JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JACS

Article

A Machine Learning Approach to Model Interaction Effects: Development and Application to Alcohol Deoxyfluorination

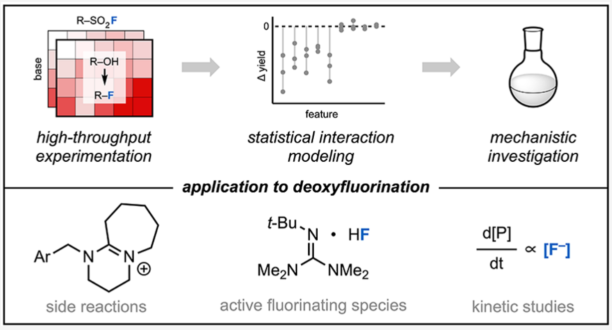
Andrzej M. Żurański,[§] Shivaani S. Gandhi,[§] and Abigail G. Doyle*

Cite This: J. Am. Chem. Soc. 2023, 145, 7898−7909

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The application of machine learning (ML) techniques to model high-throughput experimentation (HTE) datasets has seen a recent rise in popularity. Nevertheless, the ability to model the interplay between reaction components, known as interaction effects, with ML remains an outstanding challenge. Using a simulated HTE dataset, we find that the presence of irrelevant features poses a strong obstacle to learning interaction effects with common ML algorithms. To address this problem, we propose a two-part statistical modeling approach for HTE datasets: classical analysis of variance of the experiment to identify systematic effects that impact reaction yield across the experiment followed by regression of individual effects using chemistry-informed features. To illustrate this methodology, we use our previously published alcohol deoxyfluorination dataset comprising 740 reactions to build a compact, interpretable generalized additive model that accounts for each significant effect observed in the dataset. We achieve a sizeable performance boost compared to our previously published random forest model, reducing mean absolute error from 18 to 13% and root-mean-squared error from 22 to 17% on a newly generated validation set. Finally, we demonstrate that this approach can facilitate the generation of new mechanistic hypotheses, which, when probed experimentally, can lead to a deeper understanding of chemical reactivity.


R–SO$_2$F
base
R–OH
R–F
high-throughput experimentation
0
Δ yield
feature
statistical interaction modeling
mechanistic investigation
application to deoxyfluorination
Ar
N
N
side reactions
t-Bu
N • HF
Me$_2$N
NMe$_2$
active fluorinating species
d[P]/dt ∝ [F⁻]
kinetic studies

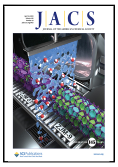
■ INTRODUCTION

The application of data-driven modeling to understand reactivity trends, predict reaction outcomes, and select optimal reaction conditions is of significant interest to the synthetic community.[1−3] For decades, chemists have used linear regression to study the impact of electronic and steric effects on reaction outcomes in the form of Hammett plots.[4] More recently, multivariate linear regression (MVLR) has been used to model the impact of a systematically varied reaction component (e.g., a catalyst) on the reaction outcome.[5] Whereas such studies generally make use of compact, *de novo* generated datasets varying a single reaction component, reaction databases such as Reaxys or the United States Patent and Trademark Office (USPTO)[6,7] contain a wealth of data on many reactions and variations of reaction components. However, reaction databases tend to bias toward high-yielding reactions, may be sparse or incomplete with respect to conditions or substrates of interest, and may lack internal consistency across reaction parameters (i.e., temperature, concentration, and stir rate).[8] These limitations notwithstanding, important advances have been made in reaction outcome and condition prediction using these databases.[9−11]

As an alternative, high-throughput experimentation (HTE) allows for rapid generation of relatively large datasets (up to a few thousand reactions) where multiple reaction components are systematically varied.[12] These datasets typically include a diverse set of substrates and conditions of practical interest, while other variables are kept fixed. HTE datasets can be used to model reactivity trends, such as identifying substrate classes that tend to be higher or lower yielding than others. More importantly, one can also model differences in performance between reagents across the substrate scope, such as conditions that are privileged for specific substrate classes, or conditions that are more selective for a particular product. This interplay between reaction components, referred to herein as interaction effects, is crucial for understanding the intricacies of reactivity.

Multiple studies—including a few of our own—have built machine learning (ML) models for yield and selectivity prediction from HTE datasets using chemically informed features and/or molecular fingerprints.[13−18] However, these models provide only minor improvements over dummy-encoded models, and there is no evidence that they can capture interaction effects.[19−22] This challenge likely arises as a consequence of the inability of ML models to consider the experimental design and structure of HTE datasets.

Received: December 8, 2022
Published: March 29, 2023



https://doi.org/10.1021/jacs.2c13093

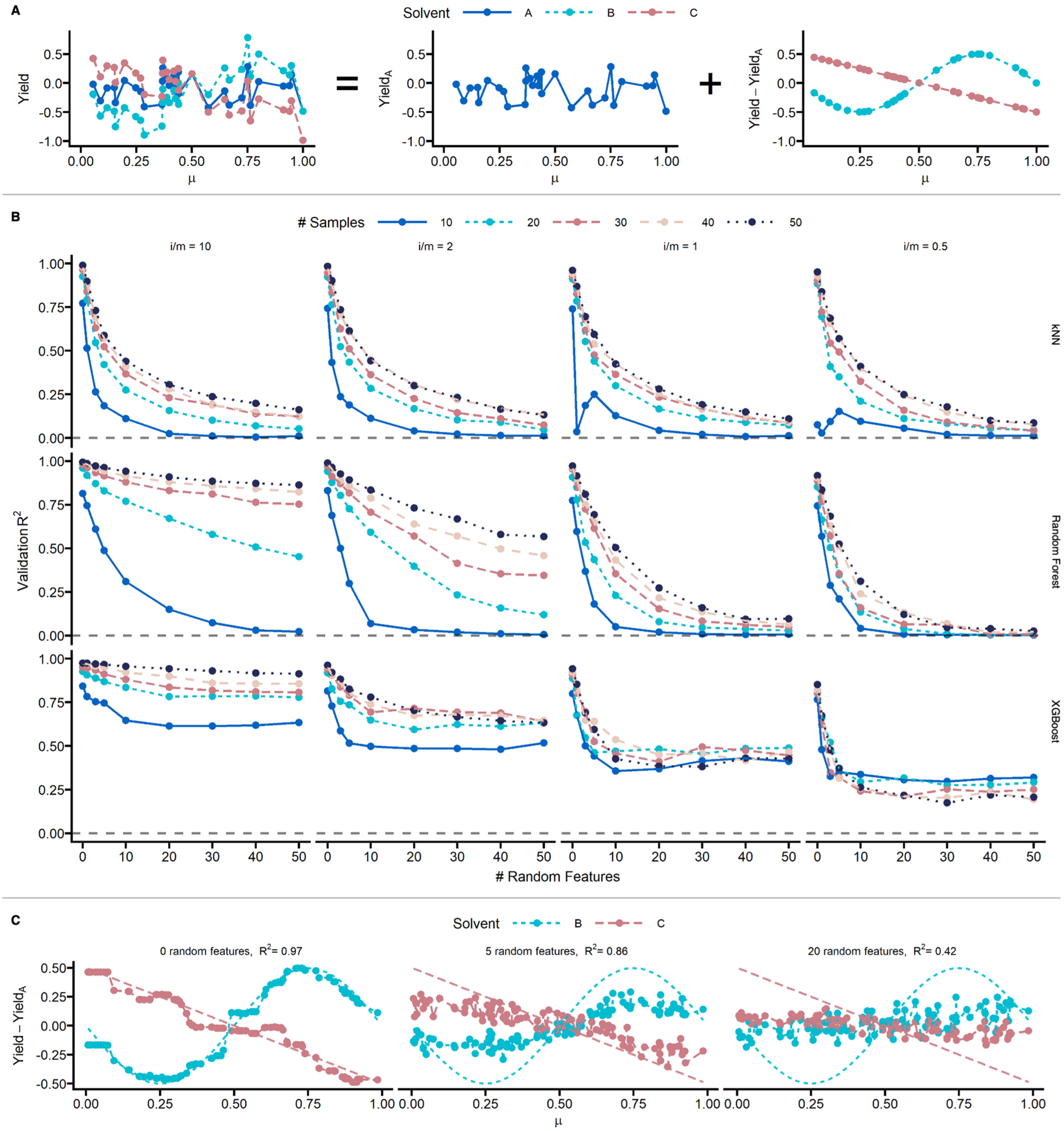


Figure 1. (A) Simulated example of a 2-factor HTE dataset with a random main effect and a substrate-dependent interaction of the same magnitude. (B) Validation R^2 of the interaction component as a function of the training sample size, the relative ratio of interaction and main effects (i/m), and the quality of the featurization. An R^2 value close to 1 represents successful learning, while an R^2 value close to 0 represents no learning. (C) A sample experiment with 30 substrates for which the interaction effect is twice the size of the main effect is used to train a random forest model. The trained model was evaluated on 100 additional substrates, and the interaction between reaction conditions was visualized.

In this manuscript, we first identify the presence of irrelevant features as a major obstacle to ML modeling of interaction effects in a simulated HTE scenario. Inadequate featurization has previously been identified as a potential obstacle to reactivity modeling,[23] but we re-evaluate it in the context of HTE. We then propose an alternative approach for modeling interactions, wherein the HTE dataset is analyzed as a control vs treatment experiment, thus explicitly taking the dataset design into account. We use ANOVA[24]—a statistical technique that separates observed variance in outcomes into systematic effects (effects that are unlikely to occur due to chance) and random effects. Using ANOVA, we identify all main and interaction effects that are significant within the experiment. Finally, we model individual effects with descriptor-based linear and nonlinear regressions, leading to an interpretable and generalizable model.

Challenges of Modeling Interaction Effects. Complex functions can be represented with ML models, but whether they can be learned from available data is an outstanding question.[25] To study the ability of ML to learn interaction effects from HTE data, we simulated an HTE dataset comprising 3 hypothetical solvents (A, B, and C) and between 10 and 50 hypothetical substrates. For the sake of the simulation, we made several assumptions about this dataset. First, we arbitrarily considered the yield in solvent A to be the main effect, though other choices (e.g., average yield over all solvents) could also be used. With this choice of the main effect, the interaction effect is the yield differential between solvent A and solvents B and C. This interaction effect was assumed to depend on a single and known property of the substrate, dipole moment μ, via the function $f(\mu)$ (eq 1). We simulated the difference in response between solvents B and A as a linear function of μ and the difference between C and A as a sinusoidal function of μ (Figure 1A), though in a real HTE dataset, these functions may not be as clearly defined.

$$y_{\mathrm{B,C}} = y_{\mathrm{A}} + f_{\mathrm{B,C}}(\mu) \tag{1}$$

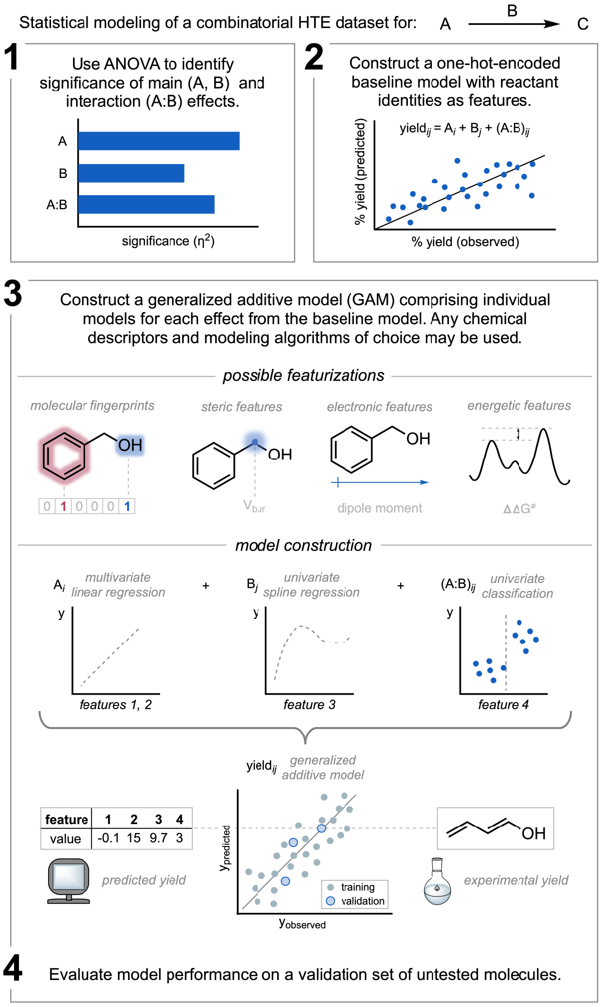
Taking these assumptions, we evaluated a selection of ML algorithms (k-nearest neighbors (kNN),[26] random forest,[27] and XGBoost[28]) for their ability to learn the interactions. We varied the following properties of the simulated HTE dataset:

a) the number of substrates (10, 20, 30, 40, and 50),
b) the relative size of the interaction effect vs the main effect (i/m = 10, 2, 1, and 0.5), and
c) the size and quality of the substrate featurization (using the relevant descriptor μ and adding 0−50 irrelevant features).

Overall, we found that the model is highly sensitive to the presence of unrelated "random" features. When the substrate featurization contains no random features, the learning succeeds (Figure 1B, no. of random features = 0). In the presence of random features, the model's ability to learn depends on the dataset size and the relative magnitude of effects. When the interaction effect dominates the main effect, both random forest (RF) and XGBoost algorithms can sift through the random features and still learn the interaction signal (Figure 1B, i/m = 10 and 2); as the interaction effect becomes less dominant, even a handful of random features prevent learning (Figure 1B, i/m = 1 and 0.5). In the latter case, the relative differences between solvents A, B, and C are lost. As illustrated in Figure 1C, with 30 substrates and an interaction effect twice as large as the main effect, the presence of 5 random features largely precludes capturing the functional shape of the interaction with RF. Therefore, chance correlations with random features, a well-known nuisance in linear modeling,[29] confuse the learning process. The simulation study suggests that using broad molecular featurization—which is certain to involve at least some number of irrelevant features—with commonplace ML algorithms is unlikely to provide a useful interaction model. Therefore, for a deeper understanding of interdependencies within HTE datasets, it is necessary to develop a modeling approach that is independent of algorithmic feature selection.

■ RESULTS AND DISCUSSION

HTE Statistical Modeling. With the goal of constructing a model that better "learns" underlying interactions in HTE datasets, we developed a novel modeling approach targeting interaction effects. Our proposed workflow is summarized in Figure 2. Though we focus on one dataset in this study, the general workflow can in theory be applied to any full factorial HTE dataset of interest:



Statistical modeling of a combinatorial HTE dataset for: A $\xrightarrow{B}$ C
1
Use ANOVA to identify significance of main (A, B) and interaction (A:B) effects.
A
B
A:B
significance (η^2)
2
Construct a one-hot-encoded baseline model with reactant identities as features.
yield$_{ij}$ = A$_i$ + B$_j$ + (A:B)$_{ij}$
% yield (predicted)
% yield (observed)
3
Construct a generalized additive model (GAM) comprising individual models for each effect from the baseline model. Any chemical descriptors and modeling algorithms of choice may be used.
possible featurizations
molecular fingerprints
steric features
electronic features
energetic features
OH
0 1 0 0 0 1
V$_{bur}$
dipole moment
ΔΔG‡
model construction
A$_i$ multivariate linear regression
+
B$_j$ univariate spline regression
+
(A:B)$_{ij}$ univariate classification
y
features 1, 2
feature 3
feature 4
yield$_{ij}$ generalized additive model
feature 1 2 3 4
value -0.1 15 9.7 3
predicted yield
y$_{predicted}$
y$_{observed}$
training
validation
experimental yield
4
Evaluate model performance on a validation set of untested molecules.

Figure 2. Our proposed modeling workflow involves ANOVA analysis, baseline and chemically informed model construction, and external validation. This figure depicts a combinatorial dataset of a hypothetical reaction between A and B.

1. Analyze the dataset with ANOVA and determine which reaction components have a significant impact on the yield. See the SI for a detailed discussion of ANOVA usage guidelines.
2. Construct a one-hot-encoded GAM model, M_0, that includes each of the effects deemed to be significant in the ANOVA analysis.
3. For individual effects in M_0, replace the explicit labels with functions of chemical features using ML. One key advantage of GAM models is their ability to model nonlinear data while retaining interpretability. In our study, we employ both univariate linear and nonlinear regressions; however, any type of model can be considered. Importantly, selected effects can be modeled, while other effects of lesser interest can remain one-hot-encoded models.
4. Evaluate the newly constructed ensemble of models against an external test set.

From the few publicly available HTE datasets,[30] we chose a three-component, full factorial alcohol deoxyfluorination

https://doi.org/10.1021/jacs.2c13093
J. Am. Chem. Soc. 2023, 145, 7898−7909

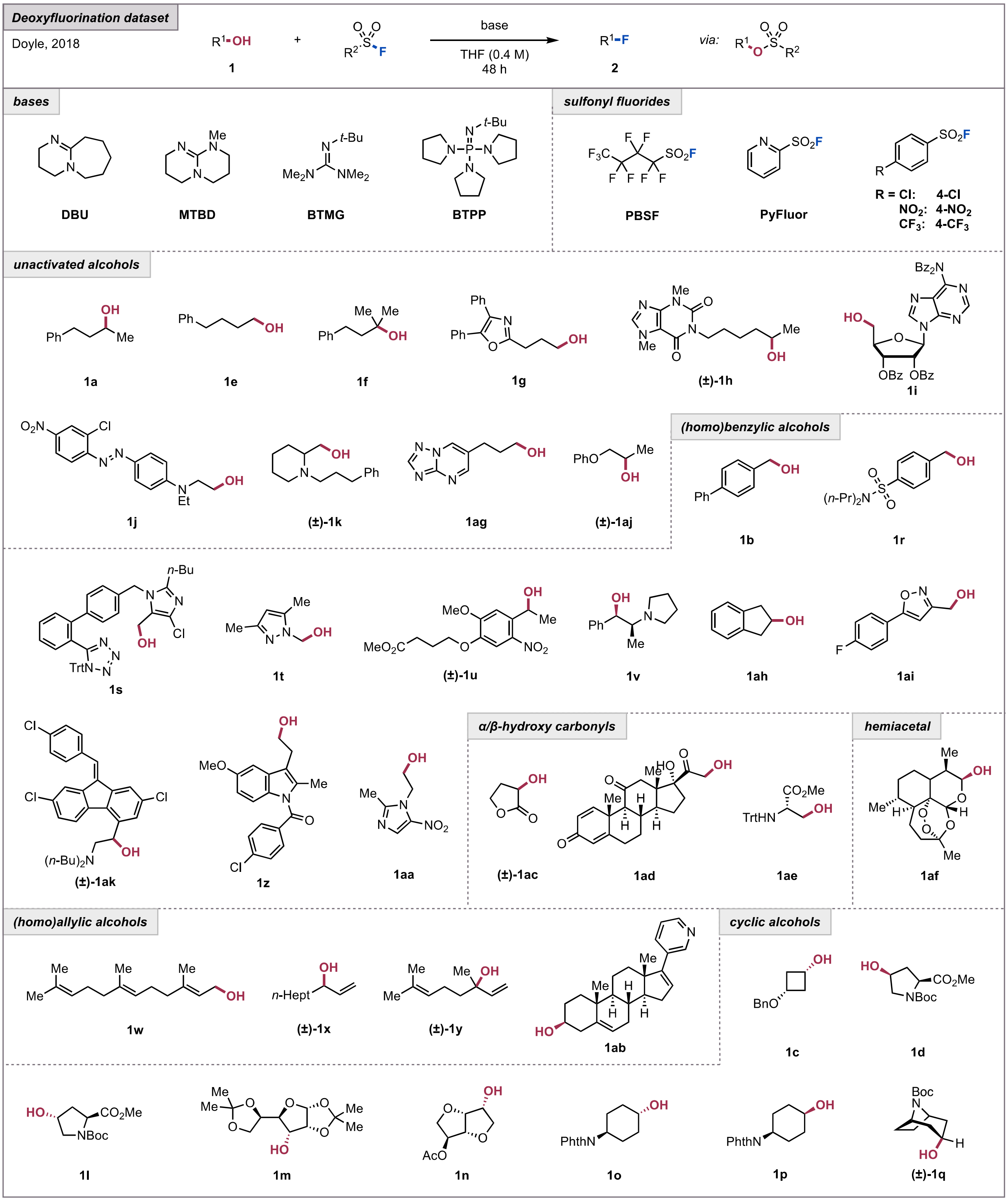


Figure 3. Previously published alcohol deoxyfluorination high-throughput experimentation dataset. Reproduced with permission from ref 14. Copyright 2018 American Chemical Society.

dataset reported by our lab in 2018 as a case study for modeling.[14] The dataset comprises 37 alcohols, including reactive primary, secondary, and benzylic alcohols, as well as unreactive strained cyclic alcohols; 5 sulfonyl fluorides of varying reactivity correlated to the leaving group ability of the corresponding sulfonate ester; and 4 strong amidine/guanidine/phosphazene bases of varying steric hindrance (Figure 3). In the original study, the dataset was used to construct an RF model for yield prediction using a combination of chemically informed features and categorical features describing the different classes of alcohols. Though the model was able to make reasonable predictions for the yields for some out-of-sample alcohols, when evaluated with leave-one-alcohol-out (LOAO) validation, the root-mean-squared error (RMSE) provided by RF did not surpass the RMSE obtained with a one-hot-encoded model that did not use any features.[22] Therefore, the RF model was unable to successfully learn the features pertinent to alcohol reactivity. Herein, we seek to apply our proposed workflow to construct a model that can not only make more accurate yield predictions but also provide mechanistic insight into interaction effects within the dataset.

The first step in our workflow is the analysis of the whole experiment with three-way ANOVA to identify which reaction components and which of their interactions significantly impact the yield. For this dataset, the significance of the main (single component) and two-way interaction (combinations of 2 components) effects can be tested. The three-way interaction (combined effect of three components) could also be tested but would require collection of systematic repetitions for every reaction.[31] The maximal GAM model M_0, which assumes that all effects are significant, can be written as

$$y_{ijk} = a_i + b_j + s_k + (ab)_{ij} + (as)_{ik} + (bs)_{jk} + \epsilon \qquad (2)$$

where a_i, b_j, and s_k are main effects of the alcohol i, base j, and sulfonyl fluoride k, respectively; $(ab)_{ij}$, $(as)_{ik}$, and $(bs)_{jk}$ are interaction effects between the respective reaction components; ϵ is a constant accounting for white noise (Figure 4A).

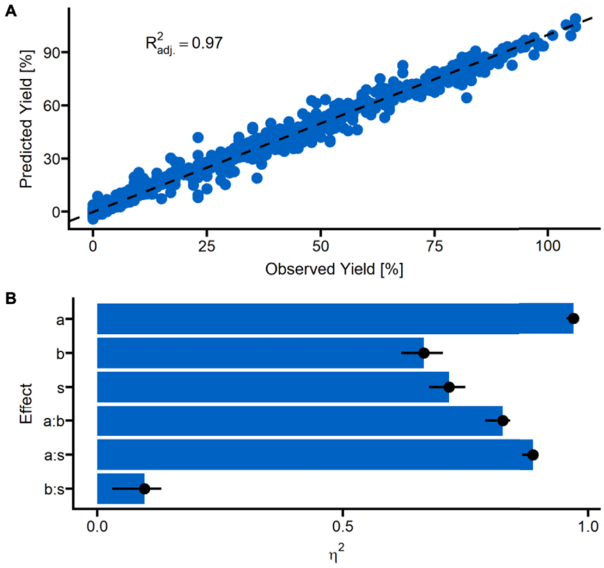


Figure 4. (A) Predicted vs observed yield for the fixed effect model M_0. (B) Partial eta-squared for each term of the model with 95% confidence intervals.

We measured the magnitude of each effect with partial eta-squared (η^2),[32] for which larger values correspond to greater significance, and observed that all main and interaction effects are significant, although the base−sulfonyl fluoride interaction effect is close to the significance threshold (Figure 4B).

Overall, M_0 sufficiently describes the data, with an adjusted R^2 of 0.97 and a residual standard error of 3.9%, consistent with the experimental error of 4.8% observed in the original study.[14] Having determined that we cannot ignore any terms from eq 1, we kept M_0 as a reference for further modeling. However, the features used in M_0 are explicit molecule labels (i.e., one-hot-encoded), such that it cannot be used for out-of-sample predictions.

The second part of the modeling workflow is to build a predictive model that can extrapolate along one or more reaction components. To accomplish this, all terms that involve these components must be replaced with functions of their respective chemical features. In this study, we built a model that extrapolates along only the alcohol component, having 37 alcohols available for model building. We did not attempt to build models that extrapolate into new bases or sulfonyl fluorides, where only 4 and 5 molecules, respectively, are available in the existing HTE dataset. However, we note that the additive nature of GAM would allow for facile incorporation of additional data, such as expansion of the base or sulfonyl fluoride dimensions, without the loss of interpretability. In M_0, three terms use the alcohol labels explicitly: the alcohol−base interaction term $(ab)_{ij}$, the alcohol−sulfonyl fluoride interaction term $(as)_{ik}$, and the alcohol main effect term a_i. Because all three terms correspond to distinct, unrelated phenomena, we modeled them independently with alcohol feature regressions (*vide infra*). To facilitate modeling, we computed DFT features of the alcohols at the M06-2X/def2-TZVP level of theory with implicit THF solvation. Steric and electronic features that may dictate the propensity of alcohols to undergo deoxyfluorination were included, as these can serve to validate or even generate mechanistic hypotheses (see the SI for a full list of features).

Interaction of Alcohol and Base. We used the predictions of the M_0 model to estimate the effect of the base on the deoxyfluorination of each alcohol using the method of estimated marginal means.[33,34] The method is analogous to a control vs treatment experiment, where we select a control base (BTPP) and regard the other bases (DBU, MTBD, and BTMG) as treatments that cause a certain change in yield relative to the control; this change can vary for each alcohol. We chose BTPP, the highest yielding base on average, as a control base to identify regions of alcohol chemical space for which other bases underperform. Though the choice of control is arbitrary, we recommend choosing the highest or lowest performing condition; this simplifies the analysis when investigating the effect of each treatment on the reaction. We then evaluated several atomic and molecular DFT-derived features of the alcohols, generated via our group's AutoQChem workflow,[35] to determine which ones correlate with the presence vs absence of a yield differential relative to the control.

This analysis led to the identification of the buried volume of the α-carbon of the alcohol (V_{bur}), for which a threshold is observed; alcohols with V_{bur} < 0.37 exhibit a large base dependence, while alcohols with V_{bur} > 0.37 show little to no dependence (Figure 5A). Examination of the chemical structures of alcohols in the dataset below the threshold reveals that DBU largely underperforms for primary, unhindered alcohols. This effect is more pronounced for benzylic alcohols within this regime, which suffer from lower yields for both DBU and MTBD. For other more sterically congested and unactivated alcohols, the effect of the base is not significant.

Based on this observation, we hypothesized that the alcohol−base interaction could arise from possible nucleophilic substitution by smaller bases DBU and MTBD. For unhindered primary and/or benzylic substrates, this side

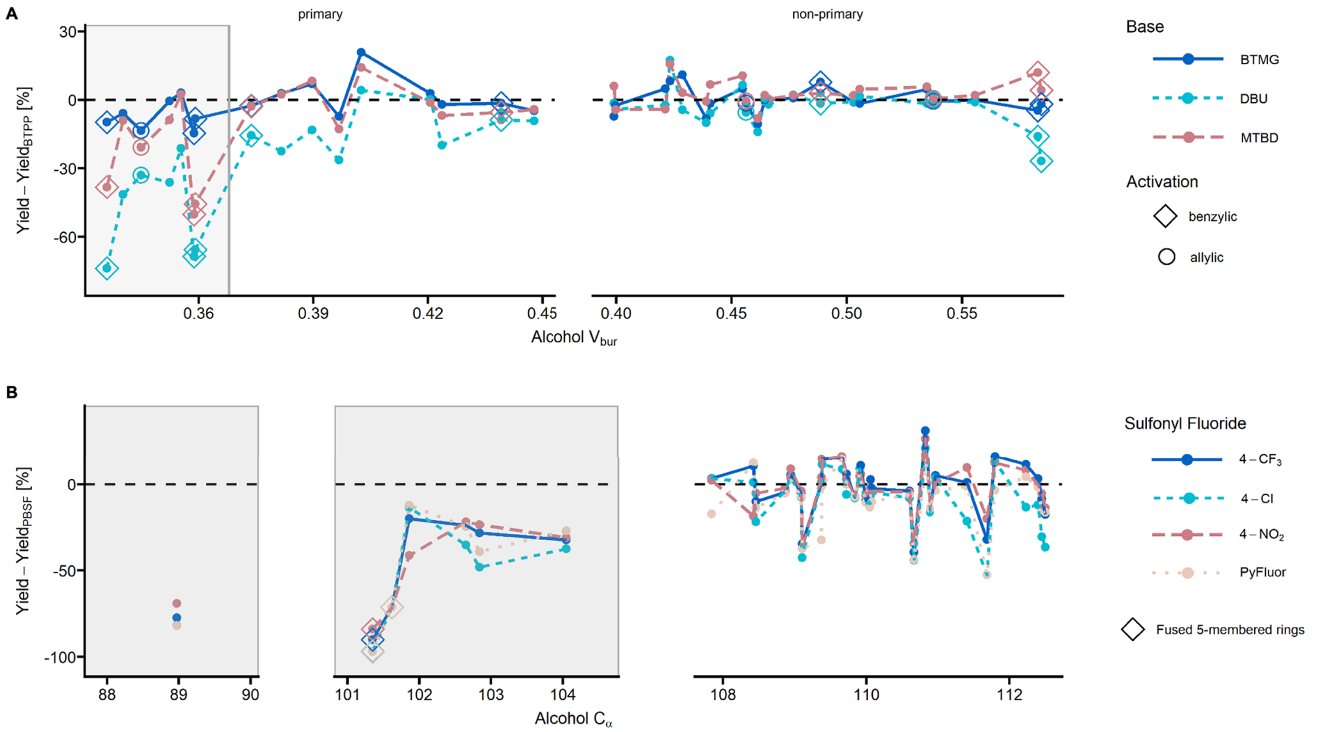
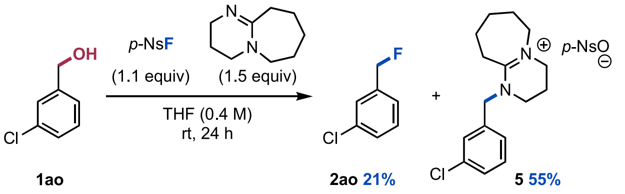


Figure 5. (A) Effect of the base identity on yield as a function of the buried volume of the α-carbon, V_{bur}. Activated alcohols are highlighted with additional markers (diamonds for benzylic, circles for allylic). The plot reveals that activated alcohols with V_{bur} < 0.37 (shaded gray box) exhibit a significant interaction effect. (B) Effect of the sulfonyl fluoride identity on yield for cyclic alcohols as a function of the ring angle measured at the α-carbon, C_α. The plot reveals that strained cyclic alcohols with C_α < 101.8 (shaded gray box) exhibit a significant interaction effect.

reaction may be competitive with the desired fluorination. To experimentally evaluate this hypothesis, we subjected unhindered, activated benzylic alcohol **1ao** to the deoxyfluorination conditions with **4-NO$_2$** and DBU. The amidinium salt **5** was observed in 55% yield, significantly outcompeting the desired benzylic fluoride **2ao** (21% yield) (Scheme 1). This finding validates our mechanistic hypothesis that was based on an alcohol−base interaction identified by the modeling approach.

Scheme 1. Reaction of 1ao with 4-NO$_2$ and DBU



Thus, the alcohol−base interaction can be modeled with base-dependent penalties for unhindered (V_{bur} < 0.37) and benzylic alcohols. Replacing the alcohol−base interaction term, $(ab)_{ij}$, from M_0 with functions of alcohol properties, we arrive at the following model:

$$b^1{}_j(V^i_{\text{bur}} < 0.37) + b^2{}_j(V^i_{\text{bur}} < 0.37\&\text{benzylic}^i) \qquad (3)$$

This case study demonstrates the utility of the modeling approach to identify interactions that suggest underlying chemical phenomena. These can inform experiments that interrogate the mechanistic basis of the identified interactions, ultimately yielding a more complete picture of the reaction of interest.

Interaction of Alcohol and Sulfonyl Fluoride. Next, we again used the M_0 model predictions to compute the effect of the sulfonyl fluoride for each alcohol with estimated marginal means. We used PBSF, the highest yielding sulfonyl fluoride on average, as the control and the others (**4-CF$_3$**, **4-Cl**, **4-NO$_2$**, and PyFluor) as treatments. Through the same correlation analysis, we identified the bond angle measured at the α-carbon position of the alcohol (α) as a chemically meaningful feature that correlates with the magnitude of sulfonyl fluoride dependence. Other features, such as sterimol L, V_{bur}, or C−O bond length, were not correlated to the observed effect (see the SI for full analysis). Closer inspection of the chemical structures revealed that cyclic substrates with significant bond angle contraction (α < 101.8), and therefore significant ring strain, are most influenced by the sulfonyl fluoride identity, with PBSF enabling higher yields of the desired products. Less strained five-membered rings (α > 101.8) also exhibit this interaction effect, albeit of smaller magnitude than the more strained rings (Figure 5B). Thus, we can model the alcohol−sulfonyl fluoride interaction, $(as)_{ik}$, using two sulfonyl fluoride-dependent penalties: for cyclic alcohols with α < 101.8° and for 5-membered rings. The resulting interaction term is

$$s^1_k(\alpha < 101.8) + s^2_k(\text{ring size}^i = 5) \qquad (4)$$

We sought to investigate the origins of the cyclic alcohol−sulfonyl fluoride interaction, initially hypothesizing that the superior leaving group ability of the perfluorobutanesulfonate anion could facilitate an S_N1 mechanism for this more challenging substrate class and explain the greater reactivity observed with PBSF. To evaluate this hypothesis, we first pursued a stereochemical study. We had previously observed that deoxyfluorination of fused bicyclic alcohol **1m** with PBSF resulted in complete inversion. However, since this stereochemical outcome could be due to either sterically restricted S_N1 substitution or an S_N2 mechanism, the same experiment was conducted with diastereomer **1an**. Inversion was again observed, in direct contradiction to our initial hypothesis and consistent with an S_N2 mechanism (Figure 6A).

To evaluate the interaction effect further, we also pursued a kinetic study. However, the deoxyfluorination mechanism presents two challenges toward kinetic analysis. First, the sulfonylation and fluorination steps cannot be decoupled; therefore, the identification of a system where sulfonylation is sufficiently rapid to allow for observation of only the rate of fluorination is imperative. Fortunately, ^{1}H NMR studies with alcohol **1an** revealed nearly instantaneous conversion to sulfonate ester **6**, with slower fluorination to deliver the desired product **2an**, such that the observed reaction rate is equal to the rate of fluorination.[36] The second limitation arises from the concomitant generation of the electrophile (sulfonate ester) and nucleophile (amidine/guanidine hydrogen fluoride); the stoichiometry of these species is therefore set as 1:1 and cannot be manipulated in isolation. To overcome this challenge, we generated the base•HF species independent of the reaction of interest. We identified 4-methoxyphenol as an appropriate sacrificial alcohol which, in the presence of PBSF and BTMG, readily undergoes sulfonylation but not fluorination, thereby generating but not consuming the active fluoride nucleophile (Figure 6B).[36] To probe the dependence of the reaction on the concentration of fluoride, we conducted studies on the reaction of **1an** with 0.8, 0.4, or 0 added equivalents of BTMG•HF (Figure 6C). A positive order dependence on the nucleophile was observed, consistent with the stereochemical data and an S_N2 mechanism. Put together, these experiments suggest that the observed interaction effect for PBSF is due not to the generation of a better leaving group for S_N1 substitution but rather to the generation of a more potent electrophile for bimolecular substitution.

In these studies, we noticed that ~10% of starting material remains unreacted despite nearly instantaneous sulfonate ester formation under standard conditions (1.1 equiv of PBSF, 1.5 equiv of BTMG, and THF (0.4 M)). Further experimentation revealed that BTMG can competitively react with PBSF (see the SI for details), though identification of the adduct has proven elusive. Nevertheless, employing trityl cation as a fluoride scavenger,[37] we observed 21% yield of trityl fluoride **7** in the presence of PBSF and BTMG, providing preliminary evidence that this adduct could also serve as a fluoride source (Figure 6D). This observation could also account for the alcohol−sulfonyl fluoride interaction by enabling access to a more nucleophilic fluoride source for the challenging substitution reaction.

With a clearer mechanistic picture of the observed interaction effect, we sought to enhance the reaction yields for more challenging cyclic substrates. We hypothesized that a larger excess of PBSF and BTMG would enable full sulfonylation of the alcohol while also increasing the concentration of competent fluoride present. Ultimately, we employed 2 equiv of PBSF and 3 equiv of BTMG, leading to a 1.5-fold increase in yield for alcohol **1q** compared to standard conditions (Figure 6E). Overall, this study highlights the role that statistical modeling can play in inspiring new mechanistic questions that facilitate deeper analyses of chemical reactivity and how we can leverage our improved understanding to enable more powerful synthetic transformations.

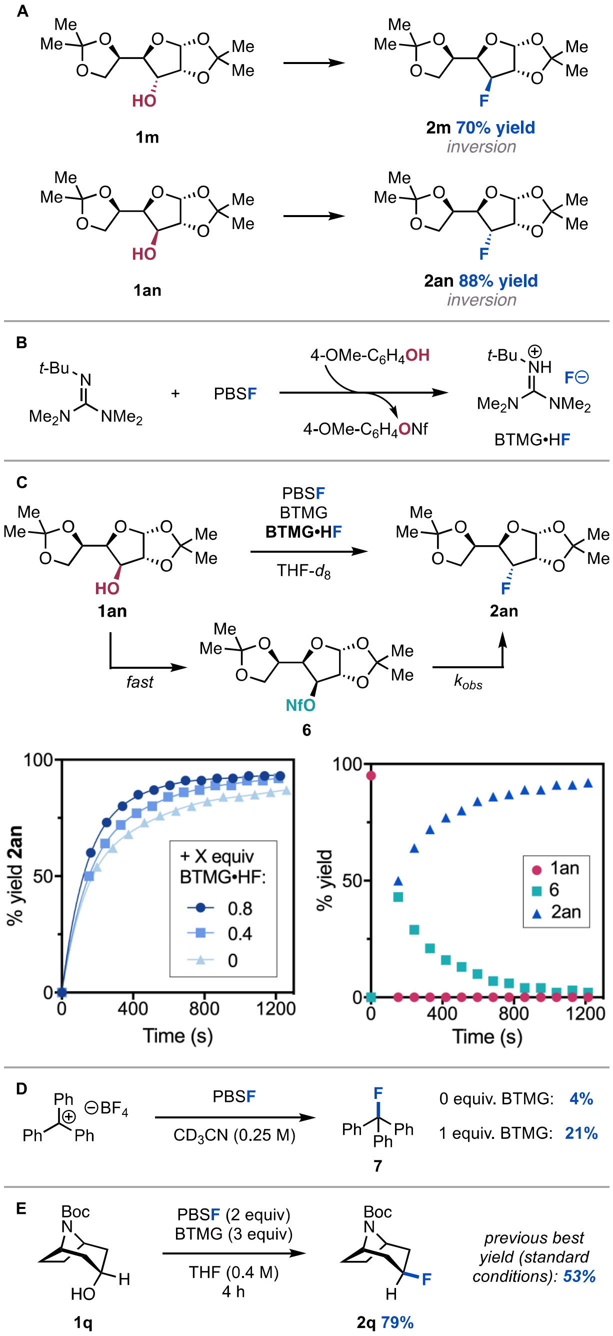


Figure 6. Standard conditions: alcohol (0.1 mmol), PBSF (1.1 equiv), BTMG (1.5 equiv), and THF (0.4 M). (A) Reactions of diastereomers **1m** and **1an** both result in inversion. (B) Full consumption of 4-methoxyphenol was observed. (C) The reaction was monitored at varied [BTMG•HF] and found to exhibit a positive-order dependence (left, PBSF (2 equiv), BTMG (3 equiv), and THF-d_8 (0.2 M)). Representative reaction profile with BTMG•HF (0.4 equiv) (right). (D) The trapping of trityl cation proceeded in 21% in the presence of PBSF and BTMG. (E) Use of PBSF (2 equiv) and BTMG (3 equiv) enabled a 26% increase in yield of product **2q**.

Alcohol Main Effect. To study the effect of the alcohol structure on reactivity, we could not use M_0 marginal means due to the presence of significant interactions with both base and sulfonyl fluoride, which would affect the averages and therefore be uninterpretable according to the principle of marginality.[38] Therefore, we examined the effect of the alcohol itself using the control levels for base (BTPP) and sulfonyl fluoride (PBSF). For primary alcohols, a strong dependence of the yield on V_{bur} was observed in all cases except for allylic alcohol **1w** (Figure 7A). When we modeled all terms determined to be important thus far, namely, V_{bur} dependence for primary alcohols and alcohol interactions with base and sulfonyl fluoride (eqs 2 and 3), as a linear function, the resulting model had an R^2 = 0.5. Thus, plenty of variance remained unexplained. Re-evaluating the dataset, we noticed that allylic alcohols **1x**, **1y**, and **1z** and homoallylic alcohol **1ab** all exhibited slightly lower yields than their less activated counterparts. Including additional penalty terms for allylic and homoallylic alcohols improved the fit of the model to R^2 = 0.65.

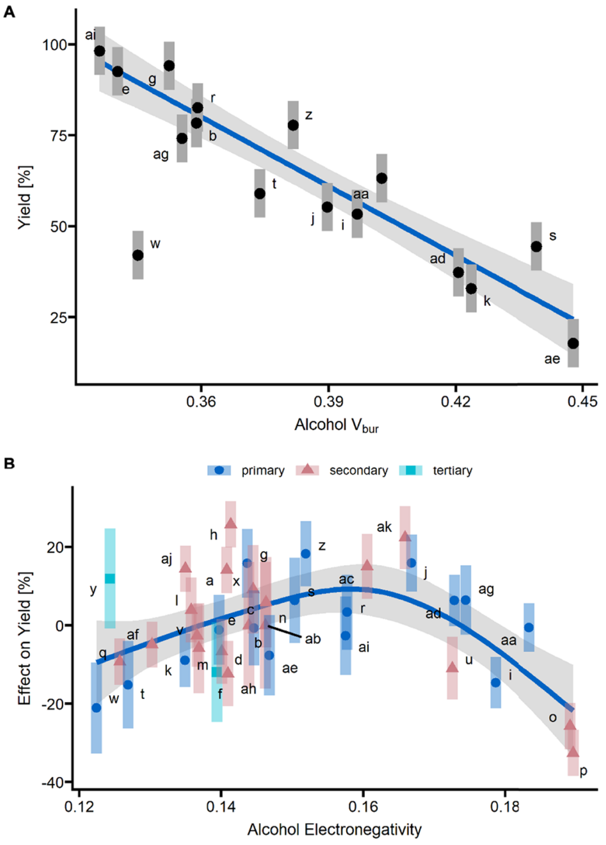


Figure 7. (A) Yield as a function of V_{bur} (linear fit) for primary alcohols with PBSF and BTPP. (B) Residual variance of the interim model as a function of alcohol electronegativity (spline fit). The point labels omit the prefixes: **1** for alcohol (*x*-axis) or **2** for the fluorinated product (*y*-axis).

Seeking further improvement, we built an interim model that added a new term for each alcohol, thereby measuring the remaining residual effect per alcohol on yield (Figure 7B). By plotting the resulting residuals against various alcohol features, we found that a correlation is only observed with electronegativity (EN) of the alcohol, defined as the negative average of HOMO and LUMO energies. Notably, we were unable to identify any other features that substantially improved the fit, prompting us to add an electronegativity term as a spline function to the alcohol main effect model.

We replaced the main effect of the alcohol, a_i, in M_0 with the appropriate functional form:

$$\text{allylic}^i + \text{homoallylic}^i + (V^i_{\text{bur}}\ \&\ \text{primary}^i) + f^2(\text{EN}) \tag{5}$$

The final model, M_1, with all alcohol-dependent terms replaced with functions of alcohol features using eqs 3−5 had R^2 = 0.83 with a standard error of 11%.

Leave-One-Alcohol-Out Validation. We evaluated the robustness and performance of the M_1 model with leave-one-alcohol-out (LOAO) validation, using the RF model built using all available features as a comparison baseline. The training and LOAO validation errors for both models are shown in Figure 8.

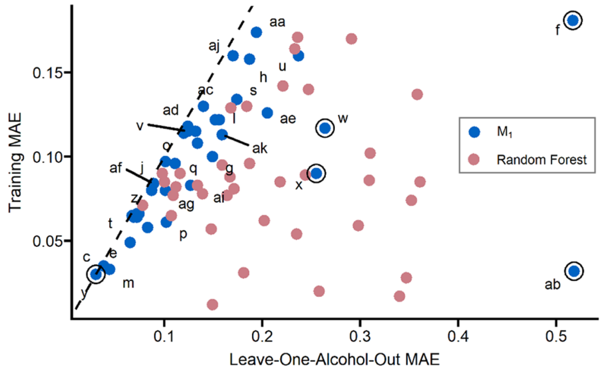


Figure 8. Comparison of leave-one-alcohol-out validation errors and training errors for the M_1 and random forest models. Circled substrates represent low yielding alcohol classes that are fitted poorly by the model (tertiary, allylic, and homoallylic alcohols).

First, we found that the mean absolute validation error (MAE) and root-mean-squared error (RMSE) are both largely reduced for M_1 (15% MAE, 17% RMSE) compared to the RF model (20% MAE, 22% RMSE). Second, the validation errors for M_1 are much closer to the training errors (difference of 5%), implying much less overfitting than the RF model, which has a much larger discrepancy between its training and validation errors (difference of 14%). Similar comparisons were made with LASSO and stepwise regression, which both underperformed compared to M_1. Furthermore, neither RF, LASSO, nor stepwise regression was able to learn the observed interaction effect (see the SI for details).

This represents a marked improvement over the original RF model: overfitting, which occurs when too much is inferred from the training data, is a known cause of poor generalization of ML models. The challenge of overfitting has been highlighted previously in the context of HTE modeling.[22]

Although M_1 performed much better in cross validation, we note that it falls short for tertiary alcohols (**1f** and **1y**) and allylic/homoallylic alcohols (**1x**, **1y**, **1w**, and **1ab**). More data are required to stabilize the model in this region of chemical space. However, the generally poor reactivity of these substrate classes presents a significant obstacle toward collecting a well-rounded dataset.

Model Generalization. Although the M_1 model adequately fits the data from the HTE, we sought to test its generalization with a new set of alcohols (Figure 9A). We first examined the overall yield prediction accuracy. Furthermore, since the M_1 model comprises three separate models (two interaction models for alcohol−base and alcohol−sulfonyl fluoride interactions and one model for the main effect of the alcohol), we also studied whether the observed interactions

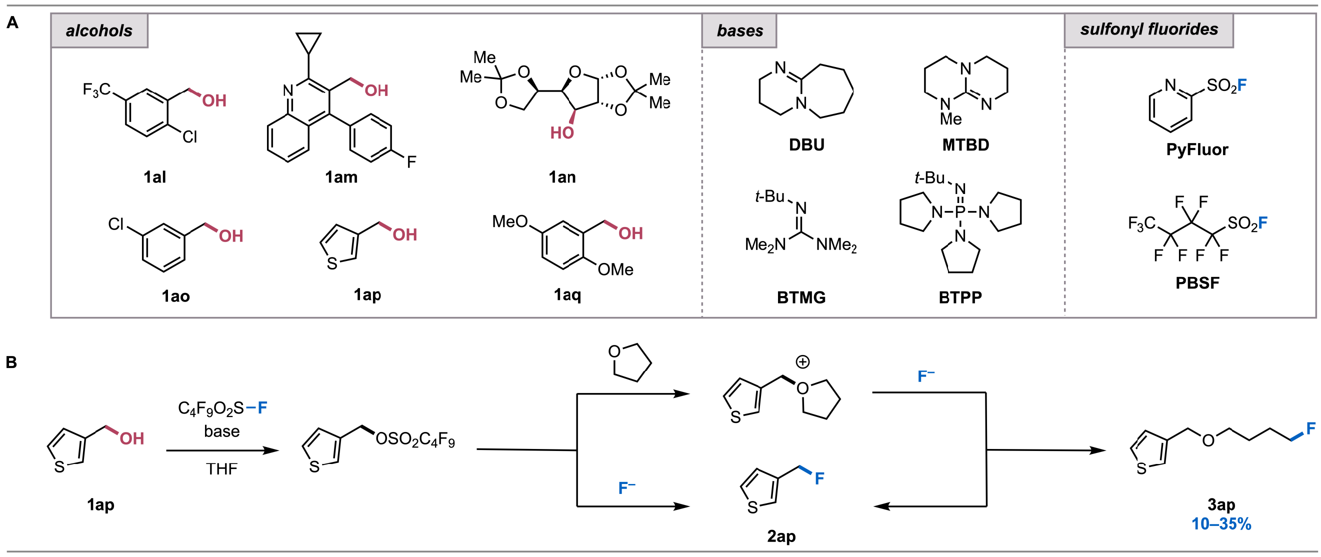


Figure 9. (A) The validation dataset includes 6 alcohols (5 benzylic and 1 cyclic), 4 bases, and 2 sulfonyl fluorides. (B) Formation of the THF-functionalized byproduct **3ap** likely results from competitive nucleophilic attack by THF. (C) Predicted and observed yield for M_1 and random forest models for the set of 7 validation alcohols. In the lower two panels, the data points for alcohols **1ap** and **1aq** correspond to the sum of desired benzylic and undesired solvent-incorporated products; the dashed line and desaturated points indicate the yield of the desired benzylic fluoride alone.

with base and sulfonyl fluoride persist for additional alcohols in the respective regions of chemical space.

Yield Prediction Accuracy. For the new validation set of alcohols, yield predictions of the M_1 model are much improved (MAE = 13%, RMSE = 17%) compared to the RF model built with all available features (MAE = 18%, RMSE = 21%). The predictions of the M_1 model (Figure 10A) are better correlated with the experimental yields, while the RF predictions (Figure 10B) center around average yield values. Nevertheless, the M_1 model still leaves room for improvement in prediction accuracy: the performance of **1al** and **1am** is underestimated, which is likely due to an unmodeled interaction between hindered primary benzylic alcohols and base (*vide infra*).

On the other hand, the performance of **1ap** and **1aq** is overestimated, especially with PBSF. Analyzing reactions of these alcohols with PBSF revealed the formation of undesired fluorinated products arising from competitive THF substitution followed by fluoride attack (Figure 9B). Interestingly, across the entire validation dataset, THF incorporation is only observed for **1ap** (10−35%) and **1aq** (5−16%), likely due to the unhindered and electronically activated nature of these alcohols.[39]

Knowledge of this side reaction was used to further improve model performance by replacing the yield of benzylic fluoride with the combined yields of benzylic (desired) and aliphatic (undesired) fluoride products. This resulted in enhanced model performance that amounted to a ~2% decrease in MAE and a ~3% decrease in RMSE for both the M_1 (Figure 10C) and RF (Figure 10D) models. This analysis highlights the challenge of using yield as the only input for an ML model: the rate and selectivity of the reaction, which are essential reactivity considerations, inevitably remain unaccounted for.

Validation of Alcohol−Base and Alcohol−Sulfonyl Fluoride Interaction Effects. For primary and benzylic alcohols (**1al**, **1am**, **1ao**, **1ap**, and **1aq**, V_{bur} < 0.37), we observed a base dependence analogous to that observed in the training data. The validation data for bicyclic alcohol **1an** are also in agreement with the original data, exhibiting a strong sulfonyl fluoride dependence. For the three *ortho*-substituted benzylic alcohols (**1al**, **1am**, and **1aq**), however, substantial base dependence is observed despite their high V_{bur} values (>0.37). Upon further analysis, we found that more electronegative benzylic alcohols (**1ao**, **1al**, and **1am**) are more sensitive to the base identity than their less electronegative counterparts (**1ap** and **1aq**) in our validation set (Figure 10E). In the original HTE dataset, two primary benzylic alcohols have V_{bur} above the hypothetical V_{bur} threshold (**1s** and **1t**), and neither of them showed base dependence. The alcohol **1t** has very low electronegativity, suggesting that electronegativity is needed for the observation of the interaction. The alcohol **1s** has high electronegativity (Figure 10F), though the lack of interaction (*vide supra*, Figure 7) could be attributed to very high V_{bur}. However, additional data are required to further elucidate the interplay between alcohol electronegativity and its V_{bur} for primary benzylic alcohols. Nevertheless, while constructing the model, we successfully identified a class of alcohols with significant interactions, as well as key features that allowed for further exploration.

CONCLUSIONS

Herein, we presented a statistical modeling approach for determining interactions within an HTE reaction dataset. First, the reaction yield was decomposed into main and interaction effects using ANOVA, and second, the individual effects were modeled with feature-based regressions. In this approach, we respect the structure of HTE datasets and aim to combine advanced modeling with chemical expertise. The first ANOVA step is trivial and can immediately be performed on any HTE dataset. However, the process of identifying key features that correlate with interaction patterns cannot be fully automated, and a level of chemical expertise is still required to build a model that is plausible from a chemical standpoint. A direct benefit of a chemist-in-the-loop approach such as ours is the ability to form mechanistic hypotheses that could explain the

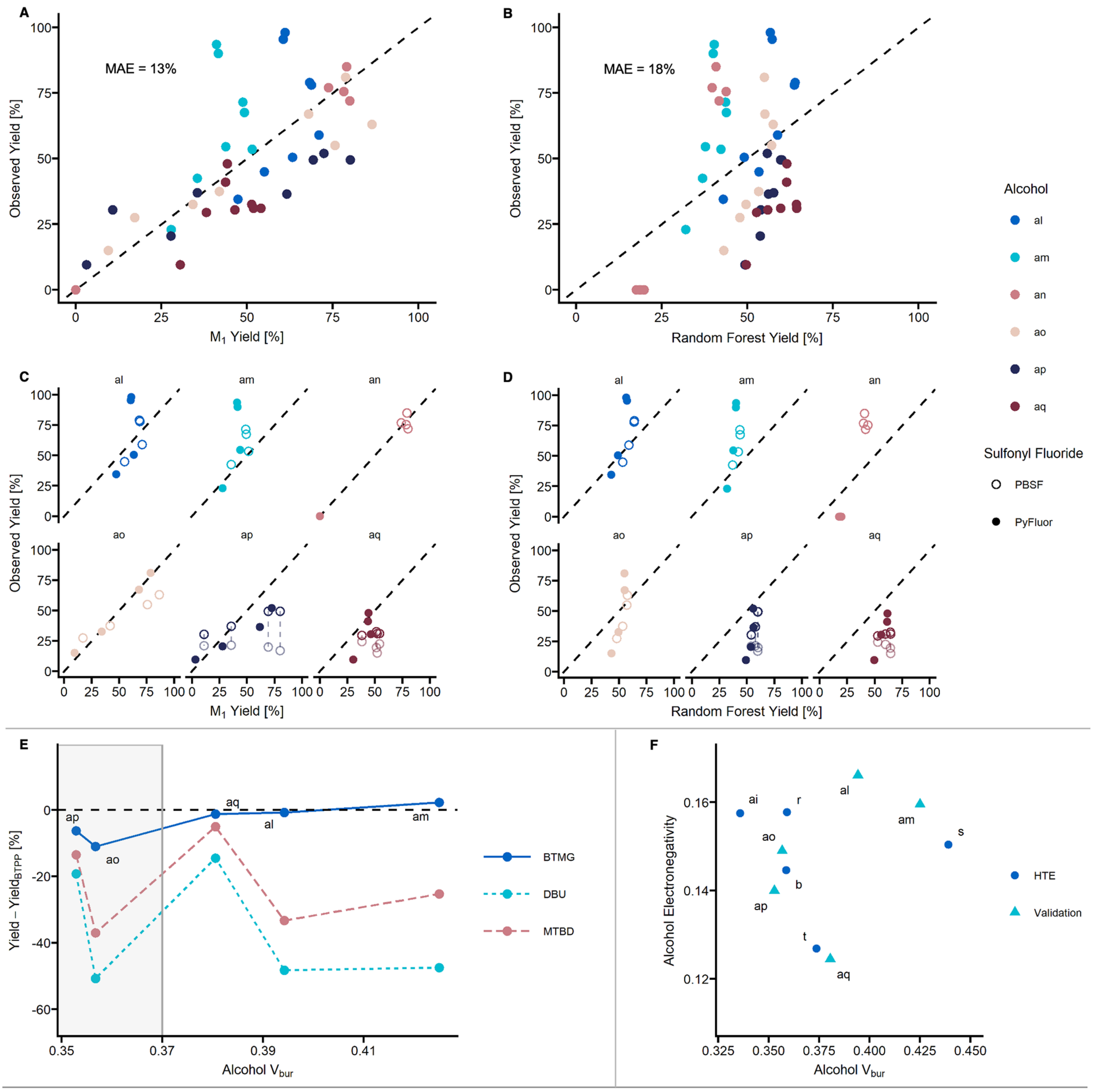


Figure 10. Predicted and observed yield is shown for the set of 7 validation alcohols for (A) M_1 and (B) random forest models. Predicted and observed yield, split by alcohol, is shown for (C) M_1 and (D) random forest models. In panels A and B, the data points for alcohols **1ap** and **1aq** correspond to the yield of the benzylic fluoride alone. In panels C and D, the data points for alcohols **1ap** and **1aq** correspond to the sum of the desired benzylic fluoride and the undesired aliphatic fluoride byproduct. (E) Effect of the base identity on yield as a function of the buried volume of the α-carbon, V_{bur}, for primary and benzylic validation alcohols. (F) Electronegativity of primary and benzylic alcohols from the HTE and validation datasets.

observed effects, thereby streamlining the process of model validation (i.e., confirming or disproving the existence of specific effects with new substrates). An indirect benefit is a more interpretable model with better generalizability to out-of-sample substrates, which we demonstrated on our previously published deoxyfluorination dataset, largely improving the MAE and RMSE compared to the original random forest algorithm (by ~25%).

We find statistical modeling to be an essential tool for studying this deoxyfluorination dataset, allowing for inspection and modeling of the two-way interactions. However, complex multiway interactions (three-way or four-way), which may be observed in larger HTE datasets, may be difficult to capture in a functional form beyond specifying that an interaction is present and significant. Nevertheless, with this initial example, we have demonstrated that decomposition of the signal into main and interaction effects could serve as a useful tool for identifying areas of chemical space where a reaction is particularly sensitive to the identity of reaction components. The identification of these sensitivities could inspire new mechanistic hypotheses, which can be tested experimentally to afford an improved understanding of the reaction of interest and chemical reactivity more broadly.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/jacs.2c13093.

> Experimental procedures, experimental data, and characterization and spectral data (PDF)

Accession Codes

CCDC 2205030−2205031 contain the supplementary crystallographic data for this paper. These data can be obtained free of charge via www.ccdc.cam.ac.uk/data_request/cif, or by emailing data_request@ccdc.cam.ac.uk, or by contacting The Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, UK; fax: +44 1223 336033.

AUTHOR INFORMATION

Corresponding Author

Abigail G. Doyle − *Department of Chemistry, Princeton University, Princeton, New Jersey 08544, United States; Department of Chemistry and Biochemistry, University of California, Los Angeles, Los Angeles, California 90095, United States;* orcid.org/0000-0002-6641-0833; Email: agdoyle@chem.ucla.edu

Authors

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

Shivaani S. Gandhi − *Department of Chemistry, Princeton University, Princeton, New Jersey 08544, United States; Department of Chemistry and Biochemistry, University of California, Los Angeles, Los Angeles, California 90095, United States;* orcid.org/0000-0003-1825-5450

Complete contact information is available at: https://pubs.acs.org/10.1021/jacs.2c13093

Author Contributions

[§]A.M.Ż. and S.S.G. contributed equally.

Notes

The authors declare no competing financial interest.

All code and raw data used for modeling can be found on GitHub via https://github.com/doyle-lab-ucla/deoxyfluorination_modeling.

ACKNOWLEDGMENTS

The authors thank Dr. Saeed Khan, Samuel Newman-Stonebraker, and Judah Raab for assistance with X-ray crystal structure determinations, Dr. Ta-Chung Ong and Dr. Robert Taylor for assistance with NMR experiments, and the mass spectrometry facility at the University of California, Irvine for HRMS measurements. This program has been funded through the generous contributions of the National Science Foundation and the CCI Center for Computer Assisted Synthesis (CHE-1565983 and CHE-1925607) and the Dreyfus Program for Machine Learning in the Chemical Sciences and Engineering. A.M.Ż. gratefully acknowledges financial support from the Schmidt DataX Fund at Princeton University, made possible through a major gift from the Schmidt Futures Foundation.

REFERENCES

(1) Baskin, I. I.; Madzhidov, T. I.; Antipin, I. S.; Varnek, A. A. Artificial Intelligence in Synthetic Chemistry: Achievements and Prospects. *Russ. Chem. Rev.* **2017**, *86*, 1127−1156.

(2) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine Learning for Molecular and Materials Science. *Nature* **2018**, *559*, 547−555.

(3) Strieth-Kalthoff, F.; Sandfort, F.; Segler, M. H. S.; Glorius, F. Machine Learning the Ropes: Principles, Applications and Directions in Synthetic Chemistry. *Chem. Soc. Rev.* **2020**, *49*, 6154−6168.

(4) Williams, W. L.; Zeng, L.; Gensch, T.; Sigman, M. S.; Doyle, A. G.; Anslyn, E. V. The Evolution of Data-Driven Modeling in Organic Chemistry. *ACS Cent. Sci.* **2021**, *7*, 1622−1637.

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

(6) Lowe, D. M. Chemical Reactions from US Patents, 1976−Sep 2016. figshare. Dataset.

(7) Lowe, D. M. Extraction of Chemical Structures and Reactions from the Literature, PhD Thesis, University of Cambridge: Cambridge, UK, 2012.

(8) Strieth-Kalthoff, F.; Sandfort, F.; Kühnemund, M.; Schäfer, F. R.; Kuchen, H.; Glorius, F. Machine Learning for Chemical Reactivity: The Importance of Failed Experiments. *Angew. Chem., Int. Ed.* **2022**, *61*, No. e202204647.

(9) 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*, 1465−1476.

(10) Maser, M. R.; Cui, A. Y.; Ryou, S.; DeLano, T. J.; Yue, Y.; Reisman, S. E. Multilabel Classification Models for the Prediction of Cross-Coupling Reaction Conditions. *J. Chem. Inf. Model.* **2021**, *61*, 156−166.

(11) Coley, C. W.; Jin, W.; Rogers, L.; Jamison, T. F.; Jaakkola, T. S.; Green, W. H.; Barzilay, R.; Jensen, K. F. A Graph-Convolutional Neural Network Model for the Prediction of Chemical Reactivity. *Chem. Sci.* **2018**, *10*, 370−377.

(12) Mennen, S. M.; Alhambra, C.; Allen, C. L.; Barberis, M.; Berritt, S.; Brandt, T. A.; Campbell, A. D.; Castañón, J.; Cherney, A. H.; Christensen, M.; Damon, D. B.; Diego, J. E.; García-Cerrada, S.; García-Losada, P.; Haro, R.; Janey, J.; Leitch, D. C.; Li, L.; Liu, F.; Lobben, P. C.; MacMillan, D. W. C.; Magano, J.; McInturff, E.; Monfette, S.; Post, R. J.; Schultz, D.; Sitter, B. J.; Stevens, J. M.; Strambeanu, I. I.; Twilton, J.; Wang, K.; Zajac, M. A. The Evolution of High-Throughput Experimentation in Pharmaceutical Development and Perspectives on the Future. *Org. Proc. Res. Dev.* **2019**, *23*, 1213−1242.

(13) 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.

(14) 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.

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

(16) Henle, J. J.; Zahrt, A. F.; Rose, B. T.; Darrow, W. T.; Wang, Y.; Denmark, S. E. Development of a Computer-Guided Workflow for Catalyst Optimization. Descriptor Validation, Subset Selection, and Training Set Analysis. *J. Am. Chem. Soc.* **2020**, *142*, 11578−11592.

(17) Schwaller, P.; Vaucher, A. C.; Laino, T.; Reymond, J.-L. Prediction of Chemical Reaction Yields Using Deep Learning. *Mach. Learn.: Sci. Technol.* **2021**, *2*, No. 015016.

(18) Probst, D.; Schwaller, P.; Reymond, J.-L. Reaction Classification and Yield Prediction Using the Differential Reaction Fingerprint DRFP. *Digital Discovery* **2022**, *1*, 91−97.

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

(20) 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*, No. eaat8763.

(21) 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.

(22) Żurański, A. M.; Alvarado, J. I. M.; Shields, B. J.; Doyle, A. G. Predicting Reaction Yields via Supervised Learning. *Acc. Chem. Res.* **2021**, *54*, 1856−1865.

(23) Skoraczyński, G.; Dittwald, P.; Miasojedow, B.; Szymkuć, S.; Gajewska, E. P.; Grzybowski, B. A.; Gambin, A. Predicting the Outcomes of Organic Reactions via Machine Learning: Are Current Descriptors Sufficient? *Sci. Rep.* **2017**, *7*, 3582.

(24) Girden, E. R. *ANOVA: Repeated Measures*. Sage Publications 1992.

(25) Domingos, P. A Few Useful Things to Know about Machine Learning. *Commun. ACM* **2012**, *55*, 78−87.

(26) Altman, N. S. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *Am. Stat.* **2012**, *46*, 175−185.

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

(28) Krishnapuram, B.; Shah, M.; Smola, A.; Aggarwal, C.; Shen, D.; Rastogi, R.; Chen, T.; Guestrin, C. XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 2016, 785−794.

(29) Topliss, J. G.; Costello, R. J. Chance Correlations in Structure-Activity Studies Using Multiple Regression Analysis. *J. Med. Chem.* **1972**, *15*, 1066−1068.

(30) Kearnes, S. M.; Maser, M. R.; Wleklinski, M.; Kast, A.; Doyle, A. G.; Dreher, S. D.; Hawkins, J. M.; Jensen, K. F.; Coley, C. W. The Open Reaction Database. *J. Am. Chem. Soc.* **2021**, *143*, 18820−18826.

(31) Replicates were performed for best performing conditions for each alcohol but not for the entire condition space.

(32) Lakens, D. Calculating and Reporting Effect Sizes to Facilitate Cumulative Science: A Practical Primer for t-Tests and ANOVAs. *Front. Psychol.* **2013**, *4*, 863.

(33) We use estimated marginal means with a treatment vs. control contrast using 'emmeans' R package (https://cran.r-project.org/web/packages/emmeans).

(34) Searle, S. R.; Speed, F. M.; Milliken, G. A. Population Marginal Means in the Linear Model: An Alternative to Least Squares Means. *Am. Stat.* **1980**, *34*, 216−221.

(35) Żurański, A. M.; Wang, J. Y.; Shields, B. J.; Doyle, A. G. Auto-QChem: An Automated Workflow for the Generation and Storage of DFT Calculations for Organic Molecules. *React. Chem. Eng.* **2022**, *7*, 1276−1284.

(36) Use of PyFluor also leads to fast sulfonylation of this alcohol, but no fluorination is observed even at elevated temperatures (up to 50 °C). This observation is consistent with the interaction effect being ascribed to the substitution reaction, not the initial sulfonylation step.

(37) Wyss, C. M.; Tate, B. K.; Bacsa, J.; Wieliczko, M.; Sadighi, J. P. Dinuclear μ-Fluoro Cations of Copper, Silver and Gold. *Polyhedron* **2014**, *84*, 87−95.

(38) Nelder, J. A. A Reformulation of Linear Models. *J. R. Stat. Soc. Ser. Gen.* **1977**, *140*, 48−63.

(39) THF incorporation was also observed for **1am** in trace yields (4−6%).