REPORT

ORGANIC CHEMISTRY

Predicting reaction performance in C–N cross-coupling using machine learning

Derek T. Ahneman,[1] Jesús G. Estrada,[1] Shishi Lin,[2] Spencer D. Dreher,[2]* Abigail G. Doyle[1]*

Machine learning methods are becoming integral to scientific inquiry in numerous disciplines. We demonstrated that machine learning can be used to predict the performance of a synthetic reaction in multidimensional chemical space using data obtained via high-throughput experimentation. We created scripts to compute and extract atomic, molecular, and vibrational descriptors for the components of a palladium-catalyzed Buchwald-Hartwig cross-coupling of aryl halides with 4-methylaniline in the presence of various potentially inhibitory additives. Using these descriptors as inputs and reaction yield as output, we showed that a random forest algorithm provides significantly improved predictive performance over linear regression analysis. The random forest model was also successfully applied to sparse training sets and out-of-sample prediction, suggesting its value in facilitating adoption of synthetic methodology.

Machine learning (ML) is the study and construction of computer algorithms that can learn from data (*1*). The ability of these algorithms to detect meaningful patterns has led to their adoption across a wide range of applications in science and technology, from autonomous vehicle control to recommender systems (*2*). ML has also been successfully applied in the biomedical sciences to enhance the virtual screening of libraries of druglike molecules for biological function (*3*–*5*). However, its application to the chemical sciences, and synthetic organic chemistry in particular, has been limited (*6*, *7*). Prior efforts have focused primarily on using ML to assist with synthetic planning via retrosynthetic pathways or to predict the products of chemical reactions given a set of reactants and conditions (*8*–*11*). Applications of ML to predict the performance of a given reaction, however, are rare. Studies in the area of heterogeneous catalysis have used ML to predict reaction performance when only a single component is varied (*12*, *13*). Two recent studies have advanced the field by evaluating predictions in multidimensional chemical space, although these studies performed a binary classification of reaction success (*14*, *15*). The use of regression-based ML to predict reaction yields in multidimensional chemical space could provide chemists with a powerful tool to navigate the adoption of synthetic methodology.

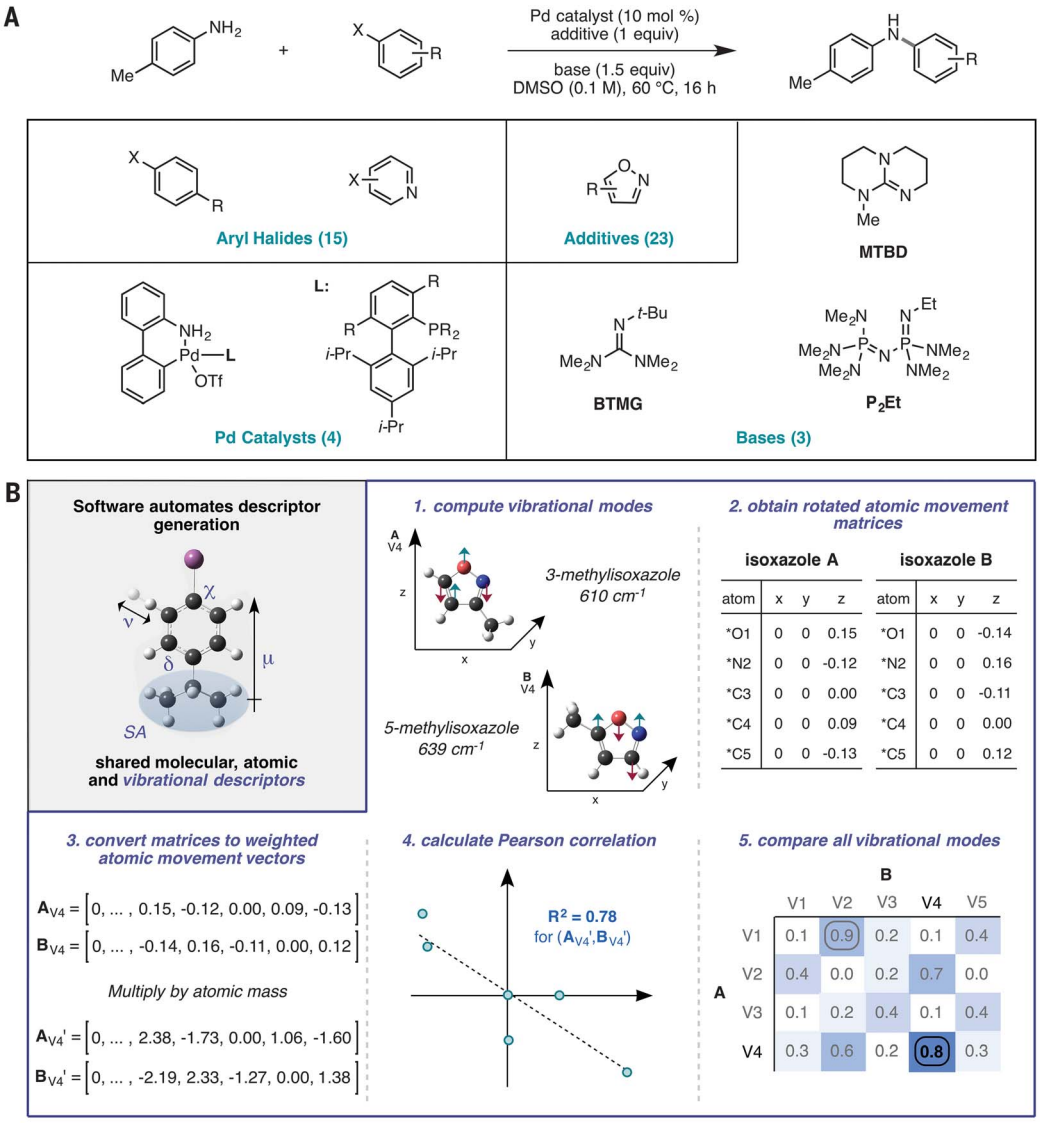
[1]Department of Chemistry, Princeton University, Princeton, NJ 08544, USA. [2]Chemistry Capabilities and Screening, Merck Sharp & Dohme Corporation, Kenilworth, NJ 07033, USA.
*Corresponding author. Email: spencer_dreher@merck.com (S.D.D.); agdoyle@princeton.edu (A.G.D.)

The many challenges in applying ML to reaction performance have previously hindered its use in the field of chemical synthesis. Implementation of these algorithms has historically been complicated for nonspecialists. Further. the amount of data required to obtain statistically meaningful results grows exponentially with the number of dimensions under study, a problem known as the "curse of dimensionality" (*1*). Given the multidimensionality of chemical structure and reactivity, it has been difficult to generate enough data or to get access to sufficiently complete and consistent data from databases to warrant implementation of these algorithms (*14*). Fortunately, over the past decade, high-throughput experimentation (HTE) has emerged as a powerful tool in industry and academia for reaction optimization and discovery (*16*, *17*). We sought to evaluate whether ML could be applied to the scale of data available to modern HTE and enable yield prediction in multidimensional chemical space.

Linear regression is the traditional tool for reaction prediction and analysis in both industry and academia (*18*). In this approach, the user assumes a linear relationship between reaction input (e.g., catalyst descriptors) and output (e.g., product selectivity) and hand-selects input variables on the basis of specific mechanistic hypotheses (*19*, *20*). A strength of linear regression is its interpretability: A good fit between reagent descriptors and output supports mechanistic inferences, such as in the seminal Hammett linear free-energy relationship (*21*).

The models obtained from linear regression analysis have also been used for prediction. Recently, Sigman and co-workers have applied multivariate linear and polynomial regression analyses to optimize reaction selectivity by predicting catalyst, ligand, and substrate effects (*22*–*24*). Predicting yield tends to be more difficult; whereas product selectivity is determined by a small number of elementary steps, many on- and off-cycle events can substantially alter reaction yield. ML approaches accept numerous input descriptors without recourse to a mechanistic hypothesis and evaluate functions with greater flexibility to match patterns in data. We postulated that ML might outperform regression analysis for yield prediction and circumvent the challenge of selecting mechanistically relevant descriptors for large and multidimensional data sets. Here, we report that a random forest ML model trained on multidimensional chemical data can be used to predict the performance of a Buchwald-Hartwig amination reaction conducted in the presence of potentially inhibitory additives and to infer underlying reactivity. We have taken steps to automate reaction parameterization and modeling with the aim of making this tool accessible to the synthetic chemistry community.

We selected the Pd-catalyzed Buchwald-Hartwig reaction as our test reaction for model development because of its broad value in pharmaceutical synthesis (Fig. 1A) (*25*). Nevertheless, the application of this reaction to complex drug-like molecules remains challenging (*26*). One limitation is the poor performance of substrates possessing five-membered heterocycles that contain heteroatom-heteroatom bonds, such as isoxazoles. These heterocycles have drug-like characteristics but are underrepresented in successful drug candidates (*27*). Thus, we sought to use ML to predict the performance of the Buchwald-Hartwig reaction in the presence of isoxazoles. Rather than evaluate the coupling of a collection of substrates directly bearing the heterocycle functionality, we pursued a Glorius fragment additive screening approach (*28*) wherein we evaluated the effects of isoxazole fragment additives on the amination of different aryl and heteroaryl halides. This method cannot always account for the full impact of a structural motif embedded within a substrate. However, the Glorius approach allowed us to test 345 diverse structural interactions between isoxazoles and aryl and heteroaryl halides. This large array would not be possible with whole molecules because of the necessity of synthesizing and isolating all possible products for quantification in this study. We conducted the coupling reactions using the ultra–high-throughput setup recently developed in the Merck Research Laboratories for nanomole-scale experimentation in 1536-well plates (*16*). Use of the Mosquito robot enabled simultaneous evaluation of more reaction dimensions than could previously be examined by classical statistical analysis. Three 1536-well plates consisting of a full matrix of 15 aryl and heteroaryl halides, 4 Buchwald ligands, 3 bases, and 23 isoxazole additives generated a total of 4608 reactions (including controls). The yields of these reactions were used as the model output. Approximately 30% of the reactions failed to deliver any product, with the



isoxazole A				isoxazole B			
atom	x	y	z	atom	x	y	z
*O1	0	0	0.15	*O1	0	0	-0.14
*N2	0	0	-0.12	*N2	0	0	0.16
*C3	0	0	0.00	*C3	0	0	-0.11
*C4	0	0	0.09	*C4	0	0	0.00
*C5	0	0	-0.13	*C5	0	0	0.12

$\mathbf{A}_{V4} = [0, \ldots, 0.15, -0.12, 0.00, 0.09, -0.13]$

$\mathbf{B}_{V4} = [0, \ldots, -0.14, 0.16, -0.11, 0.00, 0.12]$

$\mathbf{A}_{V4}' = [0, \ldots, 2.38, -1.73, 0.00, 1.06, -1.60]$

$\mathbf{B}_{V4}' = [0, \ldots, -2.19, 2.33, -1.27, 0.00, 1.38]$

A \ B	V1	V2	V3	V4	V5
V1	0.1	0.9	0.2	0.1	0.4
V2	0.4	0.0	0.2	0.7	0.0
V3	0.1	0.2	0.4	0.1	0.4
V4	0.3	0.6	0.2	0.8	0.3

Fig. 1. Application of ML to reaction prediction. (**A**) A Buchwald-Hartwig amination was used as a model reaction for data generation with simultaneous evaluation of four dimensions. The impact of 23 isoxazole additives on the amination reaction was investigated according to a Glorius fragment screening approach. Full structures are provided in fig. S1. Me, methyl; X, any halide; equiv, equivalent; DMSO, dimethyl sulfoxide; L, ligand; OTf, triflate; *i*-Pr, isopropyl; R, H or alkyl group; *t*-Bu, *tert*-butyl; BTMG, *t*-butyltetramethylguanidine; MTBD, methyltriazabicyclodecene; Et, ethyl. (**B**) Software was built to automate feature generation. Molecular, atomic, and vibrational property calculations were performed using Spartan (with density functional B3LYP and basis set 6-31G*), and these features were subsequently extracted from the resulting text files to generate a modeling data table filled with descriptors and yields. To include vibrational modes as descriptors, we compared molecular vibrations for all compounds in a class on the basis of atomic movements. To more appropriately include the movement of heavy atoms, we multiplied each atom's movement by its atomic mass. Vibrational mode vectors were compared using Pearson correlations. Only vibrational modes with $R^2 > 0.5$ and with values greater than any other entry in the same row and column were treated as matching vibrations. If the first molecule in the set (chosen arbitrarily) shared a particular matched vibration with all others in the group, that vibrational mode was considered to be conserved. In this case, the vibration's frequency and intensity were included in the modeling data table. SA, surface area; V1 through V5, vibrational modes 1 through 5; *, shared atom.

remainder quite evenly spread over the range of yields (fig. S7).

Next we turned to the selection of appropriate descriptors. In linear regression analysis, this selection is typically done by hand according to a mechanistic hypothesis, with principal component analysis sometimes being used to reduce the parameter set to an uncorrelated and statistically tractable number (*29*). For the ML model, we sought a set of descriptors that adequately characterizes the differences among the reactions without recourse to a specific hypothesis. For reasons of internal consistency and descriptor availability, calculated properties were used. To avoid prohibitively time-consuming analysis and logging of computational data, we developed software to submit molecular, atomic, and vibrational property calculations to Spartan and subsequently extract these features from the resulting text files for accessibility to a general user (Fig. 1B). The program requires only the input of reagent structures in the Spartan graphical user interface and specification of the reaction components in a Python script; it is applicable to any reaction type. The program then generates the data table that can be used for modeling. In total, 120 descriptors were extracted by the software to characterize each reaction (section III in the supplementary materials).

With these data in hand, we evaluated the predictive accuracies of linear regression and an array of ML methods using 70% of the data as a training set to predict the remaining 30% (test set) (Fig. 2A). For the linear regression models, we evaluated dimension reduction by removing correlated descriptors, as well as various regularization methods [such as LASSO (least absolute shrinkage and selection operator), ridge regression, and elastic net], but none generated good predictive performance. Turning to supervised ML models, we found that *k*-nearest neighbors, support vector machines, and a Bayes generalized linear model provided no improvement over a linear regression model. However, a single-layer neural network delivered substantial improvement over these methods. Moreover, we found that the random forest algorithm provided even better predictive performance. The test-set root mean square error (RMSE) for the random forest model was 7.8%, with a coefficient of determination R^2 value of 0.92. A significant proportion of this variation is likely attributable to experimental and analytical error. Random forest algorithms operate by randomly sampling the data and constructing decision trees, which are then aggregated to generate an overall prediction (*30*). By combining a large

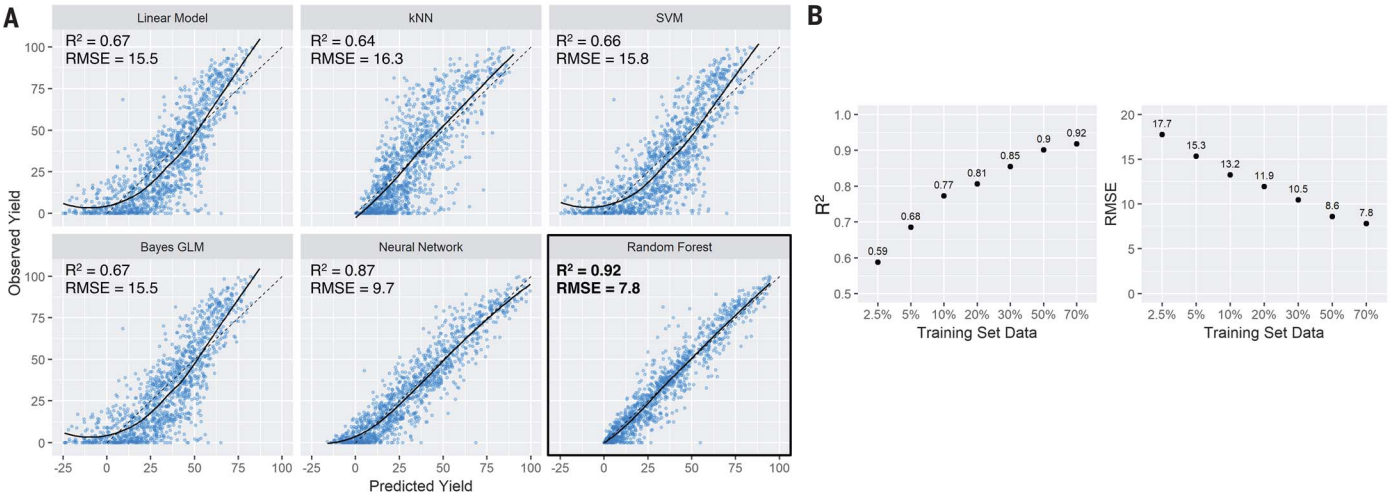


Fig. 2. Test set performance plots. (**A**) Observed versus predicted plots for various ML algorithms and linear regression analysis. For all the models, a 70/30 split of training and test data, with *k*-fold cross-validation on the training data, was performed to measure each model's generalizability to an independent data set. Only test set data are shown in plots. kNN, *k*-nearest neighbor; SVM, support vector machine; GLM, generalized linear model; dashed line, $y = x$ line; solid line, Loess best-fit curve. (**B**) Test set performance of the random forest model with sparse data. A gradual erosion in predictive accuracy occurred from 70% of the data (the entire training set) down to 2.5% of the full data set. The smaller training sets were selected randomly from the original training data.

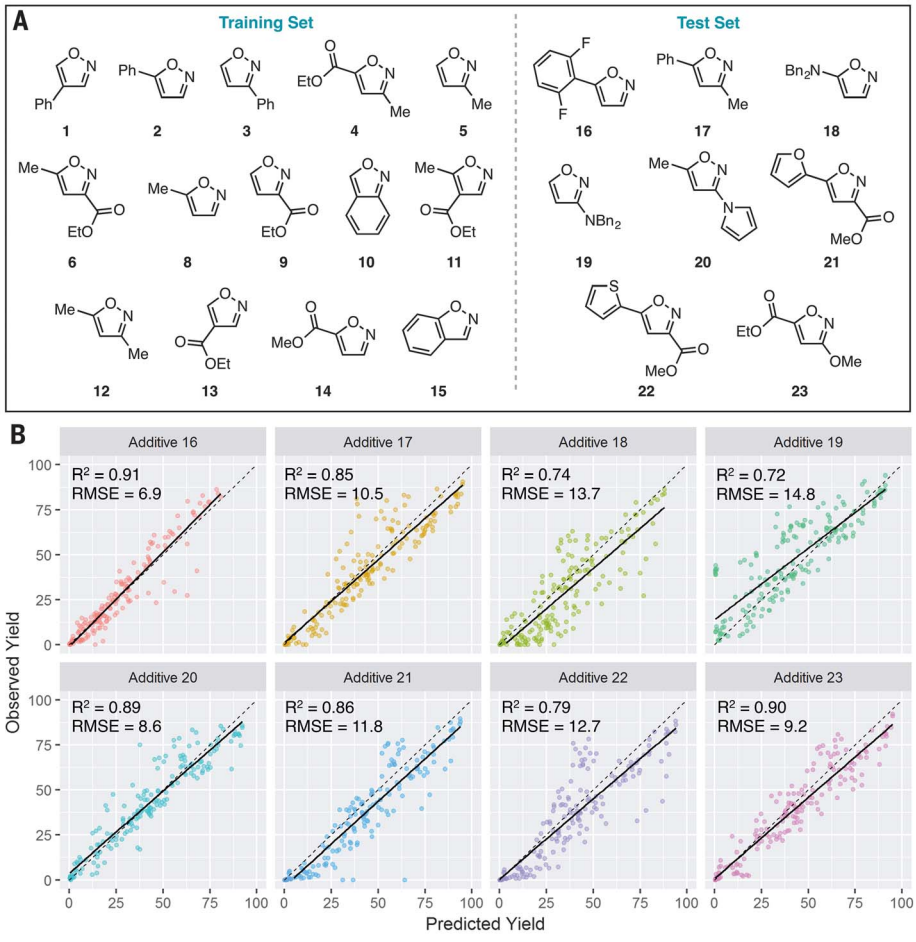


Fig. 3. Additive prediction. (**A**) Isoxazoles in the additive training set (**1** to **6** and **8** to **15**) were used to predict the performance of isoxazoles **16** to **23** in the test set. Ph, phenyl; Bn, benzyl. (**B**) Out-of-sample performance of the random forest model from (A). Test set data are shown.

number of low-precision models, the algorithm can deliver high predictive accuracy without succumbing to overfitting.

Nevertheless, ML tends to encounter predictive limitations when substantially different reaction conditions are used in the test set. This problem is exacerbated by the presence of activity cliffs, which are areas in reaction space where modest changes in chemical structure can lead to notable changes in reaction outcome (*31*). The tendency of ML algorithms to overfit and the presence of activity cliffs necessitate the collection of local reaction data (see fig. S30 for prediction of ArI and ArCl reaction outcomes from ArBr training data). One method for maximizing the extrapolative ability of a model is to use training data spread across the chemical space of interest. The ability to perform accurate prediction under sparsity effectively increases the reaction space that can be explored with the same number of experiments. For the random forest model, we were surprised to discover that enhanced predictive power over other methods could be achieved with a markedly smaller subset of the training data (Fig. 2B). With training on only 5% of the reaction data, the random forest algorithm outperformed linear regression using 70% of the same reaction data. Because 5% of the data set is only 230 experiments, these results indicate that ML can offer improvements in prediction on a scale routinely pursued in the course of reaction optimization and scope elucidation.

We next explored the ability of a random forest model to predict outcomes for reactions containing additives not included in the training data. If effective out-of-sample prediction was possible, ML could predict the effect of a new isoxazole or aryl halide structure on the outcome of a

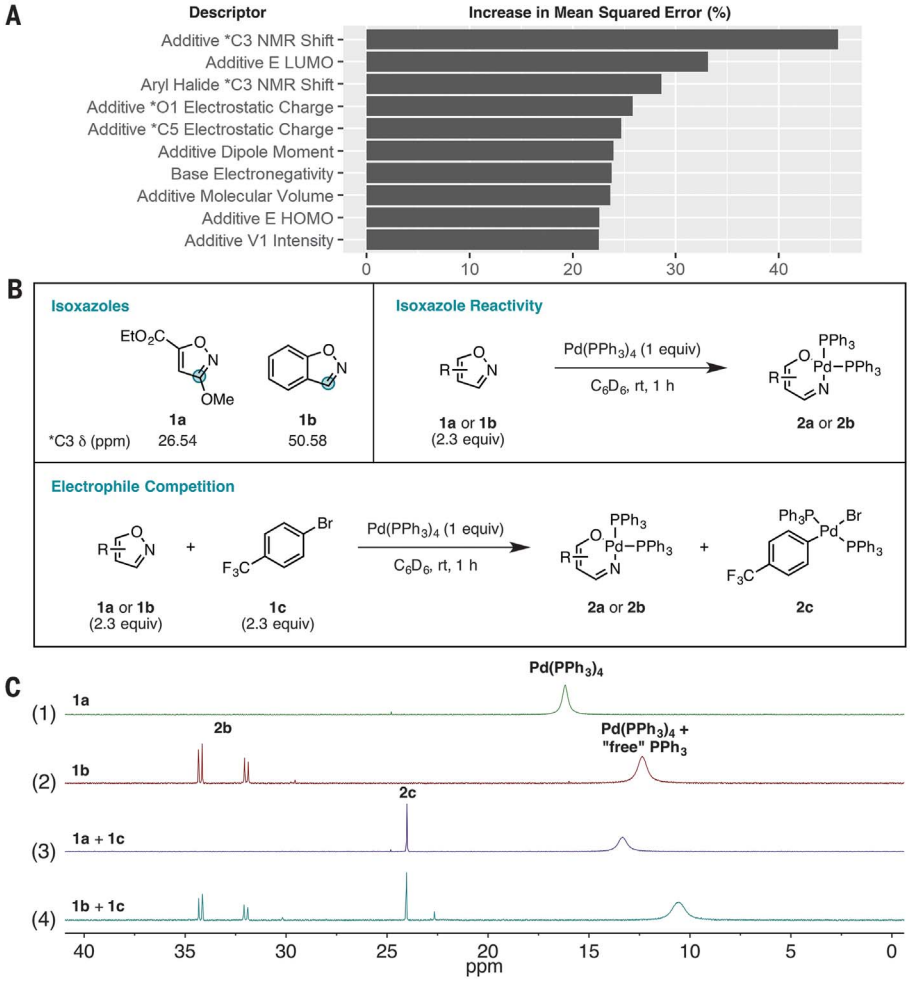


Fig. 4. Model analysis. (**A**) The 10 most important descriptors of the trained random forest model determined by measuring the percent increase in the MSE upon reshuffling of the values of a given descriptor and retraining of the model. * indicates a shared atom. E, energy; HOMO, highest occupied molecular orbital; V, vibration. (**B**) Isoxazoles and the set of reactions designed to test the hypothesis that Pd undergoes oxidative addition to certain additives, leading to diminished yield of the Buchwald-Hartwig amination. ppm, parts per million; rt, room temperature. (**C**) ^{31}P-NMR spectra for the reactions depicted in (B). Spectrum 2 shows the generation of a new Pd species, designated **2b**, upon reaction of $Pd(PPh_3)_4$ with **1b**. Species **2b** is characterized by a pair of doublets with equal integration and a coupling constant (J) consistent with two *cis* phosphines ($^2J_{PP}$ = 37 Hz, where $^2J_{PP}$ is the geminal phosphorus coupling constant). HRMS analysis of the reaction mixture indicates the presence of Pd(**1b**)$(PPh_3)_2$ (**2b**, $[M + 1]^+$ = 750.13).

Buchwald-Hartwig amination and identify the combination of base and ligand that would deliver the highest yield. To this end, we evaluated whether the results for 15 additives could be used to predict the outcomes with 8 distinct additives (Fig. 3A). On average, the out-of-sample RMSE was 11.3%, with an R^2 value of 0.83 (Fig. 3B). None of the additives created significant systematic deviations from what was predicted by the model. The high predictive ability of the model suggests that the effects of these substituents on reaction outcome were captured well by the descriptors. However, as additive consumption was not included in the output, the algorithm is likely to encounter predictive limitations when applied to substrates with embedded isoxazoles.

Having obtained a predictive model, we sought to determine whether it could be used to guide mechanistic analysis. Unlike a linear regression model, the random forest model is challenging to interpret directly. We therefore evaluated the relative importance of descriptors used to construct the model. One such measure of a descriptor's importance is the percent increase in the model's mean square error (MSE) when values for that descriptor are randomly shuffled and the model is retrained (*1*). We found that four of the five most important descriptors in predicting reaction outcomes were the additive's *C-3 nuclear magnetic resonance (NMR) shift (where the asterisk indicates a shared atom), lowest unoccupied molecular orbital (LUMO) energy, and *O-1 and *C-5 electrostatic charges (Fig. 4A). These features are not sufficient to obtain a predictive linear model (fig. S24). Taken together, the descriptors suggest that the propensity of the additive to act as an electrophile influences reaction outcomes (*32*–*34*). We hypothesized that competitive oxidative addition of the isoxazole could be a source of deleterious side reactivity. Although oxidative addition of Pd to isoxazoles is not known (*35*), such an elementary step has been reported previously for other transition metals (*36*).

To evaluate this proposal, we conducted a series of experiments with isoxazoles **1a** and **1b**, which possess the smallest and largest predicted *C-3 NMR chemical shifts of the additives in the test set, respectively (Fig. 4B). As shown in Fig. 4C, spectrum 1, isoxazole **1a** underwent no reaction with tetrakis(triphenylphosphine) palladium(0) [$Pd(PPh_3)_4$] in benzene at room temperature. On the other hand, with isoxazole **1b**, a new species was observed within 1 hour (Fig. 4C, spectrum 2). High-resolution mass spectrometry (HRMS) and spectroscopic (^{31}P, ^{13}C, and ^{1}H NMR) analyses provided strong evidence that isoxazole **1b** underwent oxidative addition at the N–O bond (section VI in the supplementary materials). Going further, we investigated how isoxazoles **1a** and **1b** performed in competition with an aryl halide. When **1a** was mixed with aryl bromide **1c**, formation of only the aryl bromide oxidative adduct (**2c**) was observed (Fig. 4C, spectrum 3). However, when isoxazole **1b** was subjected to the same competition experiment, the oxidative adducts of both the aryl bromide **1c** and isoxazole **1b** were observed in roughly equal amounts (Fig. 4C, spectrum 4). These data are consistent with the hypothesis that electrophilic isoxazole additives can undergo N–O oxidative addition to Pd(0) as a deleterious side reaction, causing diminished yields of the desired Buchwald-Hartwig aminations. Although such a hypothesis could have been obtained by alternate means, this study highlights how measuring the influence of a large collection of descriptors for their predictive ability in an ML algorithm can be used to generate hypotheses for further mechanistic inquiry. Although one should be hesitant to perform direct causal inference, this approach could be particularly enabling for larger and higher-dimensional data sets wherein it would be challenging or impossible to intuit a unified mechanism.

Vast resources and time are currently expended on the development of synthetic methods and their application to complex molecule synthesis, often in a largely ad hoc manner. Here we have shown that simple atomic, molecular, and vibrational descriptors that can be automatically extracted from the text files of Spartan calculations can be used as input for a random forest model to predict yields of multidimensional chemical data. We expect that this approach, coupled with advances in HTE and analysis with whole-molecule systems, will prove to be of broad utility in facilitating the adoption of synthetic methods by enabling prediction of a

new substrate's performance under given conditions or prediction of the optimal conditions for a new substrate.

REFERENCES AND NOTES

1. T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer, 2009).
2. M. I. Jordan, T. M. Mitchell, *Science* **349**, 255–260 (2015).
3. A. Lavecchia, *Drug Discov. Today* **20**, 318–331 (2015).
4. V. Svetnik *et al*., *J. Chem. Inf. Comput. Sci.* **43**, 1947–1958 (2003).
5. J. Ma, R. P. Sheridan, A. Liaw, G. E. Dahl, V. Svetnik, *J. Chem. Inf. Model.* **55**, 263–274 (2015).
6. M. H. Todd, *Chem. Soc. Rev.* **34**, 247–266 (2005).
7. S. Szymkuć *et al*., *Angew. Chem. Int. Ed.* **55**, 5904–5937 (2016).
8. M. A. Kayala, C.-A. Azencott, J. H. Chen, P. Baldi, *J. Chem. Inf. Model.* **51**, 2209–2222 (2011).
9. J. N. Wei, D. Duvenaud, A. Aspuru-Guzik, *ACS Cent. Sci.* **2**, 725–732 (2016).
10. C. W. Coley, R. Barzilay, T. S. Jaakkola, W. H. Green, K. F. Jensen, *ACS Cent. Sci.* **3**, 434–443 (2017).
11. B. Liu *et al*., *ACS Cent. Sci.* **3**, 1103–1113 (2017).
12. S. Kite, T. Hattori, Y. Murakami, *Appl. Catal. A Gen.* **114**, L173–L178 (1994).
13. K. Omata, *Ind. Eng. Chem. Res.* **50**, 10948–10954 (2011).
14. P. Raccuglia *et al*., *Nature* **533**, 73–76 (2016).
15. G. Skoraczyński *et al*., *Sci. Rep.* **7**, 3582 (2017).
16. A. Buitrago Santanilla *et al*., *Science* **347**, 49–53 (2015).
17. K. D. Collins, T. Gensch, F. Glorius, *Nat. Chem.* **6**, 859–871 (2014).
18. N. R. Draper, H. Smith, *Applied Regression Analysis* (Wiley, 1998).
19. M. S. Sigman, K. C. Harper, E. N. Bess, A. Milo, *Acc. Chem. Res.* **49**, 1292–1301 (2016).
20. S. E. Denmark, N. D. Gould, L. M. Wolf, *J. Org. Chem.* **76**, 4337–4357 (2011).
21. L. P. Hammett, *J. Am. Chem. Soc.* **59**, 96–103 (1937).
22. E. N. Bess, A. J. Bischoff, M. S. Sigman, *Proc. Natl. Acad. Sci. U.S.A.* **111**, 14698–14703 (2014).
23. A. Milo, E. N. Bess, M. S. Sigman, *Nature* **507**, 210–214 (2014).
24. A. Milo, A. J. Neel, F. D. Toste, M. S. Sigman, *Science* **347**, 737–743 (2015).
25. P. Ruiz-Castillo, S. L. Buchwald, *Chem. Rev.* **116**, 12564–12649 (2016).
26. P. S. Kutchukian *et al*., *Chem. Sci.* **7**, 2604–2613 (2016).
27. E. Vitaku, D. T. Smith, J. T. Njardarson, *J. Med. Chem.* **57**, 10257–10274 (2014).
28. K. D. Collins, F. Glorius, *Acc. Chem. Res.* **48**, 619–627 (2015).
29. M. Shahlaei, *Chem. Rev.* **113**, 8093–8103 (2013).
30. L. Breiman, *Mach. Learn.* **45**, 5–32 (2001).
31. M. Cruz-Monteagudo *et al*., *Drug Discov. Today* **19**, 1069–1080 (2014).
32. Another possible source of incompatibility is Pd-catalyzed isoxazole C–H arylation. However, these reactions favor electron-rich isoxazoles and typically require more forcing conditions (>100°C) than the amination. See (*33*, *34*).
33. Y. Fall, C. Reynaud, H. Doucet, M. Santelli, *Eur. J. Org. Chem.* **2009**, 4041–4050 (2009).
34. M. Shigenobu, K. Takenaka, H. Sasai, *Angew. Chem. Int. Ed.* **54**, 9572–9576 (2015).
35. Y. Tan, J. F. Hartwig, *J. Am. Chem. Soc.* **132**, 3676–3677 (2010).
36. S. Yu *et al*., *Angew. Chem. Int. Ed.* **55**, 8696–8700 (2016).

ACKNOWLEDGMENTS

We thank K. Chuang and M. Keiser of the University of California, San Francisco, for help troubleshooting the neural network implementation and K. Wu of Princeton University and R. Sheridan and Z. Peng of Merck Research Laboratories for helpful discussions. **Funding:** Financial support was provided by Princeton University, an Amgen Young Investigator Award, and a Camille Dreyfus Teacher-Scholar Award. **Author contributions:** D.T.A., J.G.E., and S.L. performed the experiments. D.T.A. wrote the code. All authors designed the experiments, analyzed the data, and wrote the manuscript. **Competing interests:** None declared. **Data and materials availability:** All code and data used to produce the reported results can be found online at https://github.com/doylelab/rxnpredict. Additional HTE yields and model analysis data are available in the supplementary materials.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/360/6385/186/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S34
Tables S1 to S4
References (*37*–*43*)

27 November 2017; accepted 1 February 2018
Published online 15 February 2018
10.1126/science.aar5169