TECHNICAL RESPONSE

ORGANIC CHEMISTRY

Response to Comment on "Predicting reaction performance in C–N cross-coupling using machine learning"

Jesús G. Estrada[1], Derek T. Ahneman[1], Robert P. Sheridan[2], Spencer D. Dreher[3]*, Abigail G. Doyle[1]*

We demonstrate that the chemical-feature model described in our original paper is distinguishable from the nongeneralizable models introduced by Chuang and Keiser. Furthermore, the chemical-feature model significantly outperforms these models in out-of-sample predictions, justifying the use of chemical featurization from which machine learning models can extract meaningful patterns in the dataset, as originally described.

In Ahneman *et al.* (*1*), we showed that a random forest (RF) algorithm built using computationally derived chemical descriptors for the components of a Pd-catalyzed C–N cross-coupling reaction (aryl halide, ligand, base, and potentially inhibitory isoxazole additive) could identify predictive and meaningful relationships in a multidimensional chemical dataset comprising 4608 reactions. Chuang and Keiser (*2*) built alternative models using random barcode features ("straw" models), wherein the chemical descriptors are replaced with random numbers selected from a standard normal distribution. One-hot encoded features, wherein each reagent acts as a categorical descriptor and is marked as absent or present, were also evaluated. Models built with either set of label features are not generalizable, meaning that they cannot make distinct predictions for new chemical entities not found in the training set. Using these alternative models, Chuang and Keiser conclude that the dataset described in our paper is insufficient to establish that models built with chemical features can generalize to new chemical entities or outperform models built with reagent-label features. However, the authors disregard the chemistry underlying the dataset, and in so doing, they base their conclusions on test sets that are poor indicators of model similarity (Plate 1 or 3) and performance (Plate 2). Here, we show that our original out-of-sample test set (Plate 3), although representative of the generalizability of the chemical descriptor model, was suboptimal in distinguishing nongeneralizable models because it was composed of primarily average-yielding additives. However, using rigorous tests of generalizability, we demonstrate that the chemical descriptor model presented in our original study is statistically distinct from and significantly outperforms models built on reagent-label features.

In our original paper, we demonstrated that a RF model delivered high predictive performance among a panel of machine learning (ML) algorithms in a 70/30 train-test split of the dataset. Chuang and Keiser show that models built with barcodes and one-hot encoding achieved near-identical predictive performances. Because a 70/30 random split of the entire data results in a test set composed of reactions with components that the model has seen in the training set at least once, a ML algorithm is capable of learning the reactivity of each reaction component. Thus, ML algorithms can perform well using a variety of representations for the reaction components, whether the representations are continuous chemical descriptors or reagent labels. For this reason, retrospective tests like those in our manuscript and Chuang and Keiser's comment can only be used to conclude that the RF algorithm outperforms other ML algorithms.

That a ML algorithm can be built with random barcodes or reagent labels does not mean that these or the chemical descriptors are meaningless, nor does it invalidate any structure-activity relationship present in a chemical descriptor model. Performing a Y-randomization test—a well-accepted control for interrogating the null hypothesis that there is no structure-activity relationship in the data—on the chemical descriptor model results in an average cross-validated R^2 value of −0.01, demonstrating that the model encodes meaningful information (*3*, *4*). Nonetheless, a model built on reagent labels as descriptors cannot be extrapolated to chemical entities not in the original set. For that, one needs some flavor of chemical descriptors. Thus, out-of-sample prediction is the appropriate test of generalizability and is the overall justification for using chemical features.

In our manuscript, we investigated the generalizability of the chemical descriptor model by using isoxazole additives on Plates 1 and 2 for training and additives on Plate 3 for out-of-sample predictions. Chuang and Keiser also investigated two alternative splits along plate lines. In so doing, they found that prediction of Plate 2 additives is poor [R^2 = 0.19, root mean square error (RMSE) = 21.7%] and conclude that the generalizability of the chemical descriptor model is more limited than we reported. However, to use the large variation in model performance across the different plate test sets to assess model generalizability, one must assume that the training sets of all three models cover a similar spread in chemical space. Figure 1A illustrates the effect of additives on yield across plate lines. Among the 23 additives examined, four additives (**10**, **11**, **13**, and **14**) serve as severe reaction poisons, resulting in substantially lower average yields than the rest. All four of these additives are located in Plate 2. Thus, a test set comprising Plate 2 additives involves a training set without any of the reaction poisons. Such a training set would be expected to result in a poorly predictive model whose performance would also be a poor indicator of generalizability (*5*). Investigating the Plate 2 predictions further, we replaced one of the reaction poisons (**13**) in Plate 2 with an average-yielding isoxazole (**2**) to afford a Plate 2′ test set, thus guaranteeing that the training set includes at least one example of a reaction poison. The model performance increased from R^2 = 0.19 to R^2 = 0.64 (Fig. 1B) (*6*).

For a more systematic evaluation of model generalizability, we turned to activity ranking for out-of-sample test set design (Fig. 1C), which is considered a better indicator of generalization than random splitting, as used in our original study (*7*). Using this method of training/test set design, we split the data into four additive out-of-sample test sets, resulting in models with an R^2 range of 0.69 ± 0.12. Whereas the observed mean performance is slightly lower than we reported based solely on Plate 3 (R^2 = 0.81), Plate 3 predictions are well within the observed range, as are Plate 1 and Plate 2′ predictions (R^2 = 0.66 and 0.64, respectively). By comparison, Plate 2 predictions (R^2 = 0.19) are significantly out of the observed range of performance. These results confirm that the chemical descriptor model has good generalizability along the additive dimension.

We next turned to the question of whether the chemical descriptor model was distinguishable from the nongeneralizable models. Chuang and Keiser show that the three models have similar aggregate test set performances for Plate 3 predictions. However, evaluating the predictions of the models for individual additives on Plate 3 reveals that the models make distinct predictions. For example, a plot of predicted yields for the chemical descriptor versus one-hot encoded models, as shown in Fig. 2A, illustrates that the chemical descriptor model makes different predictions for different out-of-sample additives, something that the one-hot and random barcode models cannot do. Furthermore, for the

[1]Department of Chemistry, Princeton University, Princeton, NJ 08544, USA. [2]Modeling and Informatics, Merck & Co., Inc., Kenilworth, NJ 07033, USA. [3]Chemistry Capabilities and Screening, Merck & Co., Inc., Kenilworth, NJ 07033, USA.
*Corresponding author. Email: spencer_dreher@merck.com (S.D.D.); agdoyle@princeton.edu (A.G.D.)

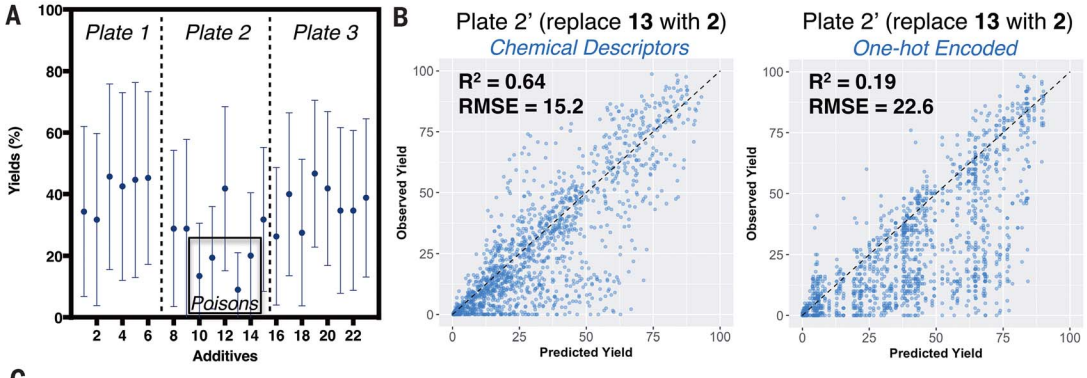
poorer-performing additives in the test set (**16** and **18**), the one-hot encoded model overpredicts yields (e.g., R^2 = 0.31 for **16**), consistent with the model's inability to predict these additives as mild reaction poisons (Fig. 2B). By contrast, the chemical descriptor model predicts that these additives will lead to diminished yields relative to the average (e.g., R^2 = 0.90 for **16**), thereby capturing chemically meaningful information in the additive dimension.

Given these findings, why are the aggregate test set performances between the various models for Plate 3 predictions similar? Using principal components analysis, we found that six of the eight additives in the test set have highly correlated outputs and are average-yielding. As such, the nongeneralizable models perform competitively with the chemical descriptor model because these models make a single prediction that is an average of all of the additives in the training set. A chemical descriptor model is expected to statistically outperform a one-hot model in test sets with a greater number of extreme outcomes. Indeed, the Plate 2′ test set described in this response exhibited a large difference between the two models, with R^2 values of 0.64 versus 0.19 (Fig. 1B). Thus, we proceeded to use Plate 2 as a template to evaluate Chuang and Keiser's null hypothesis that the two models are indistinguishable (*8*).

A total of 14 test sets were created by replacing one (four test sets), two (six test sets), or three reaction poisons (four test sets). Not surprisingly, we found the greatest difference in performance between the chemical descriptor and one-hot model for the test sets incorporating three reaction poisons (Fig. 2C). Across the 14 test sets, we observed R^2 values of 0.63 ± 0.12 for the chemical descriptor model and 0.36 ± 0.20 for the one-hot model; these values indicate that the one-hot model affords an overall worse and more variable predictive performance and that the models are distinguishable at a statistically significant level ($P < 0.01$). To evaluate the predictive value of the additive features for these 14 test sets, we also compared the chemical-feature model to a RF model built using no chemical features for the additives (R^2 = 0.54 ± 0.15) and a model built using one-hot features for the additives but chemical features for the aryl halides, bases, and ligands (R^2 = 0.36 ± 0.19) (Fig. 2D). These experiments clearly show the benefit of using chemical descriptors for out-of-sample prediction and demonstrate that the chemical descriptors used in our original study are not solely acting as reagent identifiers (*9*).

Having confirmed that the chemical descriptor model is generalizable and that the features used have chemical meaning, variable-importance analysis provides a useful tool to obtain chemical insights and guide mechanistic inquiry, as highlighted in our original study. Nonetheless, Chuang and Keiser show that RF algorithms can exhibit descriptor bias, which can skew the analysis of important features. To evaluate the impact that this might have had on our analysis, we investigated an alternative to the randomforest function: the cforest function, which has been shown to avoid descriptor selection bias (*10*). Use of the cforest algorithm resulted in cross-validation test set statistics (R^2 = 0.84) similar to those of the randomforest function. As in our original study, aryl halide and additive *C3 nuclear magnetic resonance (NMR) shifts appeared in the top five chemical descriptors, reinforcing the inference that led to the experiments designed to test whether competitive oxidative addition of the isoxazole could be a source of deleterious side reactivity. Evaluation of a decision tree (DT) model further supplemented our analysis of important descriptors. The aryl halide and additive *C3 NMR shifts appear in the first two discriminating nodes, consistent with the variable importances from the RF model discussed above. Analysis of how chemical descriptors bin reaction components in the DT model (i.e., along aryl halide electronic properties) suggests that the chemical descriptors supply the RF model with the ability to recognize chemical phenomena along the aryl halide and additive dimensions (*11*).

In summary, Chuang and Keiser's reagent-label models are valuable representations of a closed dataset and useful comparator models in tests of generalizability. Incorporation of these models into our workflow has revealed that the out-of-sample validation test in our study was not an optimal test for generalization; however, it delivered a performance representative of the



C

decreasing effect as a reaction poison →

Additives	13, 10, 11, 14, 16, 18, 9, 8, 2, 15, 1, 21, 22, 23, 17, 12, 20, 4, 5, 6, 3, 19	R^2	RMSE
Test Set 1	16 2 22 20 3	0.80	12.1
Test Set 2	14 8 21 12 6	0.77	12.9
Test Set 3	11 9 1 17 5	0.64	16.8
Test Set 4	10 18 15 23 4	0.54	17.9
	average	0.69	14.9
	std. dev.	0.12	2.9

Fig. 1. Identifying reaction poisons and their effect on prediction. (**A**) Average yields (dots) and standard deviations (error bars) of the 180 reactions involving all combinations of aryl halides (15), catalysts (4), and bases (3) of each of the 23 additives except additive **7** (control). Box highlights four additives (**10**, **11**, **13**, and **14**) with substantially lower reaction yields than the rest, indicating their characteristic as reaction poisons. (**B**) The Plate 2 out-of-sample test set was altered by replacing additive **13**, a reaction poison, with average-yielding isoxazole **2**, thereby ensuring that the training set contained one additive that served as a reaction poison. RF models were built using chemical descriptors and one-hot encoded labels for comparison. (**C**) To guarantee that the training set and test set would cover chemical space similar to that covered by the entire dataset, we designed test sets according to activity ranking. Isoxazole additives were ranked according to increasing average yields. The lowest- and highest-yielding additives were kept in all training sets to maximize chemical space. The middle 20 isoxazole additives were used to form four out-of-sample test sets by taking every fourth additive as shown; the remaining additives were used for training. Coefficient of determination (R^2) and RMSE were used to analyze model performance.

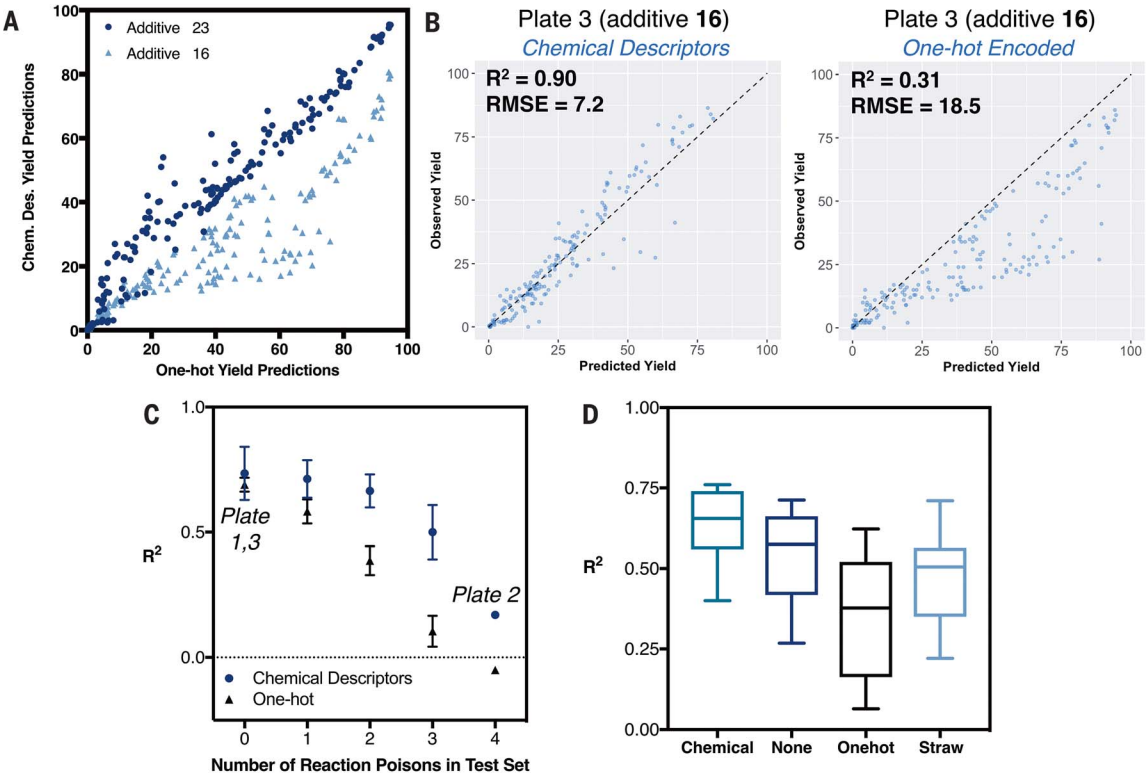


Fig. 2. Distinguishing chemical featurization from one-hot encoding. (**A**) Comparison of chemical descriptor model yield predictions versus one-hot model yield predictions for additives **16** (triangles, light blue) and **23** (circles, dark blue). A nongeneralizable model cannot make distinct predictions for out-of-sample additives. Shown are two distinct predictions made by the chemical descriptor model. (**B**) Calibration plots of observed versus predicted yields for additive **16**, a mild reaction poison. The chemical descriptor model (left) captures the effect of additive **16** causing lower reaction yields, whereas the one-hot model (right) overpredicts. (**C**) Analysis of various prospective predictions according to R^2 values. Plates 1 and 3, which contain no significant reaction poisons in the test set, show minimal differences between the chemical descriptor and one-hot models. Plate 2 contains all four reaction poisons, resulting in a poorly designed training set. Fourteen test sets were designed to incorporate various numbers of reaction poisons and were used to assess the robustness of the chemical descriptor RF model relative to a one-hot encoded RF model. Shown are R^2 averages (dots, triangles) with standard deviations. (**D**) Comparison of the 14 additive out-of-sample test performances of RF models in which the additives are described by chemical descriptors (Chemical), no features (None), one-hot features (Onehot), and straw features (Straw). Relative to absence of additive features, the use of chemical descriptors boosts model performance, whereas the use of reagent label features diminishes model performance. Shown is a box-and-whisker plot.

model's generalizability. On the other hand, our evaluation of Chuang and Keiser's conclusions highlights that an understanding of the chemical reactivity underlying a dataset is necessary in order to use the dataset and reagent-label models to assess the scope and limitations of chemical featurization for reaction prediction. Ultimately, our original conclusion that the RF model is based on meaningful and generalizable chemical features has been strengthened by this additional analysis.

Machine learning offers numerous opportunities to augment how synthetic chemists generate and use data for discovery, optimization, and adoption of synthetic methods (*12*). Its advancement and proliferation will require continued progress in the collection, analysis, and reporting of data, in the description of chemical space, and in predictive modeling. Constructive discussions among chemists, computer and data scientists, and chemical engineers will be important in making this happen.

REFERENCES AND NOTES

1. D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher, A. G. Doyle, *Science* **360**, 186–190 (2018).
2. K. V. Chuang, M. J. Keiser, *Science* **362**, eaat8603 (2018).
3. A. Tropsha, P. Gramatica, V. K. Gombar, *QSAR Comb. Sci.* **22**, 69–76 (2003).
4. All model analyses were performed in R-Studio.
5. We describe exactly in (*1*) this limitation for a model trained on aryl bromides and tested on aryl chlorides.
6. No change in performance was observed for the chemical descriptor model on the resulting Plate 1′ test set (R^2 = 0.66, RMSE = 17.6%), whereas the one-hot model exhibited a lower performance (R^2 = 0.54, RMSE = 20.6%).
7. A. Golbraikh, A. Tropsha, *Mol. Divers.* **5**, 231–243 (2002).
8. One-hot model performance according to activity ranking is R^2 = 0.61 ± 0.11.
9. Similar studies were performed using Chuang and Keiser's straw additive features (R^2 = 0.47 ± 0.15).
10. C. Strobl, A. L. Boulesteix, A. Zeileis, T. Hothorn, *BMC Bioinformatics* **8**, 25–46 (2007).
11. For the smaller base and catalyst dimensions, it is difficult to distinguish whether the RF model uses chemical features to describe meaningful chemical patterns (i.e., Base N1 electrostatic charge) or simply to label reagents.
12. For another study using a RF algorithm to predict reaction performance wherein we constructed one-hot models as comparator models and found that they delivered significantly inferior predictive ability for out-of-sample test sets relative to a chemical descriptor model; see (*13*).
13. M. K. Nielsen, D. T. Ahneman, O. Riera, A. G. Doyle, *J. Am. Chem. Soc.* **140**, 5004–5008 (2018).

ACKNOWLEDGMENTS

We thank M. K. Nielsen for initial design of the one-hot encoded dataset and helpful discussions. **Funding:** Supported by Princeton University and a Camille Dreyfus Teacher-Scholar Award. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** All data used to produce the reported results and additional analyses can be found online at https://github.com/doylelab/rxnpredict.

26 June 2018; accepted 18 September 2018
10.1126/science.aat8763

Response to Comment on "Predicting reaction performance in C–N cross-coupling using machine learning"

Jesús G. Estrada, Derek T. Ahneman, Robert P. Sheridan, Spencer D. Dreher and Abigail G. Doyle

Science **362** (6416), eaat8763.
DOI: 10.1126/science.aat8763