ACS Combinatorial Science

pubs.acs.org/acscombsci Letter

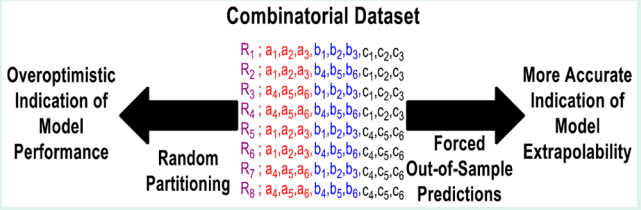
Cautionary Guidelines for Machine Learning Studies with Combinatorial Datasets

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

Cite This: *ACS Comb. Sci.* 2020, 22, 586–591 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Regression modeling is becoming increasingly prevalent in organic chemistry as a tool for reaction outcome prediction and mechanistic interrogation. Frequently, to acquire the requisite amount of data for such studies, researchers employ combinatorial datasets to maximize the number of data points while limiting the number of discrete chemical entities required. An often-overlooked problem in modeling studies using combinatorial datasets is the tendency to fit on patterns in the datasets (i.e., the presence or absence of a reactant or catalyst) rather than to identify meaningful trends between descriptors and the response variable. Consequently, the generality and interpretability of such models suffer. This report illustrates these well-known pitfalls in a case study, demonstrates the necessary control experiments to identify when this property will be problematic, and suggests how to perform further validation to assess general applicability and interpretability of models trained using combinatorial datasets.

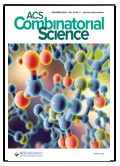


KEYWORDS: *machine learning, enantioselective catalysis*

Machine learning (ML) in organic chemistry has garnered much attention recently, with the promise to create tools capable of predicting reaction outcomes,[1−3] identifying reaction conditions,[4] planning synthetic routes,[5,6] or expediting catalyst optimization.[7,8] A requirement for machine learning studies is suitably large datasets for the construction of regression models. In the development of new catalytic reactions, generating a suitably large dataset is often prohibitive owing to the lack of available catalysts and the effort required to obtain accurate analysis of experimental outcomes. Combinatorial, high-throughput experimentation is therefore a natural pairing with ML tools in that the number of data points obtained per unique reaction component is maximized. As interest in applying ML to chemical problems increases, so will the use of combinatorial experimentation to generate sufficiently large datasets.

The desired outcome of modeling combinatorial data is to obtain a model that correctly fits chemical descriptors of reactions with experimental outcomes in a generally predictive way. It is expected that the model has developed correlations with parameters reflecting the underlying chemistry in the system of interest; however, this assumption is far from certain without careful analysis. An important aspect of combinatorial datasets that has implications in ML studies is that the data have an intrinsic pattern in which identical reaction components[9] are present many times in different reactions. When these reaction components are parameterized and concatenated combinatorially to produce reaction parameters, identical sets of descriptors will be present in many different reactions (Figure 1). Models trained on these data can fit the intrinsic pattern in the dataset (i.e., the presence or absence of a particular chemical entity) resulting from the combinatorial construction of the reactions. As a result, the correlations identified by the model may not be founded in relevant chemical information; the lack of any physical relationship between the descriptors and the regressand (e.g., yield, selectivity, etc.) can potentially hamper the utility of the model when used in an extrapolative fashion. These hidden shortcomings confound model interpretation and limit general applicability of the model. Further, when individual reaction components are present in both the training set and the external test set, it is possible that this phenomenon will carry over to the external test set, resulting in inflated accuracy and an overoptimistic indication of model performance in novel systems. Thus, two concerns must be addressed when modeling with combinatorial datasets: (1) the resulting models may be less likely to be successfully applied to novel reaction components, diminishing predictive utility, and (2) the model may lack interpretability; consequently, extraction of mechanistic information is impossible.[10] This concept, termed "data leakage", has been thoroughly investigated in the machine learning literature.[11] Additionally, many resources containing introductory level explanations are available for machine learning novices.[12−15] We feel it is imperative for experimen-

Received: June 9, 2020
Revised: September 16, 2020
Published: October 1, 2020



https://dx.doi.org/10.1021/acscombsci.0c00118
ACS Comb. Sci. 2020, 22, 586–591

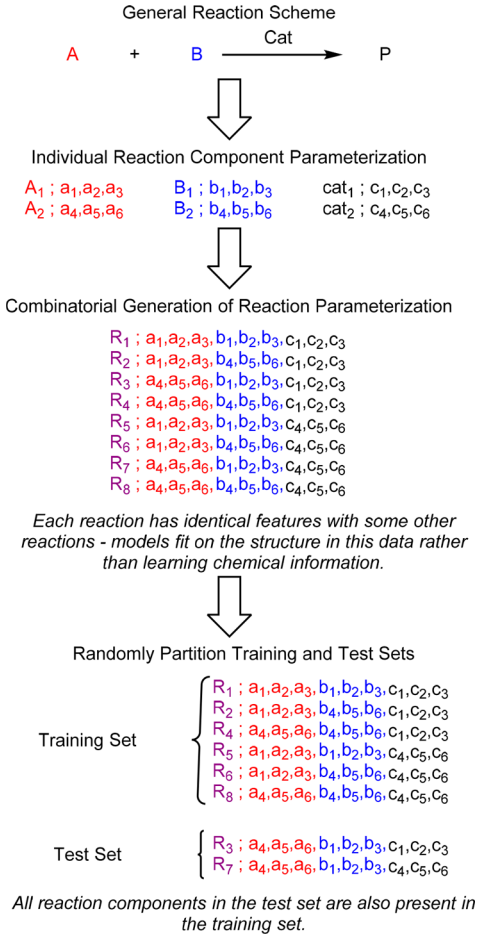


Figure 1. Combinatorial construction of reaction features with the hypothetical reaction A + B → P catalyzed by C, resulting in structured data. A and B are reactants with corresponding descriptors a and b, and cat is the catalyst represented by features c.

talists evaluating ML studies or attempting to incorporate ML studies into their own research program to familiarize themselves with these topics prior to designing ML studies. Additionally, we anticipate this study will serve as an application-specific example for experimentalists exploring this topic.

Although practitioners should be cognizant of these potential limitations, with appropriate experimental design and proper control experiments, one may avoid these potential pitfalls. In fact, this aspect of modeling with combinatorial datasets is well-known in data science and has been discussed in at least one study relating to the chemical sciences.[16] However, we feel that this concept is not well appreciated in the chemistry community, particularly with regards to machine learning in catalysis. As ML studies using combinatorial datasets continue gaining in popularity, experimentalists must be aware of these potential limitations before designing and analyzing the results of ML studies that employ such data structures. Accordingly, the work described herein investigates how data partitioning can influence hypothesis testing in studies with combinatorial datasets. To that end, this study aims to demonstrate the importance of standardized control experiments and intentional experimental design in ML endeavors to develop predictive, meaningful models using combinatorial datasets.

In this study, a dataset gathered and analyzed previously in our laboratories is examined.[17,18] In this work, the chiral Brønsted-acid-catalyzed, enantioselective formation of *N,S*-acetals was used as a model system.[19] By testing 43 different chiral catalysts with a matrix of 5 imines and 5 thiols, a total of 1075 reactions were evaluated, each of which were run in duplicate (Figure 2). The enantioselectivity of each reaction was measured, and the percent enantiomeric excess (% ee) was recorded as an average of duplicate runs. This value is then converted to the $\Delta\Delta G$ (kcal/mol) between competing transition structures, which is used as the regressand.

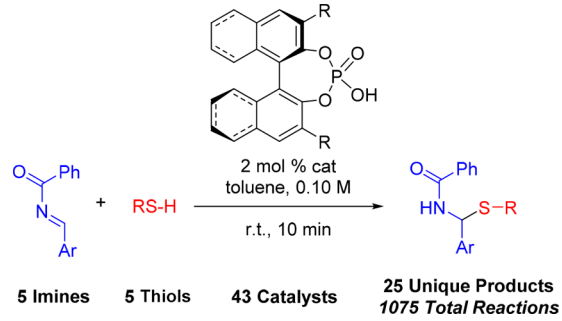


Figure 2. Dataset summary for the enantioselective formation of *N,S*-acetals.

The dataset employed to develop cross-validated ML models comprises 24 training catalysts and 16 training substrate combinations to construct a training set of 384 reactions. This set was used to train the models and to perform k-fold cross-validation. The models with the best performance were then compared in their ability to predict the outcomes of an external test set (that is, a set not involved in model training or selection in any way) of 691 reactions. In this way, every test reaction is an out-of-sample prediction. An out-of-sample prediction is a case in which at least one reaction component is not present in any reaction in the training data. An alternative phrasing of this definition is that all reaction components used to fit the model are considered in-sample components, and all other components constitute out-of-sample predictions. In this case, every imine, thiol, and catalyst used in model development constitute in-sample predictions. These are the reactions in the top left quadrant in Figure 3. It follows then that the reactions in the top right, bottom left, and bottom right quadrants are out-of-sample predictions because they include substrates or catalysts (or both) that are absent from the training data (for specific details on catalyst/substrate partitioning, see the Supporting Information). The conclusion from our previous study is that conformer-dependent molecular field descriptors produce statistically significantly more accurate models than single-conformer descriptors. In the present study, we demonstrate that different data partitioning would lead to different conclusions.

To demonstrate how data partitioning schemes can influence hypothesis testing, two different data partitioning methods were compared. The first data partitioning scheme used in this work mandates out-of-sample predictions, as described in Figure 3. Using this method of data partitioning, reaction components are intentionally omitted from the training data and only appear in the test set. The second data partitioning used random selection to identify a test set of 691 from the full 1075 reaction dataset, and the remaining 384

https://dx.doi.org/10.1021/acscombsci.0c00118
ACS Comb. Sci. 2020, 22, 586−591

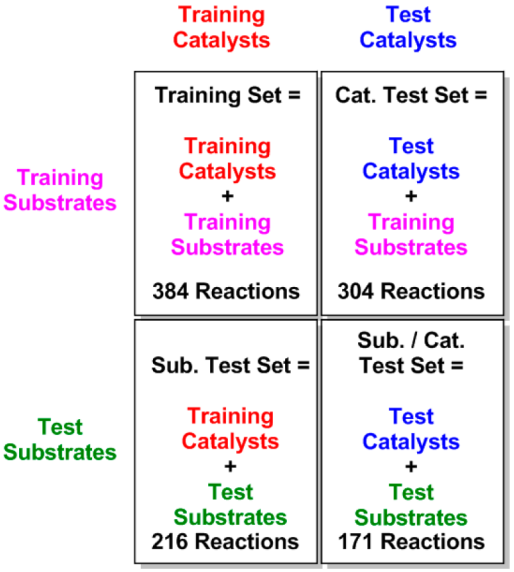


Figure 3. Data partitioning of the total 1075 member set used in ref 8. The top left quadrant is used for training and cross-validation, and the other three are external test sets. The top right quadrant has catalyst structures novel to the model but substrate combinations from the training set; the bottom left quadrant has catalyst structures present in the training data but substrate combinations novel to the model, and the bottom right quadrant has both substrate combinations and catalyst structure that are novel to the model. The three quadrants with test groups are combined to form the 671 member test set.

reactions were used for model training and validation. Note that by randomly selecting the training and test reaction sets, out-of-sample predictions in the test set are unlikely.[20] In this data partitioning scheme, we suspect test set reactions will be erroneously accurately predicted from the model simply fitting to the pattern in the data. However, in the first data partitioning scheme, this error can be mitigated by intentionally including reaction components in the external test set that are absent from the training set. Using both data partitioning methods, the performance of different descriptor classes was compared. In addition, a set of control experiments was performed by training models using randomly generated descriptors to represent discrete chemical entities and one-hot encoding of reactions (see the Supporting Information for more information regarding control experiments and descriptor types utilized in this analysis).[16]

The models discussed in this study are gradient boosting regression models trained using the 384 reaction sets. Three different types of molecular field descriptors were used in model generation: average steric occupancy (ASO), steric indicator fields (SIF), and molecular interaction fields (MIF). Additionally, two control experiments were run. In the first, individual reaction components are represented by a series of randomly generated numbers. Similarly, in the second control experiment, the reactions are parameterized with one-hot encoding. In this case, the presence of a reaction component is represented by a 1 and the absence is represented by a 0. In these control experiments, the models are clearly fitting to the pattern in the dataset and should fail when used to make extrapolative predictions for reactions with novel components (i.e., an imine, thiol, or catalyst absent from the training data) because chemical information is obviously absent from the molecular representation. If models generated using these features perform equally well to models constructed using chemical descriptors, one cannot yet determine if the chemical descriptors contain meaningful chemical information associated with reaction outcome. It is also not yet possible to assess if the model will succeed in predicting the outcome of reactions with novel components (again, in this case referring to imine, thiol, or catalyst structures not present in the training data). In this scenario, the minimum requirement for assessing model extrapolability and interpretability has not been met. To probe these concepts, a data partitioning scheme forcing out-of-sample predictions is necessary. Hypothetically, if the models derived from chemical descriptors were to demonstrate poor performance in out-of-sample predictions, it would indicate that the models have poor extrapolative performance and, consequently, will give erroneous predictions when applied to new scenarios. Further, this would indicate that the chemical descriptors do not contain information related to reaction outcome. Without forcing out-of-sample predictions, these types of conclusions would go unnoticed. In the present study, this type of analysis is necessary to draw conclusions regarding the superiority or inferiority of different chemical descriptors. Using the nonrandom data partition scheme, models calculated using chemical descriptors that perform statistically significantly better than both random featurization and one-hot encoding in making out-of-sample predictions likely have some degree of extrapolability. Consequently, the descriptors that result in models with the highest performance likely contain the most chemical information related to reaction outcome. We hypothesize that using the descriptors containing the most relevant chemical information will achieve the models with the broadest applicability domain, increasing the likelihood of making successful predictions in novel use cases (i.e., identification of a more selective catalyst, predicting a reaction outcome for a novel substrate combination, etc.). Herein, we demonstrate how the data partitioning method used influences the conclusions one would draw from this type of study.

The five different parameterizations described above (ASO, SIF, MIF, random featurization, and one-hot encoding) were used to make models using both data partitioning methods (intentionally designating out-of-sample predictions vs random partitioning). The five different models with each partitioning method are then compared on the basis of their performance in predicting reaction outcomes of the 691 reaction external test set. A summary of this analysis is provided in Figure 4.

The graphs presented in Figure 4 clearly indicate different results depending on which data partitioning method is used. First, when examining the randomly selected external test set (Figure 4, top), no chemical descriptor classes are statistically significantly different than the random featurization as determined by a one-way ANOVA with a Tukey posthoc test (see the Supporting Information for a full statistical analysis). This analysis suggests that the models are fitting to a pattern in the data, and that the descriptors are performing no better than random numbers. Accurate models can still be constructed because every catalyst or substrate present in the external test set is also present in the training set—therefore, fitting to the pattern in the data in model training carries over to the external test set. This result leads to two possible interpretations: (1) in the absence of the control models,[21] one would conclude that the three different descriptor classes are not significantly different from one another, or (2) in the presence of the control models, one would conclude that the

https://dx.doi.org/10.1021/acscombsci.0c00118
ACS Comb. Sci. 2020, 22, 586−591

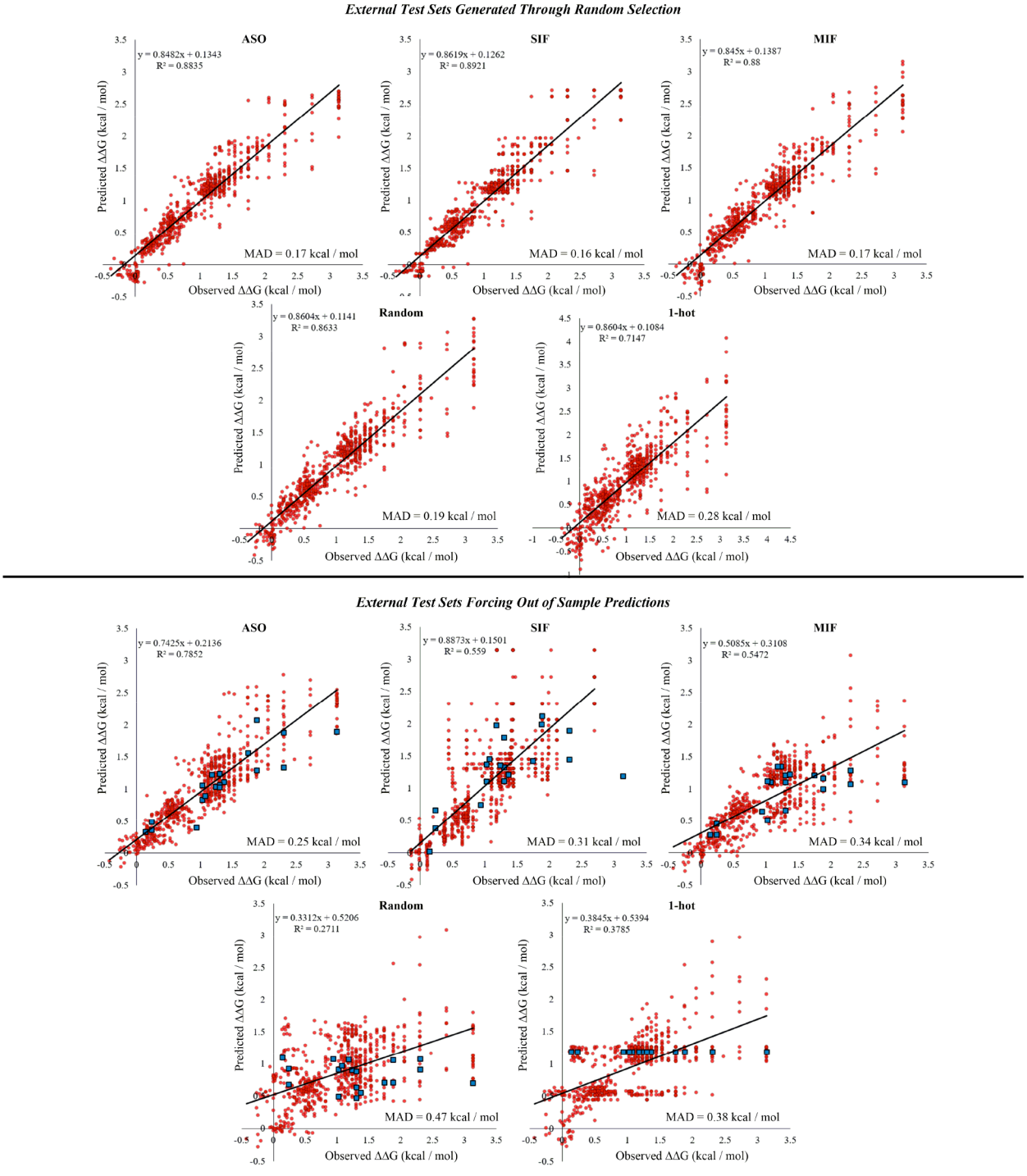


Figure 4. Comparison of models generated with different descriptor sets. The sets on the top are external test sets selected from random partitioning and contain no out-of-sample predictions, whereas the models on the bottom contain out-of-sample predictions. The blue squares in the bottom plots represent reactions in which the imine, thiol, and catalyst in that reaction do not appear in the training set.

experimental models are not necessarily learning the underlying chemistry and simply fitting to a pattern in the data.

When incorporating out-of-sample predictions in the experimental design, strikingly different results are obtained (Figure 4, bottom). In this case, all three molecular field descriptor types lead to statistically significantly more accurate models than both random features and one-hot encoding (see the Supporting Information for the complete statistical analysis and all pairwise comparisons). Further, the ASO descriptors lead to statistically significantly more accurate models than either the SIF or MIF descriptors. Clearly, a difference in data partitioning strategy changed the analysis of the experiment and allowed for better understanding of the utility of these models. If researchers were to attempt to implement the models with random partitioning of data in a novel system containing out-of-sample predictions, the predictions made would likely be significantly less accurate than expected when predicting the external test set, especially if a lower-performing descriptor class was used. In other words, common metrics such as R^2 and MAD alone are not adequate to assess

https://dx.doi.org/10.1021/acscombsci.0c00118
ACS Comb. Sci. 2020, 22, 586−591

extrapolative performance of a model. By intentionally incorporating out-of-sample predictions, a more accurate indication of how the model would perform in practical applications can be obtained. Further, higher confidence in evaluating the efficacy of different molecular representations can be obtained. This conclusion is in line with previous work exploring the concept of leave-one-cluster-out cross-validation, indicating that this concept is generally applicable to machine learning studies in the chemical sciences.[22]

An additional observation regarding the models in Figure 4 is that random features and one-hot encoding still provide some accurate predictions of reaction outcomes of the test set. In these reactions, one or more reaction components (imine, thiol, or catalyst) are present in the training data. Because not every reaction component is absent from the training data in these reactions, it is the structure in the training set that leaks over to the test set. However, the statistically significant difference between the chemically meaningful descriptors and random featurization and one-hot encoding indicates that the accuracy of those models is attributed to more than this structure alone, that is, learned chemical information. Further, there are some reactions in which none of the imine, thiol, or catalyst appears in the training data. These reactions are depicted as the blue points in Figure 4 and demonstrate the expected behavior in the control experiments (i.e., completely inaccurate predictions), whereas they are more accurately predicted with the model using chemical descriptors. This observation suggests that an experimental design designating some external test reactions in which all reaction components are out-of-sample is best for validation studies.

This study shows that validation of models trained using combinatorial datasets will benefit from intentionally including out-of-sample predictions, particularly when comparing between different descriptor sets or attempting to interpret models. Test reactions in which no reaction component is present in the training set will likely give the best indication of model performance in novel applications and model interpretability. Additionally, multiple partitions should be performed to ensure that model efficacy is not arising from fortuitous test set selection. Finally, as stated in a previous report,[16] this experimental design does not circumvent the need for additional control experiments (e.g., comparison with random featurization and one-hot encoding). Accordingly, standard control experiments and an experimental design that incorporates out-of-sample predictions are necessary for rigorous validation when modeling with combinatorial datasets.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acscombsci.0c00118.

Dataset information, modeling information, and statistical snalysis (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Scott E. Denmark − *Roger Adams Laboratory, Department of Chemistry, University of Illinois, Urbana, Illinois 61801, United States;* orcid.org/0000-0002-1099-9765; Phone: (217) 333-0066; Email: sdenmark@illinois.edu; Fax: (217) 333-3984

Authors

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

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

Complete contact information is available at: https://pubs.acs.org/10.1021/acscombsci.0c00118

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We are grateful to the W. M. Keck Foundation and the National Science Foundation for generous financial support (NSF CHE1900617). A.F.Z. is grateful to the University of Illinois for Graduate Fellowships. We thank Brennan T. Rose, William T. Darrow, and Dr. Yang Wang for experimental support. We thank Dr. Raquel Mendizábal Martell for assistance with statistical analysis. We are also grateful for the support services of the NMR, mass spectrometry X-ray crystallographic, and microanalytical laboratories of the University of Illinois at Urbana—Champaign.

■ REFERENCES

(1) Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. Predicting Reaction Performance in C-N Cross-Coupling using Machine Learning. *Science* **2018**, *360*, 186−190.

(2) Granda, J. M.; Donina, L.; Dragone, V.; Long, D.-L.; Cronin, L. Controlling an Organic Synthesis Robot with Machine Learning to Search for New Reactivity. *Nature* **2018**, *559*, 377−381.

(3) Gao, H.; Struble, T. J.; Coley, C.; Wang, Y.; Green, W. H.; Jensen, K. F. Using Machine Learning To Predict Suitable Conditions for Organic Reactions. *ACS Cent. Sci.* **2018**, *4* (11), 1465−1476.

(4) Coley, C. W.; Barzilay, R.; Jaakkola, T. S.; Green, W. H.; Jensen, K. F. Prediction of Organic Reaction Outcomes Using Machine Learning. *ACS Cent. Sci.* **2017**, *3* (5), 434−443.

(5) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI. *Nature* **2018**, *555*, 604−610.

(6) Szymkuc, S.; Gajewska, E. P.; Klucznik, T.; Molga, K.; Dittwald, P.; Startek, M.Ł; Bajczyk, M.Ł; Grzybowski, B. A. Computer-Assisted Synthetic Planning: The End of the Beginning. *Angew. Chem., Int. Ed.* **2016**, *55*, 5904−5937.

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

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

(9) In this paper, the term "reaction components" refers to the reactants (imine and thiol) or the catalyst used in the reaction.

(10) It is worth noting that there are many other factors such as colinear or confounding variables and confirmation biases that can convolute the interpretation of models. This experimental design makes model interpretation infallible; it is a minimum requirement for researchers interested in exploring model interpretation as a field of research.

(11) Kaufman, S.; Rosset, S.; Perlich, C. Leakage in Data Mining: Formulation, Detection, and Avoidance. *Proceedings of the ACM SIGKDD international Conference on Knowledge Discovery and Data Mining* **2011**, *6*, 556−563.

(12) Yildirim, S. Data Leakage in Machine Learning. https://towardsdatascience.com/data-leakage-in-machine-learning-6161c167e8ba/ (accessed 2020-09-15).

(13) Cook, A. Data Leakage. https://www.kaggle.com/alexisbcook/data-leakage/ (accessed 2020-09-15).
(14) Brownlee, J. Data Leakage in Machine Learning. https://machinelearningmastery.com/data-leakage-machine-learning/ (accessed September 15, 2020).
(15) Soni, D. Data Leakage in Machine Learning. https://towardsdatascience.com/data-leakage-in-machine-learning-10bdd3eec742/ (accessed 2020-09-15).
(16) Chuang, K. V.; Keiser, M. J. Comment on Predicting Reaction Performance in C-N Cross-Coupling using Machine Learning. *Science* **2018**, *362*, No. eaat8603.
(17) Zahrt, A. F.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. Prediction of Higher-Selectivity Catalysts by Computer-Driven Workflow and Machine Learning. *Science* **2019**, *363*, No. eaau5631.
(18) Henle, J. J.; Zahrt, A. F.; Rose, B. T.; Darrow, W. T.; Wang, Y.; Denmark, S. E. Development of a Computer-Guided Workflow for Catalyst Optimization. Descriptor Validation, Subset Selection, and Training Set Analysis. *J. Am. Chem. Soc.* **2020**, *142*, 11578−11592.
(19) Ingle, G. K.; Mormino, M. G.; Wojtas, L.; Antilla, J. C. Chiral Phosphoric Acid-Catalyzed Addition of Thiols to *N*-Acyl Imines: Access to Chiral *N,S*-Acetals. *Org. Lett.* **2011**, *13*, 4822−4825.
(20) For example, in the current dataset, a given imine or thiol is present in 20% of the reactions (one out of five). Through random selection, the probability of selecting 384 reactions that do not contain a given imine or thiol is 6.1×10^{-36}%. Similarly, the probability of not selecting a given catalyst would be 0.012%. These numbers are calculated by $p = Xn$, in which p is the overall probability, X is the probability of a single event, and n is the number of events. For example, if we were to calculate the probability that we will never select of the five imines in any of the 384 training reactions, $p = (0.8)^{384} = 6.1 \times 10^{-36}$%.
(21) This aspect is not to indicate the control experiments are optional in machine learning studies—best practices should require that these control experiments are run to avoid misinterpretation of the results.
(22) Meredig, B.; Antono, E.; Church, C.; Hutchinson, M.; Ling, J.; Paradiso, S.; Blaiszik, B.; Foster, I.; Gibbons, B.; Hattrick-Simpers, J.; Mehta, A.; Ward, L. Can Machine Learning Identify the Next High-Temperature Superconductor? Examining Extrapolation Performance for Materials Discovery. *Mol. Syst. Des. Eng.* **2018**, *3*, 819−825.

https://dx.doi.org/10.1021/acscombsci.0c00118
ACS Comb. Sci. 2020, 22, 586−591