doi:10.2533/chimia.2021.592 Chimia 75 (2021) 592–597

Leveraging Machine Learning for Enantioselective Catalysis: From Dream to Reality

N. Ian Rinehart, Andrew F. Zahrt, and Scott E. Denmark§*

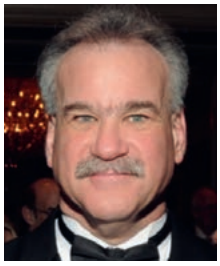
§Paracelsus Prize 2020

Abstract: Catalyst optimization for enantioselective transformations has traditionally relied on empirical evaluation of catalyst properties. Although this approach has been successful in the past it is intrinsically limited and inefficient. To address this problem, our laboratory has developed a fully informatics guided workflow to leverage the power of artificial intelligence (AI) and machine learning (ML) to accelerate the discovery and optimization of any class of catalyst for any transformation. This approach is mechanistically agnostic, but also serves as a discovery platform to identify high performing catalysts that can be subsequently investigated with physical organic methods to identify the origins of selectivity.

Keywords: Catalyst optimization · Chemoinformatic · Enantioselective catalysis · Machine learning

N. Ian Rinehart worked in the laboratories of Prof. David R. Tyler at the University of Oregon as an undergraduate and post-baccalaureate researcher investigating the synthesis of novel phosphine ligands and their transition metal complexes. He graduated with a B.S. degree in chemistry in 2016. His current research in Prof. Scott E. Denmark's laboratories at the University of Illinois Urbana-Champaign focuses on the application of machine learning to optimize chemical reactions.

Andrew F. Zahrt received a B.S. degree in both chemistry and biology at Aquinas College in 2014. He completed his PhD studies with Prof. Scott E. Denmark at the University of Illinois Urbana-Champaign in 2020. His research interests are focused around using computational tools to solve organic chemistry problems. He is a post-doctoral researcher in Prof. Klavs Jensen's laboratories at Massachusetts Institute of Technology.



Scott E. Denmark is the Reynold C. Fuson Professor of Chemistry at the University of Illinois at Urbana-Champaign. He obtained an S.B. degree from MIT in 1975 and a DSc Tech from the ETH Zürich with Albert Eschenmoser in 1980. His research interests include the synthetic, mechanistic, and stereochemical aspects of preparatively useful reactions, as well as the application of AI/machine learning to the optimization of catalysts and reactions.

1. Introduction

Efficient, catalytic, enantioselective reactions have a transformative impact on chemical synthesis, and these are important components of a synthetic chemist's toolbox. Until recently, state-of-the-art enantioselective catalyst development has relied on empiricism and the chemical intuition of a proficient chemist. This approach presents undesirable limitations, and many strategies have been developed to accelerate this process, including increasing throughput with advanced screening protocols,[1] making high-throughput computation of transition state energies feasible,[2] and using mechanism-guided correlations between Linear Free Energy Relationships (LFERs) and enantioselectivity.[3]

Over the past decade our laboratory has focused on the development of tools which merge the power of modern computing, data science, and machine learning with chemoinformatics in an effort to create models which make reliable predictions of catalyst enantioselectivity.[4] Such models address the rate limiting step in state-of-the-art enantioselective method development; finding the optimum catalyst. Because catalyst synthesis can be time consuming, exhaustively exploring new catalysts for a transformation is often infeasible. As a result, catalysts are often screened from commercially available libraries, relying on the assumption that adequate catalyst diversity is found in commercially available compounds.

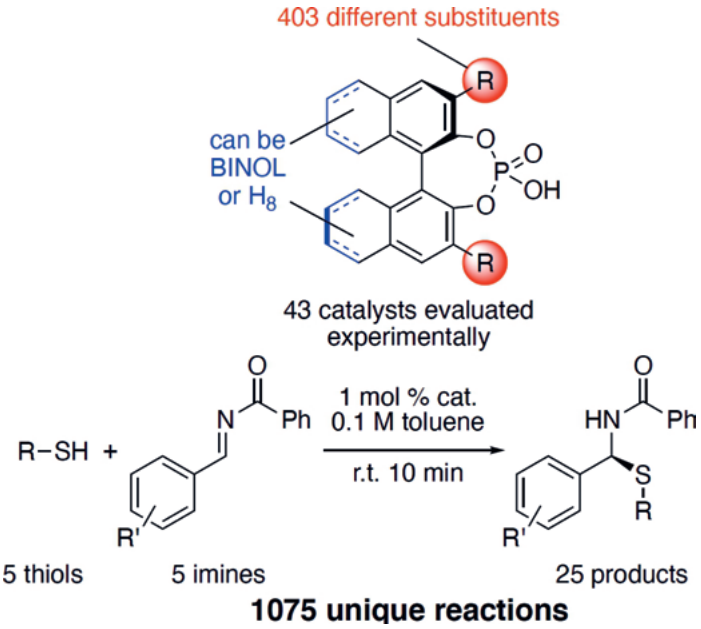
Another critical feature of our approach is to remedy the not uncommon situation in which optimization campaigns are abandoned because no efficient and selective ligand was available for rapid evaluation. Thus, we were interested in finding a way to identify the 'right' catalysts to evaluate in a reaction before moving on to a different scaffold. We also believe that enantioselective catalyst optimization is a perfect domain to apply machine learning because 'bad data' in an empirical screening campaign – which in the context of enantioselectivity can be hard from which to extract useful information – can train machine learning models from patterns too complex for a human to see. Thus, we

*Correspondence: Prof. S. E. Denmark, E-mail: sdenmark@illinois.edu, Dept. Chemistry, University of Illinois at Urbana-Champaign, Urbana, IL, 61801, United States

set out to develop a strategy which could perform the organic chemist's dream: use sub-optimal data to train a model to identify an optimized enantioselective catalyst.

2. Our Chemoinformatic Workflow

The fully chemoinformatic workflow[5] was implemented in a simulated optimization of the addition of thiols to *N*-acyl imines reported by Antilla and coworkers (Scheme 1).[6] using the chiral 3,3'-substituted BINOL-phosphoric acid (CPA)-catalysts. The workflow is depicted in Fig. 1. In the first stage, catalyst structures are translated into an *in silico* library and a representative subset is identified. The second stage involved the synthesis of this representative subset. In the third stage, that subset was used to collect data for 1,075 new reactions which were used to create models relating catalyst structure to selectivity. Finally, the optimum catalyst in the *in silico* library was identified by predicting the selectivity for every *in silico* catalyst.



Scheme. 1. Model reaction for developing the chemoinformatic workflow. Reproduced with permission from ref. [4]. Copyright 2021 American Chemical Society.

2.1 *The Role of Descriptors*

Chemoinformatics is a field which focuses on the numerical representation of chemical structures and properties. In the context of this work, calculable numerical representations of chemical properties and structures – called descriptors – are used to represent chemical entities for machine learning.

Early forays in the realm of phase transfer catalysis[7] drew inspiration from the pioneering work of Kozlowski,[8] Lipkowitz,[9] and Hirst[10] which used molecular field descriptors for Quantitative Structure Selectivity Relationships (QSSRs). An in-depth discussion of this established field is summarized in a recent review from our laboratory.[11]

The accuracy of machine-learning models for predicting the outcome of chemical reactions relies on the information encoded within the descriptors used to represent the chemical entities. The limitations of current descriptors has been noted by others.[12] On the basis of that observation and our own experience with descriptor-limited modeling, we have developed descriptors for representing chiral catalyst structures, including: (1) continuous chirality measure[13] for making QSSRs[14] and (2) the conformer-dependent quantitative quadrant descriptor,[15] both of which represent entire molecular structures. We also developed the ElectroStatic Potential Max (ESPMax, Fig. 2)[4] as a calculable electronic descriptor for representing through-bond electronic effects of catalyst substituents. It is noteworthy that this fragment-based descriptor shows a significant correlation (Fig. 2, R^2 = 0.987) with experimentally-validated Hammett parameters, though ESPMax is easily calculated and does not rely on experimental data or interpolation from the correlation in Fig. 2. This descriptor has been used by Hergenrother and coworkers in a Quantitative Structure Activity Relationship setting to understand permeability of cationic nitrogen compounds.[16]

The most important descriptor developed in our laboratory thus far is the Average Steric Occupancy (ASO) descriptor, which we see as instrumental to the success of the workflow.[4] ASO descriptors were first used to represent CPA derivatives, and the process is summarized in Fig. 3. ASOs are grid-based descriptors constructed from a steric indicator field (SIF) – meaning that they encode information at pre-determined points in a grid around a molecule, but in this case each grid point is assigned a binary 'indicator' value of 1 or 0 if it is within the van der Waals radius of an atom.

These SIF descriptors, when all candidate structures are aligned to a common orientation, encode steric occupancy at the same relative positions in space. We generate *Average* SIF, or ASO descriptors, by averaging the values at each grid point across a conformer ensemble of each catalyst. The result is a representation which looks like a 'heatmap' of steric occupancy (Fig. 3B). ASO is a high-dimensional representation of stereostructure. As a result, it can be difficult to 'see' how this describes chemical entities. Using dimensionality reduction techniques like Principal Component Analysis (PCA), we can visualize how CPAs in a diverse library are positioned in the ASO chemical space. Qualitatively, it is encouraging that in this representation different catalysts in the same class are generally grouped together (colors in 3D plot in Fig. 3C). For a more in-

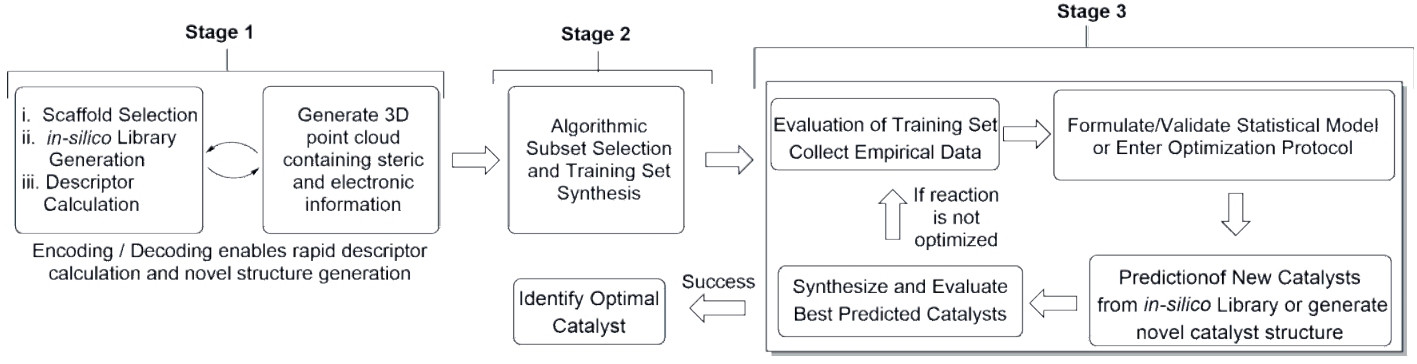


Fig. 1. Chemoinformatic workflow developed in these laboratories. Reproduced with permission from ref. [4]. Copyright 2021 American Chemical Society

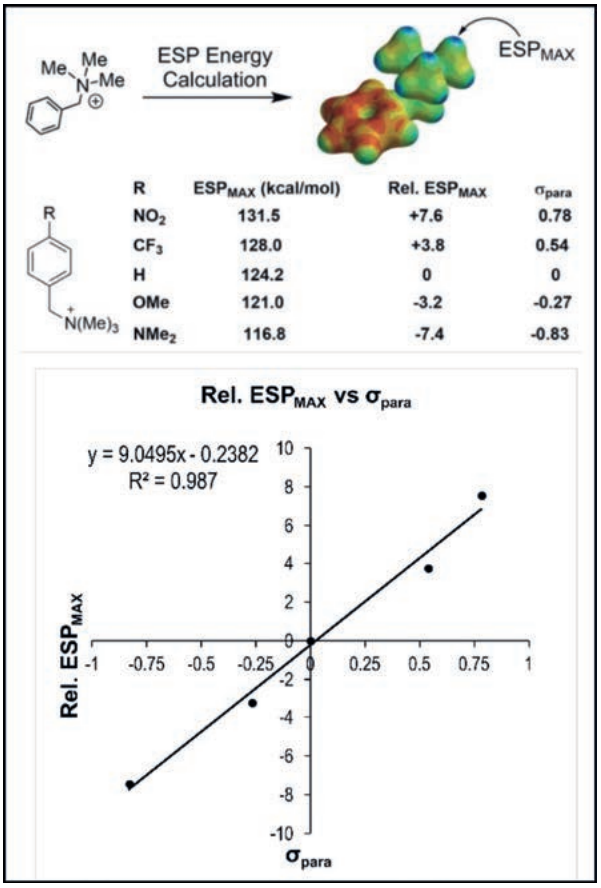


Fig. 2. ESPMax descriptor developed in our laboratory. Adapted with permission from ref. [5b]. Copyright 2020 American Chemical Society.

depth discussion of our benchmarking against other features, see the recent review.[4]

2.2 *The Role of Subset Selection*

We have learned that data used to train models must be curated to represent catalyst chemical space in a manner which is not directed by commercial availability and chemical intuition, but instead led by the directive to maximize the diversity of catalyst structures represented. By ensuring that maximum catalyst structure diversity is represented in a dataset, we have found that the predictions made from models trained on that dataset can be reliable at predicting new catalyst selectivities – a much more difficult task than predicting the selectivity of a catalyst that the model has has already seen. To accomplish this, we use algorithmic subset selection, which is an unsupervised process (meaning only catalyst features, *not* reaction data are used) in which a maximally-diverse subset of catalysts is identified from a much larger pre-defined *in silico* catalyst library, to choose an optimal training set of catalysts from which to acquire data. The process of subset selection is pictorially represented in Fig. 4. After encoding an *in silico* library of catalysts using molecular descriptors, each catalyst represents a position in chemical space. Using algorithmic selection ensures that a subset covers the breadth of that chemical space. This optimal training set of catalysts is called a 'Universal Training Set' (UTS) because it represents a particular catalyst scaffold in a manner which is agnostic to any specific transformation or reaction mechanism.

This strategy derives from the hypothesis that algorithmic subset selection from a large library should provide a better set of catalysts to optimize a reaction with than using commercially available catalysts or chemical intuition. To test this hypothesis, we devised a study that involved comparing the performance

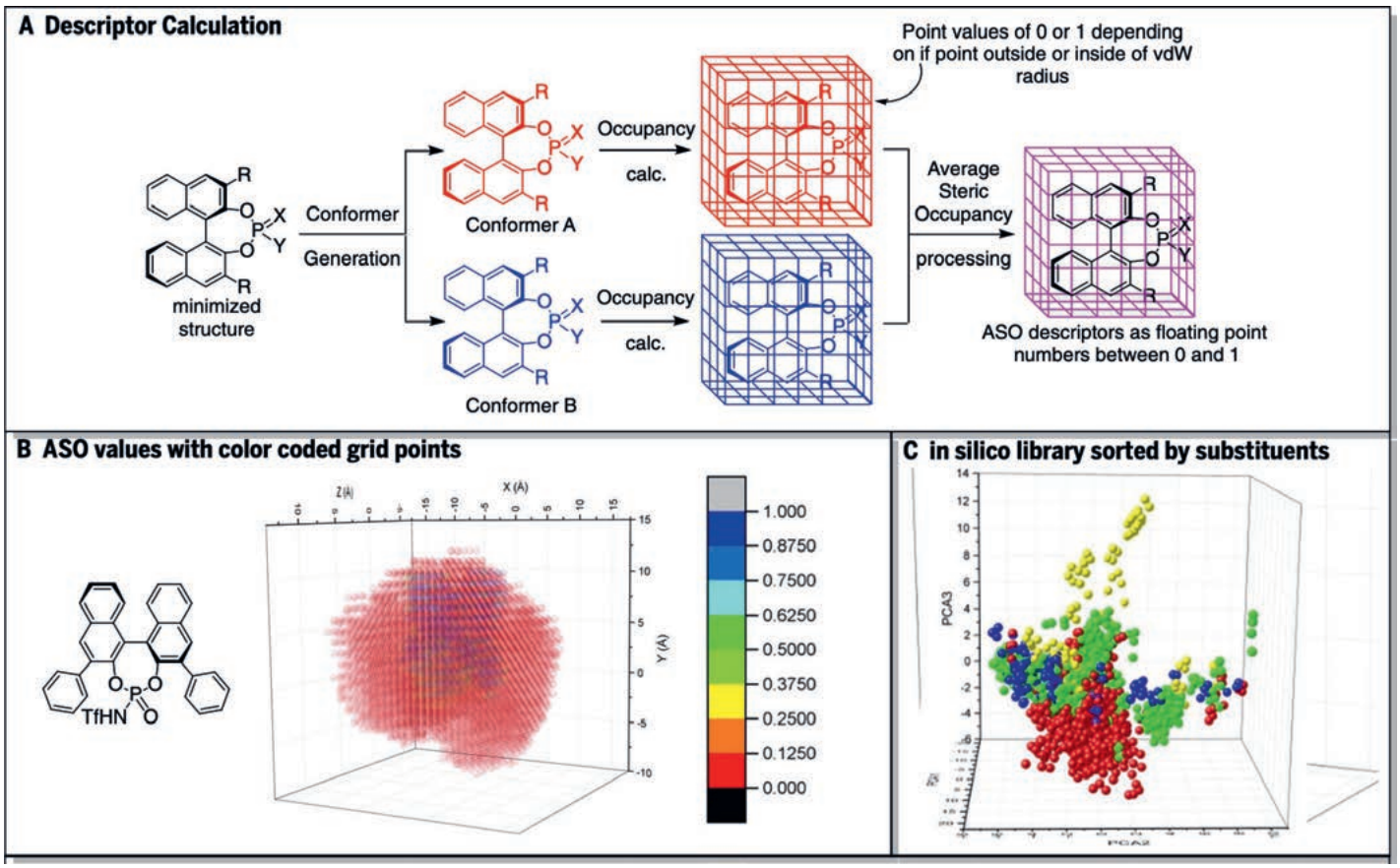


Fig. 3. A: ASO descriptor calculation, B: Heatmap of ASO descriptors of a BINOL-phosphoryltriflamide, C: Plot of first three principal components of in silico library of BINOL-phosphoryltriflamides. Adapted with permission from ref. [5a]. Copyright 2019 American Association for the Advancement of Science.

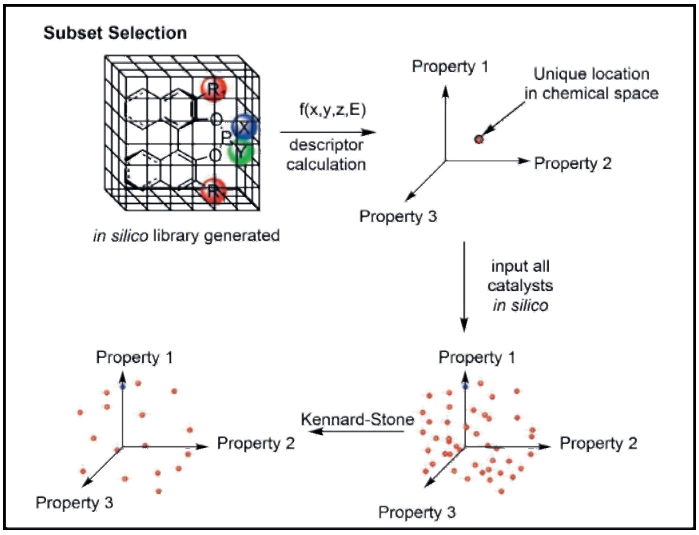


Fig. 4. Subset selection on an *in silico* library of catalysts. Adapted with permission from ref. [5a]. Copyright 2019 American Association for the Advancement of Science.

of our training set algorithmically selected *in silico* library (designed to include vast chemical diversity of synthetically accessible CPAs) to the performance of a training set consisting of commercially available CPAs.[5b] The experimental design is as follows: (1) commercially available catalysts within the *in silico* library were identified (12 in total) and used for a 'commercially available training set', (2) the top 12 CPAs from the Kennard Stone selection process were identified for comparison, and (3) the data from the 25 substrate pairs were used to train models (12 catalysts X 25 substrate pairs = 300 data points) using each catalyst ensemble, the remaining 775 data points were used as test data (*vide infra*). The results are depicted in Fig. 5. The Kennard-Stone selected catalysts showed good test set performance (Fig. 5, MAE = 0.21, RMSE = 0.26 kcal/mol, R^2 = 0.79) because it included catalysts spanning the entire catalyst chemical space in the training data. In contrast, using only commercially available catalysts to compose a training set showed diminished performance in the test set (MAE = 0.28, RMSE = 0.36 kcal/mol, R^2 = 0.53).

This result suggests that the algorithmic subset selection of training catalysts is a better way to select catalysts when designing a dataset. Our hypothesis is that the problem with the commercially available catalyst training set stemmed from its insufficient representation of the catalyst chemical space. To test this hypothesis we carried out the following experiment:[5b] (1) a clustering algorithm was used to identify groups of catalysts in the *in silico* library of CPAs, (2) using an elbow plot for k-mean clustering (k = 6) the clusters were inspected for the presence of any commercially available catalysts, (3) a representative was chosen from the one cluster which contained no commercially available catalysts and finally, (4) an *augmented* training set (now 13) was created which included data from the commercially *un*available catalyst intended to 'teach' the model about the unrepresented catalyst chemical space in the commercially available training set (Fig. 5). For new models trained on the augmented training data, the test set performance recovered significantly (MAE = 0.21, RMSE = 0.27 kcal/mol, R^2 = 0.74). This result gives us confidence that any dataset gathered from an algorithmically selected subset of catalysts will include adequate catalyst diversity to develop chemically meaningful QSSR models. Another valuable interpretation of this study is the apparently minimal cost in model performance using an augmented commercially available training set of catalysts. If a practitioner wished to avoid synthesizing an entire UTS, then commercially available catalysts could be chosen and any unrepresented clusters can be used to synthesize far fewer new catalysts.

2.3 *Validation of the Workflow*

After synthesizing the UTS of CPAs, (24 catalysts) the experimental validation campaign began by collecting enantioselectivity data for 16 substrate combinations (4 *N*-acyl imines, 4 thiols) for training data points. In addition 19 test catalysts were chosen randomly from the *in silico* library from which the test data was generated (9 different substrate combinations). After collecting 1,075 unique reaction enantioselectivities in duplicate (a total of 2,150 reactions), we began developing models relating catalyst and substrate features to enantioselectivity. We found empirically that support vector machines gave the best performance on the basis of mean absolute error (MAE) of predicted and observed selectivity values.[5a]

By design, this study involved pre-determining out-of-sample substrates and catalysts with which our models would remain naïve. We have demonstrated the importance of this design feature when working with combinatorial datasets.[17] As a consequence, we were able to independently assess the impact on test set predictions for subsets of the data which included out-of-sample substrates, catalysts, or both. The predicted vs observed plots of selectivities ($\Delta\Delta G^{\ddagger}$, kcal/mol) for the training data (384 reactions) and all three test sets (691 reactions) are depicted in Fig. 6.

These results serve to validate both the descriptors developed in that study as well as the workflow up to the penultimate step. Ultimately, our goal in developing this workflow was to identify optimal catalysts with lower selectivity data. We devised a simulation of such an optimization – necessary since this reaction was already an optimized reaction – in which data with less than 80% enantiomeric excess (Fig. 6, right chart, purple data) was used to train feed-forward neural network models which were subsequently used to predict the selectivity of catalyst/substrate combinations with known higher selectivity (red data). Our goal was to simulate a real-life situation in which an experimentalist has gathered data for a range of catalysts, but cannot break the threshold of 80% ee using chemical intuition.[5a]

The results show a predictably lower accuracy (both by R^2 and MAE) in the training and test data than with models trained on data spanning the entire range of enantioselectivity, yet the MAE for the test set (higher selectivity data) was still 0.28 kcal/mol, which is lower than errors expected for DFT calculated energies. In addition, the selectivity of various catalysts, though under-predicted, were predicted in the correct order, meaning that such a model could be used to select candidate catalysts with a higher selectivity in our simulated optimization scenario.

3. Summary and Outlook

This work is the culmination of years of trial and error and constitutes the beginning of a major research direction for our laboratory using a data-driven approach to optimize enantioselective catalysts. We have realized a workflow through the development of chemical descriptors for chiral molecules, the application of algorithmic subset selection methods to choosing catalysts for dataset acquisition, and machine learning. Readers can find a more detailed discussion of this chemoinformatic workflow in a recent review article.[4] We are currently applying our workflow to an array of catalyst optimization problems from organo-catalyzed to transition metal-catalyzed transformations.

Our future research directions include leveraging statistical modeling and 3D molecular descriptors to create and use models as an 'idea generator' for guiding mechanistic inquiry and to synthesize and use UTSs generated by this workflow for a range of privileged[1] catalyst scaffolds.

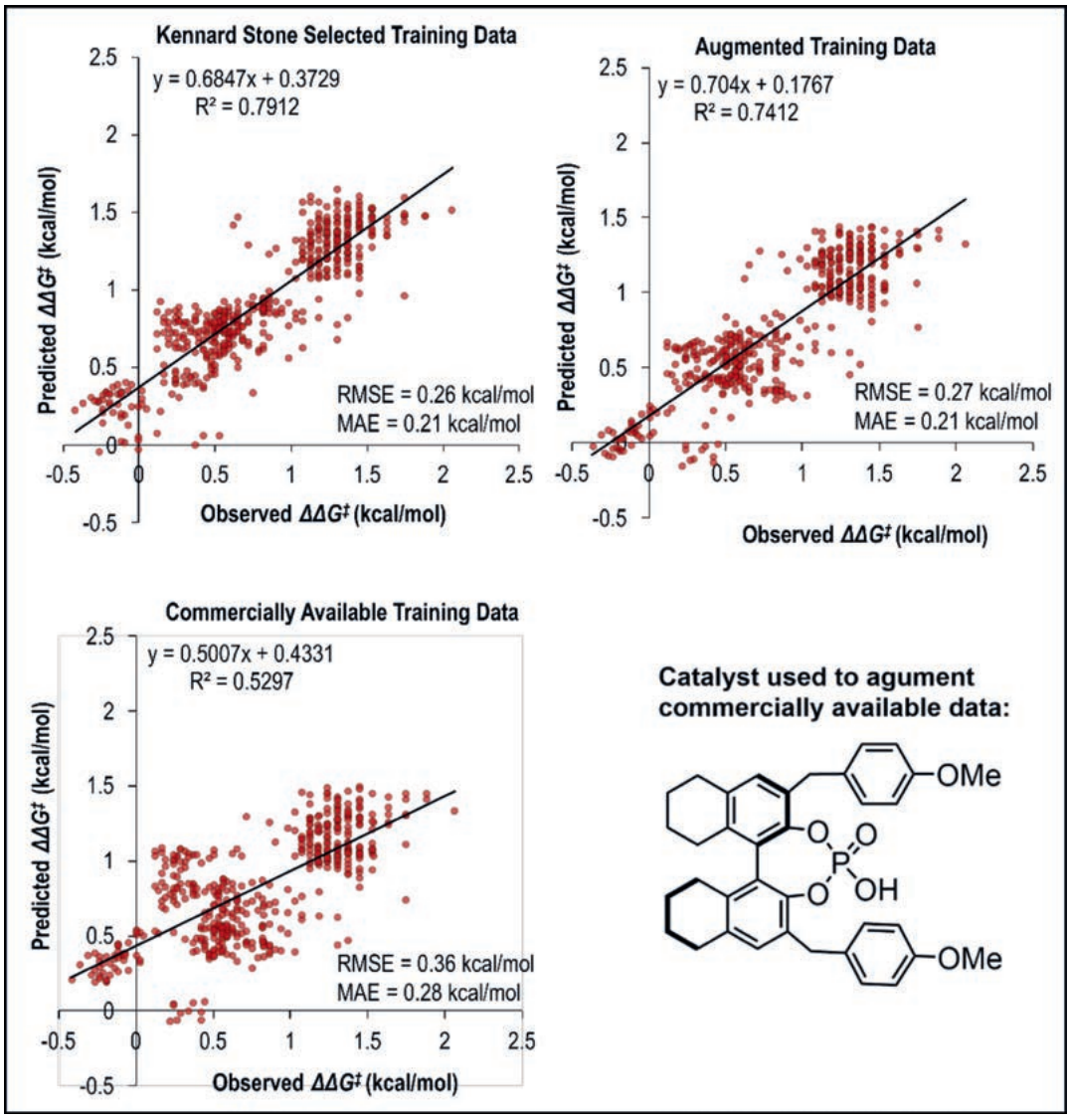


Fig. 5. Left: external test set of a model trained on algorithmically-selected catalyst training data vs. external test set of a model trained on commercially available catalyst data; Right: external test set of model trained on commercially available catalyst training data augmented by one new catalyst. Adapted with permission from ref. [5b]. Copyright 2020 American Chemical Society.

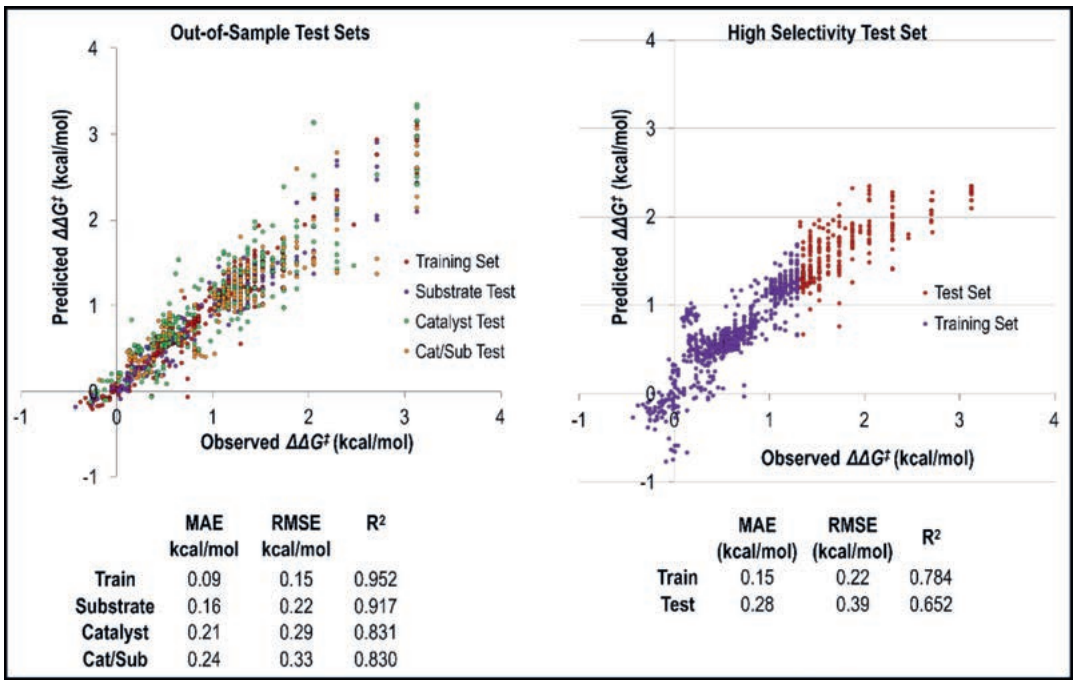


Fig. 6. Predicted vs. observed plots for training and test data described in Scheme 1. Reproduced from ref. [4] with permission. Copyright 2021 American Chemical Society.

Acknowledgements

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

Received: June 4, 2021

[1] a) C. Markert, P. Rosel, A. Pfaltz, *J. Am. Chem. Soc.* **2008**, *130*, 3234, https://doi.org/10.1021/ja0740317; b) F. Strieth-Kalthoff, C. Henkel, M. Teders, A. Kahnt, W. Knolle, A. Gomez-Suarez, K. Dirian, W. Alex, K. Bergander, C. G. Daniliuc, B. Abel, D. M. Guldi, F. Glorius, *Chem.* **2019**, *5*, 2183, https://doi.org/10.1016/jchempr.2019.06.004.

[2] E. Hansen, A. R. Rosales, B. Tutkowski, P. Norrby, O. Wiest, *Acc. Chem. Res.* **2016**, *49*, 996, https://doi.org/10.1021/acs.accounts.6b00037.

[3] a)M.S.Sigman,K.C.Harber,E.N.Bess,A.Milo,*Acc.Chem.Res.***2016**,*49*,1292, https://doi.org/10.1021/acs.accounts.6b00194; b) C. B. Santiago, J. Guo, M. S. Sigman, *Chem. Sci.* **2018**, *9* 2398, https://doi.org/10.1039/C7SC04679K.

[4] N. I. Rinehart, A. F. Zahrt, J. J. Henle, S. E. Denmark, *Acc. Chem. Res.*, **2021**, *54*, 2041, https://doi.org/10.1021/acs.accounts.0c00826.

[5] a) A. F. Zahrt, J. J. Henle, B. T. Rose, Y. Wang, W. T. Darrow, S. E. Denmark, *Science* **2019**, *363*, eaau5631, https://doi.org/10.1126/science.aau5631; b) J. J. Henle, A. F. Zahrt, B. T. Rose, W. T. Darrow, Y. Wang, S. E. Denmark, *J. Am. Chem. Soc.* **2020**, *142*, 11578, https://doi.org/10.1021/jacs.0c04715.

[6] G. K. Ingle, M. G. Mormino, L. Wojtas, J. C. Antilla, *Org. Lett.* **2011**, *13*, 4822, https://doi.org/10.1021/ol201899c.

[7] N. D. Gould, L. M. Wolf, S. E. Denmark, *J. Org. Chem.* **2011**, *76*, 4337, https://doi.org/10.1021/jo2005457.

[8] a) M. C. Kozlowski, J. C. Ianni, *J. Mol. Catal. A: Chem.* **2010**, *324*, 141, https://doi.org/10.1016/j.molcata.2010.03.030; b) M. C. Kozlowski, S. L. Dixon, M. Panda, G. Lauri, *J. Am. Chem. Soc.* **2003**, *125*, 6614, https://doi.org/10.1021/ja0293195.

[9] a) K. B. Lipkowitz, M. Pradhan, *J. Org. Chem.* **2003**, *68*, 4648, https://doi.org/10.1021/jo0267697; b) K. B. Lipkowitz, M. C. Kozlowski, *Synlett* **2003**, *10*, 1547, https://doi.org/10.1055/s-2003-40849.

[10] a) J. L. Melville, B. I. Andrews, B. Lygo, J. D. Hirst, *Chem. Commun.* **2004**, *4*, 1410, https://doi.org/10.1039/B402378A; b) J. L. Melville, K. R. J. Lovelock, C. Wilson, B. Allbutt, E. K. Burke, B. Lygo, J. D. Hirst, *J. Chem. Inf. Model.* **2005**, *45*, 971, https://doi.org/10.1021/ci0500511.

[11] A. F. Zahrt, S. V. Athavale, S. E. Denmark, *Chem. Rev.* **2020**, *120*, 1620, https://doi.org/10.1021/acs.chemrev.9b00425.

[12] G. Skoraczynski, P. Dittwalkd, B. Miasojedow, S. Szymkuc, E. P. Gajewska, B. A. Grzybowski, A. Gambin, *Sci. Rep.* **2017**, *7*, 3582, https://doi.org/10.1038/s41598-017-02303-0.

[13] H. Zabrodsky, D. Anvir, *J. Am. Chem. Soc.* **1995**, *117*, 462, https://doi.org/10.1021/ja00106a053.

[14] A. F. Zahrt, S. E. Denmark, *Tetrahedron* **2019**, *75*, 1841, https://doi.org/10.1016/j.tet.2019.02.007.

[15] A. F. Zahrt, N. I. Rinehart, S. E. Denmark, *Chem. Eur. J.* **2021**, 2343, https://doi.org/10.1002/ejoc.202100027.

[16] S. J. Perlmutter, E. J. Geddes, B. S. Drown, S. E. Motika, M. R. Lee, P. J. Hergenrother, *ACS Infect. Dis.* **2021**, *7*, 162, https://doi.org/10.1021/acsinfecdis.0c00715.

[17] A. F. Zahrt, J. J. Henle, S. E. Denmark, *ACS Comb.* **2020**, *22*, 586, https://doi.org/10.1021/acscombsci.0c00118.

[18] a) Q.-L. Zhou, 'Privileged Chiral Ligands and Catalysts', Wiley-VCH, 2011; b) T. P Yoon, E. N. Jacobsen *Science* **2003,** *299*,1691, https://doi.org/10.1126/science.1083622.