33, 2944–2959 (2014). doi: 10.1021/om400721v
35. Braiuca P, Lorena K, Ferrario V, Ebert C, Gardossi L, A three-dimensional quantitative structure-activity relationship (3D-QSAR) model for predicting the enantioselectivity of *Candida antarctica* lipase B. *Adv. Synth. Catal* 351, 1293–1302 (2009). doi: 10.1002/adsc.200900009
36. Senese CL, Duca J, Pan D, Hopfinger AJ, Tseng YJ, 4D-fingerprints, universal QSAR and QSPR descriptors. *J. Chem. Inf. Comput. Sci* 44, 1526–1539 (2004). doi: 10.1021/ci049898s; pmid: 15446810 [PubMed: 15446810]
37. Melville JL et al., Exploring phase-transfer catalysis with molecular dynamics and 3D/4D quantitative structure-selectivity relationships. *J. Chem. Inf. Model* 45, 971–981 (2005). doi: 10.1021/ci050051i; pmid: 16045291 [PubMed: 16045291]
38. Bellman RE, *Dynamic Programming* (Princeton Univ. Press, 1957).
39. Pearson K, LIII. On lines and planes of closest fit to systems of points in space. *London Edinb. Dublin Philos. Mag. J. Sci* 2, 559–572 (1901). doi: 10.1080/14786440109462720
40. Kennard RW, Stone LA, Computer aided design of experiments. *Technometrics* 11, 137–148 (1969). doi: 10.1080/00401706.1969.10490666
41. Ingle GK, Mormino MG, Wojtas L, Antilla JC, Chiral phosphoric acid-catalyzed addition of thiols to N-acyl imines: Access to chiral N,S-acetals. *Org. Lett* 13, 4822–4825 (2011). doi: 10.1021/ol201899c; pmid: 21842841 [PubMed: 21842841]
42. Steinwart I, Hush D, Scovel C, Learning from dependent observations. *J. Multivar. Anal* 100, 175–194 (2009).doi: 10.1016/j.jmva.2008.04.001
43. Simón L, Goodman JM, Theoretical study of the mechanism of Hantzsch ester hydrogenation of imines catalyzed by chiral BINOL-phosphoric acids. *J. Am. Chem. Soc* 130, 8741–8747 (2008). doi: 10.1021/ja800793t; pmid: 18543923 [PubMed: 18543923]
44. Wheeler SE, Houk KN, Through-space effects of substituents dominate molecular electrostatic potentials of substituted arenes. *J. Chem. Theory Comput.* 5, 2301–2312 (2009). doi: 10.1021/ct900344g; pmid: 20161573 [PubMed: 20161573]

45. Hansch C, Leo A, Taft RW, A survey of Hammett substituent constants and resonance and field parameters. *Chem. Rev* 91, 165–195 (1991). doi: 10.1021/cr00002a004
46. Valiev M et al., NWChem: A comprehensive and scalable open-source solution for large scale molecular simulations. *Comput. Phys. Commun* 181, 1477–1489 (2010). doi: 10.1016/j.cpc.2010.04.018
47. Pedregosa F et al., Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res* 12, 2825–2830 (2011).
48. Denmark Lab Chemoinformatics, ccheminfolib, Project ID 8113486, GitLab (2018); <https://gitlab.com/SEDenmarkLab/ccheminfolib>.
49. Chollet F, Keras: Deep learning for humans, GitHub; <https://github.com/fchollet/keras>.

Fig. 1. Summary of chemoinformatics-guided workflow.

(A) An *in silico* library of synthetically accessible catalysts is defined. For each member in the library, descriptors are calculated. (B) A representative subset is algorithmically selected on the basis of intrinsic chemical properties. (C) The representative subset is synthesized and experimentally tested. (D) The probability of identifying a highly selective catalyst in the first round of screening should be greater than that by random sampling alone. (E) The data from the training set are used to train statistical learning methods. (F) The models predict selectivity values for every member of the greater *in silico* library. (G) If successful, the model will predict the optimal catalyst for the reaction. If unsuccessful, the new data can be used as training data to make a stronger prediction in successive rounds of modeling. R, any group; X, O or S; Y, OH, SH, or NHTf; *i*-Pr, isopropyl; *t*-Bu, *tert*-butyl; Cy, cyclohexyl.

Fig. 2. Generation of ASO descriptors.

(A) Pictorial description of the ASO calculation process. (B) ASO grid points away from the catalyst have values of 0 (red), whereas grid points occupied in all conformers have a value of 1 (blue); flexible substituents can be seen in the green and yellow regions. (C) ASO discrimination of 3,3'-substituent groups: ortho-substituted arenes (red), fused-ring substituents (blue), 3,5-disubstituted arenes (yellow), and all other groups (green). (D) Bar graph representation of ASO descriptors for two different Brønsted acid catalysts. calc., calculated; vdW, van der Waals.

Fig. 3. Construction of the UTS.

(A) Subset selection with the Kennard-Stone algorithm. The algorithm then selects a representative subset of points, as qualitatively depicted. (B) Locations of the catalysts selected by the Kennard-Stone algorithm in 2D chemical space [constructed from the first two principal components (18% and 12% of variance, respectively) of the full catalyst chemical space].

Fig. 4. BINOL phosphoric acid (BPA) UTS. (A) UTS of phosphoric acid catalysts selected by the Kennard-Stone algorithm. (B) Average selectivity of training catalysts across the 16 training reactions. Ph, phenyl; Me, methyl; Et, ethyl.

A Model Reaction with Training and Test Substrate Combinations **B Test Catalysts with Averages for All Substrate Combinations** **Fig. 5. Model validation on thiol addition to *N*-acyl imines.**

(A) Model reaction screened over 16 training substrate combinations and nine test substrate combinations. (B) Test set of catalysts. Each catalyst was evaluated by the average selectivity across all 25 substrate combinations. The experimentally observed selectivity is in bold, and the predicted selectivity (reported as the free energy differential between the transition structures leading to each enantiomer) is in italics. r.t., room temperature.

Fig. 6. The averaged predicted selectivity values for external test sets plotted against observed enantioselectivity data.

(**A**) The vertical bands result from the accuracy in the analytical method, wherein the limit of detection determines enantiomeric purity to the nearest 0.5% ee. Because of the exponential relationship between ee and free energy, detectable differences in selectivity appear greater at larger free energy differentials. (**B**) The MAD data are listed in the table for each of the 10 replicate runs.

Fig. 7. Application of models from UTS.

(A) The predicted versus observed plots for the training set, substrate test set, catalyst test set, and sub-cat test set. The support vector machines method (second-order polynomial kernel, $q^2 = 0.748$ by k -fold cross validation) performs well on all external test sets, predicting reaction outcomes within 0.25 kcal/mol (MAD = 0.161, 0.211, and 0.238 kcal/mol, respectively). The vertical bands result from the limit of accuracy in the analytical method. (B) The 3D chemical space of all catalysts (from the first three principal components of the full chemical space, 13, 8, and 8% of variance, respectively). The red points are unselective catalysts, the green and yellow points are more selective, and the blue points are the most selective, with the average selectivity across all 25 reactions as a metric of catalyst selectivity. (C) Observed and predicted outcomes of reactions with substrate combinations that were not also included in the training data.

A Predicted vs. Observed: **B Average Test Catalyst Selectivity:** **Fig. 8. Reaction prediction beyond the selectivity spanned by the training set.**

(A) A model generated by using a deep feed-forward neural network simulating the optimization of an unoptimized reaction by using all data below 80% ee to train the model. The vertical bands result from the limit of accuracy in the analytical method. (B) Predicted and observed average selectivities for the eight catalysts with average enantioselectivity over 80% ee. Only the common reactions (i.e., those forming the same product) that were in the test set for each of the eight catalysts were used to calculate the average selectivities. The identity of these reactions and the predicted and observed values are available in data S3.