Reaction Chemistry & Engineering

ROYAL SOCIETY OF CHEMISTRY

PAPER

Cite this: *React. Chem. Eng.*, 2021, **6**, 694

Computational methods for training set selection and error assessment applied to catalyst design: guidelines for deciding which reactions to run first and which to run next†

Andrew F. Zahrt, Brennan T. Rose, William T. Darrow, Jeremy J. Henle and Scott E. Denmark *

Received 7th January 2021,
Accepted 8th February 2021

DOI: 10.1039/d1re00013f

rsc.li/reaction-engineering

The application of machine learning (ML) to problems in homogeneous catalysis has emerged as a promising avenue for catalyst optimization. An important aspect of such optimization campaigns is determining which reactions to run at the outset of experimentation and which future predictions are the most reliable. Herein, we explore methods for these two tasks in the context of our previously developed chemoinformatics workflow. First, different methods for training set selection for library-based optimization problems are compared, including algorithmic selection and selection informed by unsupervised learning methods. Next, an array of different metrics for assessment of prediction confidence are examined in multiple catalyst manifolds. These approaches will inform future computer-guided studies to accelerate catalyst selection and reaction optimization. Finally, this work demonstrates the generality of the average steric occupancy (ASO) and average electronic indicator field (AEIF) descriptors in their application to transition metal catalysts for the first time.

Introduction

Since the turn of the century, computational methods for enantioselective catalyst optimization have gained traction within the scientific community.[1–28] The most established method for computational catalyst design is transition state analysis with quantum chemistry or force field methods calculate the relative energy differentials leading to enantiomers which then enables more informed catalyst optimization.[2,5,6,8,10–12,29–33] A more recent alternative to these approaches is the application of quantitative structure–selectivity relationships (QSSR).[34] In this method, numerical values representing catalyst structural characteristics are correlated with an experimental observable, generating a mathematical model which can be used to evaluate new catalyst structures *in silico*. QSSR also has the advantage of being mechanistically agnostic at the outset of investigation. In enantioselective catalysis, the seminal example of QSSR was reported by Norrby and coworkers to predict ratios of isomeric products from various nucleophilic substitution reactions on palladium η^3-allyl complexes.[35] Since this initial report, this field has become an established area of research.[36–44] Of particular note, Sigman and coworkers have pioneered the application of linear free energy relationships (LFERs) for mechanistic interrogation.[45]

In our own laboratories, molecular interaction field (MIF)-based approaches have been investigated to elucidate important structural characteristics of phase transfer catalysts.[46,47] More recently, we have used additional statistical learning protocols with MIF-type descriptors to evaluate chiral catalysts, culminating in a computer-driven workflow for the optimization of enantioselective catalysts.[48,49] The aim of this workflow is to identify the most selective catalyst from a large *in silico* library of catalysts in a way that is agnostic of mechanism. However, these studies, like many studies investigating the application of machine learning to enantioselective catalysis, are proof-of-principle studies. In practice, when employing this kind of a workflow, researchers must have quantitative measures for identifying which chemical entities to synthesize and which predictions are the most reliable. Notably, some of these concepts have been explored in other applications in the chemical sciences.[50–57] However, these concepts have not been explored rigorously in the context of enantioselective catalysis. In the work described

245 Roger Adams Laboratory, Department of Chemistry, University of Illinois, 600 S. Mathews Ave, Urbana, IL 61801, USA. E-mail: sdenmark@illinois.edu; Fax: +(217) 333 3984; Tel: +(217) 333 0066

† Electronic supplementary information (ESI) available. See DOI: 10.1039/d1re00013f

herein, a variety of catalyst selection protocols are evaluated. Further, multiple metrics for error assessment are compared in multiple catalyst systems. Specifically, error assessment metrics used previously in the literature are tested to determine how best to use them in our previously published workflow. Finally, suggestions are made as to how these

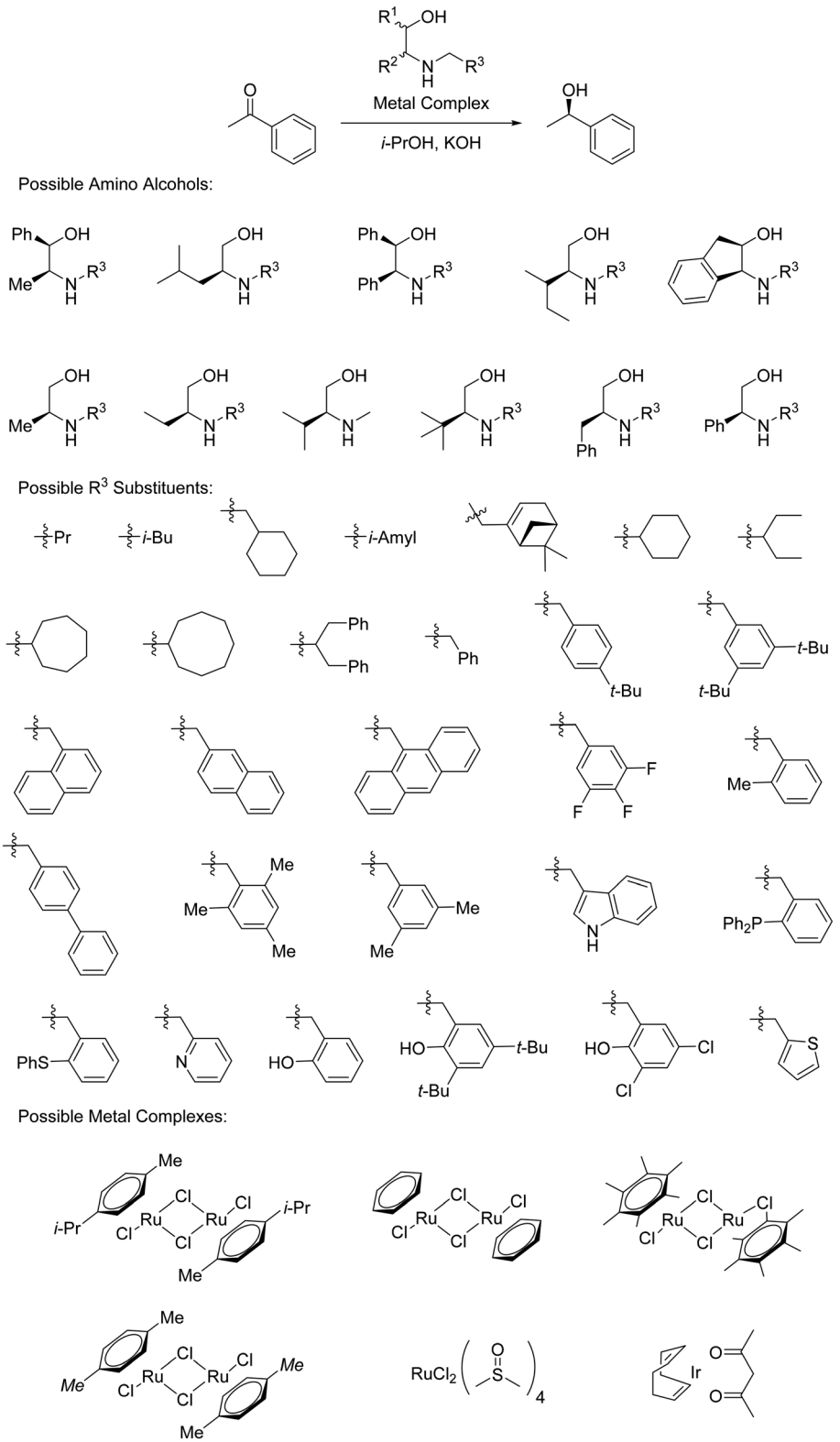

Fig. 1 Enantioselective hydrogen transfer catalyst system and possible catalyst structures.

investigations can aid decision making in ML-guided optimization campaigns.

Results and discussion

Evaluation of different training set selection methods

First, this work seeks to identify suitable algorithmic methods for subset selection (*i.e.* selection of the initial training set). Specifically, we hypothesize that algorithmic selection will reliably provide more accurate models than random selection or selection on the basis of synthetic or commercial availability. In our previous work, the Kennard–Stone algorithm was used to select an initial subset for model training and validation, and K-means clustering was used to augment training data for ML studies.[48,49] Both algorithms have been empirically successful selection methods; however, to our knowledge, no study in homogenous catalysis has investigated an array of subset selection protocols to determine which is best for selecting an initial set of experiments. To probe this hypothesis, a literature dataset of enantioselective transfer hydrogenation reactions previously used in chemoinformatic analysis will be used as a case study.[58–60] In this investigation, 330 amino alcohol ligands were combined with six transition metal complexes, which were then employed in the enantioselective transfer hydrogenation of acetophenone (Fig. 1). From this dataset, reactions providing yield and enantioselectivity values were chosen, reducing the total number of reactions to 315. The original report chose to optimize for a combination of these values, which was termed normalized performance factor (NPF). NPF is calculated as the conversion multiplied by two plus the enantiomeric excess (ee). That value is normalized to the highest performer by this metric to give NPF values for all catalysts, with the catalyst having the highest performance factor normalized to 1.

With multiple reaction outcomes to predict, this dataset was used for further exploration. As in our previous study,[49] average steric occupancy (ASO) and average electronic indicator field (AEIF) descriptors were calculated for the amino alcohol metal complexes. Although these descriptors were successfully implemented for different classes of organocatalysts, we wanted to assess the efficacy of these descriptors for representing chiral transition metal complexes. The capability to represent disparate catalyst families with the same molecular representation would indicate good generality in the molecular representation, which is a necessary requirement if comparisons between different families of catalysts is desired in future work. Additional parameters for the respective metals were also calculated (see ESI† for full computational details). When investigating methods for selecting initial subsets of compounds, the most appropriate were deemed to be those dependent on only the catalyst descriptors. The rationale for this approach is that the initial subset of catalyst structures (*i.e.* the training set) should be general for use at the outset of optimization campaigns for any application. By considering only catalyst structure descriptors, the selection process is agnostic to the specific reaction under study.

When evaluating methods to use for subset selection some considerations need to be taken in the context of our workflow. Most notably, our system for catalyst optimization first begins with the construction of a large *in silico* library containing thousands of synthetically accessible catalyst structures. The remainder of the workflow (at this stage of development) works under the assumption that this library will remain static. Thus, every sample for which a value could be measured is known at the outset of experimentation. Therefore, the ideal subset of molecules will yield models that most accurately predict reaction outcomes for the remainder of the library. Toward this end, five different methods fulfilling this criterion were used: the Kennard–Stone algorithm, K-means clustering, affinity propagation, agglomerative clustering, and mean shift clustering.[61] For each method, 33 catalysts were selected except for affinity propagation and mean shift clustering in which the number of clusters cannot be preset. For affinity propagation, 34 clusters were identified by the algorithm and for mean shift clustering 33 clusters were recommended. Thus, affinity propagation has a slight advantage over the other four methods owing to the additional cluster employed. For comparison, ten randomly selected subsets were also compiled. It is worth noting that, in general, random selection from a diverse *in silico* library likely covers more chemical space than most experimental catalyst optimization campaigns. As demonstrated in our previous work, most such campaigns over-sample a limited region of chemical space owing to commercial availability or synthetic accessibility of certain types of structures.[49] Consequently, it is likely that random selection from a diverse library would give more diverse structures than most instances of "traditional" catalyst optimization. Therefore, any selection protocol that gives consistently higher performance than random selection can be considered a particularly promising selection method.

The performance of interest in this study is the capability to determine the selectivity of every catalyst in the library *a priori*. Thus, to determine the best subset selection methods, models were trained and cross validated using only the selected catalysts and their performance compared by using all remaining catalysts in the library as an external test set. Although we have chosen this experimental design with our own workflow in mind, it is worth noting that in many cases experimentalists have an idea of which catalysts are acceptable at the outset of experimentation. The following analysis should be applicable to all of such cases. Using each initial subset (which in practice was used for model training and cross validation), an ensemble of models was generated for both the NPF and the enantioselectivity

Table 1 Summary of different selection methods for enantioselectivity models

Selection protocol	Test MAE[a] (kcal mol^{-1})	Test MAE Stdev (kcal mol^{-1})	Cross Val Score[b] (kcal mol^{-1})	Cross Val Score Stdev (kcal mol^{-1})
Random TS1	**0.198**	*0.001*	**0.297**	*0.004*
Agglomerative clustering	**0.207**	*0.003*	**0.196**	*0.005*
Kennard Stone	**0.214**	*0.014*	**0.187**	*0.017*
Affinity propagation	**0.215**	*0.013*	**0.142**	*0.008*
Random TS3	**0.217**	*0.009*	**0.181**	*0.012*
Random TS7	**0.218**	*0.012*	**0.215**	*0.002*
Random TS5	**0.220**	*0.008*	**0.284**	*0.011*
K-means clustering	**0.220**	*0.010*	**0.197**	*0.013*
Random TS4	**0.223**	*0.003*	**0.220**	*0.006*
Random TS2	**0.225**	*0.016*	**0.320**	*0.023*
Mean shift	**0.235**	*0.013*	**0.276**	*0.014*
Random TS8	**0.253**	*0.016*	**0.206**	*0.021*
Random TS10	**0.273**	*0.004*	**0.221**	*0.005*
Random TS6	**0.283**	*0.006*	**0.138**	*0.008*
Random TS9	**0.310**	*0.012*	**0.184**	*0.014*
All random	***0.242***	*0.009*	***0.227***	*0.011*
All algorithmic	***0.218***	*0.010*	***0.199***	*0.011*

[a] Average MAE for the ensemble of models. [b] Results of 3-fold cross validation with MAE as the evaluation metric.

datasets. Because limited training data was used to simulate real optimization scenarios, an ensemble of linear models was constructed (see ESI† for full detailed regarding this ensemble). The summary of the models for enantioselectivity and NPF are given in Tables 1 and 2, respectively.

The data in Tables 1 and 2 demonstrate that, in general, models derived from algorithmically selected training set outperform those selected through random selection. For both enantioselectivity models and NPF models, three of the top five models are derived from algorithmically selected training sets. Further, in both cases even the worst performing model derived from an algorithmically selected training set yields a lower MAE_{Test} than the average performance of the randomly selected training sets. Finally, all models derived from algorithmically selected training data perform with excellent accuracy as dictated by MAE_{Test} (enantioselectivity models MAE < 0.235 kcal mol^{-1}, NPF models < 0.087 NPF). This observation is encouraging in such studies, as it indicates that it is generally possible to make accurate models with limited training data, which should facilitate adoption of such methods in experimental optimization campaigns.

When comparing the performance of the different training sets, it is apparent that random training data selection results in model accuracy that is highly dependent on set selection. For the randomly selected training sets, resulting model MAE_{Test} scores ranged from 0.198 kcal mol^{-1} to 0.310

Table 2 Summary of different selection methods for NPF models

Selection protocol	Test MAE[a] (NPF)	Test MAE Stdev (NPF)	Cross Val Score[b] (NPF)	Cross Val Score Stdev (NPF)
Mean shift	**0.074**	*0.000*	**0.044**	*0.001*
Random TS7	**0.078**	*0.006*	**0.075**	*0.004*
Kennard–Stone	**0.079**	*0.001*	**0.039**	*0.002*
Agglomerative clustering	**0.081**	*0.001*	**0.053**	*0.001*
Random TS3	**0.081**	*0.002*	**0.069**	*0.003*
Random TS1	**0.082**	*0.001*	**0.048**	*0.002*
K-means clustering	**0.083**	*0.000*	**0.056**	*0.001*
Random TS9	**0.083**	*0.005*	**0.058**	*0.003*
Random TS8	**0.083**	*0.005*	**0.060**	*0.013*
Affinity propagation	**0.084**	*0.002*	**0.052**	*0.006*
Random TS10	**0.088**	*0.005*	**0.089**	*0.012*
Random TS4	**0.089**	*0.003*	**0.066**	*0.000*
Random TS5	**0.091**	*0.007*	**0.093**	*0.006*
Random TS2	**0.095**	*0.005*	**0.079**	*0.004*
Random TS6	**0.105**	*0.010*	**0.092**	*0.006*
All random	***0.088***	***0.005***	***0.073***	***0.005***
All algorithmic	***0.080***	***0.001***	***0.049***	***0.002***

[a] Average MAE for the ensemble of models. [b] Results of 3-fold cross validation with MAE as the evaluation metric.

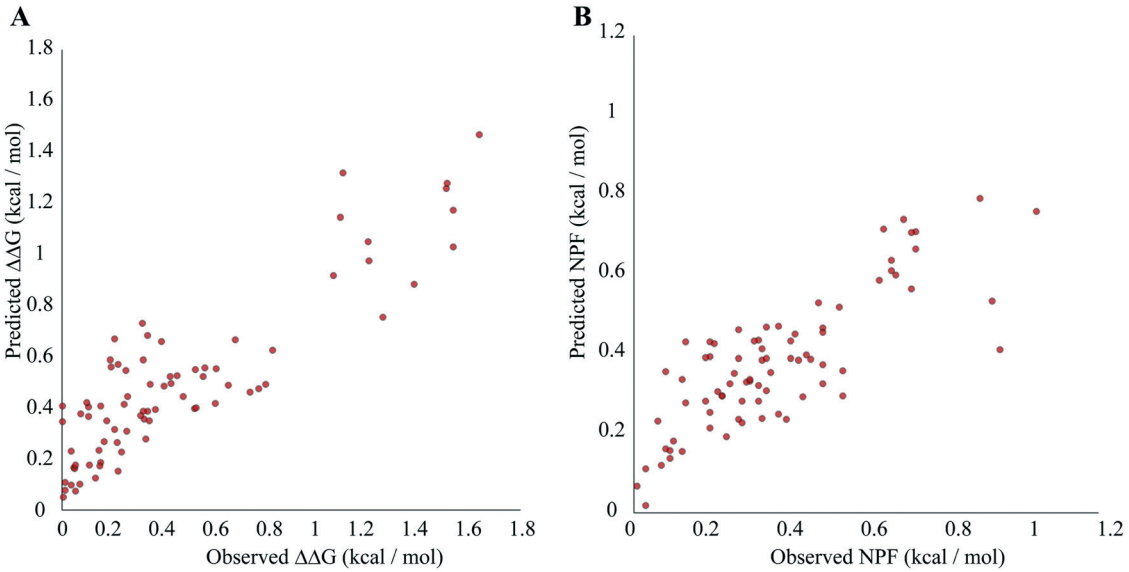

Fig. 2 (A) Test set predicted *vs.* observed for the model predicting enantioselectivity and (B) test set predicted *vs.* observed for the model predicting NPF.

kcal mol^{-1} for models predicting enantioselectivity and 0.078 to 0.105 for models predicting NPF.[62] This wide fluctuation in performance is expected when a small portion of the dataset is sampled randomly; as the number selected randomly increases, the overall variability of model accuracy decreases. However, for synthetic applications, acquiring enough datapoints to reduce this variation resulting from "fortuitous" (*e.g.* random training set 1 or random training set 7) or "unfortunate" (*e.g.* random training set 6) training set selection is undesirable and unrealistic. In this regard, all algorithmic selection methods demonstrate the advantage of avoiding an "unfortunate" random selection. It is also noteworthy that algorithmic subset selection methods result in models with lower MAEs than random sampling as indicated by comparing the MAE_{Test} of each subset selection method to the average MAE_{Test} of the random sets. In both enantioselectivity and NPF models, all algorithmic selection methods had lower MAEs than the mean MAE of the random sets. Finally, it is worth noting other datasets might have a different highest performing subset selection algorithm. Until numerous high-quality large datasets are available for benchmarking, it is not possible to determine which subset selection algorithm is the most general when applied to asymmetric catalysis. That notwithstanding, it is safe to conclude that such algorithmic methods will give a more reliable and robust selection than random selection when applied to library-based optimization protocols.

Evaluation of different error assessment metrics

Having probed suitable selection protocols for gathering initial datasets, we next sought to examine an array of error assessment metrics to inform which reaction should be run "next" in an optimization campaign. As such, the hydrogen transfer catalyst dataset was used to generate an ensemble of neural networks which was then used to evaluate different error assessment protocols. The dataset was divided into a training set of 200 reactions, a validation set of 37 reactions, and a test set of 78 reactions. A set of 2000 neural networks with randomized hyperparameters was selected, and the top 40 networks were used in the ensemble of networks. This process was repeated to create models both for predicting enantioselectivity and NPF. The external test sets for both models are depicted in Fig. 2.

Both models in Fig. 2 have excellent accuracy; enantioselectivity is predicted with MAE = 0.17 kcal mol^{-1} and NPF predicted with MAE = 0.10. This level of accuracy in itself is an interesting finding for two reasons: (1) the same conformer-dependent descriptors used to describe molecular shape in chiral Brønsted acid catalysis[48] and asymmetric phase transfer catalysis[49] have now been applied to transition metal catalysis with no modification, suggesting broad applicability of these descriptors, and (2) models have been constructed with consideration for both the enantioselectivity and conversion, demonstrating the capability of optimizing multiple reaction properties simultaneously.

As a preliminary investigation, two conceptually distinct error metrics were employed. The first type is founded on the premise that the outcome of reactions farther in feature space from the data on which the models were trained will be less reliably predicted. In this regard, four different dimensionality reduction methods were used on the total feature space. For each space, the distance between each test point and its three nearest neighbors in the training set was calculated. For unsupervised dimensionality reduction, principal component analysis (PCA) and multi-dimensional

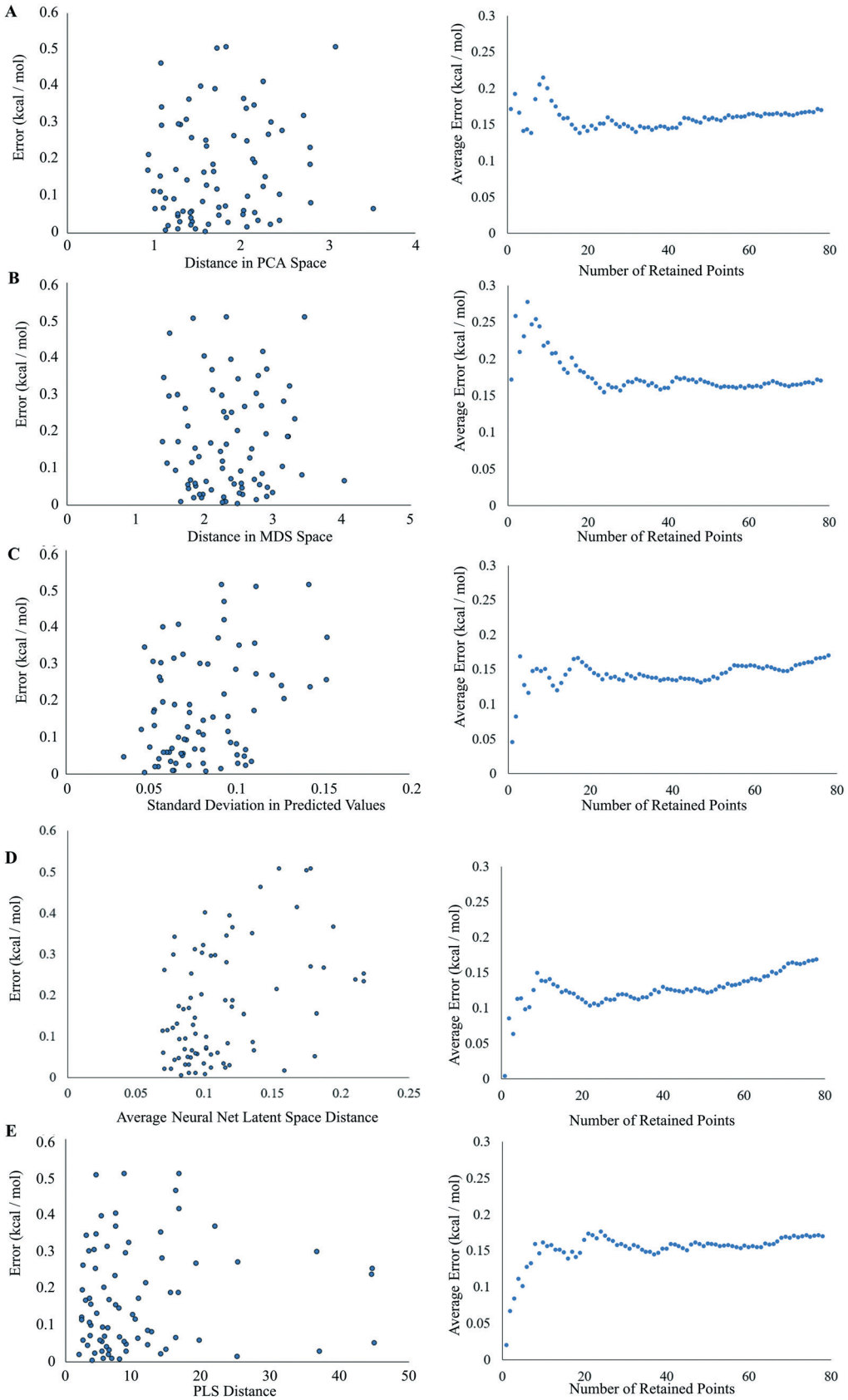

Fig. 3 Summary of different error assessment methods for enantioselectivity, including (A) distance in principal component analysis (PCA)-space from training data, (B) distance in multi-dimensional scaling (MDS)-space from training data, (C) standard deviation in predicted values of the ensemble, (D) distance in neural network latent space from the training data, and (E) distance in projection to latent structure (PLS)-space from the training data.

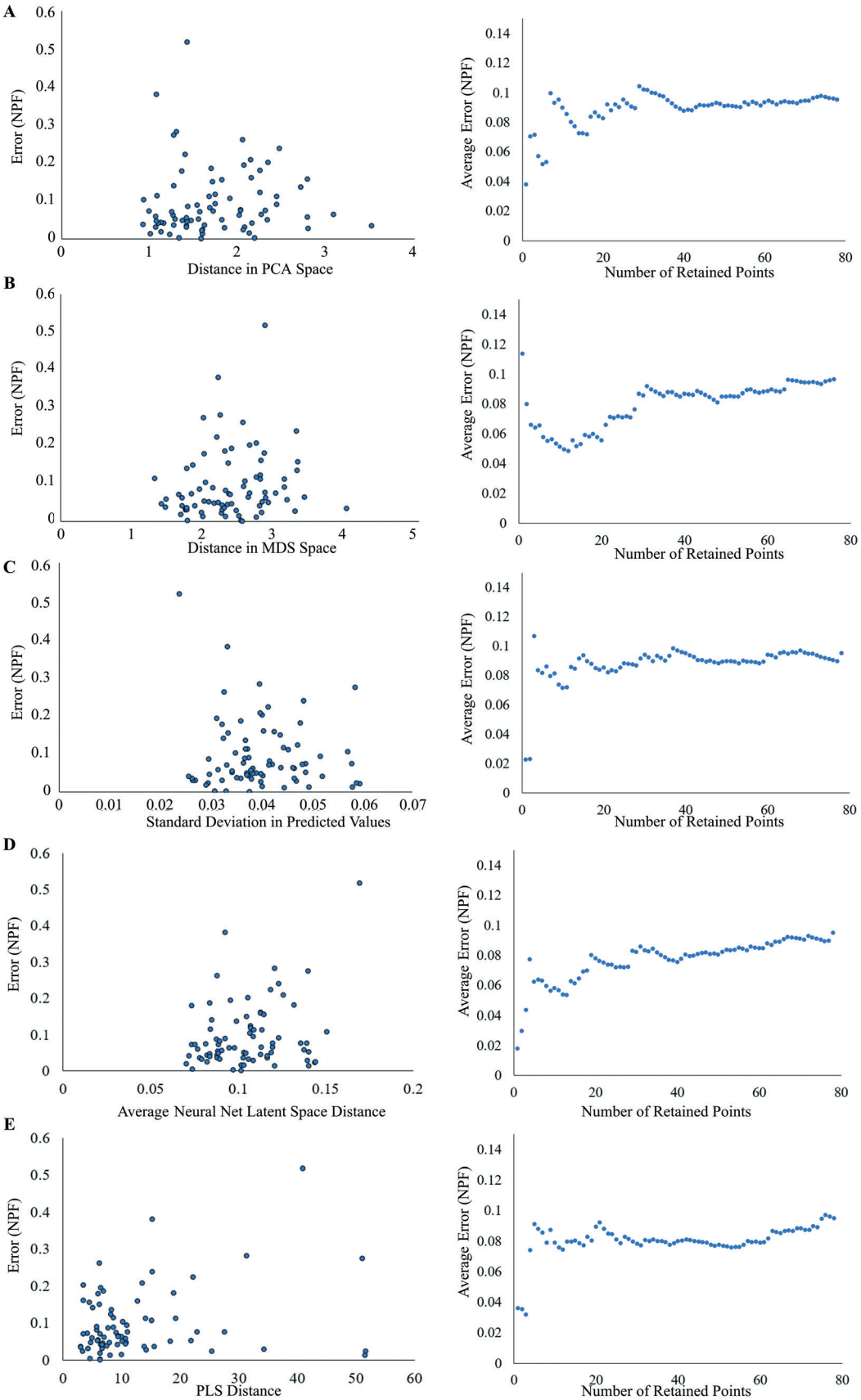

Fig. 4 Summary of different error assessment methods for NPF, including (A) distance in PCA-space from training data, (B) distance in MDS-space from training data, (C) standard deviation in predicted values of the ensemble, (D) distance in neural network latent space from the training data, and (E) distance in PLS-space from the training data.

scaling (MDS) were used to reduce the dimensionality of the input data. It is worth noting that because our workflow operates under the assumption that all future chemical entities for which predictions will be made are known and descriptors for those entities have been calculated (*i.e.* the *in silico* library is constructed and descriptors have been calculated), all samples are used in the unsupervised dimensionality reduction transforms. For supervised dimensionality reduction, distance in PLS space (commonly referred to as distance in model space, or DModX) and the average distance in the latent space[63] of the neural network were used. More specifically, because the different neural networks had different numbers of nodes in their hidden layers, the distances used are the average distances computed for the entire ensemble.

The second metric is the standard deviation in predicted values. The concept is that if predictions vary widely between different estimators, there is more uncertainty in that prediction, and it may be less reliable. Both concepts (distance in feature space and variability in predictions) have been explored previously in the chemical sciences.[50,64] All five metrics were calculated for both the enantioselectivity models (Fig. 3) and NPF (Fig. 4) models for this dataset. The different accuracy metrics can then be compared by (1) plotting error *vs.* the error metric and (2) constructing accuracy averaging curves. Accuracy averaging curves are plots in which data points (in this case test set samples) are ordered by their error metric from smallest to highest. The average error of retained points is then plotted against the number of points retained (*i.e.* the first point on the plot is the sample with the smallest error metric and its error, the last point is the entire dataset and the MAE of the dataset). In this case, as more points are retained, one would expect the average error to increase if the error metric is indeed a good indication of error. Further, a steeper curve would indicate a larger response, in turn indicated a better metric of error.

From examination of the error metrics (Fig. 3 and 4), it becomes immediately apparent that no one metric is best for each dataset. For the enantioselectivity models, average distance in neural network latent space (Fig. 3D) appears to be the best metric of prediction reliability, with the largest response in the accuracy averaging curve. Distance from the training data in PLS-space (Fig. 3E) also shows a meaningful response with regard to averaged error. Standard deviation in predicted values (Fig. 3C), distance in MDS-space (Fig. 3B), and distance in PCA space (Fig. 3A) curves are relatively flat curve, indicating less efficacious error metrics. In contrast, the NPF models have different error metrics best correlating with the residuals. As with the enantioselectivity models, average distance in neural net latent space appeared to be the best metric of error when analyzing the accuracy averaging curve (Fig. 4D). The superiority of this metric is in line with previous results.[50] However, the next greatest

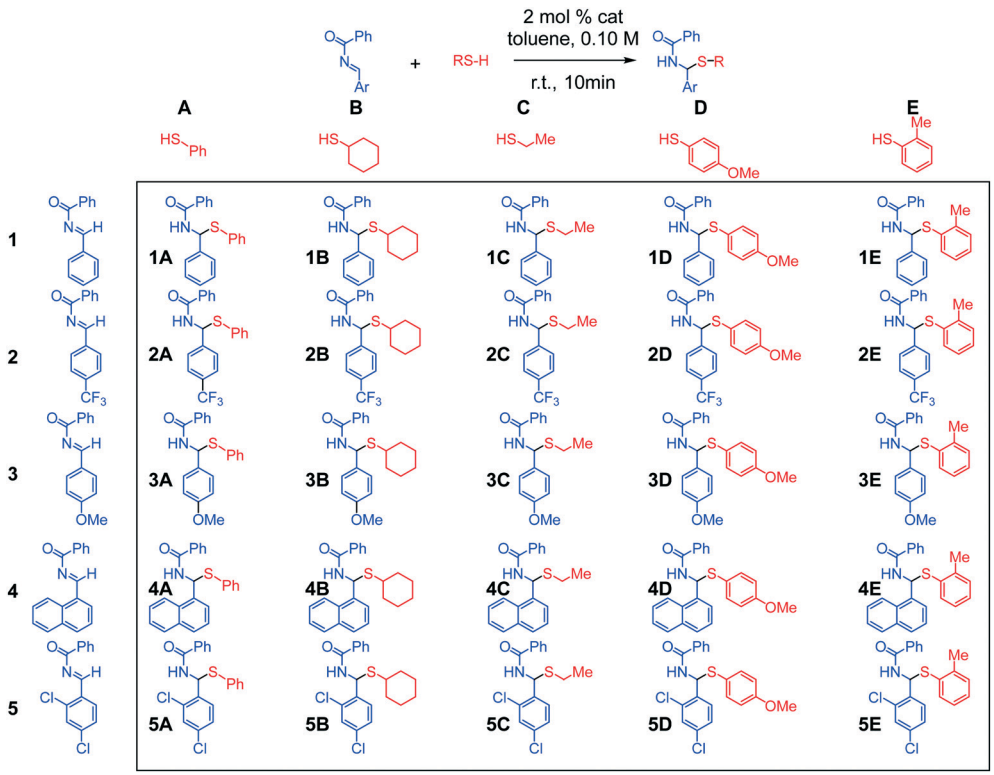

Fig. 5 Matrix of 25 different possible substrate combinations derived from imines **1**–**5** and thiols **A**–**E**. Adapted with permission from ref. 48. Copyright 2019 American Association for the Advancement of Science.

response is distance in MDS-space (Fig. 4B), all other metrics appearing to have flatter response curves.

These results suggest that the best error metric for a given application is dataset and application dependent. Despite this apparent limitation, these conclusions are useful in informing future studies. We envision that in the course of an optimization campaign the first set of models obtained will be externally validated. Each of the above metrics can be plotted against the error for each test set member to best identify which error assessment metric(s) are best for that specific application. Then, when evaluating a set of predictions to be tested experimentally, practitioners can quantitatively assess which predictions to pursue on the basis of prediction confidence. It is also worth noting that all of the curves in Fig. 3 and 4 are relatively flat. This could arise from the accuracy of the parent models; because a large portion of the total dataset is sampled and the overall accuracy of the models was very high in evaluation of the external test set, it is possible that the test points fall well within the domain of the model resulting in relatively flat response curves. In this sense, the relative response in the accuracy averaging curves is likely application dependent.

To further probe this hypothesis, our previously published dataset of BINOL phosphoric acid catalyzed additions of thiols to imines was used as an additional case study (Fig. 5).[48] Originally published by Antilla and coworkers,[65] the modularity, technical accessibility, and reproducibility of this reaction enabled collection of a dataset of 1075 reactions in duplicate runs. In this work, the dataset has been further expanded to a total of 1150 reactions in an effort to provide larger, high-quality datasets for use in ML studies. The descriptors used to represent the molecules are identical to those previously reported.[48,49] The dataset was first divided into two sets: a set of 384 reactions for training and

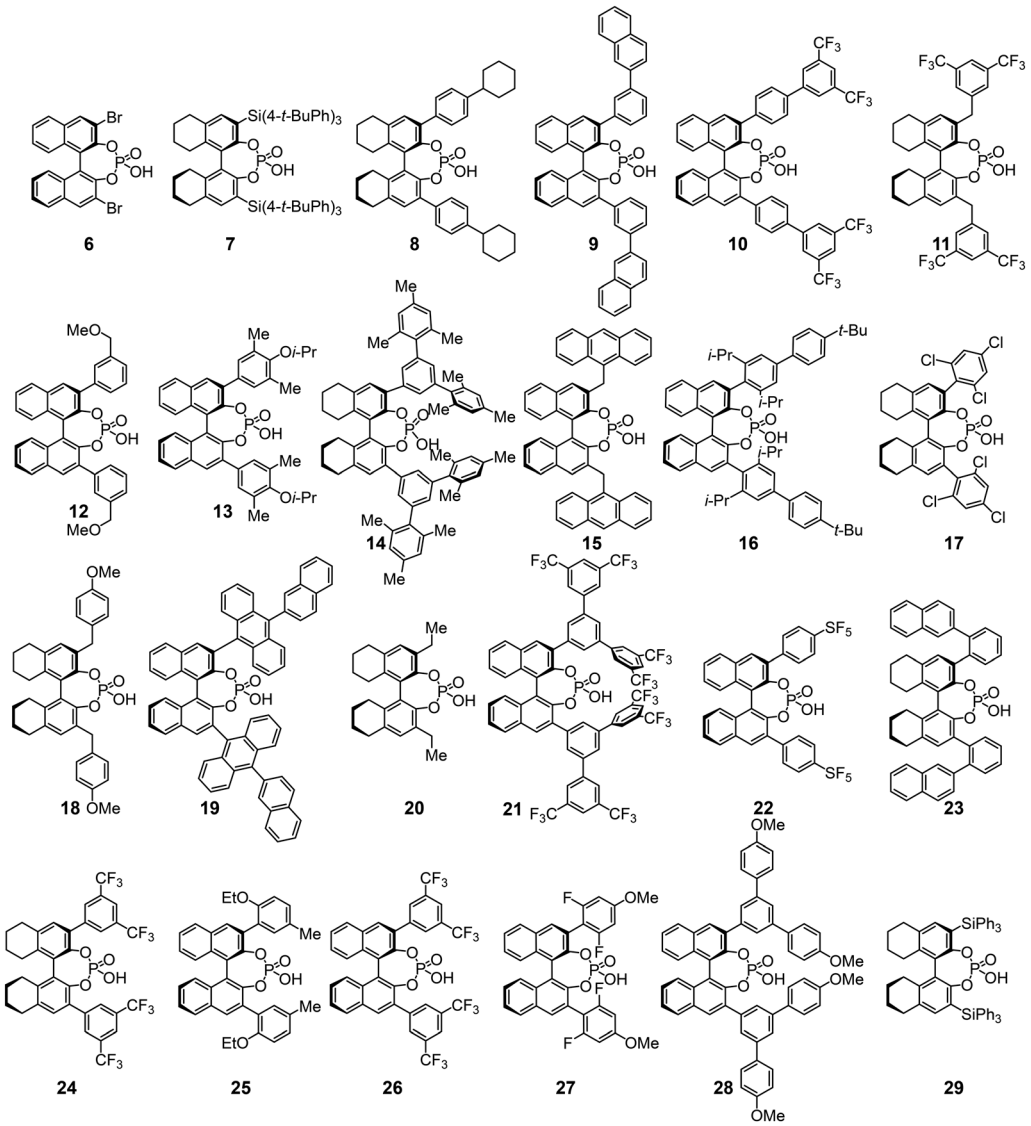

Fig. 6 The 24-member universal training set of chiral phosphoric acids.

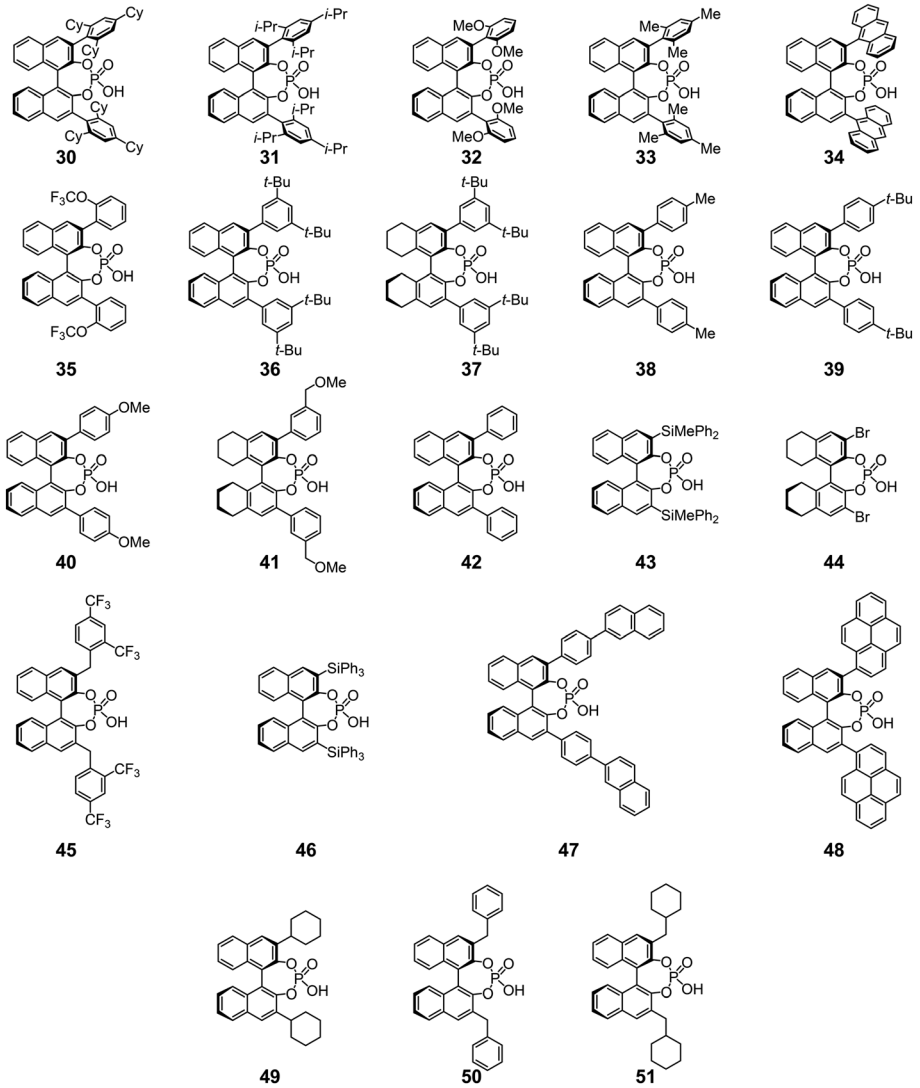

Fig. 7 The 22-member external catalyst test set of chiral phosphoric acids.

validation—24 training catalysts (Fig. 6) with 16 training substrate pairs (imines **1**–**4** and thiols **A**–**D**, Fig. 5) and the remaining 766 reactions as an external test set. The 384 training reactions include catalyst structures that were selected in our universal training set previously disclosed[48] and the remaining catalysts were selected on the basis of either their commercial availability or their qualitative diversity. In the partitioning scheme, imine **5**, thiol **E**, and catalysts 30–51 (Fig. 7) were purposefully withheld from the possible training data to force out-of-sample predictions. The 384-member set was then randomly divided into a training set of 300 members and a validation set of 84 members. Models were then generated using an ensemble of feedforward networks, which were constructed with a single hidden layer. Parameters including activation functions for each layer, number of nodes in the hidden layer, and percent dropout were optimized randomly. In total, 10 000 different hyperparameter combinations were tested and the top 100 models (determined by the performance on the validation set) were used as the final ensemble of models. The average predicted values for the model ensembles were then used as the predicted value for the purposes of this study.

The ensemble of neural networks accurately predicted the outcome of the reactions in the test set, with a MAE in test set predictions of 0.30 kcal mol^{-1} (Fig. 8). Using the absolute values of the residuals in the test set, the five different error metrics detailed above were compared by plotting the errors against each metric (metric refers to either distance in the respective dimensionality-reduced space or the standard deviation in predicted values of the ensemble). In addition, average accuracy plots were constructed for each metric (Fig. 8).

As depicted in Fig. 8, standard deviation in predicted values, distance from training data in neural network latent space, and distance from training data in PLS space all appear to be good indicators of prediction accuracy. In contrast, distance in the reduced dimensionality space for the unsupervised methods (PCA and MDS) gives a weaker response. Notably, considering both case studies, it appears

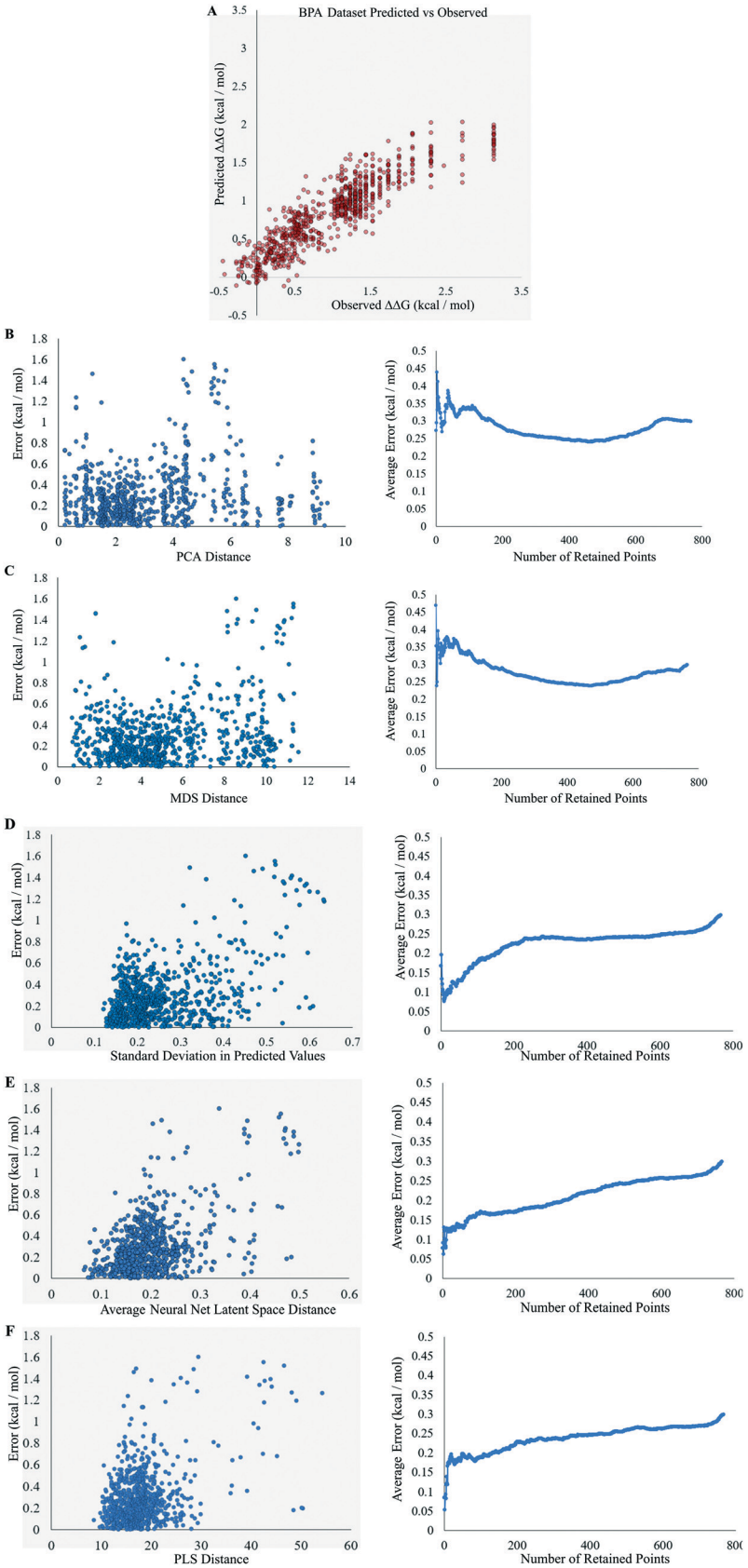

Fig. 8 (A) Predicted *vs.* observed plot for the 766-member external test set for the BPA dataset. (B) Error plot with the standard deviation in predicted values of the ensemble as the error metric (C) error plot with distance in the neural network latent space as the error metric (D) error plot with distance in PLS space as the error metric (E) error plot with distance in PCA space as the error metric (F) error plot with distance in MDS space as the error metric.

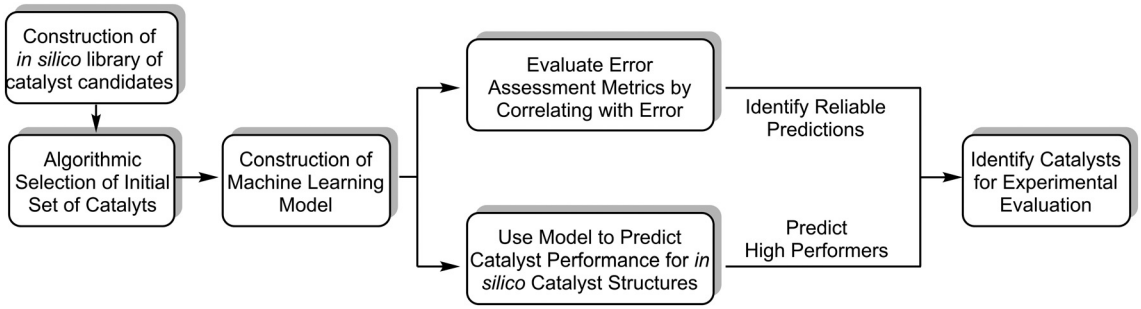

Fig. 9 Intended use of this workflow.

that metrics associated with a supervised dimensionality reduction method have best performance. With this analysis, users can identify predictions as potentially high-risk. Further, it may be possible to use multiple metrics simultaneously to better assess how much of a risk a particular prediction would be in an actual optimization campaign. Predictions that fall into multiple categories could be identified as even higher-risk predictions which may inform the end user to consider selecting a different prediction to test experimentally depending on the effort required per new data point. By using a collection of error metrics, users will make the most well-informed decisions when considering how best to use a given model.

Conclusions

This work demonstrates the successful application of ASO and AEIF descriptors to transition metal catalysts over and above their previous use for organocatalysts. The capacity to use the same 3D molecular representation across such disparate chiral catalyst families has laid the groundwork for future studies, in which comparisons across different catalyst scaffold may be desired. Further, this work demonstrates that algorithmic subset selection protocols give more reliable results and generally can be used to construct more accurate models than random selection when selecting small datasets in library-based optimization manifolds. Additionally, multiple metrics of error assessment have been investigated in this context to assist in identifying which predictions are the most reliable when attempting to use ML models in optimization campaigns, particularly those beginning with a pre-defined set of catalyst candidates. Combining these ideas will enable more efficient initialization and execution of computer-guided workflows for catalyst design. We envision that in new scenarios, practitioners can use an algorithmic selection protocol (*e.g.* informed by agglomerative clustering) to gather an initial set of catalyst structures. Next, that dataset will be acquired experimentally and used to train and validate statistical learning models. Error can then be correlated with a set of error assessment metrics to identify which metrics are best for assessing error for that particular dataset. The models can then be used to evaluate an *in silico* library of catalyst structures to identify catalysts predicted to be more selective than those in the initial set of data. When identifying which of the catalysts predicted to be more selective than the initial set should be experimentally evaluated, the decision can be informed by the error metric to guide a more reliable prediction. In other words, the more reliable predictions will be given priority over others when selecting which predictions to experimentally evaluate (Fig. 9). Further, we suspect that identification of the best error metric could also find use in an active learning campaign in which unreliable predictions could be selected as the next best reactions to improve the model. In fact, this concept has already been demonstrated in other areas of the chemical sciences.[57,66] Together, the concepts explored in this work will provide a practical guide to ML-guided optimization in catalysis.

Conflicts of interest

No conflicts of interest to declare.

Acknowledgements

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 a Graduate Fellowship. 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 K. Lipkowitz and M. Kozlowski, Understanding Stereoinduction in Catalysis via Computer: New Tools for Asymmetric Synthesis, *Synlett*, 2003, **10**, 1547–1565.
2 S. Ahn, M. Hong, M. Sundararajan, D. H. Ess and M.-H. Baik, Design and Optimization of Catalysts Based on mechanistic Insights Derived from Quantum Chemical Reaction Modeling, *Chem. Rev.*, 2019, **119**, 6509–6560.
3 E. Burello and G. Rothenberg, In Silico Design in Homogeneous Catalysis Using Descriptor Modelling, *Int. J. Mol. Sci.*, 2006, **7**, 375–404.
4 J. P. Reid and M. S. Sigman, Comparing Quantitative Prediction Methods for the Discovery of Small-Molecule Chiral Catalysts, *Nat. Rev. Chem.*, 2018, **2**, 290–305.

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

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

7 Q. Peng and R. S. Paton, Catalytic Control in Cyclizations: From Computational Mechanistic Understanding to Selectivity Prediction, *Acc. Chem. Res.*, 2016, **49**, 1042–1052.

8 C. Poree and F. A. Schoenebeck, Holy Grail in Chemistry: Computational Catalyst Design: Feasible or Fiction, *Acc. Chem. Res.*, 2017, **50**, 605–608.

9 Q. Peng, F. Duarte and R. S. Paton, Computing Organic Stereoselectivity – from Concepts to Quantitative Calculations and Predictions, *Chem. Soc. Rev.*, 2016, **45**, 6093–6107.

10 S. E. Wheeler, T. J. Seguin, T. Guan and A. C. Doney, Noncovalent Interactions in Organocatalysis and the Prospect of Computational Catalyst Design, *Acc. Chem. Res.*, 2016, **49**, 1061–1069.

11 D. J. F. Tantillo, Catalyst! React! React! Exploiting Computational Chemistry for Catalyst Development and Design, *Acc. Chem. Res.*, 2016, **49**, 1079.

12 D. Balcells and F. Maseras, Computational Approaches to Asymmetric Synthesis, *New J. Chem.*, 2007, **31**, 333–343.

13 K. N. Houk and P. H.-Y. Cheong, Computational Prediction of Small-Molecule Catalysts, *Nature*, 2008, **455**, 309–313.

14 N. Fey, A. G. Orpen and J. N. Harvey, 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.

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

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

17 A. G. Maldonado and G. Rothenberg, Predictive Modeling in Homogeneous Catalysis: a Tutorial, *Chem. Soc. Rev.*, 2010, **39**, 1891–1902.

18 A. J. Neel, M. J. Hilton, M. S. Sigman and F. D. Toste, Exploiting Non-Covalent π-Interactions for Catalyst Design, *Nature*, 2017, **543**, 637–646.

19 I. I. Baskin, T. I. Madzhidov, I. S. Antipin and A. Varnek, Artificial Intelligence in Synthetic Chemistry: Achievements and Prospects, *Russ. Chem. Rev.*, 2017, **86**, 1127–1156.

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

21 C. B. Santiago, J.-Y. Guo and M. S. Sigman, Predictive and Mechanistic Multivariate Linear Regression Models for Reaction Development, *Chem. Sci.*, 2018, **9**, 2398–2412.

22 D. J. Durand and N. Fey, Computational Ligand Descriptors for Catalyst Design, *Chem. Rev.*, 2019, **119**, 6561–6594.

23 J. E. Eksterowicz and K. N. Houk, Transition-State Modeling with Empirical Force Fields, *Chem. Rev.*, 1993, **93**, 2439–2461.

24 P. Friederich, G. dos Passos Gomes, R. De Bin, A. Aspuru-Guzik and D. Balcells, Machine Learning Dihydrogen Activation in the Chemical Space Surrounding Vaska's Complex, *Chem. Sci.*, 2020, **11**, 4584–4601.

25 F. Sandfort, F. Strieth-Kalthoff, M. Kühnemund, C. Beecks and F. Glorius, A Structure-Based Platform for Predicting Chemical Reactivity, *Chem*, 2020, **6**, 1379–1390.

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

27 M. D. Wodrich, A. Fabrizio, B. Meyer and C. Corminboeuf, Data-powered augmentation volcano plots for homogeneous catalysis, *Chem. Sci.*, 2020, **11**, 12070–12080.

28 M. Cordova, M. D. Wodrich, B. Meyer, B. Sawatlon and C. Corminboeuf, Data-Driven Advancement of Homogeneous Nickel Catalyst Activity for Aryl Ether Cleavage, *ACS Catal.*, 2020, **10**, 7021–7031.

29 E. Hansen, A. R. Rosales, B. Tutkowiski, P.-O. Norrby and O. Wiest, Prediction of Stereochemistry using Q2MM, *Acc. Chem. Res.*, 2016, **49**, 996–1005.

30 M. Burai Patrascu, J. Pottel, S. Pinus, M. Bezanson, P.-O. Norrby and N. Moitessier, From Desktop to Benchtop with Automated Computational Workflows for Computer-Aided Design in Asymmetric Catalysis, *Nat. Catal.*, 2020, **3**, 574–584.

31 A. R. Rosales, T. R. Quinn, J. Wahlers, A. Tomberg, X. Zhang, P. Helquist, O. Wiest and P.-O. Norrby, Application of Q2MM to Predictions in Stereoselective Synthesis, *Chem. Commun.*, 2018, **54**, 8294–8311.

32 A. R. Rosales, J. Wahlers, E. Limé, R. E. Meadows, K. W. Leslie, R. Savin, F. Bell, E. Hansen, P. Helquist, R. H. Munday, O. Wiest and P.-O. Norrby, Rapid Virtual Screening of Enantioselective Catalysts Using CatVS, *Nat. Catal.*, 2019, **2**, 41–45.

33 Y. Guan, V. M. Ingman, B. J. Rooks and S. E. Wheeler, AARON: An Automated Reaction Optimizer for New Catalysts, *J. Chem. Theory Comput.*, 2018, **14**, 5249–5261.

34 A. F. Zahrt, S. V. Athavale and S. E. Denmark, Quantitative Structure–Selectivity Relationships in Enantioselective Catalysis: Past, Present, and Future, *Chem. Rev.*, 2020, **120**, 1620–1689.

35 J. D. Oslob, B. Åkermark, P. Helquist and P.-A. Norrby, Steric Influences on the Selectivity in Palladium-Catalyzed Allylation, *Organometallics*, 1997, **16**, 3015–3021.

36 M. C. Kozlowski, S. L. Dixon, M. Panda and G. Lauri, 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.

37 P.-W. Phuan, J. C. Ianni and M. C. Kozlowski, 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.

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

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

40 M. Kozlowski and J. Ianni, Quantum Molecular Interaction Field Models of Substrate Enantioselection in Asymmetric Processes, *J. Mol. Catal. A: Chem.*, 2010, **324**, 141–145.

41 K. Lipkowitz and M. Pradhan, 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.

42 J. Chen, W. Jiewu, L. Mingzong and T. You, 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.

43 M. Hoogenraad, G. M. Klaus, N. Elders, S. M. Hooijschuur, B. McKay, A. A. Smith and E. W. P. Damen, Oxazaborolidine Mediated Asymmetric Ketone Reduction: Prediction of Enantiomeric Excess Based on Catalyst Structure, *Tetrahedron: Asymmetry*, 2004, **15**, 519–523.

44 J. B. Ven der Linden, I.-J. Ras, S. M. Hooijschuur, G. M. Klaus, N. T. Luchters, P. Dani, G. Verspui, A. A. Smith, E. W. P. Damen, B. McKay and M. Hoogenraad, Asymmetric Catalytic Ketone Hydrogenation: Relating Substrate Structure and Product Enantiometic Excess Using QSPR, *QSAR Comb. Sci.*, 2005, **24**, 94–98.

45 M. S. Sigman, K. C. Harper, E. N. Bess and A. Milo, The Development of Multidimensional Analysis Tools for Asymmetric Catalysis and Beyond, *Acc. Chem. Res.*, 2016, **49**, 1292–1301.

46 S. E. Denmark, N. D. Gould and L. M. A. Wolf, 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.

47 S. E. Denmark, N. D. Gould and L. M. A. Wolf, 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.

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

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

50 J. P. Janet, C. Duan, T. Yang, A. Nandy and H. J. A. Kulik, Quantitative Uncertainty Metric Controls Error in Neural Network-Driven Chemical Discovery, *Chem. Sci.*, 2019, **10**, 7913–7922.

51 W. Wang, T. Yang, W. H. Harris and R. Gómez-Bombarelli, Active Learning nad Neural Network Potentials Accelerate Molecular Screening of Ether-Based Solvate Ionic Liquids, *Chem. Commun.*, 2020, **56**, 8920–8923.

52 A. Tropsha, P. Gramatica and V. K. Gombar, The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models, *QSAR Comb. Sci.*, 2003, **22**, 69–77.

53 J. P. Janet, L. Chan and H. J. Kulik, Accelerating Chemical Discovery with Machine Learning: Simulated Evolution of Spin Crossover Complexes with an Artificial Neural Network, *J. Phys. Chem. Lett.*, 2018, **9**, 1064–1071.

54 A. Nandy, J. Zhu, J. P. Janey, C. Duan, R. B. Getman and H. J. Kulik, Machine Learning Accelerates the Discovery of Design Rules and Exceptions in Stable Metal–Oxo Intermediate Formation, *ACS Catal.*, 2019, **9**, 8243–8255.

55 V. Sunder and L. Colwell, The Effect of Debiasing Protein–Ligand Binding Data on Generalization, *J. Chem. Inf. Model.*, 2020, **60**, 56–62.

56 A. Golbraikh and A. Tropsha, Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection, *J. Comput.-Aided Mol. Des.*, 2002, **16**, 357–369.

57 J. P. Janet, S. Ramesh, C. Duan and H. J. Kulik, Accurate Multiobjective Design in a Space of Millions of Transition Metal Complexes with Neural-Network-Driven Efficient Global Optimization, *ACS Cent. Sci.*, 2020, **6**, 513–524.

58 N. Vriamont, B. Govaerts, P. Grenouillet, C. de Bellefon and O. Riant, Design of a Genetic Algorithm for the Simulated Evolution of a Library of Asymmetric Transfer Hydrogenation Catalysts, *Chem. – Eur. J.*, 2009, **15**, 6267–6278.

59 Q.-Y. Zhang, D.-D. Zhang, J.-Y. Li, Y.-M. Zhou and J. Xu, Virtual Screening of a Combinatorial Library of Enantioselective Catalysts with Chirality Codes and Counterpropagation Neural Networks, *Chemom. Intell. Lab. Syst.*, 2011, **109**, 113–119.

60 Q.-Y. Zhang, D.-D. Zhang, J.-Y. Li, Y.-M. Zhou and J. Xu, Prediction of Enantiomeric Excess in a Catalytic Process: A Chemoinformatics Approach Using Chirality Codes. MATCH Commun. Math, *Comput. Chem.*, 2012, **67**, 773–786.

61 L. Buitnick, *et al.*, API design for machine learning software: experiences from the scikit-learn project, *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 2013, pp. 108–122.

62 Note that NPF is normalized to the highest performing catalyst, so the value is unitless. All NPF values fall in the range 0 to 1.

63 It is noteworthy that this metric is different than the distance in latent space in a single neural network (such as in ref. 49). This is the average distance in latent space across the entire

ensemble of networks. The effect of averaging both distance and predicted value means that the relation between prediction confidence and distance may not be closely related as one would expect in a single neural network. The use of the distance in latent space of a single neural network as an error metric has already be demonstrated in ref. 49.

64 J. S. Smith, B. Nebgen, N. Lubbers, O. Isayev and A. E. Roitberg, Less is more: Sampling Chemical Space with Active Learning, *J. Chem. Phys.*, 2018, **148**, 241733.

65 G. K. Ingle, M. G. Mormino, L. Wojtas and J. C. Antilla, Chiral Phosphoric Acid-Catalyzed Addition of Thiols to N-Acyl Imines: Access to Chiral N,S-Acetals, *Org. Lett.*, 2011, **13**, 4822–4825.

66 N. S. Eyke, W. H. Green and K. F. Jensen, Iterative experimental design based on active machine learning reduces the experimental burden associated with reaction screening, *React. Chem. Eng.*, 2020, **5**, 1963–1972.

