ACCOUNTS of chemical research

pubs.acs.org/accounts

Article

Predicting Reaction Yields via Supervised Learning

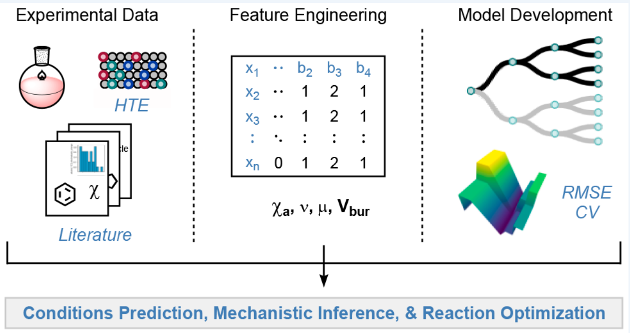
Published as part of the Accounts of Chemical Research special issue "Data Science Meets Chemistry".

Andrzej M. Żurański, Jesus I. Martinez Alvarado, Benjamin J. Shields, and Abigail G. Doyle*

Cite This: *Acc. Chem. Res.* 2021, 54, 1856−1865

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



CONSPECTUS: Numerous disciplines, such as image recognition and language translation, have been revolutionized by using machine learning (ML) to leverage big data. In organic synthesis, providing accurate chemical reactivity predictions with supervised ML could assist chemists with reaction prediction, optimization, and mechanistic interrogation.

To apply supervised ML to chemical reactions, one needs to define the object of prediction (e.g., yield, enantioselectivity, solubility, or a recommendation) and represent reactions with descriptive data. Our group's effort has focused on representing chemical reactions using DFT-derived physical features of the reacting molecules and conditions, which serve as features for building supervised ML models.

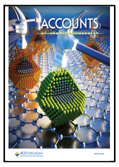
In this Account, we present a review and perspective on three studies conducted by our group where ML models have been employed to predict reaction yield. First, we focus on a small reaction data set where 16 phosphine ligands were evaluated in a single Ni-catalyzed Suzuki−Miyaura cross-coupling reaction, and the reaction yield was modeled with linear regression. In this setting, where the regression complexity is strongly limited by the amount of available data, we emphasize the importance of identifying single features that are directly relevant to reactivity. Next, we focus on models trained on two larger data sets obtained with high-throughput experimentation (HTE). With hundreds to thousands of reactions available, more complex models can be explored, for example, models that algorithmically perform feature selection from a broad set of candidate features. We examine how a variety of ML algorithms model these data sets and how well these models generalize to out-of-sample substrates. Specifically, we compare the ML models that use DFT-based featurization to a baseline model that is obtained with features that carry no physical information, that is, random features, and to a naive non-ML model that averages yields of reactions that share the same conditions and substrate combinations. We find that for only one of the two data sets, DFT-based featurization leads to a significant, although moderate, out-of-sample prediction improvement. The source of this improvement was further isolated to specific features which allowed us to formulate a testable mechanistic hypothesis that was validated experimentally. Finally, we offer remarks on supervised ML model building on HTE data sets focusing on algorithmic improvements in model training.

Statistical methods in chemistry have a rich history, but only recently has ML gained widespread attention in reaction development. As the untapped potential of ML is explored, novel tools are likely to arise from future research. Our studies suggest that supervised ML can lead to improved predictions of reaction yield over simpler modeling methods and facilitate mechanistic understanding of reaction dynamics. However, further research and development is required to establish ML as an indispensable tool in reactivity modeling.

KEY REFERENCES

- Wu, K.; Doyle, A. G. Parameterization of Phosphine Ligands Demonstrates Enhancement of Nickel Catalysis via Remote Steric Effects. *Nat. Chem.* **2017**, *9*, 779−784.[1] *In this study, we demonstrate that, when using a small set of 16 ligands in a Ni-catalyzed cross-coupling reaction, a multiple linear regression model can be constructed with a small set of steric and electronic ligand features.*
- 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] *We found that a random forest algorithm could be utilized to model a high-throughput experimentation (HTE) data set (3955 reactions) to predict reaction*

Received: November 18, 2020
Published: March 31, 2021



https://doi.org/10.1021/acs.accounts.0c00770

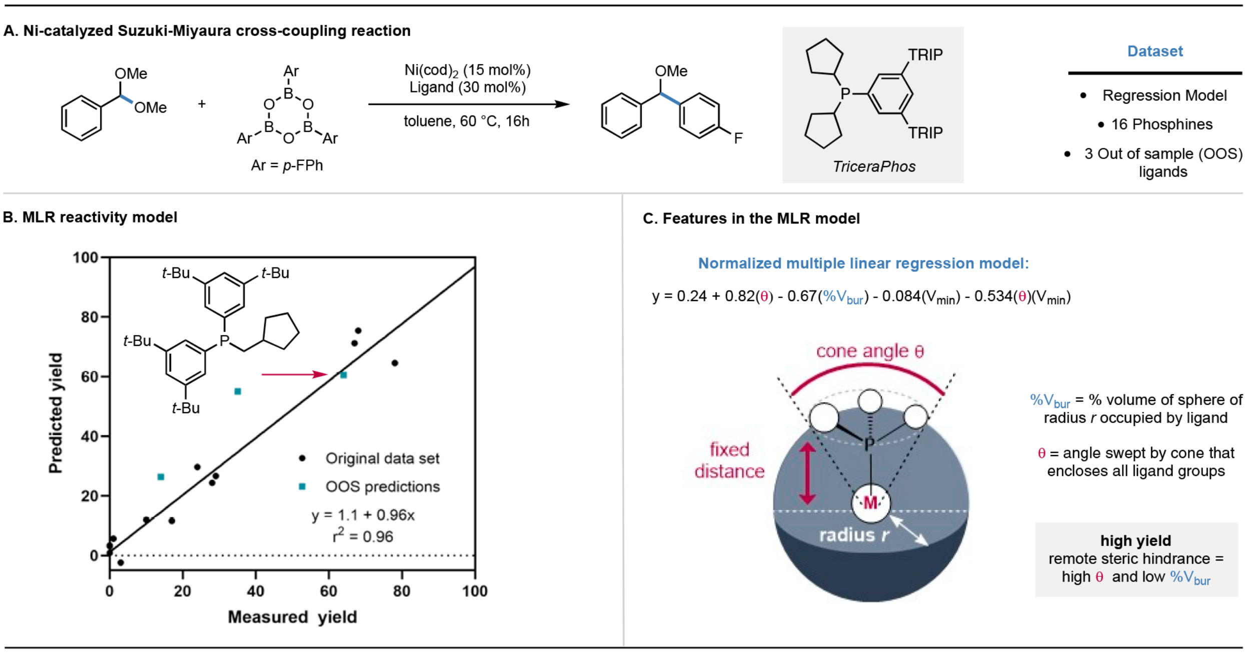

A. Ni-catalyzed Suzuki-Miyaura cross-coupling reaction
OMe
OMe
+
Ar
B
O O
Ar B O B Ar
Ar = p-FPh
Ni(cod)2 (15 mol%)
Ligand (30 mol%)
toluene, 60 °C, 16h
OMe
F
TRIP
P
TRIP
TriceraPhos
Dataset
• Regression Model
• 16 Phosphines
• 3 Out of sample (OOS) ligands
B. MLR reactivity model
t-Bu
t-Bu
t-Bu
P
t-Bu
Predicted yield
Measured yield
0
20
40
60
80
100
● Original data set
■ OOS predictions
y = 1.1 + 0.96x
r2 = 0.96
C. Features in the MLR model
Normalized multiple linear regression model:
y = 0.24 + 0.82(θ) - 0.67(%Vbur) - 0.084(Vmin) - 0.534(θ)(Vmin)
cone angle θ
fixed distance
P
M
radius r
%Vbur = % volume of sphere of radius r occupied by ligand
θ = angle swept by cone that encloses all ligand groups
high yield
remote steric hindrance = high θ and low %Vbur

Figure 1. Study of ligand dependence on a Ni-catalyzed Suzuki−Miyaura cross-coupling. (A) Cross-coupling reaction between benzaldehyde dimethyl acetal and *para*-fluorophenyl boroxine. (B) MLR reactivity model. (C) Definition of the cone angle and buried volume features.

performance in a Pd-catalyzed aryl amination. Featurization provided mechanistic insight.

- Estrada, J. G.; Ahneman, D. T.; Sheridan, R. P.; Dreher, S. D.; Doyle, A. G. Response to Comment on "Predicting Reaction Performance in C-N Cross-Coupling Using Machine Learning". *Science* **2018**, *362*, eaat8763.[3] *Technical analysis of the previously published random forest model and its ability to predict reactivity for unseen molecules.*
- Nielsen, M. K.; Ahneman, D. T.; Riera, O.; Doyle, A. G. Deoxyfluorination with Sulfonyl Fluorides: Navigating Reaction Space with Machine Learning. *J. Am. Chem. Soc.* **2018**, *140*, 5004−5008.[4] *A random forest algorithm was applied to a HTE data set (740 reactions) to predict optimal deoxyfluorination conditions with respect to the base and sulfonyl fluoride.*

■ INTRODUCTION

In chemistry, the origins of data-driven modeling can be traced back to the Hammett equation (1937)[5] and the development of quantitative structure−activity and property relationships[6−8] (QSAR/QSPR) that it inspired. The Hammett equation is a linear free energy relationship (LFER) that relates chemical structure, originally represented by quantitative experimental parameters or descriptors, to reactivity. Subsequent developments included the use of multivariate as well as univariate relationships and the incorporation of computationally generated descriptors. While linear regression has been the dominant algorithm used to relate structure to activity, other machine learning methods, like clustering analysis (1972),[9,10] neural networks (1973),[11] and random forests (2003)[12] have also been explored shortly after their development. Nevertheless, the broad adoption of these more complicated algorithms by the synthetic community has been limited. With the growing availability of molecular properties data sets[13,14] and reaction data sets,[15] and improved access to computing power, ML techniques are receiving renewed attention with applications in retrosynthesis planning and reactivity prediction.[16−23]

In the field of organic synthesis, one must know not only the forward sequence of steps to construct a molecule but also the reaction conditions to execute each step with high yield and selectivity. Because reaction space is highly dimensional, identification of these conditions is both time- and resource-consuming. More importantly, understanding the mechanistic origins of chemical interactions in highly dimensional reaction data (e.g., substrate and catalyst structure−reactivity relationships) is challenging even for highly experienced chemists. Thus, as a group, we became interested in using ML for predicting and understanding reaction outcomes in multidimensional space. Combined with the ongoing developments in the field, we anticipate that these efforts will afford useful tools that can guide the generation and use of data for reaction discovery, reaction optimization, and the synthesis of complex molecules.

In the following sections, we present how data was collected for our previous studies and how unsupervised ML may be used to reduce bias in the generation of new reaction data sets. We subsequently discuss the modeling results from three of our supervised ML studies, augmenting them with analyses not originally reported. We specifically focus on model selection and comparisons, model training, and chemical interpretability of the resulting models.

■ EXPERIMENTAL DATA

Experimental reaction data can either be data-mined from publicly available databases[15,24] or generated *de novo*. A common challenge with the first approach is that the published data can be incomplete or inconsistently reported.[25] Importantly, available reaction data often explore a narrow scope of conditions and are biased toward positive results, making it more appropriate for tasks such as automated retrosynthesis. Nevertheless, we are working in collaboration with MIT, Google, Merck, and Pfizer to address this limitation with the development of an open access reaction database.[26] On the other hand, *de novo* generation of data sets allows for custom exploration of the chemical space, but this approach is limited by experimental

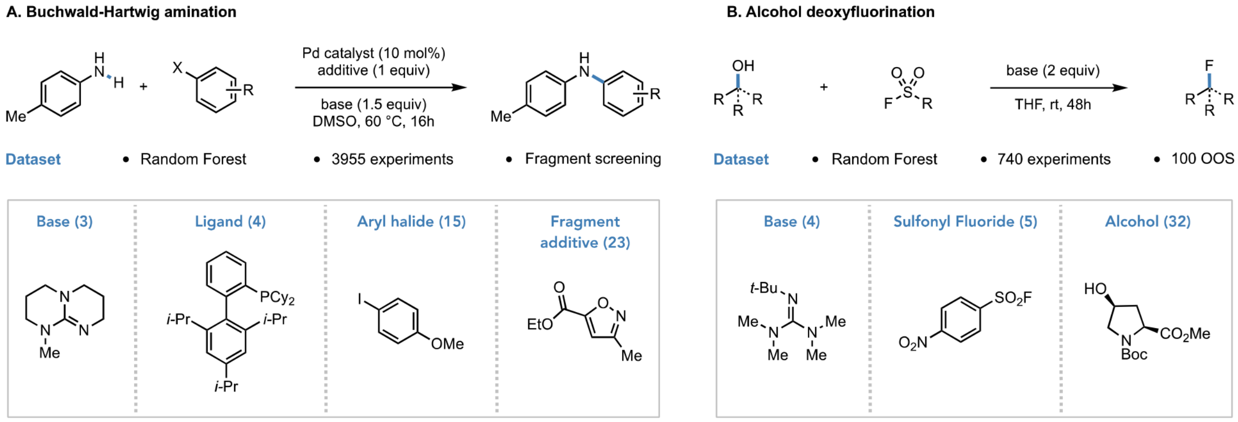


Figure 2. (A) Buchwald–Hartwig amination data set. (B) Alcohol deoxyfluorination data set. A representative example from each reaction component is shown in each data set.

capacity. In our studies, *de novo* generated data has consisted of a chemist selecting substrates and conditions for a given reaction based on existing data, mechanistic understanding, and previous experience, much in the same way that reaction optimization and scope delineation is often conducted for a new method. This can result in certain reactions being statistically overleveraged, that is, having disproportionately large influence on model parameters, if they happen to be isolated in the chemical space.[27] A systematic approach to data set design using unsupervised machine learning, such as clustering based on specific featurization, offers an interesting alternative and is already actively being explored by us and others.[28−32]

FEATURE ENGINEERING

Molecule representation for supervised ML can be broadly categorized into nonlearned and learned representations.[33] The former encompasses deriving feature vectors from computable descriptors, which have been the workhorse for QSAR/QSPR studies.[34] Efforts to obtain learned representations starting from a simple input, such as a SMILES string or atom coordinates, and transforming them using an algorithm are well under way. Notwithstanding, the published learned representation approaches for predicting properties of individual molecules[35,36] or reactivity[37] have been restricted to the use of large data sets.

When experimentally generating a custom data set for a specific transformation, the data set size accessible is insufficient for a learned representation approach. Hence, we have used density functional theory (DFT) computations to obtain compact and comprehensible featurizations of substrates, catalysts, and reagents, to facilitate finding reactivity relationships in our data sets. A drawback of DFT-based featurization is the high computational cost that depends on the level of theory and basis set selection, as well as the size and flexibility of the reaction components. Other static representations such as molecular fingerprints[38,39] or Mordred descriptors[40] are available as alternatives. They reduce computational cost, potentially at the expense of interpretability and applicability in small- to medium-size data sets owing to the large quantity of features involved. A limitation of all nonlearned representations is the supposition that the selected descriptors are sufficient to model the data, a choice that often requires knowledge of chemistry or a mechanism *a priori*.

As an example, in our group's study of Ni-catalyzed Suzuki–Miyaura cross-coupling of benzaldehyde-derived acetals with aryl boroxines[1] (see Figure 1), we found that selecting relevant features was crucial to the success of the modeling effort. Specifically, we were unable to obtain a good multiple linear regression (MLR) model using only electronic or steric features of the phosphine ligand. Initially, 13 phosphine ligands were screened, including two new phosphines designed specifically for Ni: the DinoPhos ligands TriceraPhos and TyrannoPhos. We observed that the Tolman cone angle (θ),[41] a steric feature, was insufficient to explain the variance in reactivity on its own. Through feature engineering, we found that addition of a second steric feature, percent buried volume ($\%V_{\text{bur}}$),[42] together with an electronic feature, the minimum electrostatic potential (V_{min}), resulted in an improved model. The model was then successfully validated with three out-of-sample (OOS) ligands not included in model development.

Our featurization effort allowed us to contextualize the DinoPhos ligands within a new steric regime, which we refer to as remote steric effects. In this regime, there is minimal crowding near the metal center but high crowding at the periphery. This type of modeling pipeline, which involves computationally generated and interpretable features followed by fitting a MLR model, had been previously utilized to model enantioselectivity[43−45] as well as catalytic activity[46] by other groups. Our study extends this pipeline to model reaction yield. Unlike modeling enantioselectivity or catalytic activity, which directly measure energy differences, there are multiple factors that contribute to reaction yield. This may make the prediction task more challenging. However, since yield measurements are more accessible and prevalent, we believe that this type of modeling is a worthwhile task. For example, because reaction yield is bound between 0% and 100%, the MLR model may make unphysical predictions, such as <0% or >100% yield, owing to an incorrect assumption on the normality of the model residuals. However, when data is not heavily skewed toward a single boundary, as is the case in this data set, the resulting bias is typically negligible.

MODELING HIGH DIMENSIONAL HTE DATA SETS

We then sought to model multicomponent reaction data sets, that is, data sets where multiple reagents and reactants are varied, with ML. We designed two data sets focused on reactions of value for the synthesis of bioactive compounds that were amenable to HTE.[47] The data sets included a Buchwald–Hartwig (BH) amination data set[2] generated in collaboration with colleagues at Merck Research Laboratories comprising 3955 reactions that span four reaction components and a deoxyfluorination data set[4] with 740 reactions spanning three

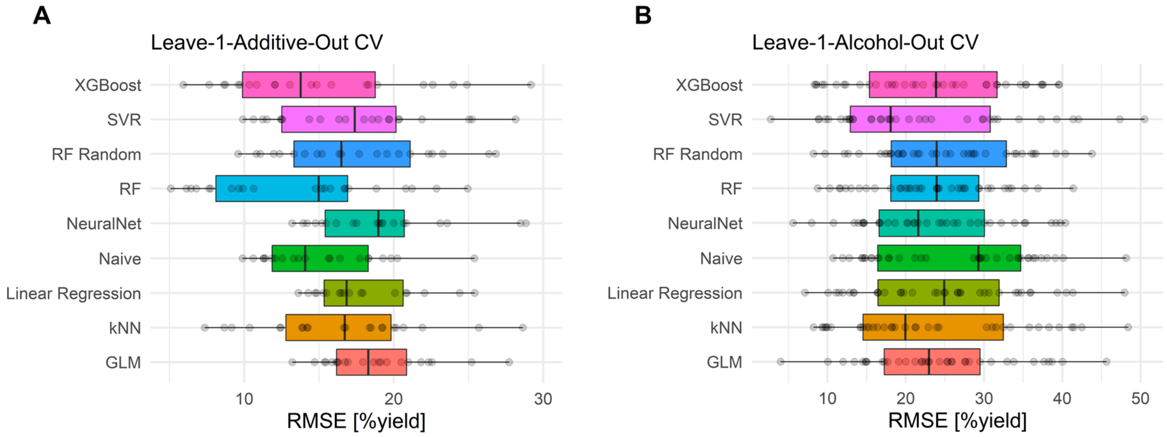


Figure 3. (A) Leave-one-molecule out cross-validation RMSE for the BH amination data set. (B) Leave-one-molecule out cross-validation RMSE for the deoxyfluorination data set. Each point represents validation RMSE for a single additive or alcohol. The bars show 25%−75% interquantile ranges of the RMSE distributions.

reaction components, Figure 2. For the BH amination, we adapted the Glorius fragment additive screening approach[48] by evaluating the effects of isoxazole additives on different aryl and heteroaryl halide couplings rather than coupling substrates bearing isoxazole functionality directly. Both data sets were built by exhaustively evaluating the combinatorial reaction space defined by the components in each reagent category.

With these larger data sets, we set to explore the accuracy of ML modeling in predicting reactivity for out-of-sample reagents and determine whether the trained ML models could guide mechanistic interpretation via their ability to select individual features. Since multiple reaction components are varied simultaneously in these data sets, to represent the reactions, we concatenated individual feature vectors for each component. The BH amination reactions were described with 120 DFT-based numeric features, while the deoxyfluorination reactions had 7 DFT-based numeric features and 16 binary features. Recently, similar modeling studies on multicomponent reaction data sets have also been performed to predict enantioselectivity,[29] regioselectivity,[49] and reactivity.[50]

Modeling HTE data sets is a special case of multi-input ML, where features of substrates and conditions are combined to make a yield prediction. An analogy, although imperfect, could be made to rainfall prediction (yield prediction) in various geographic locations (substrates) under different conditions: time of year, humidity, and cloud coverage (base, solvent, and catalyst). The predictions for nearby locations (similar substrates) might be closely related, but they do not need to be if, for example, a mountain range separates them (activity cliff[51]). The same conditions can also have a different effect on rainfall depending on the location. Therefore, an HTE reaction data set requires a model that can efficiently approximate reactivity with smooth surfaces and divide the chemical space if such separations are justified by data. Therefore, we included several universal function approximation algorithms, such as random forests or neural networks, as candidates to model these data sets.

■ MODEL SELECTION

To select an appropriate ML model, candidate models are developed (trained) on training data and their generalization error is measured on test data. Then, the model with the best estimated generalization error is typically chosen. A common approach is to split the data randomly to generate the training and test sets.[52] This approach was taken in our previously published studies. However, with a random split of an HTE data set, the training set sees every compound and its performance multiple times while only leaving specific combinations of reaction components to the test set. Such an estimate is therefore an optimistic estimate of how the model would perform on unseen (out-of-sample) molecules. If the goal is to predict the yield for an unseen catalyst or to select the highest yielding set of reaction conditions for a new substrate, a more use-inspired test of generalization is valuable.[53] Thus, building on a valuable exchange with Chuang and Keiser,[54] we have turned to a different approach to estimate generalization error with leave-one-molecule out cross-validation, and advocate that the community do so as well because it is more representative of a synthetic chemists' use. *In the following sections, we re-evaluate model selection using the leave-one-molecule out cross-validation for both HTE reaction data sets as follows*:

- We designated the additive in the BH amination data set and alcohol in the deoxyfluorination data set as the reaction component from which to leave a single molecule out. For example, when an isoxazole is left out from the BH amination data set, all 180 reactions that use that additive comprise the validation set. Our data sets contain 22 isoxazole additives and 37 alcohols, creating 22 and 37 validation folds, respectively. In our modeling experience, special care needs to be taken to build the model for each fold independently from models for other folds to simulate real-life scenarios. If model independence across folds is not met, for example, features are fixed using the entire data set, the cross-validation results are optimistically biased and do not correspond to the generalization ability of the model.[55] Therefore, the fold models generally vary in their feature selections and how they model the reactivity surface. This variance reflects the stability of the overall modeling effort with respect to small changes in training data, that is, substituting an additive or alcohol with another one, which in turn impacts model generalizability.

- ML algorithms under consideration are linear regression, GLM (generalized linear model), SVR (support vector regression), kNN (k-nearest-neighbors), RF (random

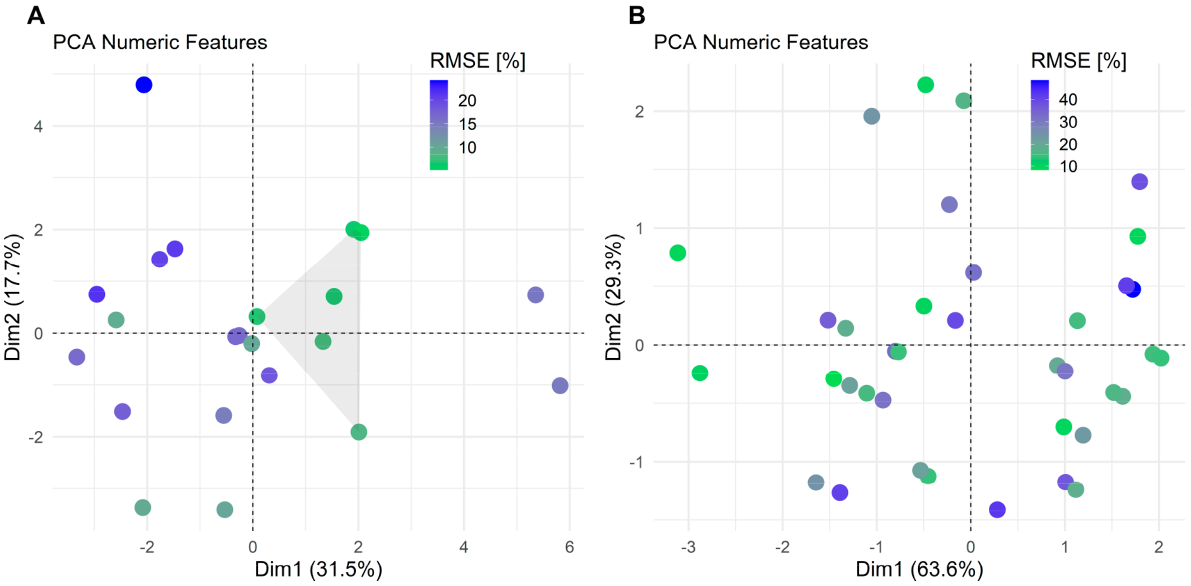


Figure 4. Leave-one-molecule out cross-validation RMSE with respect to molecular featurization. The explained variance in each principal component is shown in parentheses. (A) BH amination data set additives. A region where the RMSE is below 8% is highlighted. (B) Alcohols in the deoxyfluorination data set.

forest), XGBoost, and a feed forward NN (neural network).

- Two baseline models are also included in the comparison. Since ML models always provide a numeric performance estimate, the models can only be judged by comparing their generalizability to other models. For our reaction data sets, we built the following baseline models:
    - ML model with a RF algorithm but with random features that carry no physical information other than labeling the molecules.[3]
    - Non-ML naive model where a prediction for a reaction yield is made by averaging yields of all training set reactions that share the same substrates and conditions, for example, when predicting a deoxyfluorination reaction yield for a new alcohol, but with a base and sulfonyl fluoride that has been used for a range of alcohols, we average all reaction yields for the same base and sulfonyl fluoride and all alcohols from the training set.

Model hyperparameters were optimized using nested cross-validation[56] (See SI Table S1 for details), except for the naive and linear regression models that have no hyperparameters. We further analyze the leave-one-molecule out cross-validation results, Figure 3, with statistical tests:[57]

- Models perform differently on the BH amination data (ANOVA p-value $< 5 \times 10^{-5}$), while all models perform similarly on the deoxyfluorination data (ANOVA p-value of 0.8) indicating that for deoxyfluorination the generalization error does not significantly depend on the model among the 9 candidates considered.
- For BH amination only the RF model performs significantly better than either of the two baseline models (paired Wilcoxon test p-values < 0.01).

These results indicate that the RF is the best model for the BH amination data, consistent with our original study. For the deoxyfluorination data set, where we did not evaluate other models besides RF in the original study, there is no best model.

The prediction accuracy varies substantially between the test molecules for both reaction data sets. For the RF model, the root mean square error (RMSE) is between 5% and 25% for the additives from the BH amination data set and between 9% and 41% for the alcohols from the deoxyfluorination data set. This indicates that the yield cannot always be accurately predicted for an arbitrary new molecule, prompting us to consider how useful these models are for out-of-sample prediction and how can they be improved. One hypothesis is that there might be an area of chemical space in which the model works well. In Figure 4, the prediction RMSE for each individual molecule is visualized using the top two principal components of their molecular featurization. For the BH amination additives, we qualitatively identify a central region of feature space where prediction errors are 8% or lower on average, while for the deoxyfluorination alcohol set, no such region can be identified. For the deoxyfluorination model, these observations suggest that either the DFT-based featurization does not capture information that relates structure to activity or the algorithm is not capable of doing so with the available data. In the two data sets, we sampled a comparable number of distinct compounds (22 isoxazole additives for the BH amination and 37 alcohols for the deoxyfluorination) to represent their respective chemical spaces. However, the diversity and size of the alcohol chemical space results in a much sparser alcohol selection as compared to isoxazole additives selection in their chemical space. This likely makes generalization to out-of-sample alcohols more challenging and may explain the limitations of the deoxyfluorination modeling.

■ MODEL TRAINING

While the generalization error is a critical performance metric of a ML model, it is also useful to verify how well the generalization error corresponds to the model's fit to the training data. The proximity of the training fit is controlled with hyperparameters, which are chosen to optimize the generalization error. However, how well this hyperparameter tuning prevents overfitting generally varies from algorithm to algorithm.

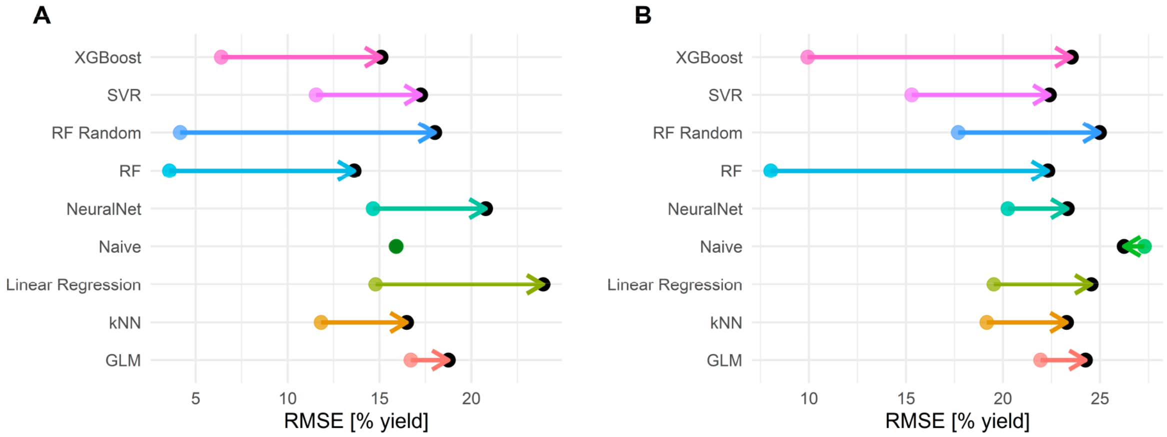


Figure 5. Leave-one-molecule out training and CV error comparison. Color points represent mean training error; black points represent their respective CV mean error. (A) BH amination data set validated over additives. (B) Deoxyfluorination data set validated over alcohols.

As shown in Figure 5, the nested leave-one-molecule out cross-validation did not prevent the models from overfitting. The tree-based models, that is, RF and XGBoost, overfit the most, while structurally simpler models, such GLM and kNN overfit to a lesser extent. The naive model does not overfit; however, no regression is used when obtaining this model's predictions. A large gap between training and validation performance hinders the usage of the trained model surface for mechanistic interpretation.

Our best performing RF algorithm employs tree bagging as one of the means to prevent overfitting, that is, the individual trees are fit to bootstrap samples[58] of the training data. In a reaction data set, however, a bootstrap sample exposes each tree to all training molecules with high probability, although only a subset of reactions. As illustrated in Figure 6, for the deoxyfluorination data set, every alcohol is sampled multiple times in a bootstrap sample. Reaction yields that share the same molecule(s) are inherently correlated; therefore, the resulting trees are expected to be correlated as well, which in turn enhances overfitting.

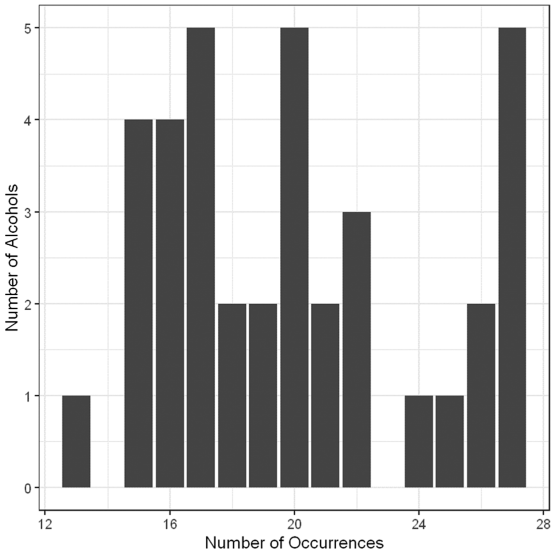


Figure 6. Number of occurrences of alcohols in a bootstrap sample taken from the deoxyfluorination data set. A maximum number of distinct reactions for each alcohol is 20 if every base and sulfonyl fluoride combination is present.

An alternative formulation in which bootstrap samples are taken over the sets of molecules can decorrelate the trees, though no such implementation exists to date. Given that HTE reaction data sets pose unique challenges to ML algorithms, as illustrated in the above example, we believe researching new or modifying existing algorithms to better model these data sets is of high priority.

■ MODEL INTERPRETATION

For the BH amination data set, the prediction to an out-of-sample additive using the DFT based featurization and the RF model leads to a 2−6% (95% confidence interval) improvement over the baseline RF model with random features. While an improvement of this magnitude will not make the model more useful for prediction, it is attributable to DFT-based featurization and can be interpreted mechanistically. While the RF prediction mechanism is difficult to understand, feature metrics, such as Gini importance or permutation importance, are often used as indirect measures of feature importance.[59] In our case, to study the importance of a molecular feature or a set of features, we replace them with random numbers (noise them up), to ensure that they are ineffective and repeat the cross-validation with models trained on partially noised up feature sets. From a chemists' perspective, we are interested in whether feature interactions across different reaction components play an important role. Thus, we noise up all features from a single reaction component, while keeping features of all other components unchanged. The resulting RMSE increases are presented in Figure 7A. Only noising up features of the additive results in a significant drop in model performance, implying that modeling interactions between components is not augmented by DFT features as compared to random features in this data set.

We further noise up each of the additive features and measure the increase of the cross-validation performance (see Figure 7B). Individual feature's contributions are small and mostly insignificant; however, noising up a single feature does not extract its total importance. This is due to correlations with other features within the same reaction component. Among the additive features, those describing the C3 atom, which in our featurization is the carbon next to the nitrogen in the isoxazole ring, achieve highest importance, particularly the C3 NMR shift. We use a partial dependence plot[60] to visualize marginal dependence of the model, trained on all data, on the C3 NMR

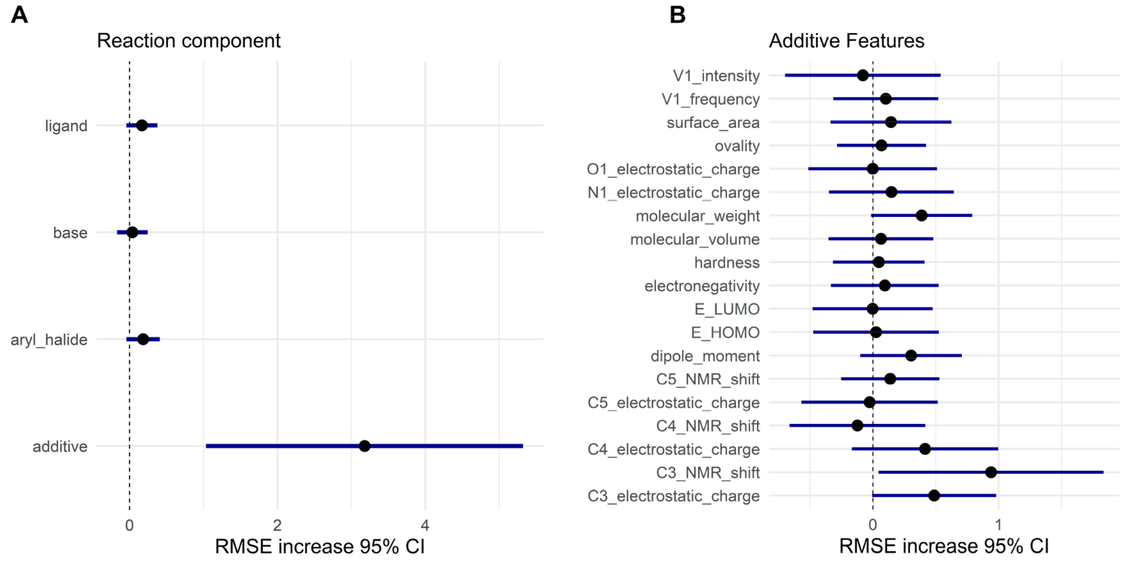
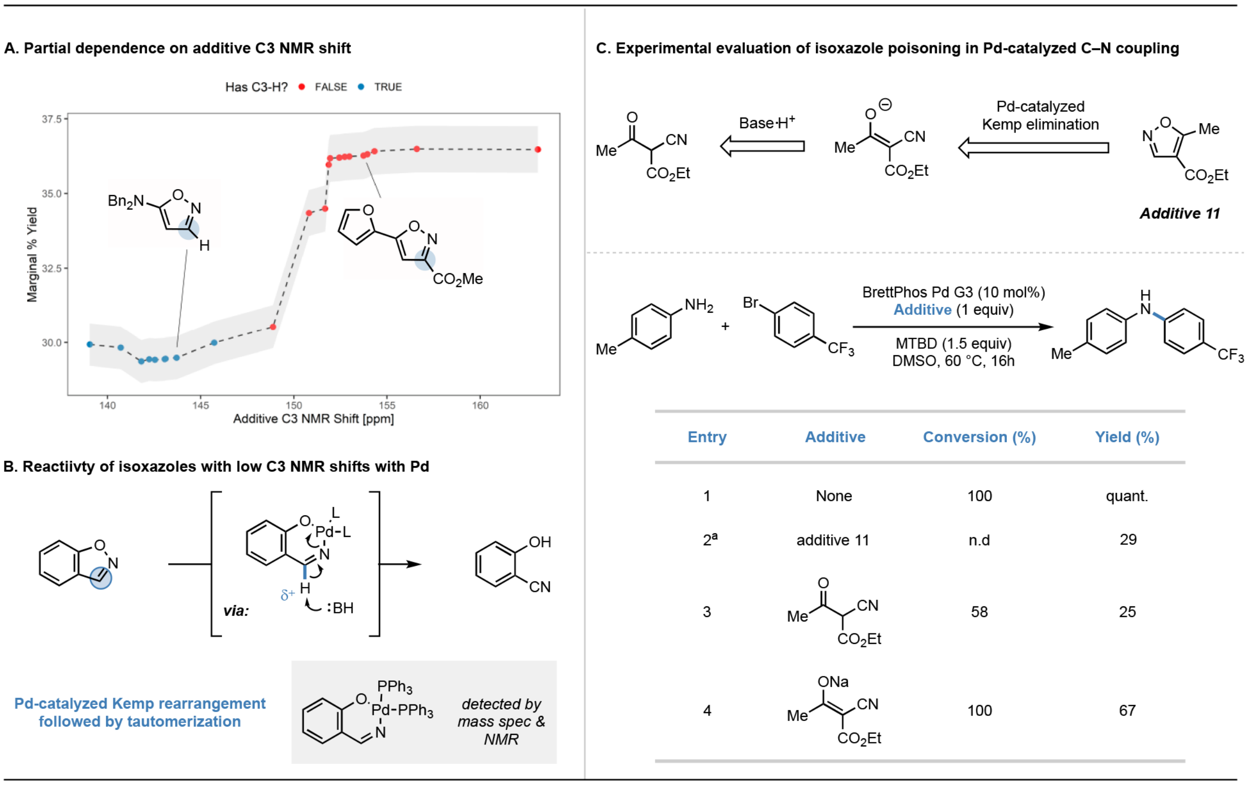


Figure 7. Leave-one-molecule out cross-validation RMSE increase resulting from noising up features. (A) All features of a single reaction component. (B) Individual features of the additives. The bands are 95% CI of the paired Wilcoxon test with respect to the model with original features.


A. Partial dependence on additive C3 NMR shift
Has C3-H? ● FALSE ● TRUE
37.5
35.0
32.5
30.0
Marginal % Yield
140
145
150
155
160
Additive C3 NMR Shift [ppm]
Bn₂N
O
N
H
O
O
N
CO₂Me
B. Reactiivty of isoxazoles with low C3 NMR shifts with Pd
O
N
O
L
Pd-L
N
δ+
H
:BH
via:
OH
CN
Pd-catalyzed Kemp rearrangement followed by tautomerization
PPh₃
O
Pd-PPh₃
N
detected by mass spec & NMR
C. Experimental evaluation of isoxazole poisoning in Pd-catalyzed C–N coupling
O
Me
CN
CO₂Et
Base·H⁺
O⊖
Me
CN
CO₂Et
Pd-catalyzed Kemp elimination
N
O
Me
CO₂Et
Additive 11
Me
NH₂
+
Br
CF₃
BrettPhos Pd G3 (10 mol%)
Additive (1 equiv)
MTBD (1.5 equiv)
DMSO, 60 °C, 16h
Me
H
N
CF₃

Entry	Additive	Conversion (%)	Yield (%)
1	None	100	quant.
2[a]	additive 11	n.d	29
3	O Me CN CO₂Et	58	25
4	ONa Me CN CO₂Et	100	67

Figure 8. (A) Partial dependence of the BH amination activity model as a function of additive C3 NMR shift obtained with out-of-bag portion of the random forest training data. Individual points represent individual additives. The shaded band corresponds to ±2 standard error on the prediction. (B) Kemp-type rearrangement in the presence of Pd for an example isoxazole that has a C3−H bond. C3 position is indicated with a shadowed circle above. (C) [a]*t*-Bu-BrettPhos was used as the ligand. Reaction poisons were used to identify the source of detrimental reactivity. Note, all ligands studied resulted in lower than 29% yield.

shift feature. In Figure 8, a step-like increase in yield is observed for additives with C3 NMR shift >150 ppm.

Looking at the chemical structures, it is evident that additives with C3 NMR shift <150 ppm primarily have a C3−H bond and additives with C3 NMR shift >150 ppm have a fully substituted C3. The fact that reaction poisons tend to have a C3−H bond led us to propose and investigate a mechanism for isoxazole decomposition via C3−H deprotonation. Namely, isoxazoles bearing a C3−H could undergo a Pd-catalyzed Kemp-type rearrangement to form α-cyano ketones and aldehydes after N−O oxidative addition (Figure 8b). In our previous study, we showed that in the absence of palladium, no isoxazole rearrangement was observed, even upon heating. Importantly via mass spectrometry and NMR analysis, we were able to identify an oxidative adduct between palladium and benzo[*d*]-isoxazole, suggesting that the adduct may be responsible for the isomerization process depicted in Figure 8. While it is unclear whether solely the oxidative adduct or the subsequent rearranged product results in a severe poison for the BH amination, we found that the 3-ketobutyric methyl ester derivative that would result from the Pd-catalyzed rearrangement of additive **11** severely inhibits reactivity (Figure 8c, entries 1−4). It is noteworthy that the partial dependence plot did not explicitly suggest that the rearranged product was the reaction poison. Therefore, while these observations may have been identified independent of modeling, this analysis highlights how ML can be used as a tool for developing an experimentally testable hypothesis to gain some mechanistic understanding of chemical processes.

CONCLUSIONS

ML is a rapidly developing field of research and its potential in reactivity prediction is being systematically explored. In the examples studied in our lab, we observe that ML can achieve quite similar generalization accuracy using physics-agnostic features as with quantitative physical features. In the BH data set, we see statistically significant improvements in generalization with DFT features indicating that it provides transferable chemical insight and allowed us to learn about the underlying mechanism. Nevertheless, in both the BH and deoxyfluorination data sets, the generalization error as measured by more rigorous and realistic tests than previously reported suggests that improvement is still needed in increasing the reliability and efficiency of ML tools in chemical over-the-arrow prediction. We expect that progress will require improvements in areas such as featurization and algorithm development.

Given the statistical aspect of ML modeling, we advocate that results need to be clearly communicated with a focus on generalization error comparisons to baseline models or non-ML models. Model validations and hyperparameter tuning should be performed with use-case motivated resampling, such as leaving molecules out. Finally, model hyperparameters should always be reported.

We hope that this work encourages others to research ML techniques and evaluate their application to organic synthesis problems. With advances in HTE and increased access to high-quality data, we are optimistic that predictive and interpretable reactivity models could become common place in many aspects of organic synthesis.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.accounts.0c00770.

Optimized hyperparameter table (PDF)

AUTHOR INFORMATION

Corresponding Author

Abigail G. Doyle − *Department of Chemistry, Princeton University, Princeton, New Jersey 08544, United States;* orcid.org/0000-0002-6641-0833; Email: agdoyle@princeton.edu

Authors

Andrzej M. Żurański − *Department of Chemistry, Princeton University, Princeton, New Jersey 08544, United States*

Jesus I. Martinez Alvarado − *Department of Chemistry, Princeton University, Princeton, New Jersey 08544, United States*

Benjamin J. Shields − *Department of Chemistry, Princeton University, Princeton, New Jersey 08544, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.accounts.0c00770

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

Biographies

Andrzej M. Żurański is a Data Science Fellow in the Chemistry Department at Princeton University. He obtained his Ph.D. in physics at Princeton University in 2014. He joined the Chemistry Department in 2019 and is a member of NSF Center for Computer Assisted Synthesis (C-CAS). He is interested in exploring and developing machine learning methods with applications in organic synthesis.

Jesus I. Martinez Alvarado received a B.S. in chemistry from the University of North Carolina at Chapel Hill in 2016, where he conducted research in the lab of Prof. Jeffrey S. Johnson. He is currently a graduate student in the Doyle lab at Princeton University and is interested in incorporating data science techniques broadly in organic chemistry.

Benjamin J. Shields is a postdoctoral researcher in chemistry and computer science at Princeton University. He studied chemistry and mathematics at the University of North Carolina, Asheville, and obtained his masters and Ph.D. in chemistry from Princeton University. Ben is interested in chemical synthesis, photophysics, computer-aided drug design, and machine learning.

Abigail G. Doyle is the A. Barton Hepburn Professor of Chemistry in the Chemistry Department at Princeton University. She obtained her Ph.D. in catalysis and physical organic chemistry at Harvard University in 2008 under the direction of Prof. Eric Jacobsen after receiving her A.B. and A.M. in chemistry and chemical biology from Harvard in 2002. She joined the faculty at Princeton University in 2008 and is currently a co-PI for the NSF CCI Center for Computer Assisted Synthesis (C-CAS) and the DOE EFRC Bioinspired Light-Escalated Chemistry (BioLEC). The Doyle laboratory is interested in developing new approaches to chemical synthesis and catalysis, with a focus on Ni-catalyzed cross coupling and nucleophilic fluorination methodology.

ACKNOWLEDGMENTS

We are grateful to the past and present members of our research group for their contributions to this work and to our collaborators from Merck for their intellectual contributions and access to HTE resources. This program has been funded through the generous contributions of the National Science Foundation (CHE-1565983) and under the CCI Center for Computer Assisted Synthesis (CHE-1925607). We gratefully acknowledge financial support from the Schmidt DataX Fund at Princeton University made possible through a major gift from the Schmidt Futures Foundation.

REFERENCES

(1) Wu, K.; Doyle, A. G. Parameterization of Phosphine Ligands Demonstrates Enhancement of Nickel Catalysis via Remote Steric Effects. *Nat. Chem.* **2017**, *9* (8), 779−784.

(2) 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* (6385), 186−190.

(3) Estrada, J. G.; Ahneman, D. T.; Sheridan, R. P.; Dreher, S. D.; Doyle, A. G. Response to Comment on "Predicting Reaction Performance in C-N Cross-Coupling Using Machine Learning". *Science* **2018**, *362* (6416), No. eaat8763.

(4) Nielsen, M. K.; Ahneman, D. T.; Riera, O.; Doyle, A. G. Deoxyfluorination with Sulfonyl Fluorides: Navigating Reaction Space with Machine Learning. *J. Am. Chem. Soc.* **2018**, *140* (15), 5004−5008.

(5) Hammett, L. P. The Effect of Structure upon the Reactions of Organic Compounds. Benzene Derivatives. *J. Am. Chem. Soc.* **1937**, *59* (1), 96−103.

(6) Hansch, C.; Fujita, T. P -*σ*-*π* Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc.* **1964**, *86* (8), 1616−1626.

(7) Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, I. I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; Consonni, V.; Kuz'min, V. E.; Cramer, R.; Benigni, R.; Yang, C.; Rathman, J.; Terfloth, L.; Gasteiger, J.; Richard, A.; Tropsha, A. QSAR Modeling: Where Have You Been? Where Are You Going To? *J. Med. Chem.* **2014**, *57* (12), 4977−5010.

(8) Muratov, E. N.; Bajorath, J.; Sheridan, R. P.; Tetko, I. V.; Filimonov, D.; Poroikov, V.; Oprea, T. I.; Baskin, I. I.; Varnek, A.; Roitberg, A.; Isayev, O.; Curtalolo, S.; Fourches, D.; Cohen, Y.; Aspuru-Guzik, A.; Winkler, D. A.; Agrafiotis, D.; Cherkasov, A.; Tropsha, A. QSAR without Borders. *Chem. Soc. Rev.* **2020**, *49* (11), 3525−3564.

(9) Hansch, C.; Unger, S. H.; Forsythe, A. B. Strategy in Drug Design. Cluster Anlysis as an Aid in the Selection of Substituents. *J. Med. Chem.* **1973**, *16* (11), 1217−1222.

(10) Kowalski, B. R.; Bender, C. F. Pattern Recognition. Powerful Approach to Interpreting Chemical Data. *J. Am. Chem. Soc.* **1972**, *94* (16), 5632−5639.

(11) Hiller, S. A.; Golender, V. E.; Rosenblit, A. B.; Rastrigin, L. A.; Glaz, A. B. Cybernetic Methods of Drug Design. I. Statement of the Problem—The Perceptron Approach. *Comput. Biomed. Res.* **1973**, *6* (5), 411−421.

(12) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comp. Sci.* **2003**, *43* (6), 1947−1958.

(13) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum Chemistry Structures and Properties of 134 Kilo Molecules. *Sci. Data* **2014**, *1* (1), 140022.

(14) Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2012**, *52* (11), 2864−2875.

(15) Lowe, D. M. Extraction of Chemical Structures and Reactions from the Literature, Doctoral thesis, University of Cambridge, 2012, DOI: DOI: 10.17863/CAM.16293.

(16) Wei, J. N.; Duvenaud, D.; Aspuru-Guzik, A. Neural Networks for the Prediction of Organic Chemistry Reactions. *ACS Cent. Sci.* **2016**, *2* (10), 725−732.

(17) Coley, C. W.; Green, W. H.; Jensen, K. F. Machine Learning in Computer-Aided Synthesis Planning. *Acc. Chem. Res.* **2018**, *51* (5), 1281−1289.

(18) de Almeida, A. F.; Moreira, R.; Rodrigues, T. Synthetic Organic Chemistry Driven by Artificial Intelligence. *Nat. Rev. Chem.* **2019**, *3* (10), 589−604.

(19) Davies, I. W. The Digitization of Organic Synthesis. *Nature* **2019**, *570* (7760), 175−181.

(20) Haywood, A. L.; Redshaw, J.; Gaertner, T.; Taylor, A.; Mason, A. M.; Hirst, J. D. Machine Learning for Chemical Synthesis. *Machine Learning in Chemistry: The Impact of Artificial Intelligence*; Royal Society of Chemistry, 2020; pp 169−194, .

(21) Coley, C. W.; Eyke, N. S.; Jensen, K. F. Autonomous Discovery in the Chemical Sciences Part I: Progress. *Angew. Chem., Int. Ed.* **2020**, *59* (51), 22858−22893.

(22) Coley, C. W.; Eyke, N. S.; Jensen, K. F. Autonomous Discovery in the Chemical Sciences Part II: Outlook. *Angew. Chem., Int. Ed.* **2020**, *59* (52), 23414−23436.

(23) Shen, Y.; Borowski, J. E.; Hardy, M. A.; Sarpong, R.; Doyle, A. G.; Cernak, T. Automation and Computer-Assisted Planning for Chemical Synthesis. *Nat. Rev. Methods Primers* **2021**, *1* (1), 23.

(24) Reaxys. https://new.reaxys.com/.

(25) Tetko, I. V.; Engkvist, O.; Koch, U.; Reymond, J.; Chen, H. BIGCHEM: Challenges and Opportunities for Big Data Analysis in Chemistry. *Mol. Inf.* **2016**, *35* (11−12), 615−621.

(26) Open Reaction Database. https://docs.open-reaction-database.org.

(27) Jia, X.; Lynch, A.; Huang, Y.; Danielson, M.; Lang'at, I.; Milder, A.; Ruby, A. E.; Wang, H.; Friedler, S. A.; Norquist, A. J.; Schrier, J. Anthropogenic Biases in Chemical Reaction Data Hinder Exploratory Inorganic Synthesis. *Nature* **2019**, *573* (7773), 251−255.

(28) Martin, T. M.; Harten, P.; Young, D. M.; Muratov, E. N.; Golbraikh, A.; Zhu, H.; Tropsha, A. Does Rational Selection of Training and Test Sets Improve the Outcome of QSAR Modeling? *J. Chem. Inf. Model.* **2012**, *52* (10), 2570−2578.

(29) 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* (26), 11578−11592.

(30) Shields, B. J.; Stevens, J.; Li, J.; Parasram, M.; Damani, F.; Alvarado, J. I. M.; Janey, J. M.; Adams, R. P.; Doyle, A. G. Bayesian Reaction Optimization as a Tool for Chemical Synthesis. *Nature* **2021**, *590* (7844), 89−96.

(31) Christensen, M.; Yunker, L.; Adedeji, F.; Häse, F.; Roch, L.; Gensch, T.; Gomes dos, G. P.; Zepel, T.; Sigman, M.; Aspuru-Guzik, A.; Hein, J. Data-Science Driven Autonomous Process Optimization. *ChemRxiv* **2020**, DOI: 10.26434/chemrxiv.13146404.v1.

(32) Reymond, J.; Ruddigkeit, L.; Blum, L.; van Deursen, R. The Enumeration of Chemical Space. *WIREs Comput. Mol. Sci.* **2012**, *2* (5), 717−733.

(33) Staker, J.; Marques, G.; Dakka, J. Machine Learning in Chemistry: The Impact of Artificial Intelligence. *Theor. Comput. Chem. Ser.* **2020**, 372−397.

(34) Dehmer, M.; Varmuza, K.; Bonchev, D. Statistical Modelling of Molecular Descriptors in QSAR/QSPR. *Quantitative and Network Biology* **2018**, DOI: 10.1002/9783527645121.

(35) Wieder, O.; Kohlbacher, S.; Kuenemann, M.; Garon, A.; Ducrot, P.; Seidel, T.; Langer, T. A Compact Review of Molecular Property Prediction with Graph Neural Networks. *Drug Discovery Today: Technol.* **2020**, DOI: 10.1016/j.ddtec.2020.11.009.

(36) Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K.-R. SchNet - A Deep Learning Architecture for Molecules and Materials. *J. Chem. Phys.* **2018**, *148* (24), 241722.

(37) Gao, H.; Struble, T. J.; Coley, C. W.; 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.

(38) Senese, C. L.; Duca, J.; Pan, D.; Hopfinger, A. J.; Tseng, Y. J. 4D-Fingerprints, Universal QSAR and QSPR Descriptors. *J. Chem. Inf. Comp. Sci.* **2004**, *44* (5), 1526−1539.

(39) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742−754.

(40) Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; Takagi, T. Mordred: a molecular descriptor calculator. *J. Cheminf.* **2018**, *10*, 4.

(41) Tolman, C. A. Phosphorus Ligand Exchange Equilibriums on Zerovalent Nickel. Dominant Role for Steric Effects. *J. Am. Chem. Soc.* **1970**, *92* (10), 2956−2965.

(42) Clavier, H.; Nolan, S. P. Percent Buried Volume for Phosphine and N-Heterocyclic Carbene Ligands: Steric Properties in Organometallic Chemistry. *Chem. Commun.* **2010**, *46* (6), 841−861.

(43) Rosales, A. R.; Ross, S. P.; Helquist, P.; Norrby, P.-O.; Sigman, M. S.; Wiest, O. Transition State Force Field for the Asymmetric Redox-Relay Heck Reaction. *J. Am. Chem. Soc.* **2020**, *142* (21), 9700−9707.

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

(45) Reid, J. P.; Proctor, R. S. J.; Sigman, M. S.; Phipps, R. J. Predictive Multivariate Linear Regression Analysis Guides Successful Catalytic Enantioselective Minisci Reactions of Diazines. *J. Am. Chem. Soc.* **2019**, *141* (48), 19178−19185.

(46) De Jesus Silva, J.; Ferreira, M. A. B.; Fedorov, A.; Sigman, M. S.; Copéret, C. Molecular-Level Insight in Supported Olefin Metathesis Catalysts by Combining Surface Organometallic Chemistry, High Throughput Experimentation, and Data Analysis. *Chem. Sci.* **2020**, *11* (26), 6717−6723.

(47) Shevlin, M. Practical High-Throughput Experimentation for Chemists. *ACS Med. Chem. Lett.* **2017**, *8* (6), 601−607.

(48) Pitzer, L.; Schäfers, F.; Glorius, F. Rapid Assessment of the Reaction-Condition-Based Sensitivity of Chemical Transformations. *Angew. Chem., Int. Ed.* **2019**, *58* (25), 8572−8576.

(49) Li, X.; Zhang, S.; Xu, L.; Hong, X. Predicting Regioselectivity in Radical C-H Functionalization of Heterocycles through Machine Learning. *Angew. Chem., Int. Ed.* **2020**, *59* (32), 13253−13259.

(50) Sandfort, F.; Strieth-Kalthoff, F.; Kühnemund, M.; Beecks, C.; Glorius, F. A Structure-Based Platform for Predicting Chemical Reactivity. *Chem.* **2020**, *6* (6), 1379−1390.

(51) Maggiora, G. M. On Outliers and Activity CliffsWhy QSAR Often Disappoints. *J. Chem. Inf. Model.* **2006**, *46* (4), 1535−1535.

(52) Hastie, T.; Tibshirani, R.; Friedman, J. *Elements of Statistical Learning, Data Mining, Inference, and Prediction*; Springer, 2009; DOI: 10.1007/978-0-387-84858-7.

(53) Zahrt, A. F.; Henle, J. J.; Denmark, S. E. Cautionary Guidelines for Machine Learning Studies with Combinatorial Datasets. *ACS Comb. Sci.* **2020**, *22*, 586−591.

(54) Chuang, K. V.; Keiser, M. J. Comment on "Predicting Reaction Performance in C-N Cross-Coupling Using Machine Learning.". *Science* **2018**, *362* (6416), No. eaat8603.

(55) Smialowski, P.; Frishman, D.; Kramer, S. Pitfalls of Supervised Feature Selection. *Bioinformatics* **2010**, *26* (3), 440−443.

(56) Stone, M. Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society. Series B (Methodological)* **1974**, *36*, 111−147.

(57) Heumann, C.; Schomaker, M.; Shalabh *Introduction to Statistics and Data Analysis, With Exercises, Solutions and Applications in R*; Springer, 2016; DOI: 10.1007/978-3-319-46162-5.

(58) Efron, B. Bootstrap Methods: Another Look at the Jackknife. *Ann. Statistics* **1979**, *7* (1), 1−26.

(59) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45* (1), 5−32.

(60) Goldstein, A.; Kapelner, A.; Bleich, J.; Pitkin, E. Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation. *J. Comput. Graph. Stat.* **2015**, *24* (1), 44−65.