Additional Data and Analyses for

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

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

*Corresponding authors. Email: spencer_dreher@merck.com (S.D.D.) and agdoyle@princeton.edu (A.G.D.)

This PDF file includes:

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

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

*Corresponding authors. Email: spencer_dreher@merck.com (S.D.D.) and agdoyle@princeton.edu (A.G.D.)

Additional Data and Analyses

Contents

I. General Information

All data handling and subsequent modeling were done in R-studio. The normalized feature-response matrix from our original report was used to create all training and test sets. Unless otherwise noted, random forest models were built using the implementation in the R caret package using default parameters. An alternative random forest model was evaluated using the `cForest()` from the R party package, as it has been shown to avoid descriptor selection bias. Model performances were evaluated using the coefficient of determination (R^2) and Root-Mean Squared Error (RMSE). In our original manuscript, reported R^2 values were based on squared Pearson's R coefficients of the observed vs. predicted yield plots. In this response we have moved to using the more general original definition of R^2. Principal Component Analysis (PCA) was performed in MATLAB. Data files used in this response are available in Github.

II. Y-randomization test

A well accepted method of model validation and control that tests whether a model is built on Structure-Activity Relationships (SAR) or on chance correlation is the *Y-randomization test*. Obtaining high R^2 values after the response variable has been randomly shuffled is an indication that the model is not learning from data but rather modeling intricacies in the dataset (e.g. structural redundancies), as suggested by Chuang and Keiser in their Technical Comment. The yield column in the original feature-response matrix was randomly shuffled using the `sample()` function in R, thus decorralating chemical descriptors and yields. The resulting dataset was then partitioned using a 70/30 split for training and test sets. A random forest model was trained and the test set performance measured. An R^2 range of -0.01 ± 0.01 was obtained by performing the randomization 5 times. A low R^2 value is an indication that the model is making meaningful structure activity relationships from the data. Whether the description of chemical structure in an SAR model is represented by generalizable features (e.g. chemical descriptors) or non-generalizable features (e.g. reagent identifiers) is a different issue and requires different tests of generalizability.

III. Activity Ranking

In their Technical Comment, Chuang and Keiser evaluate out-of-sample prediction for the three plates in the data set, including Plate 3 from our original report. After observing large variability in predictive performances (e.g. Plate 2 $R^2 = 0.19$ vs Plate 3 $R^2 = 0.81$) they conclude that the generalizability of our RF model is more limited than we report. However, in order to use the variability in model performance obtained from the three different plates in the dataset to assess generalization, one must assume that a similar spread in chemical space is covered by the training sets of all three models. In other words, if the plates are internally

consistent (e.g. similar spread of chemical space), then splitting along plate lines would constitute a form of random splitting, a training/test set design that is often used in model validation. However, as shown in Fig. 1 of this response, evaluating out-of-sample predictions for Plate 2 does not constitute a form of random splitting because the training set used to build the random forest model does not contain a single example of a severe reaction poison, and is therefore expected to perform poorly on predicting other poisons. As a result, the three random forest models built to predict the reactivity of out-of-sample additives from different plates are not similar enough and performance variability cannot be used to conclude that the models are not generalizable.

An alternative and commonly used form of designing training and test sets is based on activity ranking (Fig. S1). Additives were ranked based on increasing average observed yields of the reactions (180 reactions/additive) they were in. Increasing reaction yields corresponds to decreasing effect of the additives as a reaction poison. The lowest yielding additive (**13**) and the highest yielding additive (**19**) were kept in all training sets in order to cover the largest chemical space in the dataset and to facilitate even splitting. The remaining 20 additives were then used to create 4 test sets by taking every 4th term from the ranked set of additives as shown below. Chemical descriptor and one-hot encoded models were then built and model test performances evaluated (Table S1).

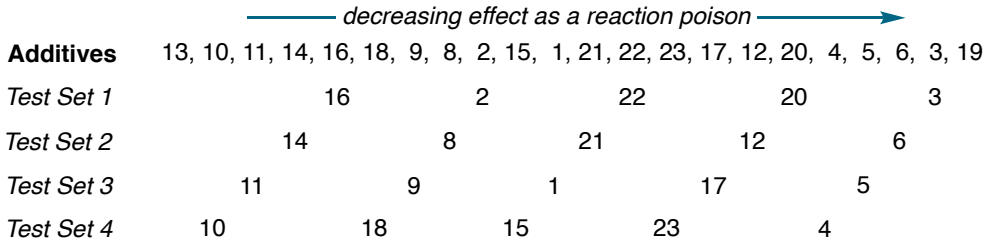


Figure S1. Designing training and test sets based on activity ranking.

Table S1. Activity ranking test set model performances.

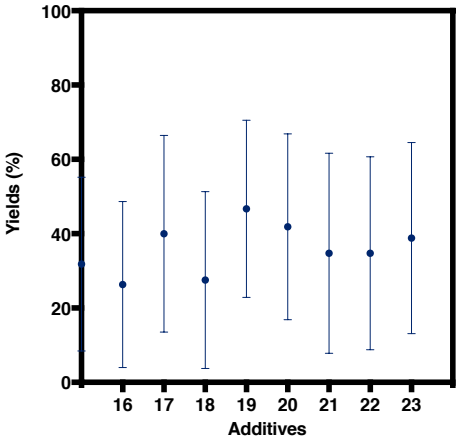
			Chemical Descriptors		One-hot encoded	
Test Set Number	**Training Set**	**Test Set**	**R^2**	**RMSE**	**R^2**	**RMSE**
1	1, 4-6, 8-15, 17-19, 21, 23	2, 3, 16, 20, 22	0.80	12.1	0.71	14.6
2	1-5, 9-11, 13, 15-20, 22, 23	6, 8, 12, 14, 21	0.77	12.9	0.69	15.1
3	2-4, 6, 8, 10, 12-16, 18-23	1, 5, 9, 11, 17	0.64	16.8	0.51	19.5
4	1-3, 5, 6, 8, 9, 11-14, 16, 17, 19-22	4, 10, 15, 18, 23	0.54	17.9	0.51	18.5
		average	0.69	14.9	0.61	16.9
		std. dev.	0.12	2.9	0.11	2.4

The chemical descriptor model achieved a performance with an R^2 range of 0.69 ± 0.12, while the one-hot encoded model exhibited a slightly lower performance of 0.61 ± 0.11. Note that additive poison **13** was included in all training sets, guaranteeing that all random forest

models were trained on at least one reaction poison. Test sets 2, 3, 4 all contain one reaction poison each (**14**, **11** and **10**, respectively) and test set 1 contains no reaction poisons. While the difference in performance between the chemical descriptor model and one-hot encoded model based on activity ranking is small ($\Delta R^2 = 0.08$), they are distinguishable according to a *paired 2-tailed t-test* at a 95% confidence ($p = 0.03$).

IV. Plate 3 Analysis

Even though the chemical descriptor model and the one-hot encoded model share similar performances for the Plate 3 test set ($R^2 = 0.81$ vs 0.71, respectively), we wondered if the Plate 3 predictions from our original report alone could be used to distinguish against a one-hot encoded model. We began by performing a more detailed look into the effect of Plate 3 additives on reaction yields. As shown in the figure below (Fig. S2 Left), additives **16** and **18** lead to diminished reaction yields relative to the other additives in the set. We proceeded to gather additional insights on the additives by applying two methods of unsupervised learning, Covariance Analysis and Principal Component Analysis (PCA). A correlation matrix of the observed yields was obtained by calculating Pearson correlations for all pairs of additives in the plate (Fig. S2 Right). As shown below, additives **16** and **18**, as well as additive **19**, have a clearly distinct reactivity compared to the remaining additives in the test set as evident by the lower correlation coefficients (light blue). Note that these correspond to the two lowest-yielding additives (**16** and **18**) as well as the highest-yielding additive (**19**) in the set.



	16	17	18	19	20	21	22	23
16	1.00							
17	0.60	1.00						
18	0.74	0.61	1.00					
19	0.50	0.81	0.51	1.00				
20	0.59	0.97	0.58	0.80	1.00			
21	0.67	0.85	0.69	0.69	0.80	1.00		
22	0.70	0.84	0.74	0.67	0.81	0.95	1.00	
23	0.64	0.92	0.67	0.72	0.88	0.91	0.91	1.00

Figure S2. (Left) Average observed yields (dots) with standard deviations (ticks) plotted for additives in Plate 3. (Right) Correlation matrix of the observed yields for all additives in Plate 3 using Pearson correlations.

Using the `pca` function in MATLAB, the additives in Plate 3 were further analyzed according to their observed yields. We found that 94% of the variation in the observed yields is accounted for by the first two Principal Components (PCs), with the first PC accounting for a substantially large portion of the overall variance (89%). The coefficients (loadings) of the first 2 PCs were used to explore additive reactivity (Fig. S3 Left). For clarity, a varimax rotation was performed using the `rotatefactors` function. A plot of the coefficients for the two principal components is shown below (Fig. S3 Right). Again, additives **16** and **18** form a cluster of additives with distinct reactivity (PC2) compared to the other cluster comprised of the remaining 6 additives (PC1). Taken together, 6 out of the 8 additives in the Plate 3 test set behave very similarly to each other, while only two of the additives (**16** and **18**) have a distinct and lower reactivity from the average, and as such are moderate reaction poisons.

	Coefficients		Coefficients*	
Additive	PC1	PC2	PC1	PC2
16	0.290	0.521	-0.102	0.588
17	0.385	-0.312	0.496	-0.001
18	0.313	0.582	-0.122	0.649
19	0.320	-0.398	0.499	-0.108
20	0.359	-0.327	0.485	-0.029
21	0.391	0.054	0.270	0.288
22	0.378	0.123	0.217	0.333
23	0.377	-0.096	0.354	0.162

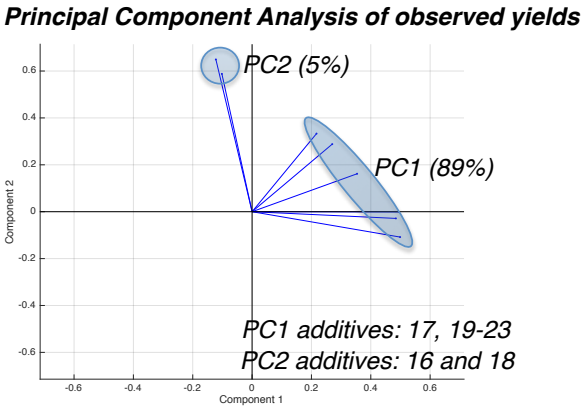


Figure S3. (Left) Coefficients for the first 2 principal components. *Coefficients after a varimax rotation. (Right) Coefficients plotted for the first 2 PCs showing clustering of additives.

We proceeded to compare yield *predictions* obtained from the chemical descriptor model and the one-hot encoded model. As shown in the plot below (Fig. S4), the chemical descriptor model makes distinct and lower yield predictions for additive **16** (teal circles) and **18** (black diamonds), while it predicts similar yields for the remaining six additives (blue triangle with standard deviation ticks), additives that behave almost identically as per the study above. Since one-hot encoding cannot make distinct predictions, it forms a prediction based on the average observed yield of all the additives in the training set. The one-hot encoded yield prediction happens to be reasonable for six out of the eight additives in the test set since these are average-yielding. It is therefore not surprising that a one-hot model performs rather well when evaluated on the Plate 3 test set since it only overpredicts in two cases. However, the chemical descriptor model's ability

to recognize that two additives in the test set are moderate reaction poisons results in improved performance. More importantly, the chemical descriptor model performed distinct predictions for different classes of additives (e.g. average yielding additives versus moderate reaction poisons), a clear indication that the chemical features are not simply acting as reagent identifiers.

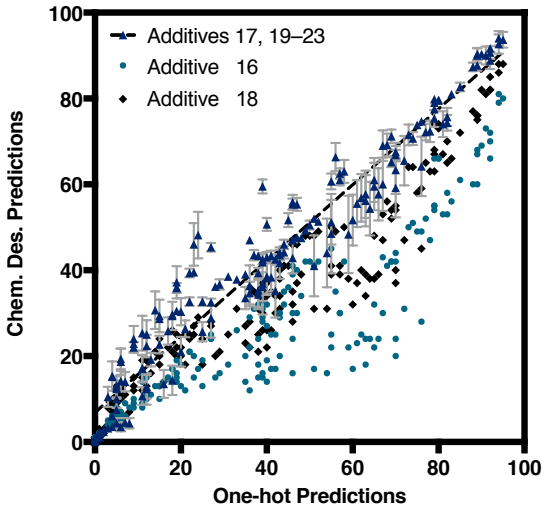


Figure S4. Chemical descriptor model yield predictions versus one-hot encoded model yield predictions for additives in Plate 3.

V. New Test Sets

A. Replacing additive **13** with additive **2** (Plate 2' test set)

After observing that the training set corresponding to the Plate 2 test set contained no severe reaction poisons, we evaluated the effect of replacing one of the low-yielding additives (**13**) from Plate 2 with an average yielding additive (**2**) from Plate 1. As shown below (Fig. S5), the chemical descriptor model when tested on the new Plate 2' test set exhibited a substantial increase in performance as compared to the original Plate 2 test set reported by Chuang and Keiser (R^2 from 0.19 to 0.64). We attribute the improved model performance to the inclusion of a severe poison in the training set thereby increasing the chemical space covered by the model. The performance of the Plate 2' test set is also consistent with the observed performance based on activity ranking (0.69 ± 0.12). The one-hot encoded model, however, still delivered a poor predictive performance (R^2 = 0.19), the largest difference in performance relative to the chemical descriptor model. The effect of the additive replacement was also evaluated by measuring the model performances on the resulting Plate 1' test set. The chemical descriptor model had no change in performance with the new Plate 1' test set (R^2 = 0.66, RMSE = 17.6), consistent with a generalizable model capable of predicting reaction poisons. On the other hand, the one-hot encoded model gave a lower performance (R^2 = 0.54, RMSE =

20.6), a result stemming from the one-hot encoded model's inability to predict the lower reactivity of additive poison **13**.

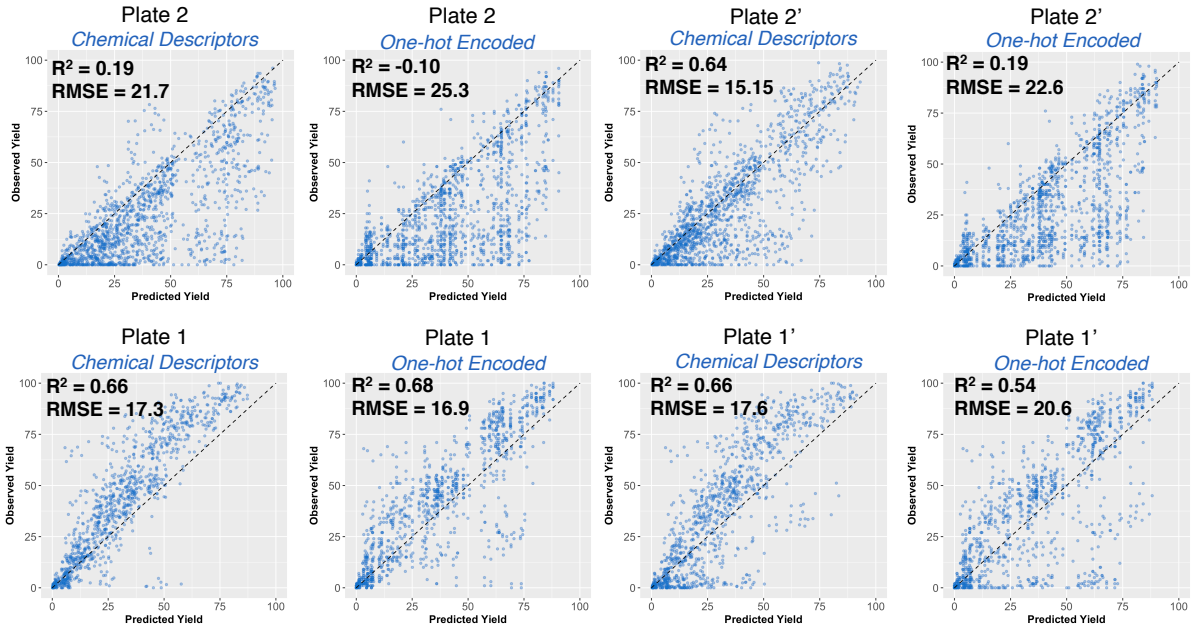


Figure S5. Test set performances for chemical descriptor and one-hot encoded models for Plates 1 and 2, as well as the resulting test sets after additive **13** was replaced by **2**.

B. Additional Test Sets

Since the largest difference in performance between the chemical descriptor model and the one-hot encoded model was observed for the Plate 2' test set containing 3 severe reaction poisons, we proceeded further to compare the two models on a variety of test sets with more extreme behaving additives (reaction poisons) by using Plate 2 as a template. A total of fourteen test sets were created by replacing the four low-yielding additives (**10**, **11**, **13** and **14**) in groups of one, two and three reaction poisons with average-yielding additives from Plate 1 (Table S2). Average-yielding additives **2**, **4**, and **6**, were chosen as they correspond to the first three additives in the dataset (the first three entries in Plate 1 are additives **2**, **4** and **6**).

Table S2. Model performances for the fourteen designed test sets.

Plate 2 Test Set	8, 9, 10, 11, 12, 13, 14, 15
Additive Poisons	10, 11, 13, 14
Replace with	2

			Chemical Descriptors		**One-hot encoded**		**Straw-man**	
Test Set Number	**Training Set**	**Test Set**	**R^2**	**RMSE**	**R^2**	**RMSE**	**R^2**	**RMSE**
5	1, 3-6, 10, 16-23	2, 8, 9, 11, 12, 13, 14, 15	0.54	17.1	0.11	23.6	0.21	22.36
6	1, 3-6, 11, 16-23	2, 8, 9, 10, 12, 13, 14, 15	0.43	19.2	0.07	24.6	0.30	21.22
7	1, 3-6, 13, 16-23	2, 8, 9, 10, 11, 12, 14, 15	0.64	15.2	0.19	22.6	0.51	17.57
8	1, 3-6, 14, 16-23	2, 8, 9, 10, 11, 12, 13, 15	0.40	19.5	0.05	24.5	0.22	22.18
		average	0.50	17.7	0.10	23.8	0.31	20.8
		std. dev.	0.11	2.0	0.06	0.9	0.14	2.2

Plate 2 Test Set	8, 9, 10, 11, 12, 13, 14, 15
Additive Poisons	10, 11, 13, 14
Replace with	2, 4

			Chemical Descriptors		One-hot encoded		Straw-man	
Test Set Number	Training Set	Test Set	R^2	RMSE	R^2	RMSE	R^2	RMSE
9	1, 3, 5, 6, 10, 11, 16-23	2, 4, 8, 9, 12, 13, 14, 15	0.74	13.8	0.36	21.5	0.49	19.23
10	1, 3, 5, 6, 10, 13, 16-23	2, 4, 8, 9, 11, 12, 14, 15	0.73	13.8	0.47	13.4	0.65	15.60
11	1, 3, 5, 6, 10, 14, 16-23	2, 4, 8, 9, 11, 12, 13 15	0.67	15.4	0.35	21.5	0.46	19.60
12	1, 3, 5, 6, 11, 13, 16-23	2, 4, 8, 9, 10, 12, 14, 15	0.61	16.9	0.42	20.5	0.50	19.07
13	1, 3, 5, 6, 11, 14, 16-23	2, 4, 8, 9, 10, 12, 13, 15	0.57	17.8	0.31	22.5	0.38	21.38
14	1, 3, 5, 6, 13, 14, 16-23	2, 4, 8, 9, 10, 11, 12, 15	0.67	15.3	0.41	20.4	0.58	17.30
		average	0.66	15.5	0.39	20.0	0.51	18.7
		std. dev.	0.07	1.6	0.06	3.31	0.10	2.0

Plate 2 Test Set	8, 9, 10, 11, 12, 13, 14, 15
Additive Poisons	10, 11, 13, 14
Replace with	2, 4, 6

			Chemical Descriptors		One-hot encoded		Straw-man	
Test Set Number	Training Set	Test Set	R^2	RMSE	R^2	RMSE	R^2	RMSE
15	1, 3, 5, 10, 11, 13, 16-23	2, 4, 6, 8, 9, 12, 14, 15	0.76	13.5	0.62	17.0	0.57	18.06
16	1, 3, 5, 10, 11, 14, 16-23	2, 4, 6, 8, 9, 12, 13, 15	0.73	14.4	0.52	19.4	0.54	19.04
17	1, 3, 5, 10, 13, 14, 16-23	2, 4, 6, 8, 9, 11, 12, 15	0.75	13.8	0.62	17.0	0.71	14.86
18	1, 3, 5, 11, 13, 14, 16-23	2, 4, 6, 8, 9, 10, 12, 15	0.60	17.7	0.57	18.3	0.56	18.57
		average	0.71	14.8	0.58	17.9	0.59	17.6
		std. dev.	0.08	1.9	0.05	1.14	0.08	1.89

The results of the fourteen test sets were then combined for each of the two models as well as the straw-man models used by Chuang and Keiser (Table S3). The three spread in performances were then compared using *paired 2-tailed t-tests*. The chemical descriptor model delivers performances that are statistically distinct from the one-hot models and the straw-man models at a >99.99% confidence according to both model performance metrics (R^2 and RMSE).

Table S3. Combined model performances with statistical analyses.

	Chemical Descriptors		One-hot encoded		Straw-man	
Test Set Number	R^2	RMSE	R^2	RMSE	R^2	RMSE
5	0.54	17.1	0.11	23.6	0.21	22.4
6	0.43	19.2	0.07	24.6	0.30	21.2
7	0.64	15.2	0.19	22.6	0.51	17.6
8	0.40	19.5	0.05	24.5	0.22	22.2
9	0.74	13.8	0.36	21.5	0.49	19.2
10	0.73	13.8	0.47	13.4	0.65	15.6
11	0.67	15.4	0.35	21.5	0.46	19.6
12	0.61	16.9	0.42	20.5	0.50	19.1
13	0.57	17.8	0.31	22.5	0.38	21.4
14	0.67	15.3	0.41	20.4	0.58	17.3
15	0.76	13.5	0.62	17.0	0.57	18.1
16	0.73	14.4	0.52	19.4	0.54	19.0
17	0.75	13.8	0.62	17.0	0.71	14.9
18	0.60	17.7	0.57	18.3	0.56	18.6
average	0.63	16.0	0.36	20.5	0.48	19.0
std. dev.	0.12	2.1	0.20	3.2	0.15	2.3
p value			<0.0001	<0.0001	<0.0001	<0.0001

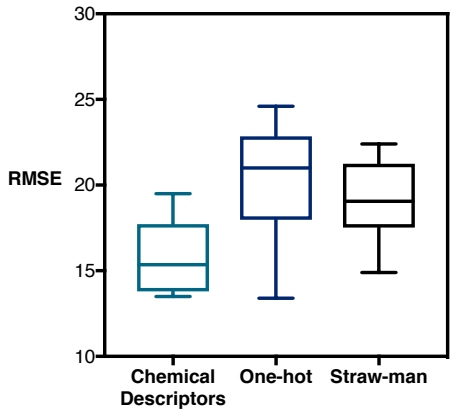


Figure S6. Box and Whisker plot of the RMSE values of the 14 test sets evaluated on the chemical descriptor models and the two comparator models.

C. Comparing different featurization along additive dimension

Since the 14 designed out-of-sample test-sets described above are along the additive dimension, we also evaluated the predictive value of the different forms of additive featurization. RF models were built using 1. Additive chemical features (Chemical) 2. No additive chemical features (None) 3. Additive one-hot features (Onehot) and 4. Additive straw-man features (Strawman) while the remaining aryl halide, base and catalyst dimensions were described by chemical descriptors. The RF model with no additive features exhibited a model performance with $R^2 = 0.54 \pm 0.15$. Inclusion of the additive chemical features boosts the model performance to $R^2 = 0.63 \pm 0.12$, consistent with the ability of the chemical descriptor RF model to predict the diminished yields of reaction poisons and perform overall better predictions (Fig. S7). On the other hand, inclusion of one-hot additive features and straw-man additive features lower the model performance to $R^2 = 0.36 \pm 0.19$ and $R^2 = 0.47 \pm 0.15$ respectively, clearly showing that the chemical descriptors allow the RF model to detect meaning chemical patterns along the additive dimension and are not acting as reagent labels.

Table S4. Combined model performances with statistical analyses.

	Chemical Descriptor		No Additive Features		One-hot Features		Straw-man Features	
Test Set Number	**R^2**	**RMSE**	**R^2**	**RMSE**	**R^2**	**RMSE**	**R^2**	**RMSE**
5	0.53	17.1	0.35	20.2	0.13	23.4	0.22	22.2
6	0.43	19.2	0.29	21.4	0.09	24.3	0.28	21.5
7	0.64	15.2	0.44	18.8	0.17	22.8	0.51	17.5
8	0.40	19.5	0.27	21.5	0.06	24.3	0.22	22.2
9	0.74	13.8	0.55	18.1	0.35	21.7	0.50	19.2
10	0.73	13.8	0.66	15.5	0.48	19.1	0.65	15.7
11	0.67	15.4	0.54	18.2	0.34	21.6	0.46	19.6
12	0.61	16.9	0.61	16.9	0.41	20.7	0.47	19.6
13	0.57	17.8	0.49	19.4	0.33	22.2	0.37	21.5
14	0.67	15.3	0.60	16.8	0.40	20.6	0.57	17.4
15	0.76	13.5	0.71	14.9	0.61	17.4	0.56	18.4
16	0.74	14.4	0.63	17.1	0.51	19.7	0.53	19.2
17	0.75	13.8	0.71	14.8	0.62	16.9	0.71	14.8
18	0.60	17.7	0.68	15.9	0.56	18.6	0.54	18.9
average	0.63	16.0	0.54	17.8	0.36	21.0	0.47	19.1
std.dev	0.12	2.1	0.15	2.2	0.19	2.4	0.15	2.3

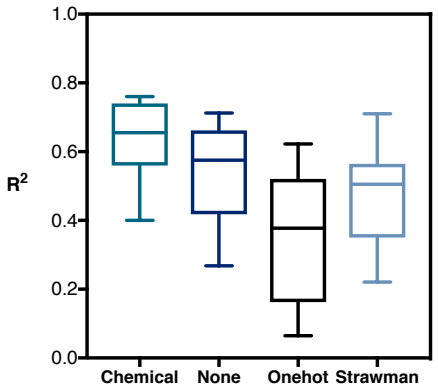


Figure S7. Box and Whisker plot of the R^2 values of the 14 additive out-of-sample test sets evaluating different forms of additive featurization.

VI. Random forest models and descriptor selection bias

Having established that the random forest model is performing meaningful structure-activity relationships according to the Y-randomization test and that the chemical descriptor models are generalizable then the chemical features are serving more than just acting as reagent identifiers. As such, variable importances can be used to obtain chemical insights and guide mechanistic analyses. Any descriptor selection bias, a result of algorithm implementations, can skew descriptor analysis. Often the reasons an algorithm may exhibit selection bias is due to descriptors not being normalized prior to modeling or if features differ in number of classes *(Response Ref9)*. Since our dataset was normalized during data handling, any selection bias is most likely due to the number of features per dimension being different (i.e. additives, aryl halides, base and catalyst).

The random forest algorithm that is implemented in the R caret package can exhibit descriptor selection bias. To avoid such bias, Strobl *et al*. developed an alternative function, the `cForest` function available in the R party package. Application of the `cForest` algorithm to our dataset using a 70/30 split training to test set resulted in a model with good predictive performance (R^2 = 0.84) compared to the standard randomforest algorithm (R^2 = 0.92). Variable importances were then calculated using the `importance()` function. As shown below (Fig. S8), three out of the top five predictors are aryl halide electronic descriptors, as well as the additive C3 NMR shift feature, just as we observed in our original report and used to guide our mechanistic experiments.

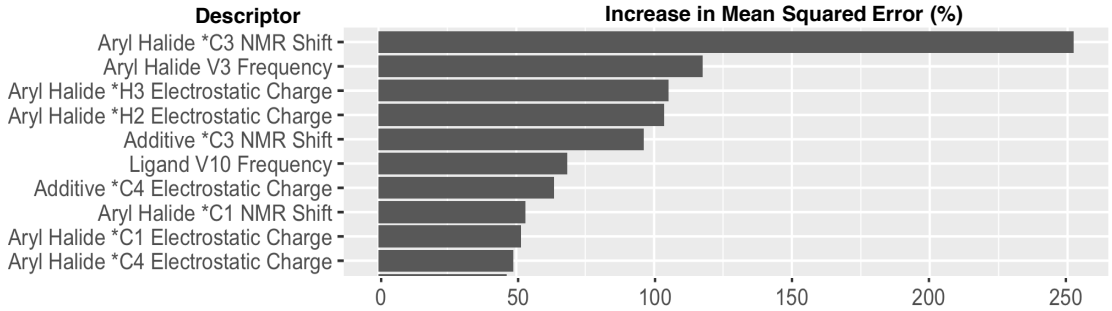


Figure S8. Variable importance plot for the random forest model built using the cForest function.

It is important for chemists employing machine-learning algorithms and techniques to refer back to the chemistry throughout the process. A look at the reactivity based on average observed yields of each component along all four dimensions (Fig. S9) shows that the greatest variability in yields is observed along the aryl halide and additive dimensions. A variable importance plot showing many aryl halide electronic descriptors as well as some isoxazole electronic features would be consistent with this observation. While there are base and catalyst effects, the dataset was not designed to study in detail the effect of base and catalyst on the Buchwald-Hartwig amination; hence why few bases and ligands were included.

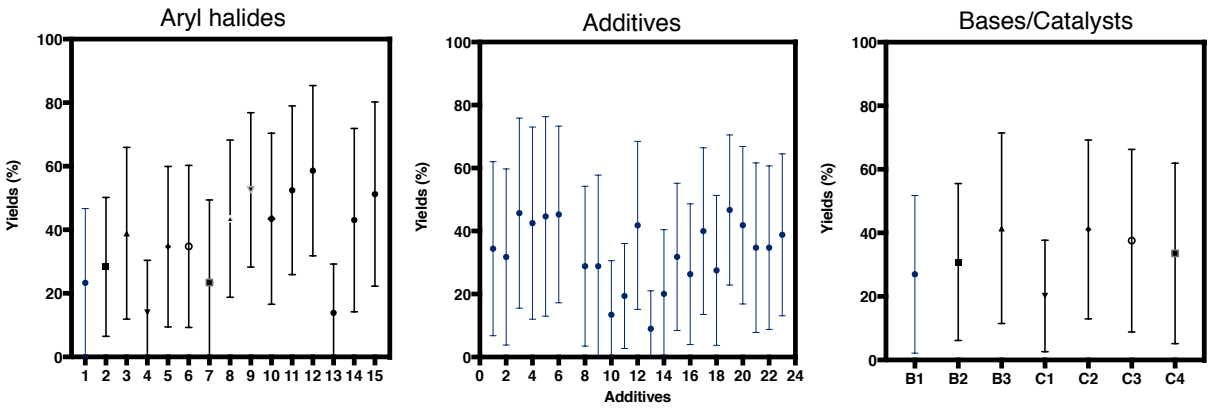


Figure S9. Average yields (dots) and standard deviations (ticks) of the observed yields of each reagent.

VII. Decision Trees for Descriptor analysis

While the cForest function was designed to avoid feature selection bias, we decided to supplement our model analysis and evaluation of important variables through the use of individual Decision Trees (DT). Using the `rpar()` function in R, a decision tree was built using the entire unscaled dataset (Fig. S10). Not surprisingly, the first discriminating node is the aryl halide C3 NMR shift, consistent with the variable importance plot obtained from the `cForest` function, which showed the same feature as the most important descriptor. The next two nodes are the Aryl Halide H2 electrostatic charge and the Additive C3 NMR Shift, which also appear in the top 5 descriptors. The fact that the first three discriminating node features of the Decision Tree appear in the top 5 descriptors of the `cForest` function is consistent with these three features being important in predicting reaction success.

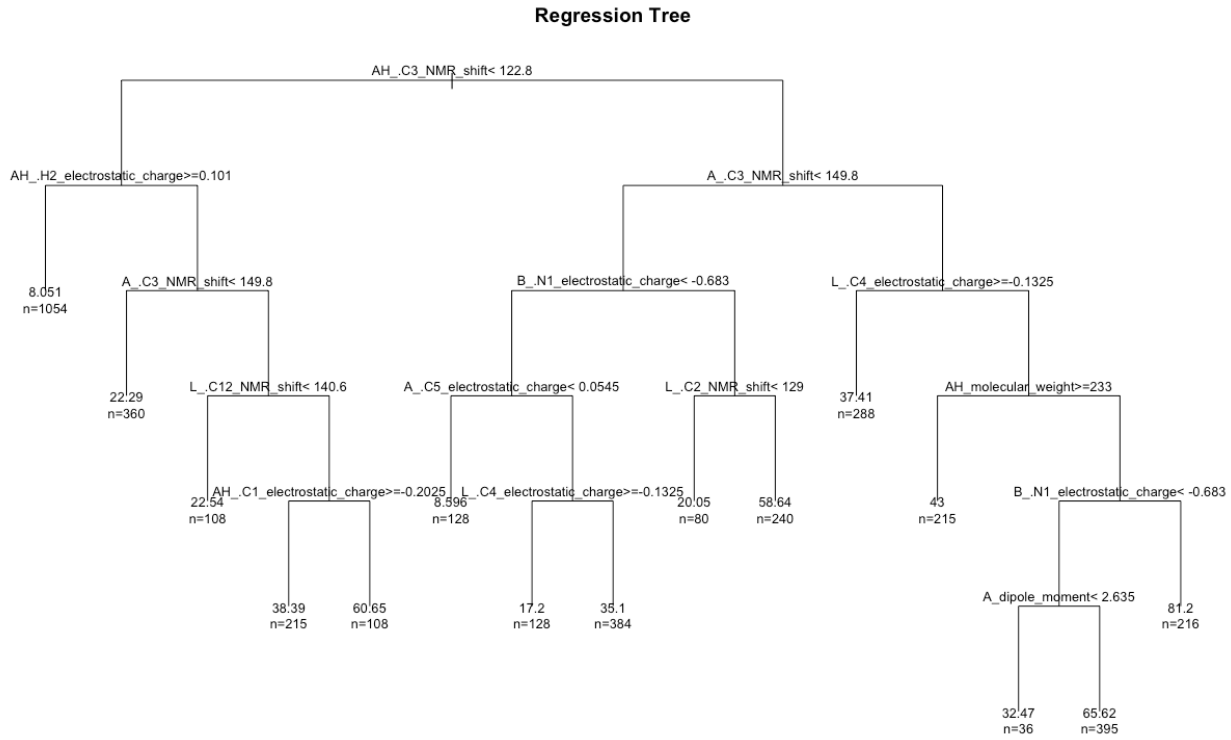


Figure S10. Decision tree modeling unscaled dataset. AH = Aryl Halide, A = Additive, B = Base, L = Ligand

Looking at the Decision Tree model in more detail by determining how each node bins reaction components reveals that the Aryl Halide C3 NMR shift is able to distinguish *chlorides and bromides*, the poorly reactive substrates, from *iodides and the 2-pyridyl series*, the highly reactive substrates. The Aryl Halide H2 Electrostatic Charge further divides the *chlorides* from the *bromides*, ultimately predicting very low yields for the *phenyl chlorides*. The additive C3 NMR shift divides the isoxazole additives in half in terms of their effect as reaction poisons. The Base N1 Electrostatic Charge separates *MTBD*, the best base, from the other two bases (*P2Et and BTMG*). The three ligand features (C2 NMR Shift, C12 NMR shift, C4 Electrostatic Charge) all have the same effect of separating *XPhos*, the worst ligand, from the other three ligands (*tBuXPhos, tBuBrettPhos, AdBrettPhos*). Nodes resulting in singling out reaction components, as observed for the base (MTBD) and catalyst (XPhos) dimensions, are often evidence of overfitting decision trees and models learning the reactivity of individual components. However, as mentioned earlier, the dataset was not designed to evaluate these dimensions in detail (only few examples in each of these dimension), and for the bases and catalysts used the yield does not

vary substantially. Expanding the base and catalyst dimension could reveal important properties of these classes of compounds (e.g. is the Base N1 electrostatic charge important in predicting reaction success?). Taken together, the decision tree model aims to explain yield variability through the following chemical phenomena: 1. The electronic properties of the aryl halide substrate, 2. The effect of reaction poisoning by the isoxazole additive 3. Whether MTBD (best base) and/or XPhos (worst catalyst) were used in the reactions.

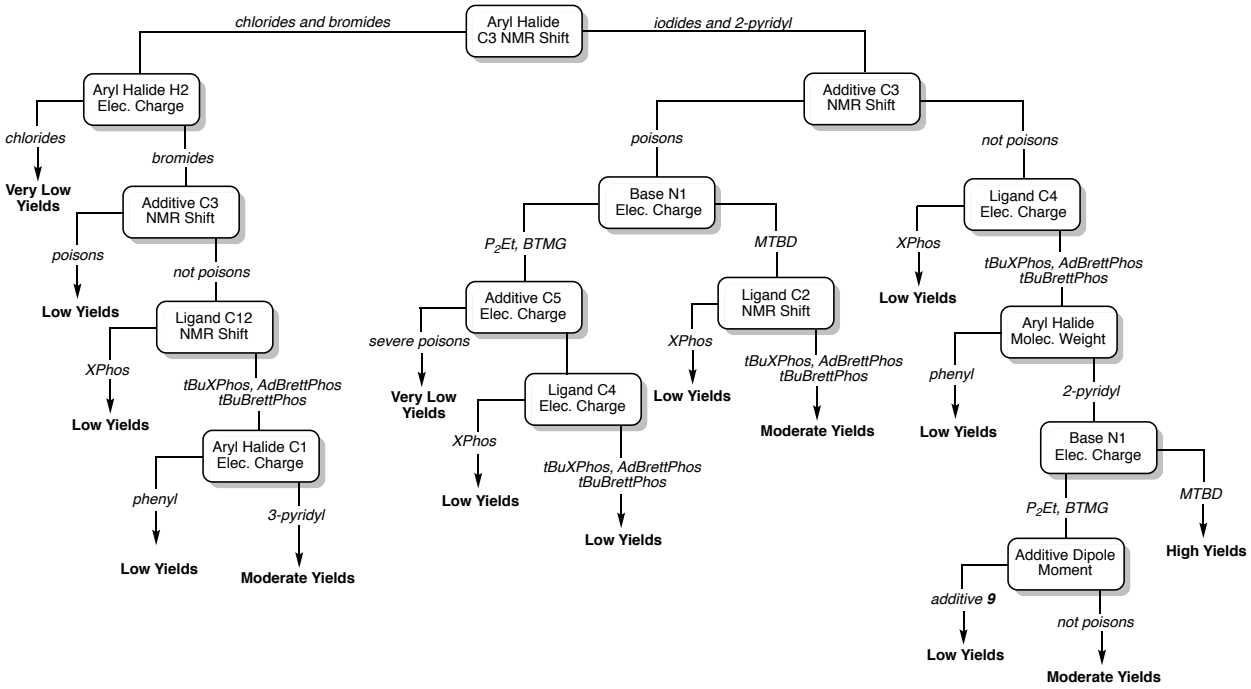


Figure S11. Summarized Decision Tree showing split of reaction components by each node. 0-9% = Very Low Yields, 10-39% = Low Yields, 40-69% = Moderate Yields, 70-100% = High Yields.

As shown in the summarized decision tree above, the aryl halide features work together to describe the electronic properties of the aryl halides (e.g. chlorides vs iodides, phenyl vs pyridyl). Since ease of Pd oxidative addition to the aryl halide largely dictates reaction success, aryl halide electronic features will be important in explaining reaction performance. The role of the additive C3 NMR shift, which appeared in both analyses in this response and in our original report, in describing isoxazole impact on reaction success is not as obvious. Initial observation of the additives ranked by average observed yields and their corresponding C3 NMR shifts showed that this node splits the additives in half. Looking at the chemical structures of the additives revealed that additives with C3 NMR shift <150 ppm have a C3–H bond and tend to behave as reaction

poisons. Similarly, additives with C3 NMR shift > 150 ppm have a fully substituted C3 position and tend to achieve higher yields.

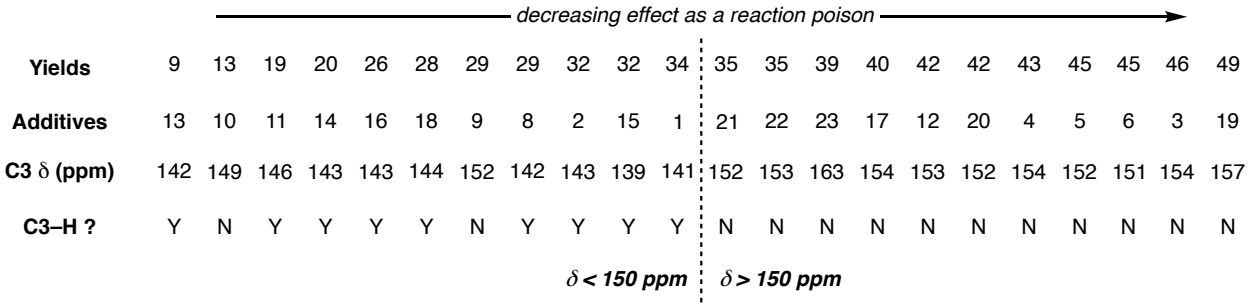


Figure S12. Additives with corresponding average observed yields, C3 NMR shift and whether the additive contains a C3–H bond.

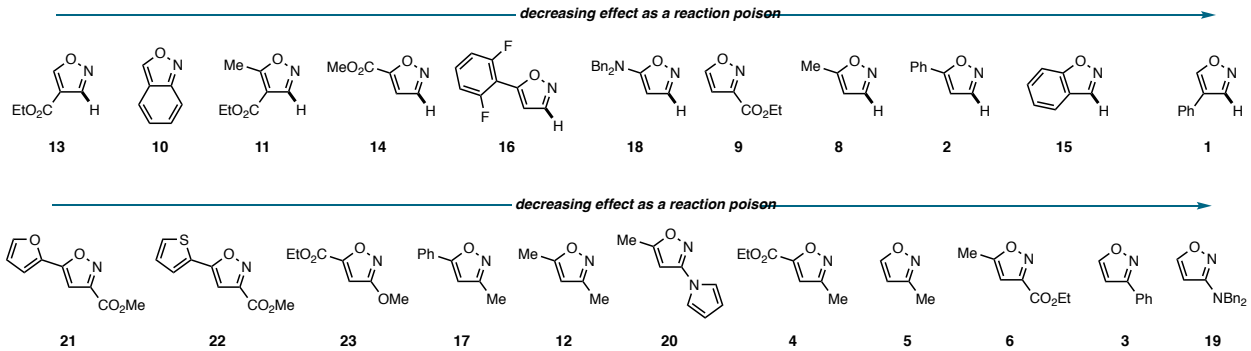


Figure S13. Chemical structures of isoxazole additives ranked by reaction poisoning effect.

The fact that reaction poisons tend to have a C3–H bond is consistent with isoxazole decomposition observed during our original mechanistic investigation. Isoxazoles bearing a C3–H undergo a Pd-catalyzed Kemp-type rearrangement to form α-cyano ketones and aldehydes after N–O oxidative addition. In the absence of Pd, no isoxazole rearrangement was observed, even in the presence of heat. These decomposition products could act as reaction poisons to the palladium catalyst and the acidic proton of the isoxazole could quench the base thus resulting in diminished yields of the cross-coupled product. The additive C3 NMR shift could then be acting as both an electronic continuous variable as observed for the aryl halide dimension, describing the relative ease of oxidative addition to Palladium. It could also act as a categorical variable describing reaction poisoning by isoxazole decomposition products leading to diminished yields. The fact that a continuous variable can behave as a categorical variable, does not imply that the

feature is not generalizable or chemically meaningless. In our original report, we mention that no linear correlation was observed between yields and the additive C3 NMR shift. The ability of a descriptor to act as both a continuous variable as well as a categorical variable speaks to the advantages of the more flexible ML algorithms like Random Forests.

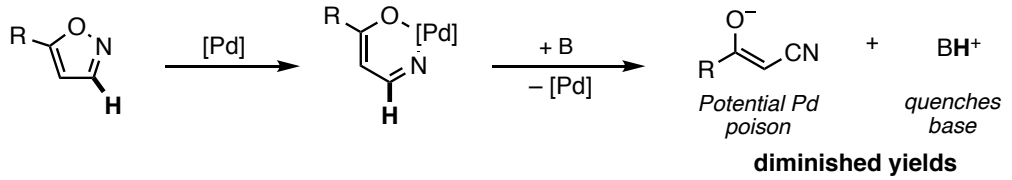


Figure S14. Kemp-type isoxazole rearrangement as a possible source of reaction poisoning.

VIII. Leave-One-Out and Leave-Multiple-Out Cross-validation

An additional form of cross-validation that evaluates model generalizability is the "leave-reagent-out" method. The simplest and often used version is the Leave-One-Out (LOO) protocol in which each reagent is left out one at a time, and a RF model trained on the remaining reaction components. The RF model is then tested on the left out reagent. This is done iteratively for all 44 reaction components and the cross-validated correlation coefficient (q^2) evaluated. Caution must be taken however, since the test set sizes differ depending on the reagent that is left out. On average the chemical descriptor model outperforms the one-hot encoded model (Table S5). However, based on a LOO cross-validation, it is difficult to distinguish against a one-hot encoded model (R^2 0.73 for the chemical descriptor model vs 0.69 for the one-hot encoded model).

Alternatively, Leave-Multiple-Out could be performed by removing several different reagents (e.g. additive 1, additive 4, aryl halide 1, base 2) at a time from the training set, and testing on all reactions that contain those reagents. Many runs (trials) are carried out and the overall performance obtained from the average. According to a LMO cross-validation for 20 trials, the difference in performance between the chemical descriptor and the one-hot encoded model increases compared to the LOO method, as would be expected (Table S6). However, care must be taken when performing LMO cross-validation on datasets with dimensions containing only a small number of reagents since it could result in a *test set containing all* reagents from a given dimension. Our dataset contains three bases and four catalysts, and are relatively smaller dimensions compared to the aryl halide (15 examples) and additive

dimensions (22 examples). A Leave-Four-Reagents-Out could result in a test set containing all three bases or all four catalysts.

Table S5. Combined model performances for Leave-One-Out with statistical analyses.

		Chemical Descriptors		One-hot Encoded	
REAGENT	**Test Set Size**	**R^2**	**RMSE**	**R^2**	**RMSE**
additive 1	179	0.68	20.7	0.56	19.5
additive 2	180	0.94	10.5	0.70	16.7
additive 3	178	0.82	15.9	0.74	18.1
additive 4	180	0.78	18.3	0.91	11.5
additive 5	178	0.85	17.4	0.76	17.9
additive 6	180	0.77	22.4	0.93	11.9
additive 8	180	0.88	9.2	0.71	16.9
additive 9	180	0.57	24.6	0.49	23.1
additive 10	180	0.51	25.9	0.48	30.8
additive 11	180	0.80	14.2	0.67	24.3
additive 12	180	0.85	10.8	0.88	11.1
additive 13	180	0.62	20.9	0.31	35.8
additive 14	180	0.87	8.6	0.62	24.1
additive 15	180	0.68	17.5	0.77	14.2
additive 16	180	0.92	6.6	0.74	17.8
additive 17	180	0.93	7.4	0.87	10.4
additive 18	180	0.62	15.5	0.74	16.9
additive 19	180	0.73	15.2	0.74	17.3
additive 20	180	0.92	7.3	0.85	11.8
additive 21	180	0.95	6.4	0.92	8.1
additive 22	180	0.96	5.2	0.93	7.3
additive 23	180	0.94	6.6	0.90	9.0
P2Et	1320	0.51	21.8	0.50	21.6
BTMG	1317	0.83	12.9	0.78	12.1
MTBD	1318	0.77	19.2	0.77	19.5
A. Halide 1	264	0.45	11.3	0.24	35.1
A. Halide 2	263	0.61	13.2	0.65	22.1
A. Halide 3	263	0.74	18.0	0.62	17.1
A. Halide 4	264	0.10	20.0	0.02	46.4
A. Halide 5	264	0.66	13.3	0.75	22.1
A. Halide 6	264	0.72	9.9	0.75	15.5
A. Halide 7	262	0.04	38.6	0.02	43.1
A. Halide 8	264	0.85	21.0	0.82	13.2
A. Halide 9	264	0.89	15.0	0.80	20.3
A. Halide 10	264	0.86	20.3	0.82	13.6
A. Halide 11	264	0.91	11.3	0.81	20.5
A. Halide 12	264	0.87	16.4	0.80	23.5
A. Halide 13	264	0.22	34.1	0.17	35.7
A. Halide 14	264	0.66	29.3	0.69	17.4
A. Halide 15	263	0.88	16.7	0.77	21.0
Xphos	990	0.57	25.7	0.55	25.8
tBuXPhos	986	0.81	20.0	0.86	12.8
tBuBrettPhos	989	0.93	9.2	0.93	8.8
AdBrettPhos	990	0.90	9.2	0.88	10.6
	average	0.73	16.21	0.69	19.36
	std. dev.	0.22	7.47	0.23	9.00

Table S6. Combined model performances for Leave-Multiple-Out with statistical analyses.

		Chemical Descriptors		One-hot Encoded	
Trial	**Test Set Size**	**R^2**	**RMSE**	**R^2**	**RMSE**
1	792	0.51	22.1	0.35	24.0
2	1313	0.67	17.4	0.76	13.9
3	539	0.75	13.7	0.76	13.9
4	598	0.87	14.3	0.60	20.6
5	1314	0.85	11.8	0.74	14.9
6	1260	0.54	22.7	0.54	24.2
7	684	0.30	25.2	0.36	22.7
8	598	0.64	18.2	0.20	33.9
9	683	0.30	25.8	0.40	23.6
10	2109	0.58	21.9	0.51	20.6
11	1608	0.58	19.1	0.47	21.2
12	540	0.64	17.3	0.57	19.0
13	1670	0.42	23.5	0.35	25.6
14	598	0.60	23.3	0.74	16.2
15	1558	0.72	18.1	0.56	20.9
16	599	0.74	16.5	0.68	17.4
17	600	0.71	16.4	0.48	22.4
18	1382	0.33	23.6	0.30	25.9
19	1314	0.35	23.4	0.51	26.5
20	598	0.85	16.3	0.80	18.4
	average	0.60	19.52	0.53	21.29
	std. dev.	0.18	4.10	0.17	4.90