nature research

https://doi.org/10.1038/s41586-021-03213-y

## Supplementary information

# Bayesian reaction optimization as a tool for chemical synthesis

In the format provided by the authors and unedited

# Supplementary Information

## Bayesian Reaction Optimization as A Tool for Chemical Synthesis

Shields, Benjamin J.[‡]; Stevens, Jason[†]; Li, Jun[†]; Parasram, Marvin[‡]; Damani, Farhan[⁑], Martinez Alvarado, Jesus[‡]; Janey, Jacob[†]; Adams, Ryan P.[⁑*]; Doyle, Abigail G.[‡*]

[†]Chemical Process Development, Bristol-Myers Squibb, 1 Squibb Dr, New Brunswick, New Jersey 08901, USA. [‡]Department of Chemistry, Princeton University, Princeton, New Jersey 08544, USA [⁑]Department of Computer Science, Princeton University, Princeton, New Jersey 08544, USA

correspondence to: agdoyle@princeton.edu and rpa@princeton.edu

## Table of Contents

## List of Figures

## List of Tables

## I. Software

Two software packages and one web application were written to support this work. The first, *auto-qchem*, was written to facilitate high-throughput computational chemistry and reaction featurization. The second, *edbo*, was written as a user-friendly implementation of Bayesian optimization. The web application, *EvML*, was written to collect user data for comparisons with human expert performance. Table S1 delineates where to find each component required to reproduce this work.

Note: All GitHub pages *will be* made public upon publication. Prior to publication, these resources are available upon request.

| Project Component | Repository | Page Link |
|---|---|---|
| Automated DFT calculations | *auto-qchem* | Main page |
| Generating input files | *auto-qchem* | Gaussian input |
| Raw DFT output files | *auto-qchem* | Data |
| Processing DFT output | *auto-qchem* | Descriptor generation |
| Extended buried volume | *auto-qchem* | Occupied volume |
| Bayesian reaction optimization | *edbo* | Bayesian Optimization |
| Data sets & representations | *edbo* | Data |
| Examples & experiments | *edbo* | Experiments |
| Reaction Optimization Game | *EvML* | Game & Data |

**Table S1.** Index of software and data resources. Note: All GitHub pages will be made public upon publication.

*auto-qchem*: Quantum mechanical descriptor generation software including: (1) an automation framework for running conformational analysis and generating Gaussian jobs from SMILES strings, coordinate files, or ChemDraw files. (2) Descriptor extraction scripts to process Gaussian LOG files for relevant chemical features. (3) Descriptor computation code which generates new descriptors from the computational output. This was first implemented in *Mathematica*[1] and then

translated to an open-source python package.[2] See GitHub for usage instructions, source code, documentation, and examples.

*edbo*: Experimental Design via Bayesian Optimization. Bayesian optimization software written in python.[3] Enables the iterative selection of experiments in a way that balances exploitation of information and exploration of the search space. A general implementation of BO geared towards optimization of chemical reactions. See GitHub for installation instructions, source code, documentation, and examples. For best performance in your application see the reaction representation section for information about priors optimized for different encodings.

*EvML*: Expert versus Machine Learning. Reaction optimization game written as an R/shiny application and hosted on an Amazon Web Services server.[4] The reaction optimization game was used to compare the performance of BO to that of human experts in industry and academia. See GitHub for server setup instructions, source code, and tabulated results.

## II. Development Data

Palladium-catalyzed cross coupling data for Suzuki-Miyaura[5] and Buchwald-Hartwig[6] reactions (**2a**–**e**) were used as objectives for optimizer development. The raw literature data consisted of excel files containing the identities of reaction components in each experiment and were provided in the corresponding supplementary information files. These data were pre-processed to include only experiments relevant to optimization (e.g. exclude controls) and re-tabulated as csv files containing SMILES strings for each molecular component and the corresponding reaction yields. All cleaned data and reaction encodings can be found in the examples on GitHub.[3]

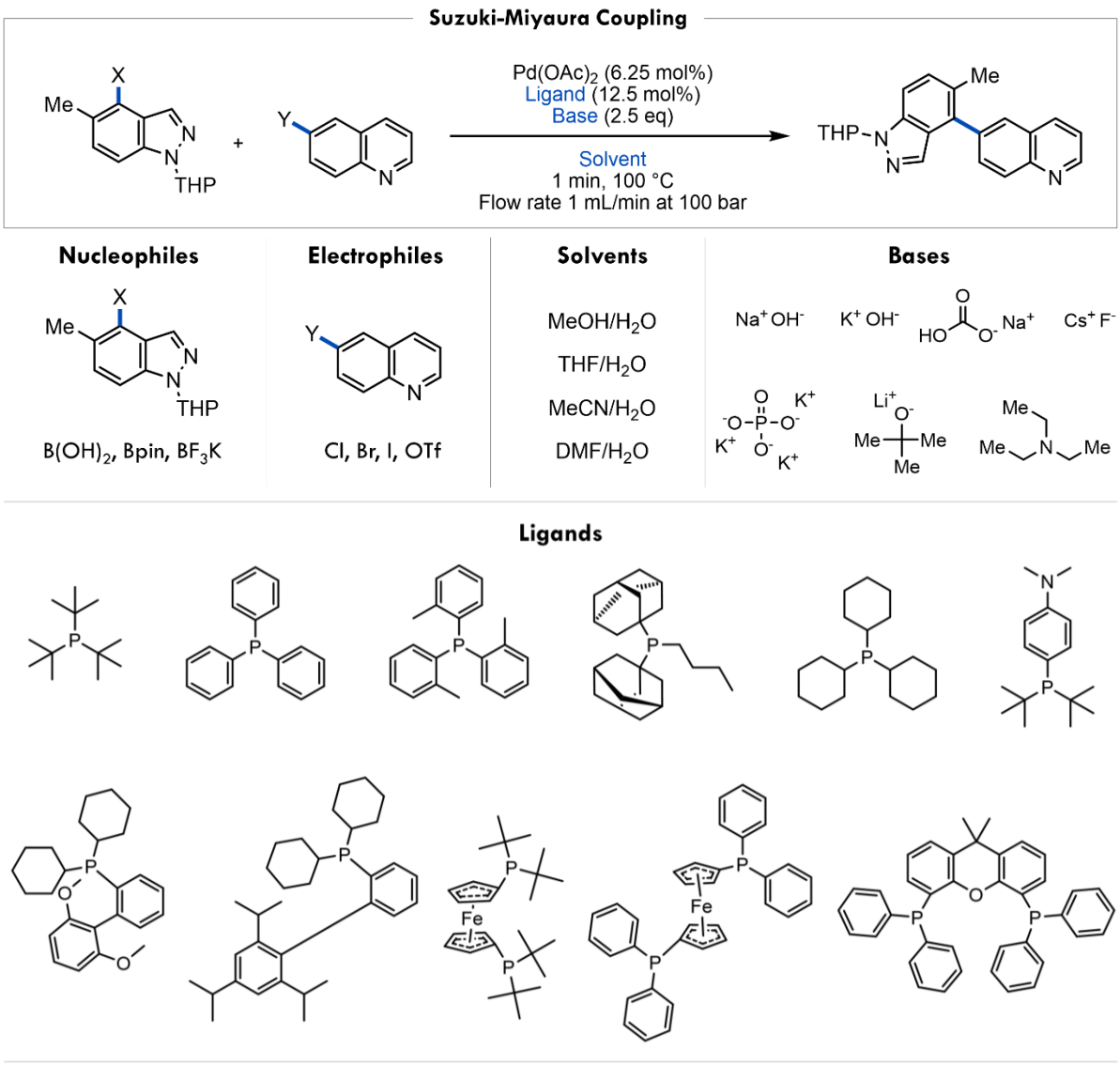



**Figure S1.** Reaction scheme for Suzuki reaction **1**.

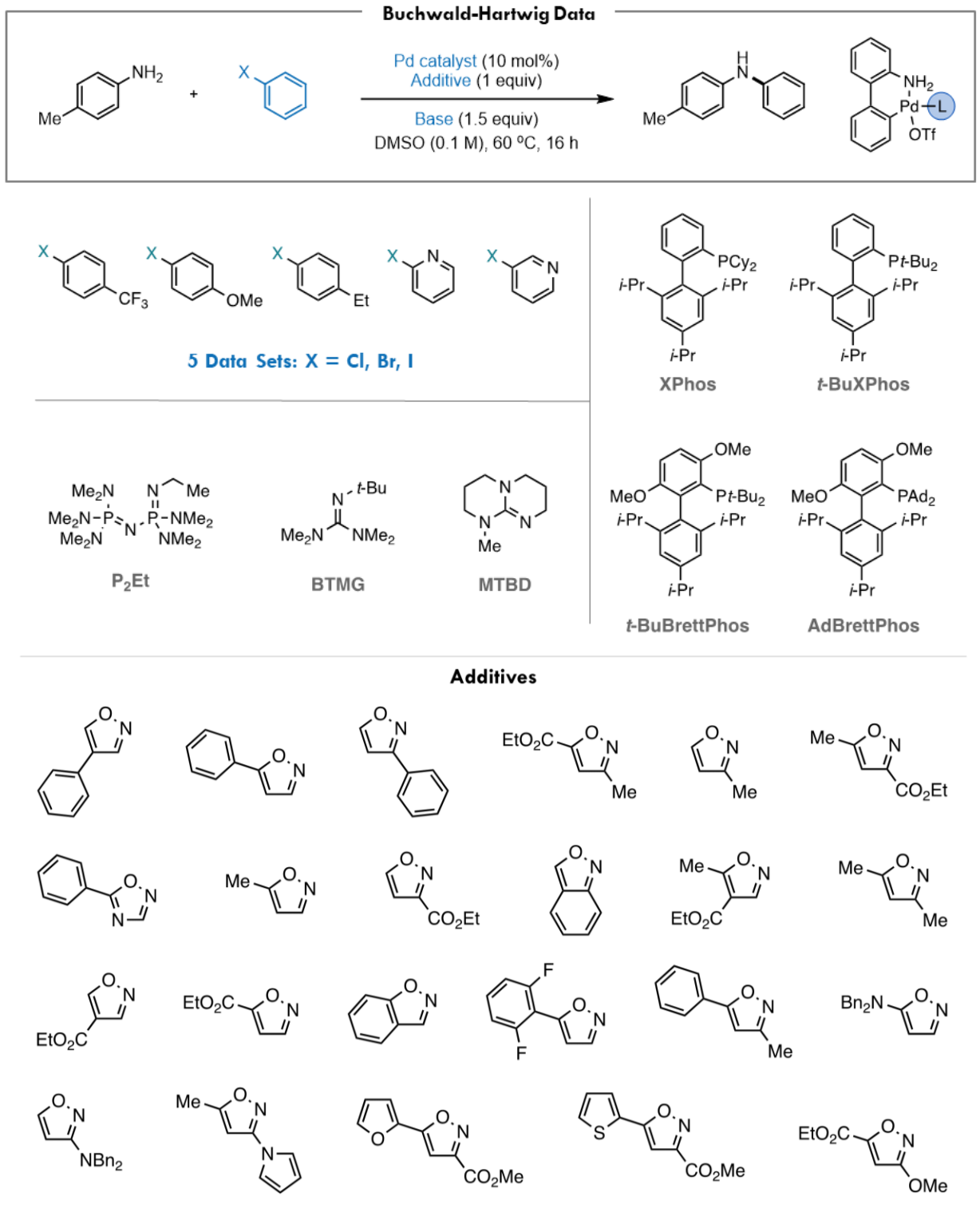



**Figure S2.** Reaction schemes for aryl amination reactions **2a-e**.

## III. Reaction Representations

Descriptor sets. All descriptor sets utilized in this work can be found on the *edbo* GitHub site.[3] Source code and descriptor computation/extraction scripts can be found on the *auto-QChem* GitHub site.[1]

DFT output files. Gaussian 16 output files generated via *auto-QChem* can be found on the GitHub page under examples.[1]

Representations. It is well established that the art of feature engineering can play a critical role in the development of predictive models.[7] However, rather than developing bespoke descriptors for each reaction we imagined that a general, fully automated, approach to featurization would render BO more straightforward to apply by non-experts. Thus, we investigated reaction representations based on Density Functional Theory (DFT), Mordred fingerprints, and one-hot-encoding (OHE).

Feature selection. While explicit feature selection and transformation can offer significant improvements, here we focus on treatments which enable the automatic selection of relevant features.[7] We believe that abstraction of this process, via implicit feature selection, facilitates less biased representations in the absence of *a priori* knowledge and a greater degree of accessibility to users with no modeling experience. However, we emphasize that in the wild, tailored feature development is entirely compatible with our approach. The specifics of encoding and feature selection are left entirely in the hands of users.

Descriptor processing pipeline. Each descriptor set was treated by (1) removing or one-hot-encoding non-numeric features, (2) decorrelation by removing highly correlated features (Pearson correlation coefficient > 0.95) with any other selected descriptor, and (3) normalizing each feature on the unit hypercube. The standardization of responses (zero mean and unit variance) is handled internally by the software. Normalization of descriptors is important for GP model performance, for example because the priors were optimized for the [0,1] interval (*vide infra*). However, it is not important for the RF surrogate model.

One-hot-encoding. One-hot-encoding (OHE) is the simplest way to represent categorical reaction data. Here the presence or absence of a given component is represented by a 1 or a 0, respectively (Figure S3). We independently optimized GP priors for OHE representations using multi-objective self-optimization (see surrogate model section). In table S2 we present the optimized GP priors for this representation. We provide functions for generating OHE representations in the feature utilities module of *edbo*.

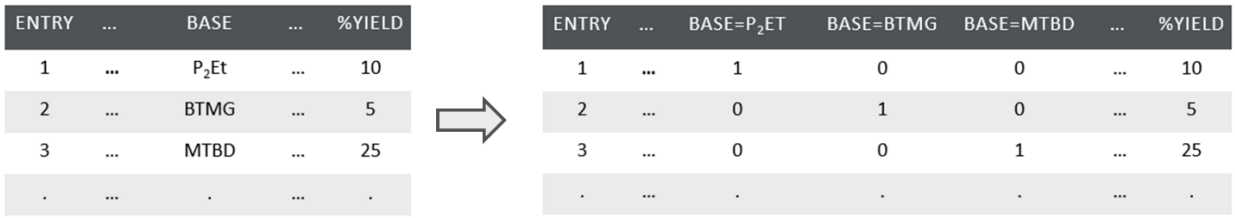



**Figure S3.** One-hot-encoding.

| Hyperparameter | Prior Distribution | Concentration | Rate | Mode |
|---|---|---|---|---|
| length scale | Gamma | 3.0 | 1.0 | 2.0 |
| output scale | Gamma | 5.0 | 0.2 | 20.0 |
| noise | Gamma | 1.5 | 0.1 | 5.0 |

**Table S2.** Recommended GP priors for OHE reaction representations.

Mordred Descriptors**.** The open-source molecular descriptor-calculation software Mordred was used to calculate > 1800 two- and three-dimensional descriptors for each reaction component.[8] Mordred calculates descriptors implemented in RDKit in addition to its own implementations. We independently optimized GP priors for Mordred representations using multi-objective self-optimization (see surrogate model section). We provide functions for generating Mordred representations from SMILES strings in the feature utilities module of *edbo*.

| Hyperparameter | Prior Distribution | Concentration | Rate | Mode |
|---|---|---|---|---|
| length scale | Gamma | 2.0 | 0.1 | 10.0 |
| output scale | Gamma | 2.0 | 0.1 | 10.0 |
| noise | Gamma | 1.5 | 0.1 | 5.0 |

**Table S3.** Recommended GP priors for Mordred reaction representations.

Density Functional Theory Descriptors. As a third reaction representation we investigated the use of quantum mechanical modeling for the development of chemical descriptors. In principal, salient structural and chemical features of inputs can be learned in terms of electronic and steric properties. In our prior work, we employed computational chemistry to estimate the global and local electronic and steric properties of reaction components.[6] Building on this work, in this study we developed an automation framework compatible with parallel high-performance computing clusters which enabled us to run hundreds of calculations simultaneously. In addition, we developed a framework for the automated extraction and generation of chemical descriptors for the resulting computational output. First, we wrote software in *Mathematica*.[1] However, we have also translated the software to an open-source python package called *auto-qchem*.[2] Altogether, our software computed > 1000 descriptors for each reaction.

The software generates starting geometries for density functional theory (DFT) calculations from SMILES strings or ChemDraw files using Open-Babel 2.4.1.[9] Where applicable, the software also uses Open-Babel for conformational analysis to generate a diverse set of up to 30 starting structures. Then DFT calculations (*vide infra*) were carried out to collect a set of geometries corresponding to stationary points (local optima) and unique conformers were selected according to root-mean-square distance > 0.1 Å. Then with unique optimal structures in hand, descriptor sets were generated by choosing the lowest energy conformer and/or carrying out Boltzmann analysis.

Quantum mechanical modeling was carried out using Gaussian 16 A.03 software.[10] For all DFT calculations, the B3LYP hybrid exchange-correlation functional was used. Gas-phase geometry optimization and frequency calculations were carried out using the 6-31G* basis set and LANL2DZ basis set for heavy atoms not supported by 6-31G*. Time-dependent DFT calculations were carried out on the gas-phased optimized geometry at the same level of theory. For each stationary point, harmonic analysis (frequency calculation) of the optimized geometry was used to check for convergence according to the following criteria: no negative Hessian eigenvalues, maximum force < 0.000450, root-mean-square force < 0.000300, maximum displacement < 0.001800, and root-mean-square displacement < 0.001200. Alternatively, if maximum force < 0.00001 and root-mean-square force < 0.00001 the stationary point was accepted. Calculations which did not satisfy convergence criteria were resubmitted. After calculations were completed

and satisfied convergence criteria (in most cases), DFT and TD-DFT chemical and physical descriptors were extracted for each molecule. When conformational analysis was carried out descriptors for the minimum and maximum energy structures, Boltzmann averaged descriptors, and summary statistics (mean and standard deviation) were computed.

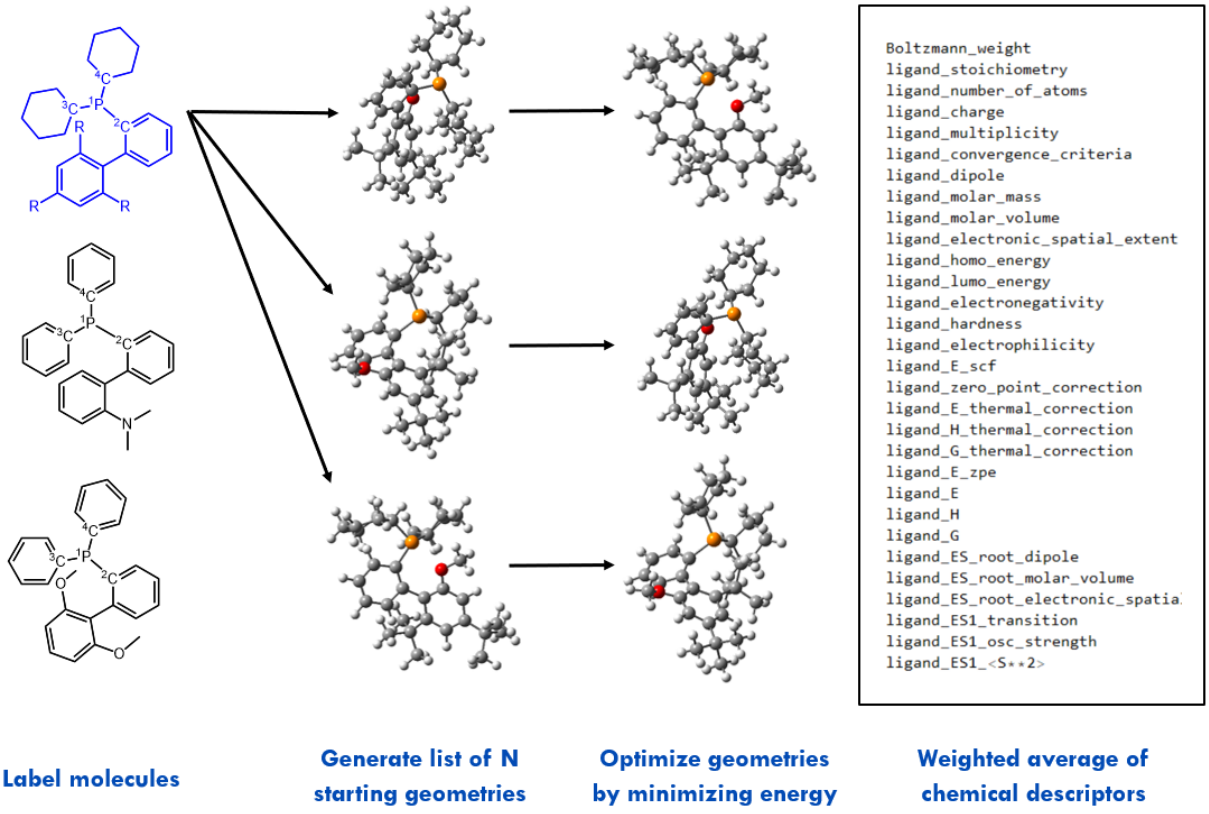



**Figure S4**. Descriptor generation pipeline: (1) label molecules in ChemDraw, (2) use software to generate starting geometries, (3) run quantum mechanical calculations, (4) extract and process chemical descriptors.

After extracting data from the DFT output, the initial set of descriptors contained global (e.g. LUMO energy and dipole moment) and local (e.g. atomic NMR shift and charge for labeled atoms) electronic and global steric (e.g. molar volume) descriptors. We supplemented this set with the vibrational descriptors highlighted in our previous study.[6] Then we added descriptors corresponding to atoms of the minimum and maximum computed charge (hot spot analysis); we expected that such descriptors could capture potential sites of reactivity. We posited that this set still lacked descriptors which captured local steric properties of molecules. Thus, we elected to develop a new descriptor function which acts on the cartesian coordinates of optimized molecular

geometries and returns the % of a sphere of a set radius which is occupied by the Van der Waals radii of the surrounding atoms (Figure S5 and S6). The central atoms were taken to be overlapping structural elements of each input class (e.g. for phosphine ligands phosphorus). Note, the Mathematica and python functions are a direct analogy of SambVca, a tool to calculate the buried volume of organometallic ligands.[11]

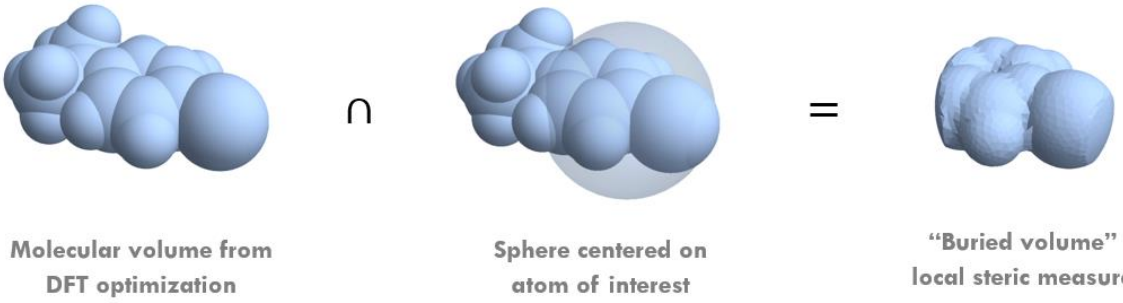



**Figure S5.** Computing site specific buried volume.

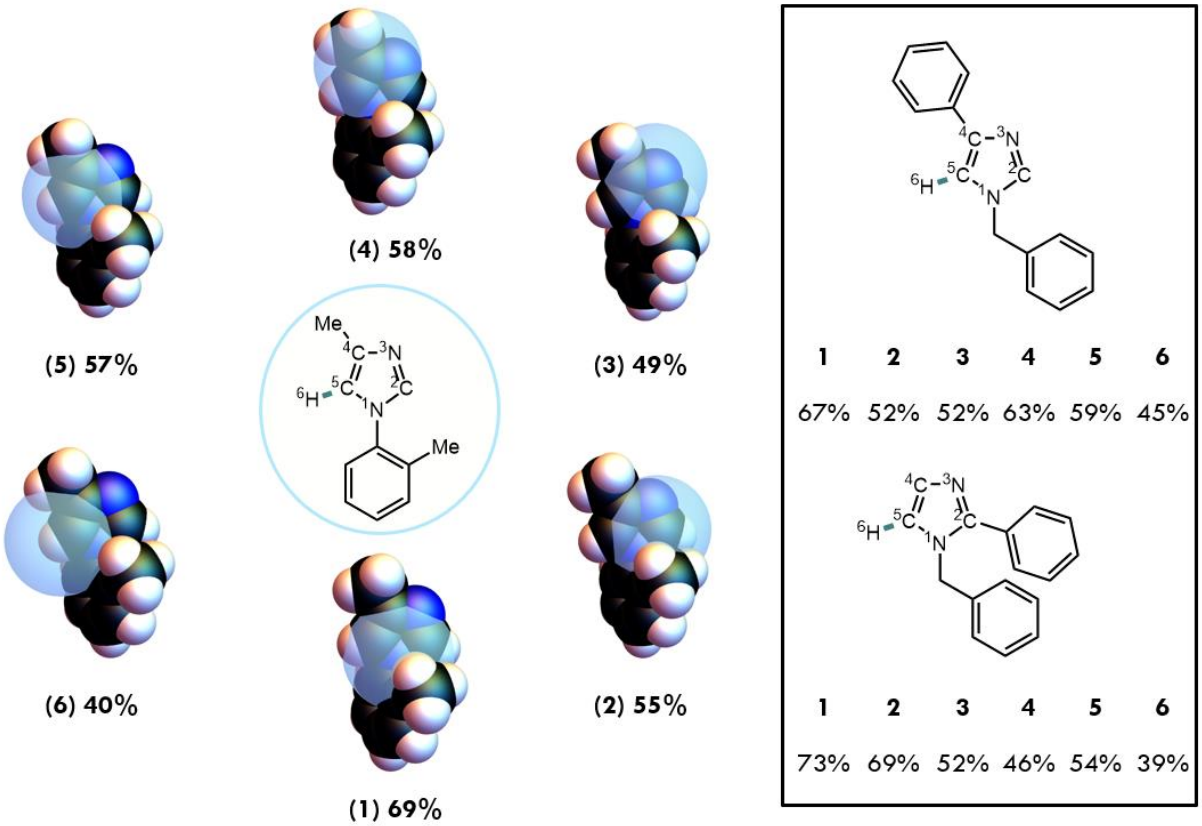



**Figure S6.** Example of steric fingerprint for overlapping cores of substituted imidazole's.

We independently optimized GP priors for DFT representations using multi-objective self-optimization (see surrogate model section). In table S4 we present the optimized GP priors for this representation. We provide functions for generating DFT representations in *auto-QChem*. In addition, we provide descriptor extraction and conformational analysis workflows in the examples.

| Hyperparameter | Prior Distribution | Concentration | Rate | Mode |
|---|---|---|---|---|
| length scale | Gamma | 2.0 | 0.2 | 5.0 |
| output scale | Gamma | 5.0 | 0.5 | 8.0 |
| noise | Gamma | 1.5 | 0.5 | 1.0 |

**Table S4.** Recommended GP priors for DFT reaction representations.

Descriptor Performance. While not the focus of this study, we were interested in investigating the differential performance of the DFT, cheminformatics, and binary encodings. Thus, following the same self-optimization procedure we identified model configurations tuned to make data efficient predictions for each encoding (*vide infra*). In GP regression with individually optimized priors we were surprised to find that all featurization approaches offered similar performance (Figure S7). However, on the margin DFT descriptors appeared to give improved learning curves. In BO experiments with GP models we found that DFT encoding tended to give more consistent results. For example, when comparing loss after 10 batches of 5 experiments we found that DFT encodings tended to have comparable or improved average performance and fewer outliers (no optimizations which failed to converge within a close margin of the optimal yield) (Figure S8). Other things equal we favor the use of structure-based reaction encodings to facilitate chemically meaningful interpretations of the surrogate models. Still these results suggest that in similar search spaces, simple binary encodings may in general be enough to effectively differentiate reactions in optimization. Given their simplicity, this finding is quite intriguing.

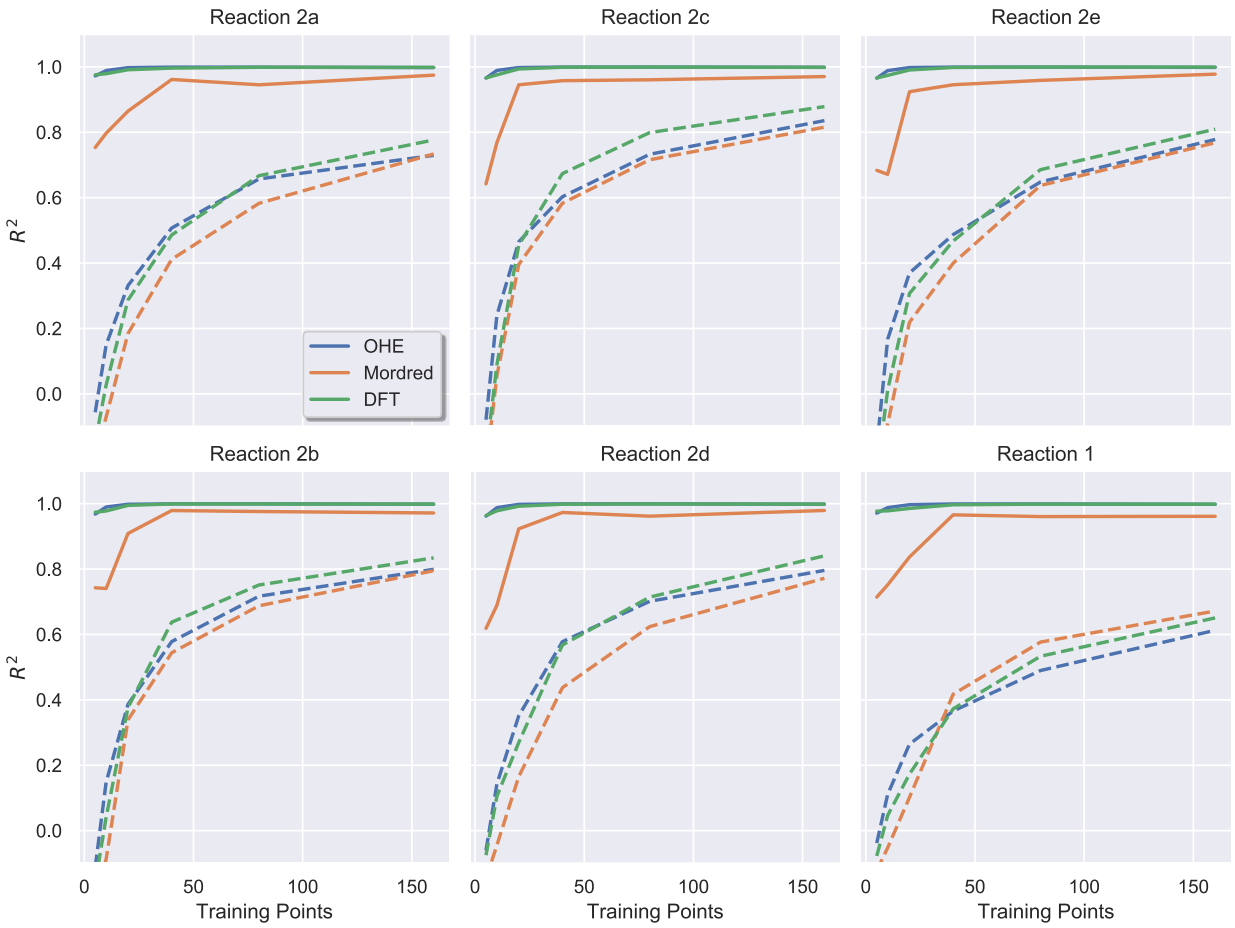



**Figure S7.** Comparison of learning curves for OHE, Mordred, and DFT encodings with gaussian process models. Average (N=20) model training (solid) and test (dashed) scores with varying amount of data. Training was carried out on random subsets of each data set with test performance measured on the remainder of the data

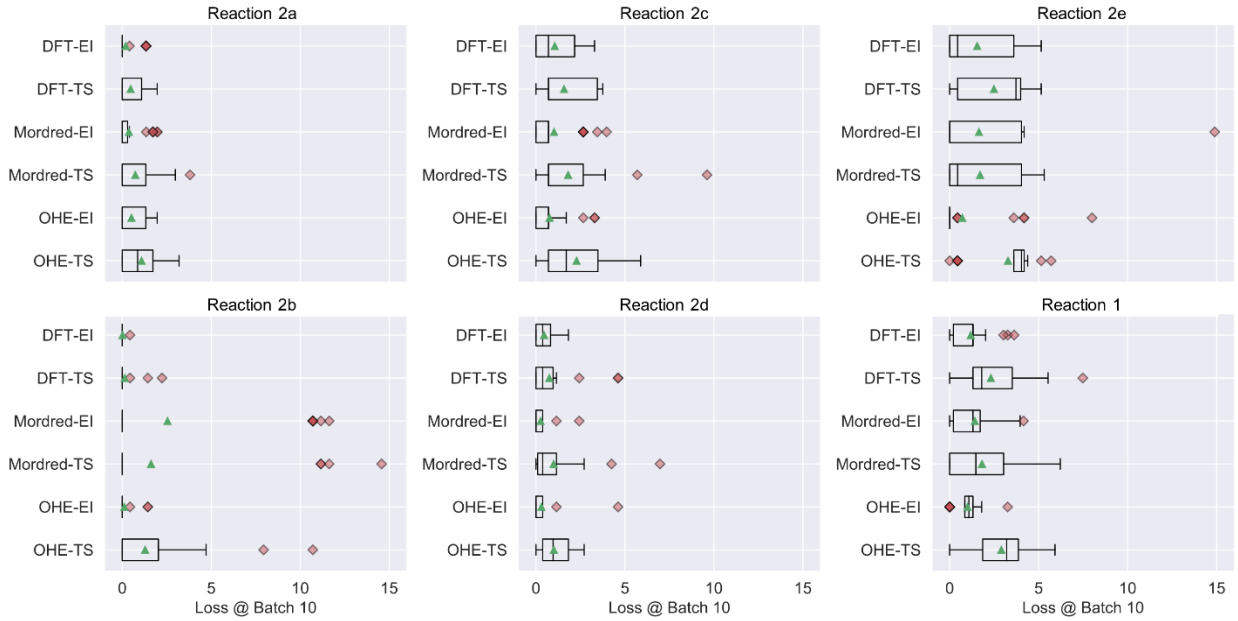



**Figure S8.** Summary of BO performance with different reaction representations. Box and whisker plots showing performance after 10 batches of experiments using 30 random initializations, a Gaussian Process surrogate model, and batch size = 5. Green triangles and red diamonds denote the mean and outliers, respectively.

## IV. Initialization Methods

One goal of Bayesian reaction optimization is to enable the identification of optimal conditions in a way that does not depend on the initial points selected. That is, we seek an optimizer with convergence which is largely invariant to the choice of initial points. Thus, we elected to carry out our experiments in the manuscript body using random selection. However, other methods such as design of experiments (DOE) or transfer learning could be employed to make better initial choices. Alternatively, the initial choice of experiments may be best left in the hands of chemists, providing an additional avenue for the incorporation of domain expertise into a BO campaign. This decision is left entirely in the hands of users. In this section we present a machine learning approach to the selection of initial experiments using clustering. Specifically, we use a clustering algorithm to identify reactions which are close together in experimental space (as determined by reaction encoding and a given distance metric). Then, we select the central points of the clusters (or the experiment closes to it) as the initial experiments. Figure S9 shows an example of using *k*-medoids clustering to select 5 initial experiments on a uniform 2D grid. Notice that the exemplars in this example nicely sample the space. We were curious if such an approach could offer a more ideal distribution of information in the first few iterations of optimization and perhaps lead to faster convergence.

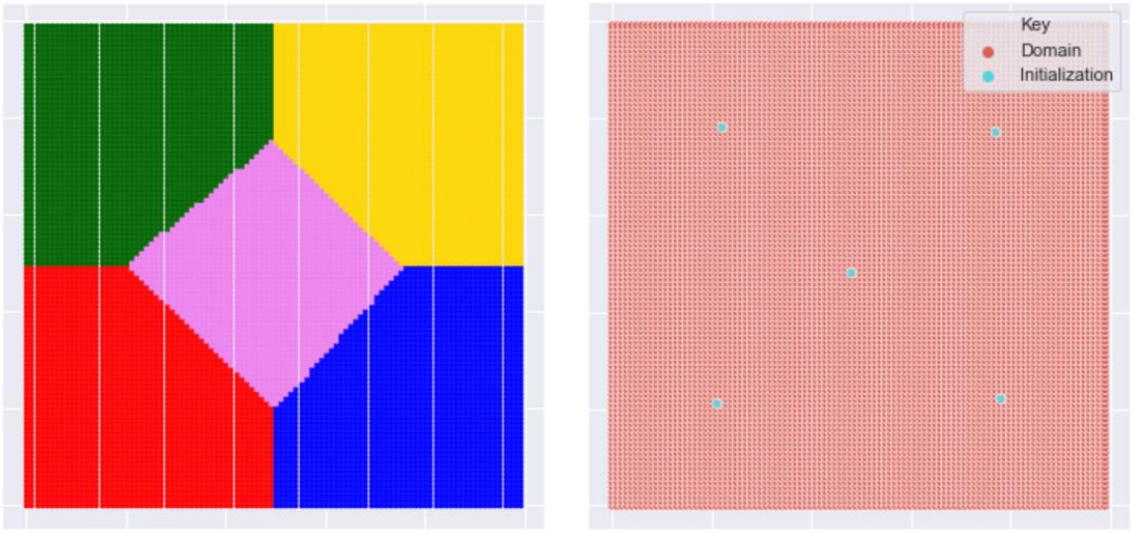



**Figure S9**. Selecting initial points via clustering. Here we show an example using *k*-medoids to select 5 initial experiments on a 2D grid of points: (left) clusters identified via PAM and (right) exemplars taken as the initial points.

*k*-Means. *k*-Means clustering[12] is a method for partitioning unlabeled data into groups that attempts to minimize squared Euclidean distance between points in a cluster and its designated center. In the *k*-means algorithm cluster centers (or centroids) are selected which may or may not correspond to real experiments in sparse reaction space. Thus, for a given search space *k*-means clustering was used to identify *k* experiments by choosing experiments which are closest to the centroids. We explored two approaches. First, we select *k* such that it is the number of experiments desired (as in Figure S9). In this case *k* may not best represent the data. Alternatively, we select *k* by optimizing the average silhouette width[13], a measure of how similar experiments in each cluster are to each other versus the other clusters. Note that we require the constraint that $k \geq N$, where $N$ is the number of desired experiments. Then, the algorithm randomly selects $N$ experiments from the $k$ experiments closest to the centroids. In practice, for the development data sets we saw little difference between these two approaches. *k*-Means clustering can be used for initialization in *edbo* using the initialization method argument.

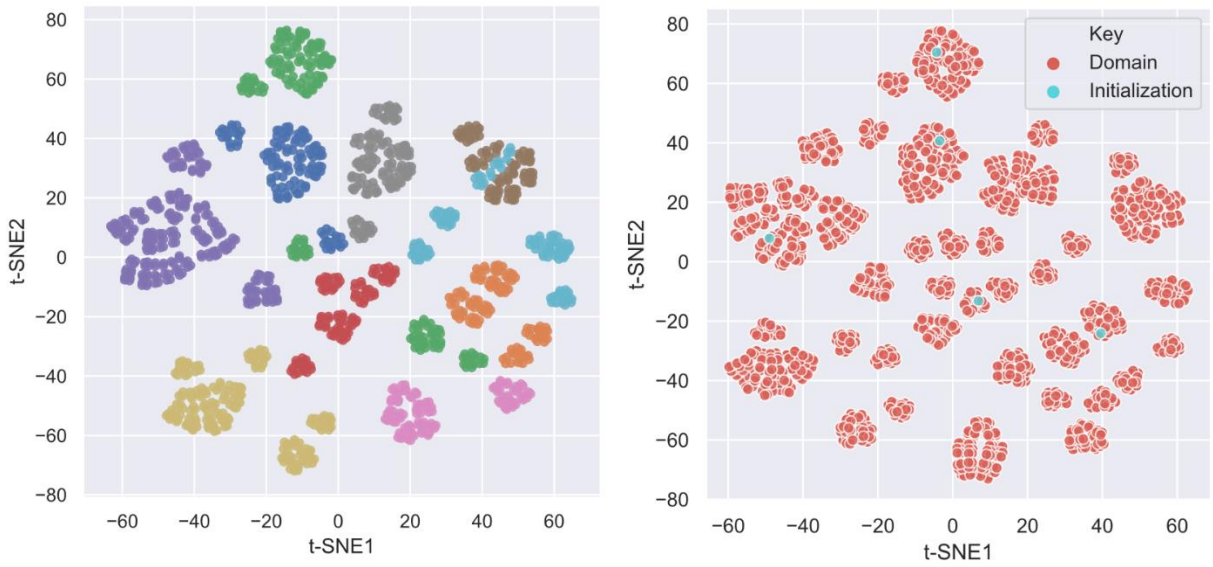



**Figure S10**. Selecting experiments via *k*-means clustering. Here we show an example using *k*-means to select 5 initial experiments from reaction **1** (Suzuki data): (left) 10 clusters identified via k-means and optimizing silhouette width and (right) experiments taken as the initial points.

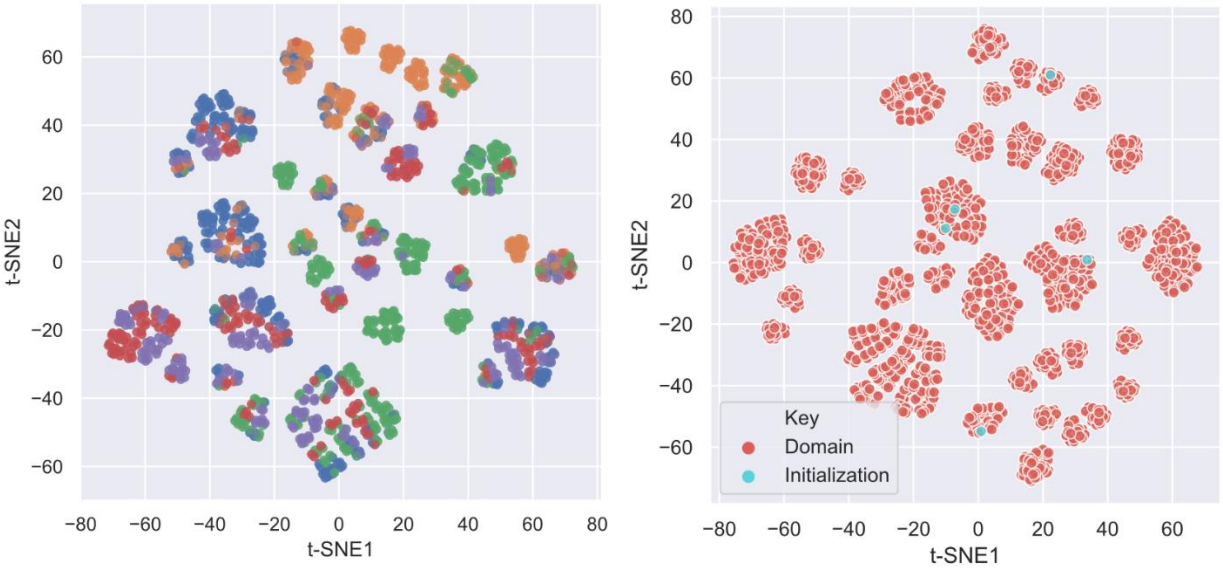



**Figure S11**. Selecting experiments via *k*-medoids clustering. Here we show an example using *k*-medoids to select 5 initial experiments from reaction 1 (Suzuki data): (left) 5 clusters identified via PAM (right) exemplars taken as the initial points. Here we select *k* as the number of experiments desired rather than optimizing silhouette width.

*k*-Medoids. *k*-Medoids clustering[12] also attempts to minimize distance between points in a cluster and its center. *k*-Medoids uses experiments as center points rather than centroids as in *k*-means clustering. This gives the benefit that the algorithm identifies "exemplars" which can be understood as the prototypical experiment for a cluster. For a given search space the partitioning around medoids (PAM) algorithm was used to identify *k* experiments, where *k* is the number of exemplars to be identified. Here we use Gower distance as a distance metric for mixed data types.[14,15] *k*-Medoids clustering can be used for initialization in *edbo* using the initialization method argument.

Bayesian optimization performance. We investigated optimization performance of *k*-means and *k*-medoids compared to random sampling by running simulations on the 6 development objectives. For both clustering algorithms, statistics were generated by randomly choosing initial centroids (*k*-means) and exemplars (*k*-medoids). In addition, when clusters were selected by optimizing silhouette width with $k > N$, cluster centers were selected at random. Our hypothesis was selecting initial points via clustering could give a better distribution of information leading to better global surrogate model fits. In turn, we anticipated that initialization via clustering may give more consistency in BO and improved intermediate and long-term convergence. We first analyzed the BO losses over 30 random restarts for each initialization method after 5 batches of 5 experiments. While, the average loss tended to be lower when initialized via clustering, the difference was marginal (Figure S12). We next investigated BO loss after 10 batches of 5 experiments. Here we saw no significant difference between the initialization methods (Figure S13). Overall, we found that for these data sets this approach does not significantly impact optimization performance for the development objectives.

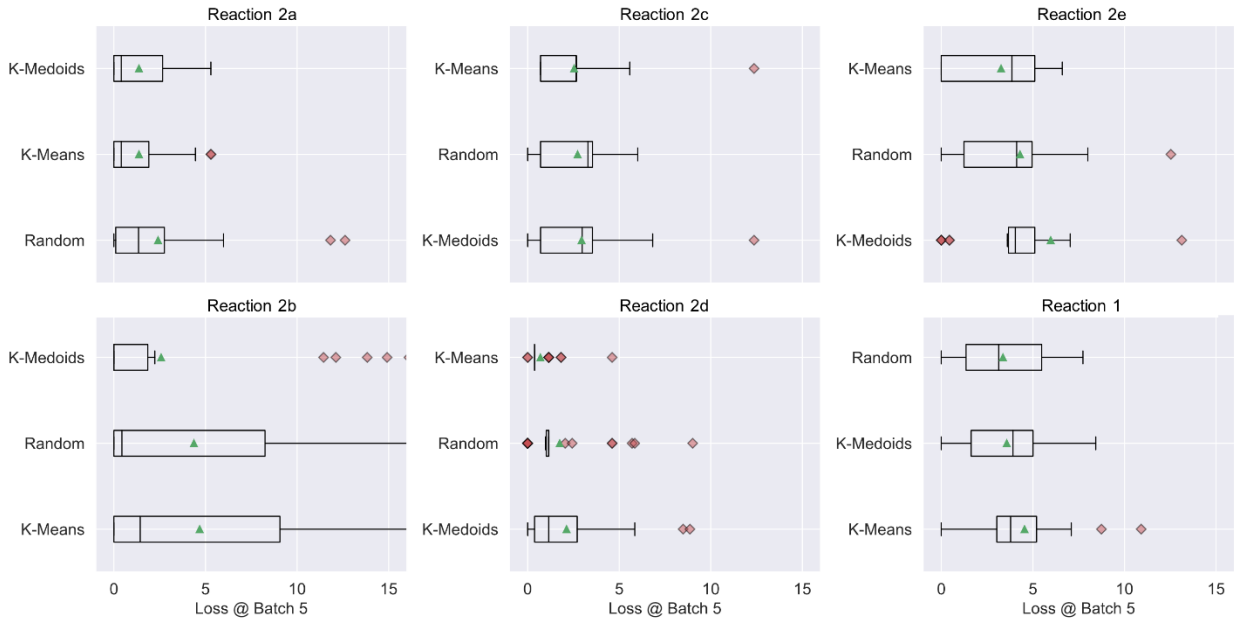



**Figure S12.** Summary of BO performance with different initialization methods. For *k*-means we select $k \geq N$, where $N$ is the number of desired experiments, by optimizing silhouette width. Box and whisker plots showing performance after 5 batches of experiments using 30 random initializations, a Gaussian Process surrogate model, and batch size = 5. Green triangles and red diamonds denote the mean and outliers, respectively.

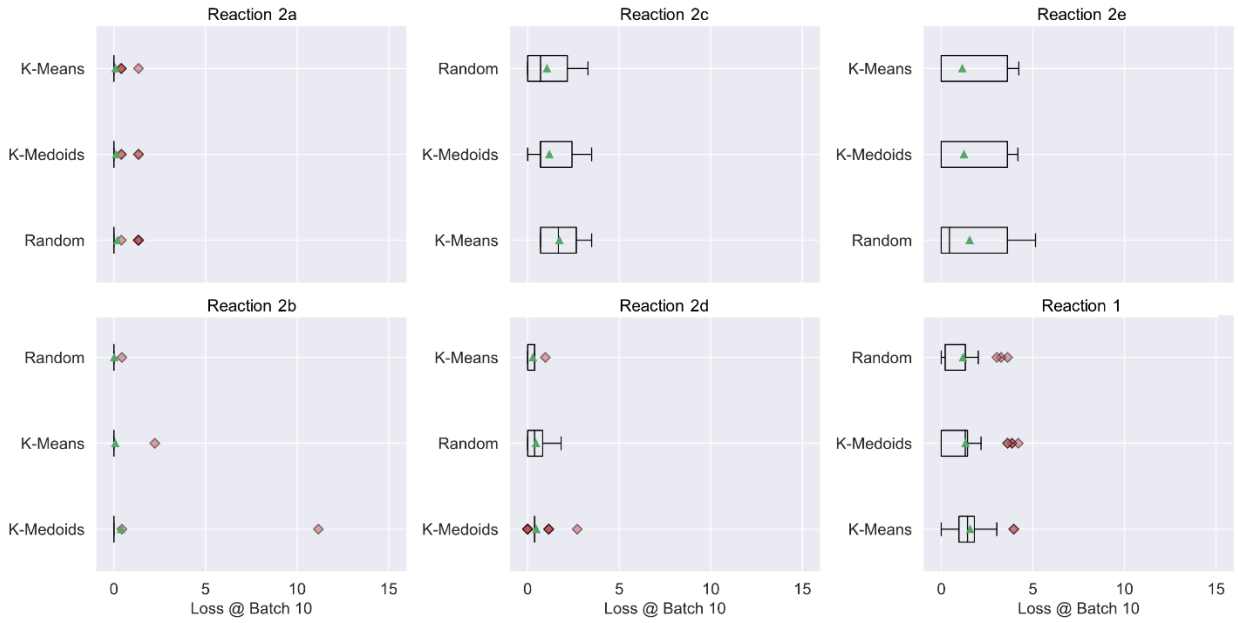



**Figure S13.** Summary of BO performance with different initialization methods. For k-means we select $k \geq N$, where $N$ is the number of desired experiments, by optimizing silhouette width. Box and whisker plots showing performance after 10 batches of experiments using 30 random initializations, a Gaussian Process surrogate model, and batch size = 5. Green triangles and red diamonds denote the mean and outliers, respectively.

## V. Surrogate Models

Background. For a given search space, Bayesian reaction optimization begins by collecting initial reaction outcome data via an experimental design (DOE) or drawing from existing results. These data are used to train a probabilistic surrogate model which is constructed by conditioning a prior over functions, capturing our assumptions about the reaction's response surface (smoothness, experimental noise, etc.) and making it possible to infer the position of global optima. The most basic requirements of an effective surrogate model are the ability to make predictions and estimate variance. Thus, in principal BO can be implemented with many different models depending on the problem. For example, for optimizations over discrete or mixed domains, the response surface may be best characterized by a random forest (RF) model.[16] A RF is an ensemble of decision trees which represents a set of piecewise functions trained by randomly subsampling the data. While RF models do not provide an estimate of variance, in practice one can use the empirical variance across the trees in the ensemble.[17] RF models tend to make good predictions close to the training data but are poor extrapolators. Thus, far away from the training data, the individual trees may give similar predictions resulting in the underestimation of variance.[17] In turn this can impede exploration and result in the optimizer getting trapped in local maxima. For continuous domains, it is common to assume the unknown function is sampled from a Gaussian process (GP), sometimes called Kriging.[18] A GP is a distribution over functions defined by its prior mean and covariance kernel. In GP regression, the choice of kernel determines the shape of functions in the distribution. Notably, for GPs principled predictions and variance estimates can be computed analytically.[18]

Gaussian process models. For an excellent treatment of gaussian process regression (GPR) see *Gaussian Processes for Machine Learning*.[18] Here we focus only on the aspects of the model which were important considerations in tuning GPR performance for chemical reaction data. We utilize the python package *GPyTorch*[19], a scalable and GPU compatible implementation of GPs which is well suited to our task. Using the *GPyTorch* framework we investigated varying GP configurations including different kernels, priors, and learning parameters.

Making predictions and estimating variance. A GP is specified completely by its mean function and covariance function. Here we model chemical reactions using a constant mean. Thus, the GP mean prediction $\mu$ at point $\boldsymbol{x}$ is given by

$$\mu(\boldsymbol{x}) = k_\theta(\boldsymbol{x})^T (K_\theta + \sigma_n^2 I)^{-1} \boldsymbol{y} \qquad (1)$$

where $k_\theta(\boldsymbol{x})$ is the vector of covariances between $\boldsymbol{x}$ and the training points, $K_\theta$ is the matrix of covariances between all training points, $\sigma_n^2$ is estimated noise variance, $I$ is the identity matrix, and $\boldsymbol{y}$ is a vector of responses corresponding to the training data. GP variance at $\boldsymbol{x}$ is then given by

$$V[f(\boldsymbol{x})] = k_\theta(\boldsymbol{x}, \boldsymbol{x}) - k_\theta(\boldsymbol{x})^T (K_\theta + \sigma_n^2 I)^{-1} k_\theta(\boldsymbol{x}) \qquad (2)$$

where $k_\theta(\boldsymbol{x}, \boldsymbol{x})$ is the covariance of $\boldsymbol{x}$ with itself. Notice that predictions are a linear combination of observations (we restrict the GP posterior to contain only those functions which agree with the observed data points within noise) and variance does not depend on the observed responses. Both functions suggests poor scaling with increasing data due to the matrix inversion. Importantly, *GPyTorch* has features for enhanced scalability which will become useful in reaction optimization problems with large search spaces.[19]

Covariance function. Typically, in supervised learning similarity (distance) between points is employed for making predictions under the assumption that similar points will have related responses. In GPR, the covariance function (or kernel) defines the notion of similarity between two points. After initial experimentation, we focused on the Matérn[18] class of kernels (equation 3) which enable the effective modeling of complex nonlinear relationships found in experimental data. The Matérn class of covariances between two points $k$, distance $\boldsymbol{r}$ from each other, is given by

$$k(\boldsymbol{r}) = \alpha \frac{2^{1-\nu}}{\Gamma(\nu)} \left( \frac{\sqrt{2\nu}\, \boldsymbol{r}}{l} \right)^{\nu} K_\nu \left( \frac{\sqrt{2\nu}\, \boldsymbol{r}}{l} \right) \qquad (3)$$

where $\alpha$ is an output scale parameter, $\Gamma$ is the gamma function, $\nu$ is a shape parameter, $K_\nu$ is the modified Bessel function, and $l$ is the length scale parameter. One of the primary machine learning problems of GPR is estimating $l$. This is because the shape of functions drawn from the GP on a given interval depends on $l$. From $l$'s placement in the denominators of (3) one can see that length scale acts to marginalize distance between points. That is, small and large values of $l$ tend to imply larger and smaller variations on a given interval, respectively. This relationship can be clearly seen in figure S14 which shows samples drawn from a GP prior using Matérn52 covariance with $l$ set to different values. Here, it can also be seen that larger values of $l$ will make the corresponding function distribution approach linear on a fixed interval.

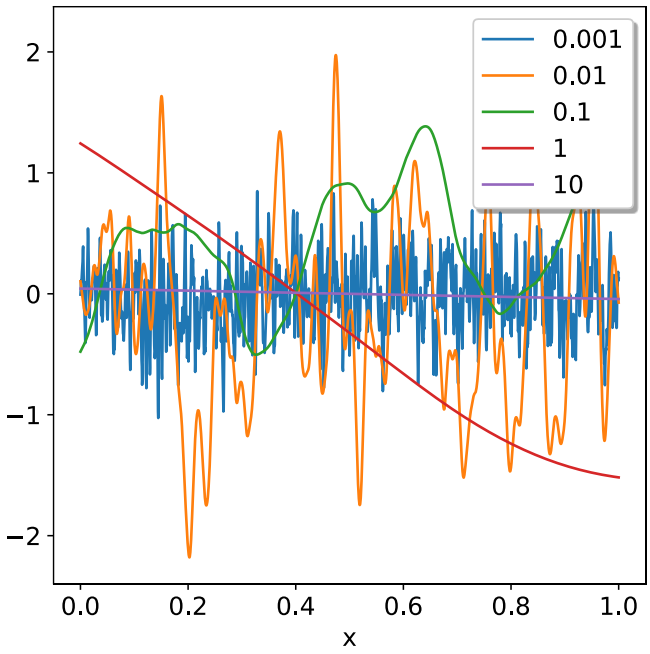



**Figure S14.** Samples drawn from a constant mean GP with a Matérn52 kernel set to different length scales.

The parameter $\nu$ also impacts the shape of functions. For example, as $\nu$ increases, the GP function distribution becomes smoother and in the limit of $\nu = \infty$ the Matern kernel reduces to the standard radial basis function kernel. However, this parameter is not estimated in model training. Rather it is selected by the user. In this work, regression experiments with reactions **1** and **2a–e** suggested that $\nu = 5/2$ is optimal. In which case, equation 3 reduces to the following:

$$k_{\nu=5/2}(\boldsymbol{r}) = \alpha(1 + \frac{\sqrt{5}\,\boldsymbol{r}}{l} + \frac{5\boldsymbol{r}^2}{3l^2})e^{-\frac{\sqrt{5}\,r}{l}} \qquad (4)$$

Optimization. In GPR, the primary machine learning task is to estimate kernel hyperparameters ($\theta$), the noise level ($\sigma_n^2$) and output scale $\alpha$. Here this is accomplished by optimizing the log marginal likelihood

$$\log p(y|X, \theta, \sigma_n^2, \alpha) = -\frac{1}{2}\boldsymbol{y}^{\mathrm{T}}K_y^{-1}\boldsymbol{y} - \frac{1}{2}\log\left|K_y\right| - \frac{n}{2}\log(2\pi) \qquad (5)$$

where $K_y = K + \sigma_n^2 I$. In equation 5 the first, second, and third terms correspond to data fit, a complexity penalty, and a normalization constant respectively. Model training consisted of maximizing the log marginal likelihood using a gradient based optimizer. Here we use Adam[20] as implemented in *PyTorch*.[21]

Automatic relevance determination. For DFT and cheminformatics encodings, which contain hundreds of features for a given reaction (after decorrelation), the automatic identification of important dimensions is critical to model performance. This issue is exacerbated in the low data regime where, the model needs to learn the variation in each dimension from only a few data points. While the RF ensemble naturally identifies important features and discourages overfitting via random sampling, these components needed to be explicitly included in the GP model. In particular, learning a length scale per dimension can be critical for achieving good performance with high-dimensional encoded reaction data. This is because many dimensions are likely to be irrelevant to reaction outcome. Here we utilize automatic relevance determination (ARD) to learn a length scale per dimension.[18]

Priors. We hypothesized that in the absence of sufficient data, the inclusion of a parameter per dimension could lead to overfitting. Thus, we assumed that most of the dimensions are irrelevant, impose this belief via a gamma prior centered on longer length scales, and estimate parameters via Bayesian inference. In particular, a prior can be included in the GPR hyperparameter estimation procedure by adding the log probability for the hyperparameter value specified by the prior distribution to equation 5. Together, after estimation, this procedure results in belief biased models

which capture general trends in the low data regime and data biased models which capture response surface nuances when more evidence is available.

Self-optimization. We identified GPR models tailored for OHE, Mordred, and DFT encodings using our software to optimize itself. For each encoding we focused on searching for model architectures which (1) balanced results on all 6 development data sets and (2) maximized average performance in the low data regime. To do this we formulated a multi-objective optimization as the sum of average (N=20) test $R^2$ values, where each model was trained with 25 data points and performance was measured in predicting the remainder of the space, for all development data sets. Notice that since $R^2 \leq 1$, the multi-objective maximum is 6, which denotes perfect predictions across all 6 data sets. Figure S15 shows an example of *edbo* convergence during Bayesian optimization of average cumulative regression performance using a GP surrogate model and expected improvement as an acquisition function. Over the course of our optimization studies, we identified several parameters which were important for achieving good model performance across all data sets (Table S5).

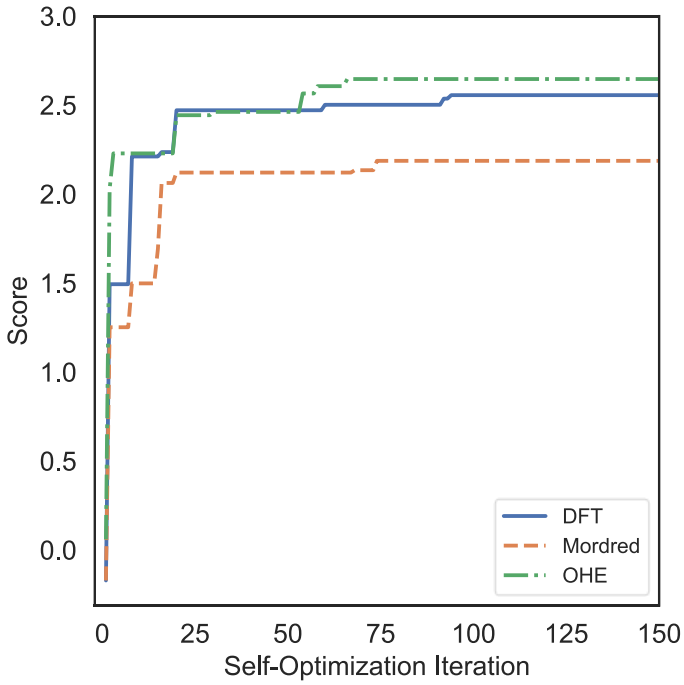



**Figure S15.** Example of self-optimization curves for different reaction encodings. Bayesian optimization of model performance using GPR and expected improvement as an acquisition function. See figure S7 for learning curves with parameters optimized for each encoding.

Model (de)optimization. Over the course of our optimization studies we found that DFT encoding, the Matérn52 kernel, a gamma length scale prior with a mode of 5.0, a gamma output scale prior with a mode of 8.0, a gamma noise prior with a mode of 1.0, and automatic relevance determination to learn a length scale per dimension gave balanced model performance in the low data regime. In (de)optimization experiments we were surprised to find that in these experiments OHE gave a small improvement in cumulative score over DFT encoding. For DFT and cheminformatics encodings, which contain hundreds of features for a given reaction (after decorrelation), the automatic identification of important dimensions is critical to model performance. In this case the models are tasked with learning hundreds of hyperparameters from only 25 data points. Thus, it is not surprising that the performance trends roughly with the number of descriptors. However, in more detailed learning curve experiments (See below, Figure S26) we found that DFT encoding tended to outperform OHE above some threshold amount of training data (typically between 25 and 50 experiment). Next, we investigated the choice of the kernel's shape parameter (Matérn, ν). We found that the choice of this parameter was moderately significant and that smoother functions tended to be better. This complements the observed preference for priors centered on longer length scales. When little data is available the prior will carry more weight than the training data and lead to smoother functions. Indeed, we found that the length scale prior was the most critical component to model success and removing it led to a dramatic decrease in cumulative regression performance. The inclusion of priors on output scale and noise also impacted model performance, albeit to a lesser extent. Finally, in these experiments, ARD was moderately significant. However, as the amount of training data increased the GPs capacity for modeling response surface nuances also increased, leading to a greater gap between single length scale and ARD models (Figure S16). Again, this observation is consistent with the prior which imposes similar length scales when little data is available.

It was observed that length scale priors had a significant impact on GP regression performance (Table S5). To probe further, we carried out additional experiments for each featurization, varying the selected priors in order to investigate the sensitivity of model performance to our selected parameters. By varying the concentration and rate parameters in the gamma distribution for each component we collected multi-objective scores (*vide supra*) for a series of priors. Figures S17–S25 show the relationship between prior mode and observed score. Intriguingly, we observed a

"goldilocks" effect in the relationship between length scale prior mode and model performance. That is, both above and below some prior mode threshold model performance dropped sharply. In keeping with our findings, the peaks were centered on the parameter values identified via self-optimization. Importantly, the optimal length scale prior modes trend with the dimensionality of the representations OHE < DFT < Mordred. Indeed, according to our hypothesis as the number of features increases, we anticipate that the average significance of each dimension will decrease, which manifests in a longer average optimal length scale. While less significant for model performance, we also investigate the sensitivity of output scale and noise priors. We found that, each encoding displayed a different relationship between output scale prior mode and regression score. From equation 1 we expect that the learned noise variance $\sigma_n^2$ should not depend on the reaction representation. Indeed, the noise prior sensitivity is largely independent of the reaction encoding.

| ENTRY | ENCODING | $\nu$ | P($l$) (R, C, M) | P($\alpha$) (R, C, M) | P($\sigma_n^2$) (R, C, M) | ARD | SCORE |
|---|---|---|---|---|---|---|---|
| **1** | **DFT** | **5/2** | **(2.0, 0.2, 5.0)** | **(5.0, 0.5, 8.0)** | **(1.5, 0.5, 1.0)** | **True** | **2.45** |
| **2** | Mordred | 5/2 | (2.0, 0.2, 10.0) | (2.0, 0.1, 10.0) | (1.5, 0.1, 5.0) | True | 1.99 |
| **3** | OHE | 5/2 | (3.0, 1.0, 2.0) | (5.0, 0.2, 20.0) | (1.5, 0.1, 5.0) | True | 2.57 |
| **4** | DFT | 3/2 | (2.0, 0.2, 5.0) | (5.0, 0.5, 8.0) | (1.5, 0.5, 1.0) | True | 2.45 |
| **5** | DFT | 1/2 | (2.0, 0.2, 5.0) | (5.0, 0.5, 8.0) | (1.5, 0.5, 1.0) | True | 2.09 |
| **6** | DFT | $\infty$ | (2.0, 0.2, 5.0) | (5.0, 0.5, 8.0) | (1.5, 0.5, 1.0) | True | 2.38 |
| **7** | DFT | 5/2 | None | (5.0, 0.5, 8.0) | (1.5, 0.5, 1.0) | True | 0.96 |
| **8** | DFT | 5/2 | (2.0, 0.2, 5.0) | None | (1.5, 0.5, 1.0) | True | 2.13 |
| **9** | DFT | 5/2 | (2.0, 0.2, 5.0) | (5.0, 0.5, 8.0) | None | True | 2.43 |
| **10** | DFT | 5/2 | (2.0, 0.2, 5.0) | (5.0, 0.5, 8.0) | (1.5, 0.5, 1.0) | False | 2.11 |

**Table S5.** Lessons learned in GP optimization. *P(R,C,M)* denote the selected gamma priors on model hyperparameters where *R*, *C*, and *M* correspond to the rate, concentration, and mode of the distributions respectively.

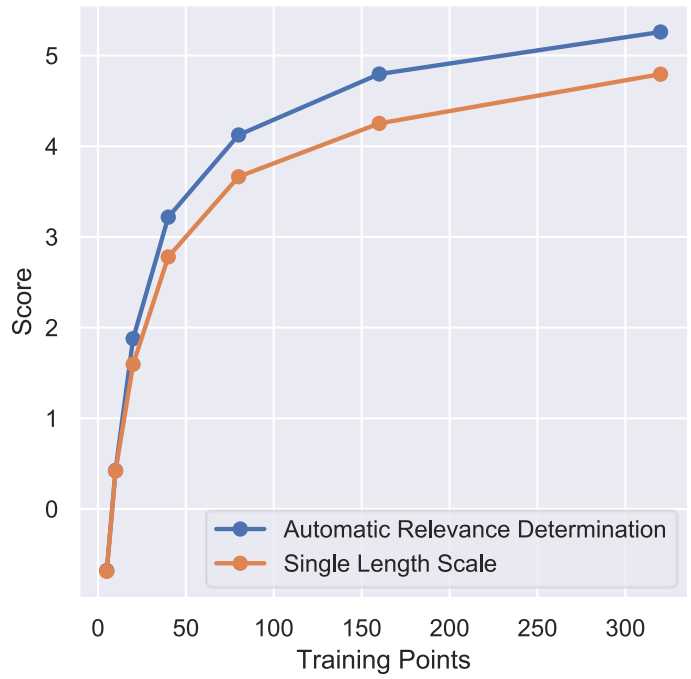



**Figure S16.** Multi-objective learning curves for GPR with and without ARD. Cumulative average (N=10) curve for GPR configuration 1 and 10 (Table S5).

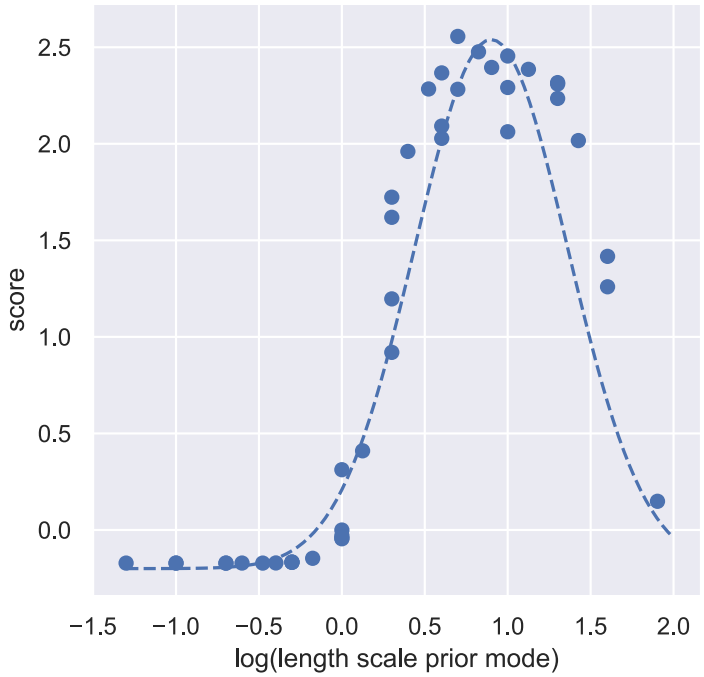



**Figure S17.** Length scale prior sensitivity for DFT encoded reactions **1** and **2**. Score refers to multi-objective, cumulative $R^2$ score.

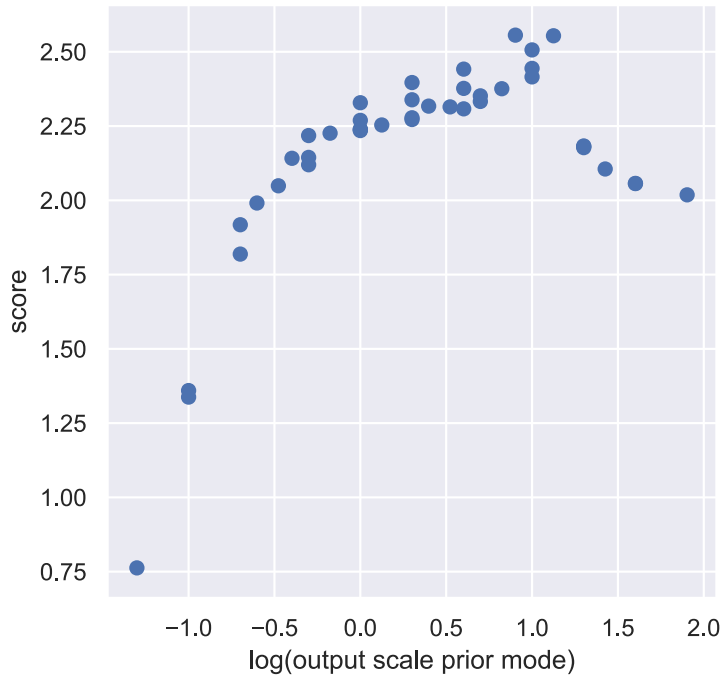



**Figure S18.** Output scale prior sensitivity for DFT encoded reactions **1** and **2**. Score refers to multi-objective, cumulative $R^2$ score.

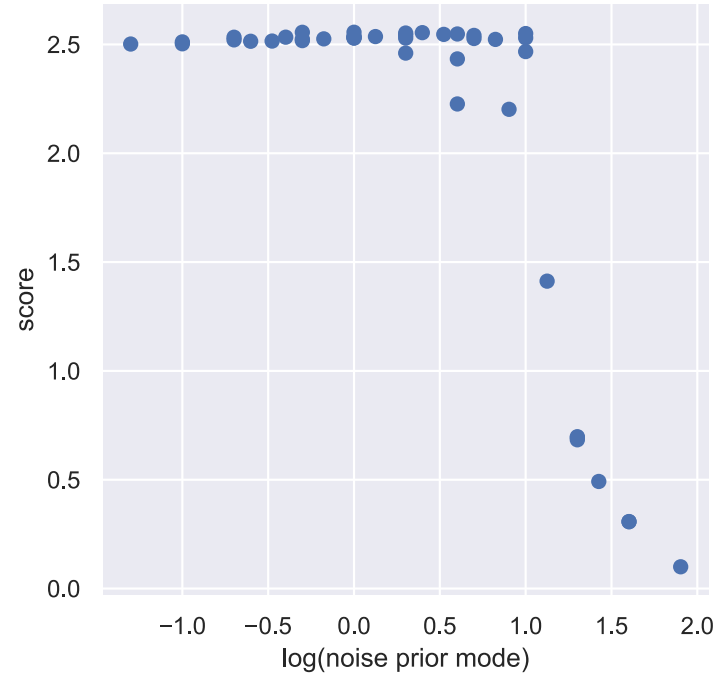



**Figure S19.** Noise variance prior sensitivity for DFT encoded reactions **1** and **2**. Score refers to multi-objective, cumulative $R^2$ score.

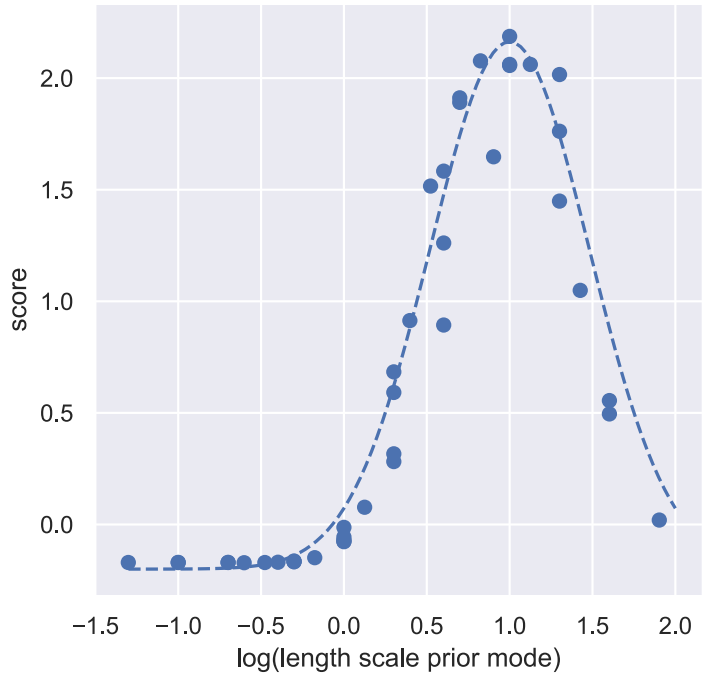



**Figure S20.** Length scale prior sensitivity for Mordred encoded reactions **1** and **2**. Score refers to multi-objective, cumulative $R^2$ score.

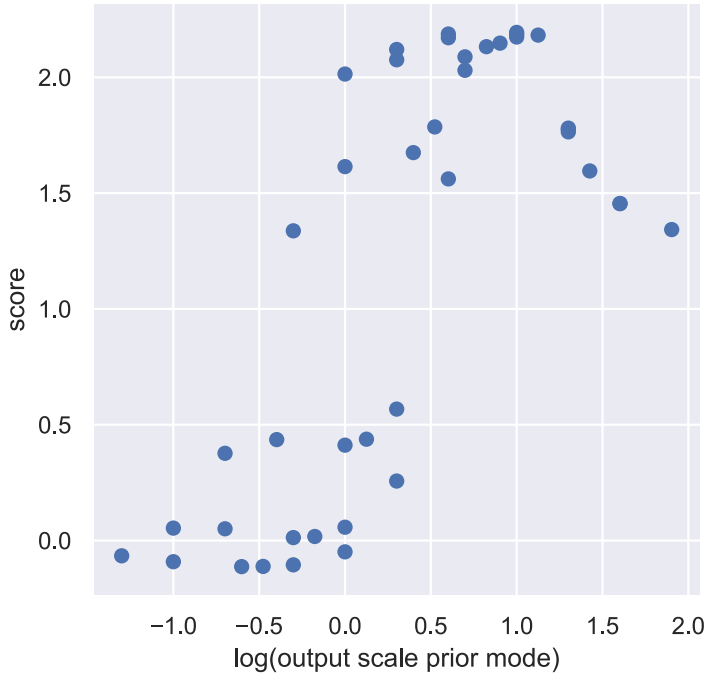



**Figure S21.** Output scale prior sensitivity for Mordred encoded reactions **1** and **2**. Score refers to multi-objective, cumulative $R^2$ score.

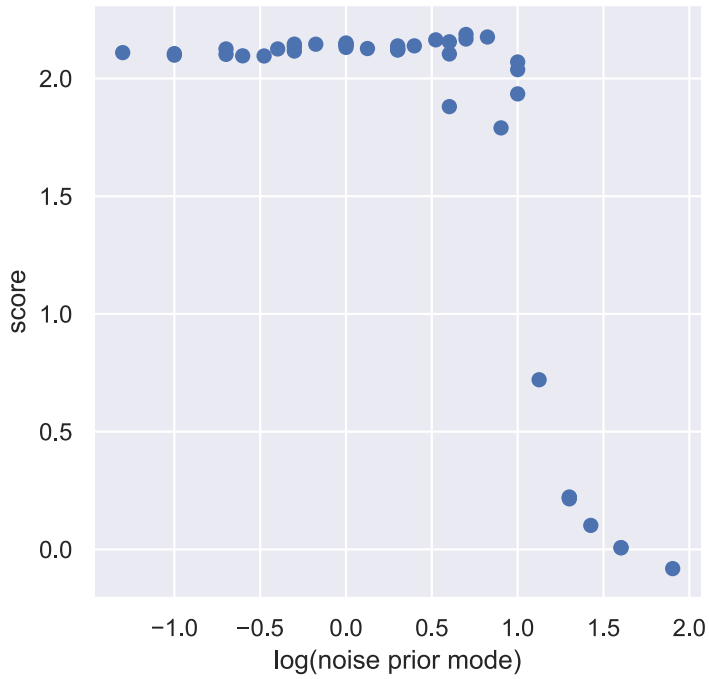



**Figure S22.** Noise variance prior sensitivity for Mordred encoded reactions **1** and **2**. Score refers to multi-objective, cumulative $R^2$ score.

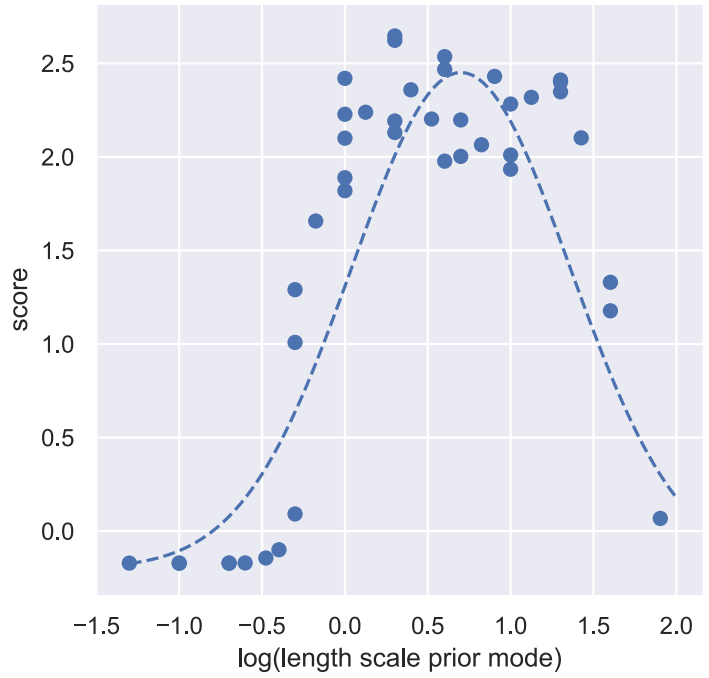



**Figure S23.** Length scale prior sensitivity for OHE representations of reactions **1** and **2**. Score refers to multi-objective, cumulative $R^2$ score.

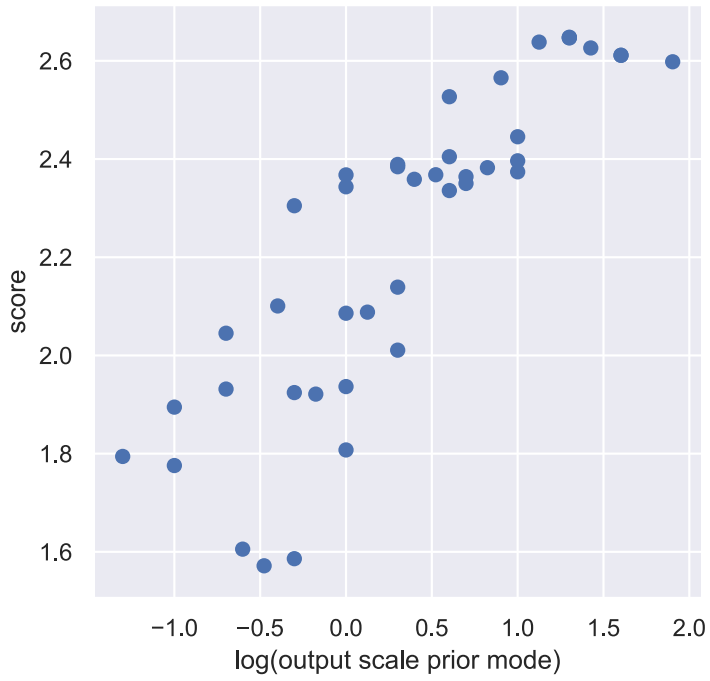



**Figure S24.** Output scale prior sensitivity for OHE representations of reactions **1** and **2**. Score refers to multi-objective, cumulative $R^2$ score.

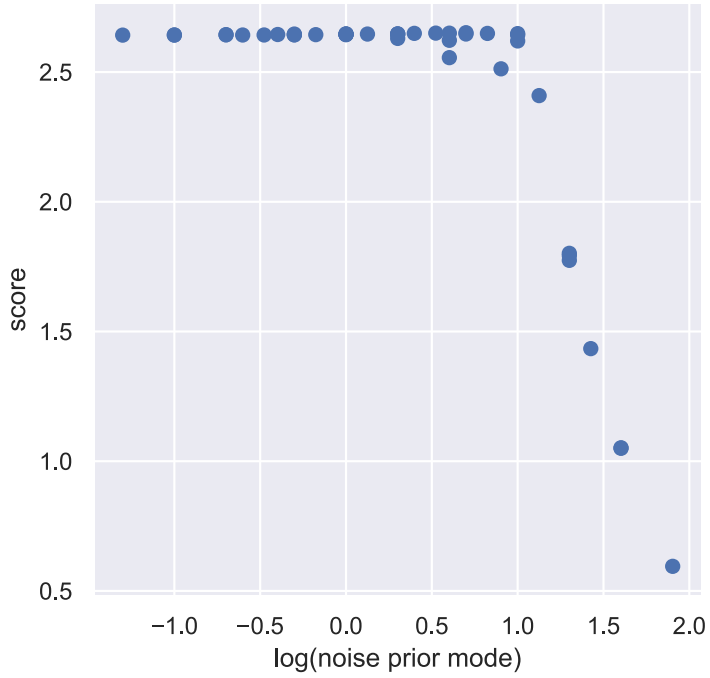



**Figure S25.** Noise variance prior sensitivity for OHE representations of reactions **1** and **2**. Score refers to multi-objective, cumulative $R^2$ score.

Regression Performance. When little data is available, our GP model emulates regularized linear regression, shrinking the significance of most dimensions, to mitigate overfitting. Thus, we sought to compare Bayesian linear regression (BLM) (equivalent to a GP with a linear kernel), to the non-linear RF and GP regressors. To draw a direct analogy to GP regression, we carried out BLM regression using ARD to prune irrelevant features (See GitHub for implementation). To probe the learning potential of the surrogate models we investigated their ability to predict reaction outcomes with varying amounts of training data (Figure S26). In general, we found that GP regression offered good performance in the very low data regime (< 25 experiments) and on average made more accurate predictions with the same training data. Intriguingly, it was observed that in the intermediate-low data regime (25<75 experiments) linear and non-linear models offered somewhat similar performance, albeit RF and GP regression tend to overfit the training data as evidenced by their near unity average training $R^2$ scores. However, as the amount of data increased the linear model's performance fell flat, indicating non-linear relationships between reaction parameters and yield of desired product.

Learning curves. We investigated each model's performance with varying amounts of training data using learning curves (Figure S26). For a given reaction, we randomly sampled *N* training data points, trained each model, and compute training $R^2$ score. Then using the trained models, we predicted the remainder of the data points and compute the test $R^2$ score compared to the observed responses. For each training set size *N*, this process was repeated 25 times to give the average learning curves.

Interpreting Models. While not the focus of this study, we were intrigued by the prospect of interpreting the surrogate models to guide mechanistic inference. Thus, we set out to probe the relationship between DFT descriptors and reaction outcome captured by each model class. We decided to investigate Suzuki reaction 1, as reaction 2 has been extensively studied.[6] As a thought experiment, we began by selecting a subset of the > 1300 computed DFT descriptors which could be most easily interpreted on a mechanistic basis (Figure S27). Next, we used 320 (<10%) randomly selected data points to train each model and evaluated their performance in predicting the remainder of the space. This corresponds to the last point in the learning curve in figure S26. Indeed, the BLM ($R^2$ = 0.51) performed more poorly than the GP ($R^2$ = 0.73) and RF ($R^2$ = 0.72)

models, indicating non-linear relationships between reaction parameters and yield of desired product.

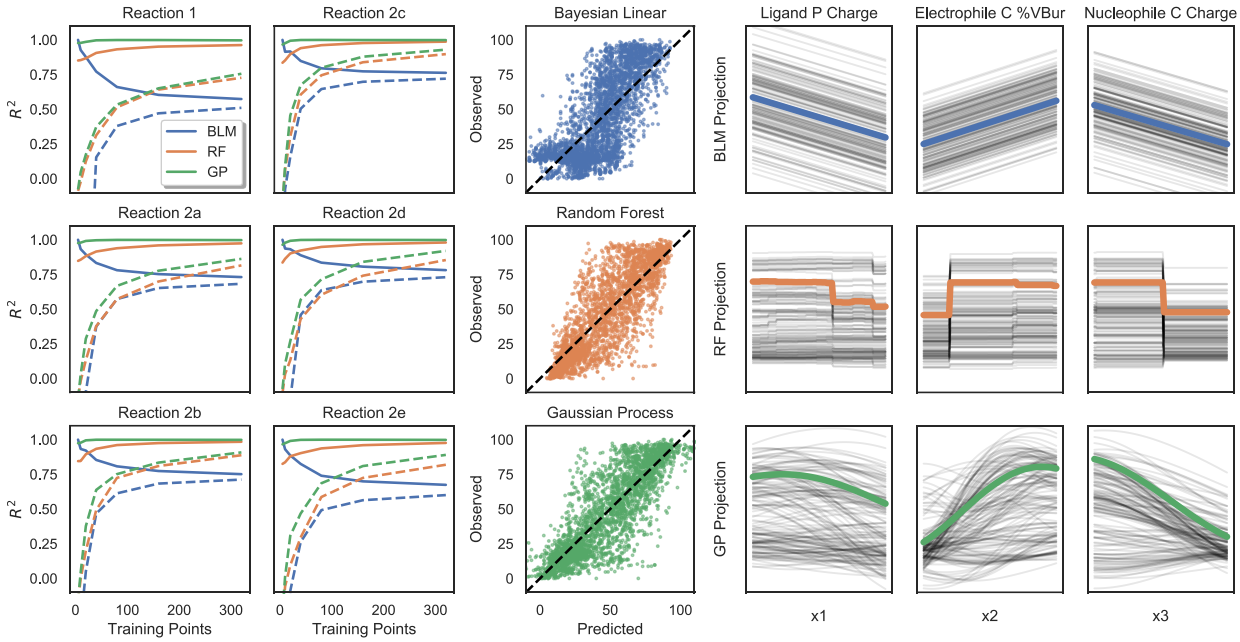



**Figure S26.** Surrogate models: (**left**) Learning curves for development objectives with Bayesian linear (blue), random forest (orange), and gaussian process (green) models. Training and test scores are solid and dashed lines, respectively. Each curve point represents the average of 25 random splits. (**center**) Test set performance for reduced descriptor set after training with 320 (<10%) data points from reaction 1 (Suzuki coupling). Test set performance: BLM $R^2$ = 0.51, RF $R^2$ = 0.72, GP $R^2$ = 0.73. (**right**) One dimensional projections of trained surrogate models generated by varying dimension $x_i$ while holding all others constant. Thin lines show surrogate samples drawn randomly from the domain and thick lines show the surrogate projection with all other dimensions set to their mean.

Figure S26 shows one dimensional projections of each model fit to 1, the DFT encoded Suzuki reaction data. The projections are generated by computing the marginal impact of one descriptor with the others held constant at their average (thick blue line) or at randomly selected domain point values (thin gray lines). Notice that independent of the characteristic shape of each model, projections capture similar trends. For example, each model suggests that increasing the donor capacity of ligands (more negative charge on phosphorus) tends to lead to improved yields.

Traditionally, one of the benefits of linear over non-parametric models has been interpretability. However, to the degree that we can estimate the magnitude and direction of a features impact the BL, RF, and GP models can all be interpreted similarly. Functions visualizing 1D projections can be found in the *edbo* plot utilities.

```
electrophile = ['electrophile_atom1_NPA_charge', 'electrophile_atom2_NPA_charge', 'electrophile_atom1_NMR_shift',
        'electrophile_lumo_energy', 'electrophile_hardness', 'electrophile_electrophilicity', 'electrophile_dipole',
        'electrophile_atom1_%VBur', 'electrophile_atom2_%VBur', 'electrophile_molar_volume', 'electrophile_molar_mass',
        'electrophile_electronic_spatial_extent']

nucleophile = ['nucleophile_atom1_NPA_charge', 'nucleophile_atom2_NPA_charge', 'nucleophile_atom1_NMR_shift',
        'nucleophile_homo_energy', 'nucleophile_lumo_energy', 'nucleophile_hardness', 'nucleophile_electrophilicity',
        'nucleophile_dipole', 'nucleophile_atom1_%VBur', 'nucleophile_atom2_%VBur', 'nucleophile_molar_volume',
        'nucleophile_molar_mass', 'nucleophile_electronic_spatial_extent']

base = ['base_atom1_NPA_charge', 'base_homo_energy', 'base_hardness', 'base_dipole', 'base_atom1_%VBur',
        'base_molar_volume', 'base_molar_mass', 'base_electronic_spatial_extent', 'base_c_min_atom=F',
        'base_c_min_atom=N', 'base_c_min_atom=O']

ligand = ['ligand_c_max_NPA_charge_Boltz', 'ligand_c_max_NPA_valence_Boltz', 'ligand_homo_energy_Boltz', 'ligand_lumo_energy_Boltz',
        'ligand_hardness_Boltz', 'ligand_dipole_Boltz', 'ligand_c_max_%VBur_Boltz', 'ligand_c_max_%VBur_MAXG', 'ligand_c_max_%VBur_MING',
        'ligand_molar_volume_Boltz', 'ligand_electronic_spatial_extent_Boltz', 'ligand_molar_mass_Boltz', 'ligand_c_min_atom=C_Boltz',
        'ligand_c_min_atom=N_Boltz', 'ligand_c_min_atom=O_Boltz', 'ligand_c_min_%VBur_Boltz', 'ligand_molar_mass_Boltz']

solvent = ['solvent_c_min_NPA_charge', 'solvent_lumo_energy', 'solvent_homo_energy', 'solvent_hardness', 'solvent_dipole',
           'solvent_c_min_%VBur', 'solvent_molar_mass', 'solvent_c_min_atom=C', 'solvent_c_min_atom=O']

suz.data = suz.data[electrophile + nucleophile + base + ligand + solvent + ['yield']]
suz.uncorrelated(threshold=0.95)
```

**Figure S27.** Screen shot of descriptors selected for each component of the Suzuki reaction.

Bayesian optimization performance. To probe further, we investigated optimization performance of each model type by running simulations on the 6 development objectives (Figure S28). For each model, statistics were generated by randomly initializing optimizations at different points in the search space. Overall, we found that BO performance was largely consistent with regression results. Importantly, both RF and GP models showed promising performance. While GPs gave marginally improved prediction performance it is worth considering RFs if encodings deviate significantly from those investigated in this work. This is because RF regressions are more robust to changes in encoding, dimensionality, and scaling. GPs in contrast work best with purposefully scaled features and parameters tuned for a given data type.

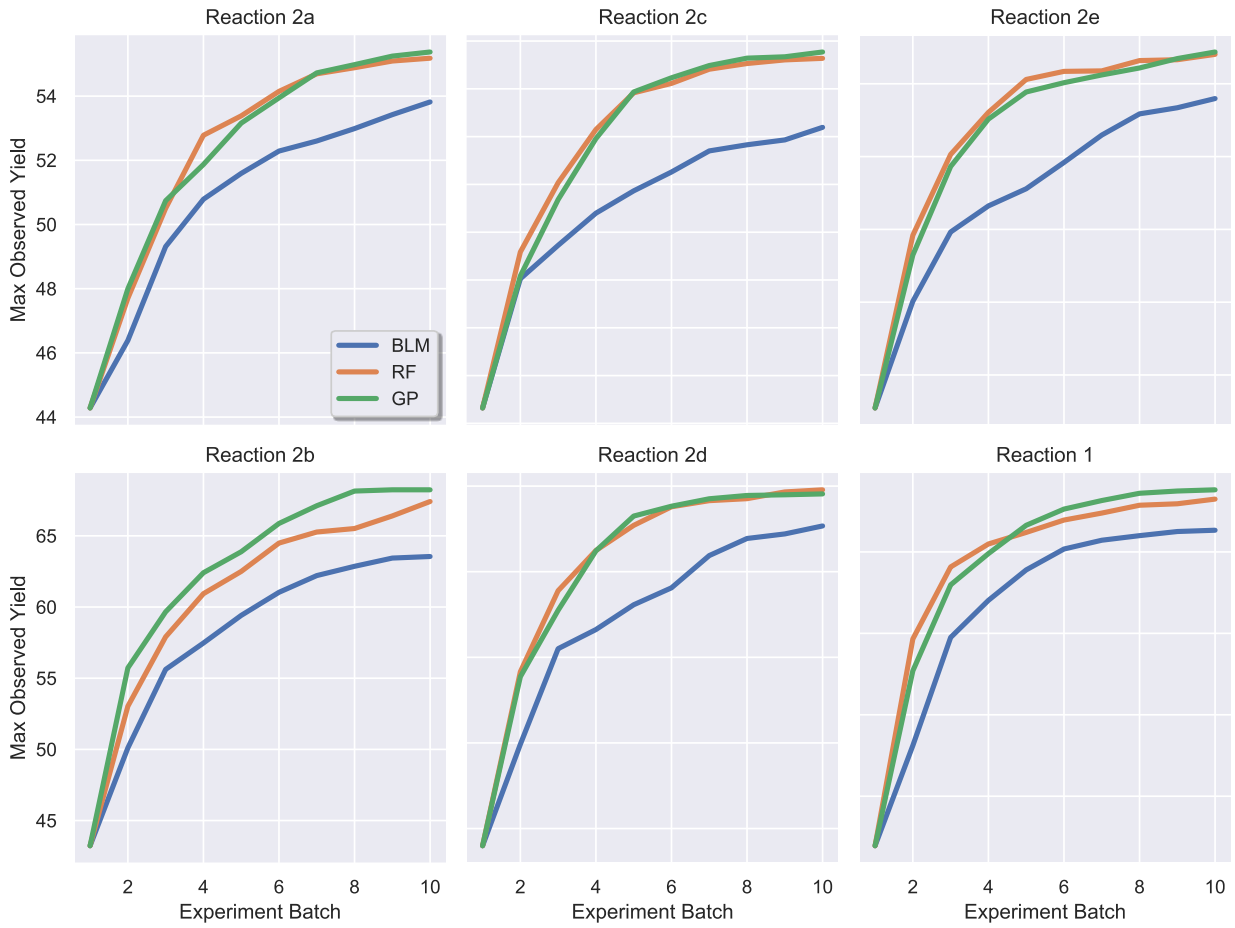



**Figure S28.** Surrogate model BO performance. Optimization curves for development objectives with Bayesian linear (blue), random forest (orange), and gaussian process (green) models. Plots show average performance over 30 random initializations, with EI as an acquisition function, and batch size = 5.

## VI. Acquisition Functions

Functions. Acquisition function selection was carried out entirely on an experimental basis, by comparing BO performance on the 6 development objectives. Over the course of our studies we investigated popular acquisition functions from the BO literature[17] including expected improvement (EI)[22], probability of improvement (PI)[23], upper confidence bound (UCB)[24], and Thompson sampling (TS).[25] For comparison, we also included acquisition functions which either were purely exploitative (mean maximization), purely exploratory (variance maximization), balanced exploration and exploitation via random selection of the next point with probability ε (ε-Greedy), or selected points at random. See GitHub for implementation details.

Parallel acquisition. We sought to investigate methods for the selection of arbitrary batches of experiments to run in parallel. TS naturally lends itself to selecting batches of experiments via the sampling of *N* candidate response surfaces from the posterior predictive distribution of a GP surrogate model (Figure S29). To achieve parallel decision making for analytic base functions (EI, PI, UCB), we iteratively fantasize the experiments which maximize the acquisition function, at each step taking the previous iterations surrogate model on faith and factoring its prediction into the next selected experiment (Figure S30). This approach, sometimes called the Kriging believer algorithm, can be thought of as a chess player thinking ahead multiple moves, at each step inferring the probable consequences of their actions. In addition, we consider the construction of hybrid acquisition functions to encourage batch diversity. For hybrid functions, the first experiment was selected via an analytic acquisition function and the remainder were selected via TS.

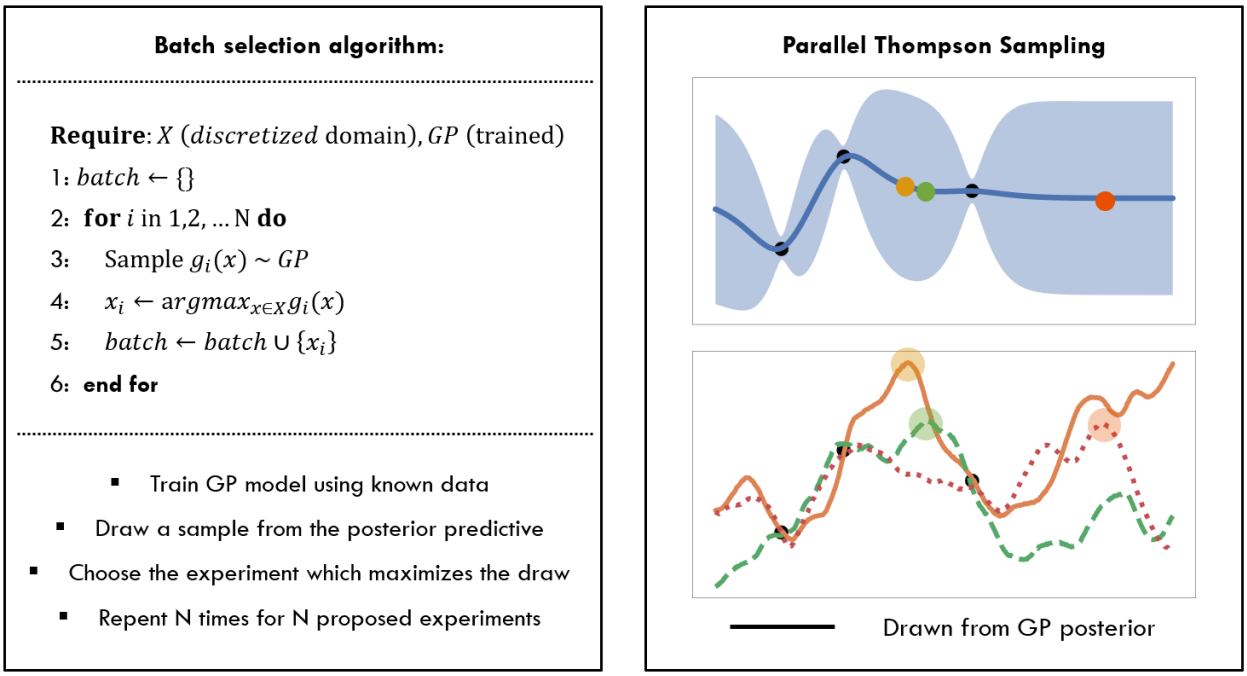



**Figure S29.** Batch selection via parallel Thompson sampling.

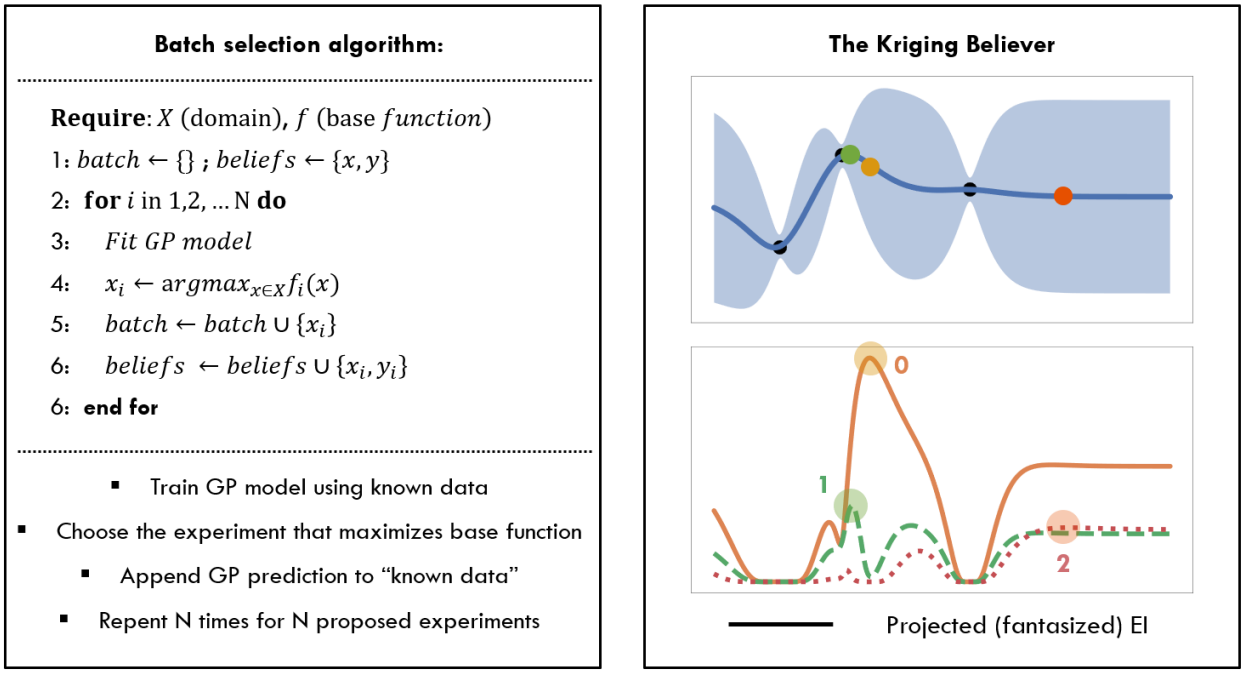



**Figure S30.** Batch selection via the Kriging believer algorithm.

Bayesian optimization performance. We investigated optimization performance of each acquisition function by running simulations on the 6 development objectives (Figure S31–S36). For each acquisition function, statistics were generated by randomly initializing optimizations at different points in the search space. Overall, we found that EI with the kriging believer algorithm gave the most consistent results. This can be best seen in figure S37, which displays the distribution of results after 10 batches of experiments as a box plot of losses. Next, we investigated the effect of batch size on the two basic approaches, EI and TS. We found that both parallel EI and TS gave impressive performance, on average finding (near)optimal configurations in only 50 experiments including 5 random initialization points (figures S38). However, between the two we found that the average performance of parallel EI was the most promising.

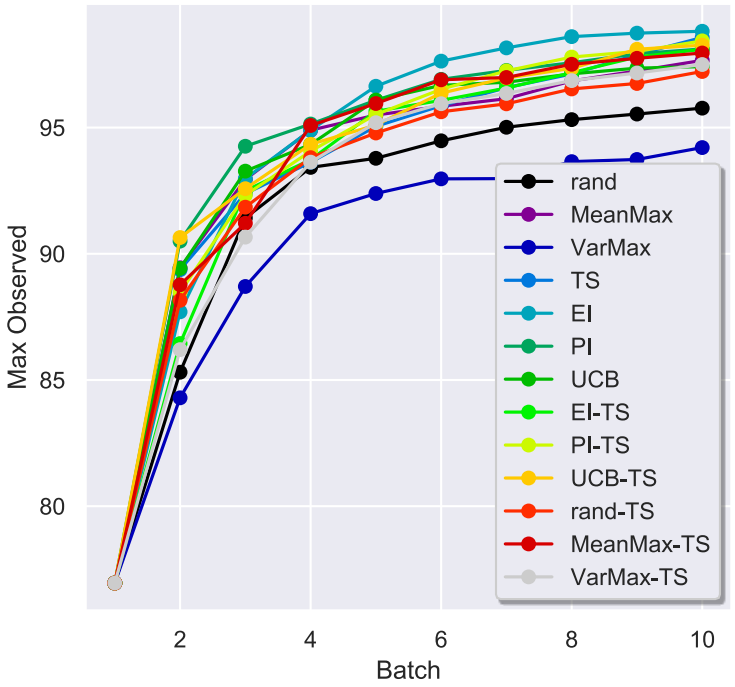



**Figure S31**. Acquisition function performance on development objective **1**. Average of 30 random initializations with a GP surrogate model, DFT encoding, and batch size = 5.

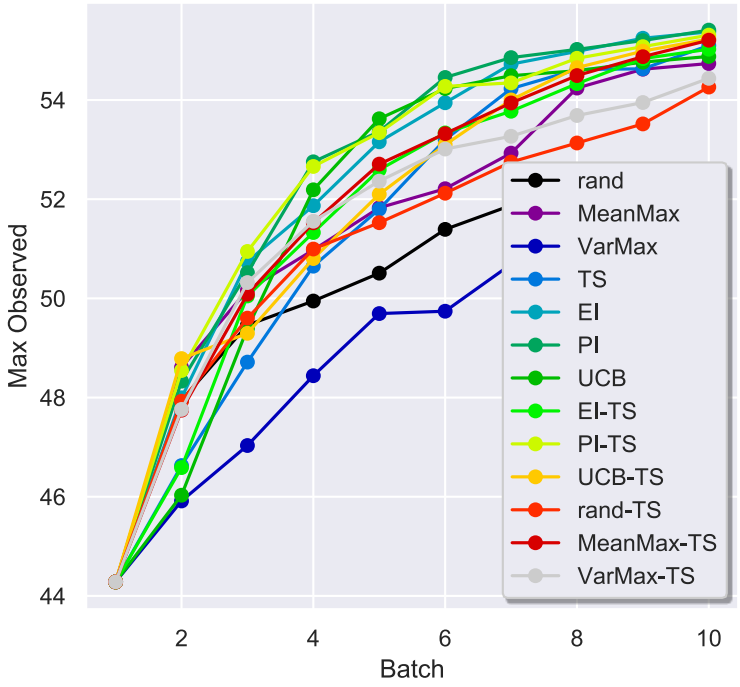



**Figure S32**. Acquisition function performance on development objective **2a**. Average of 30 random initializations with a GP surrogate model, DFT encoding, and batch size = 5.

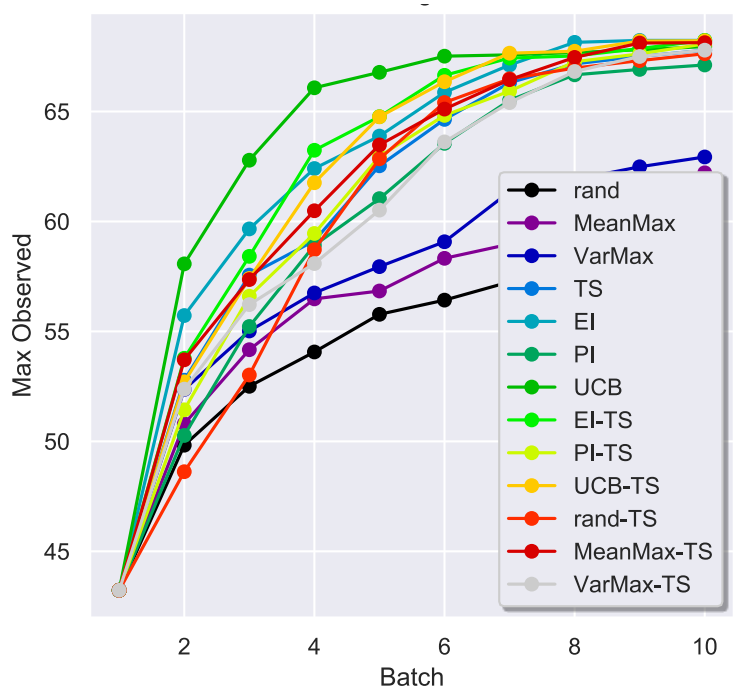



**Figure S33**. Acquisition function performance on development objective **2b**. Average of 30 random initializations with a GP surrogate model, DFT encoding, and batch size = 5.

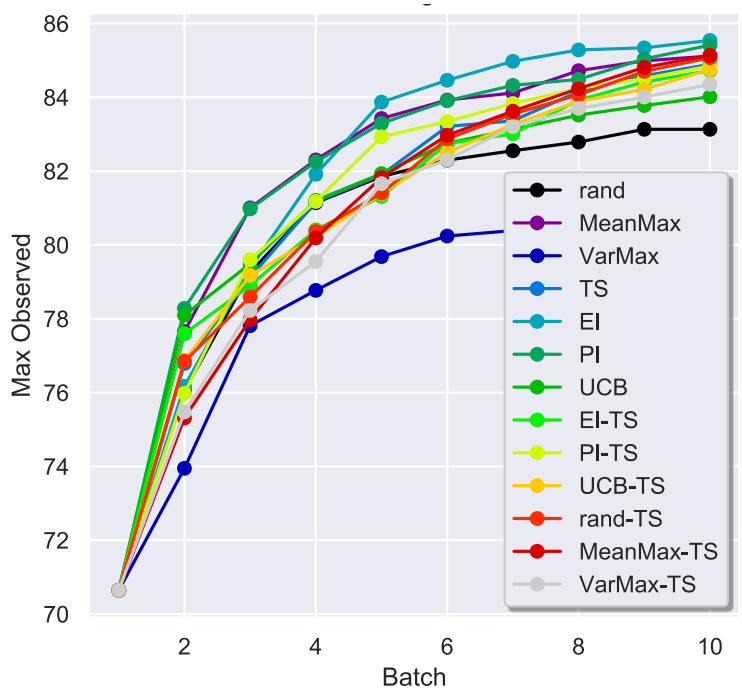



**Figure S34**. Acquisition function performance on development objective **2c**. Average of 30 random initializations with a GP surrogate model, DFT encoding, and batch size = 5.

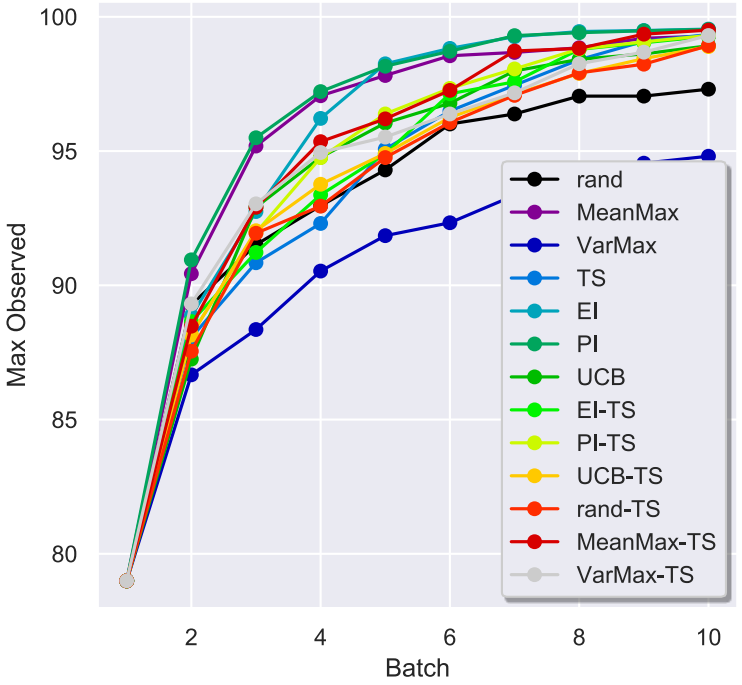



**Figure S35**. Acquisition function performance on development objective **2d**. Average of 30 random initializations with a GP surrogate model, DFT encoding, and batch size = 5.

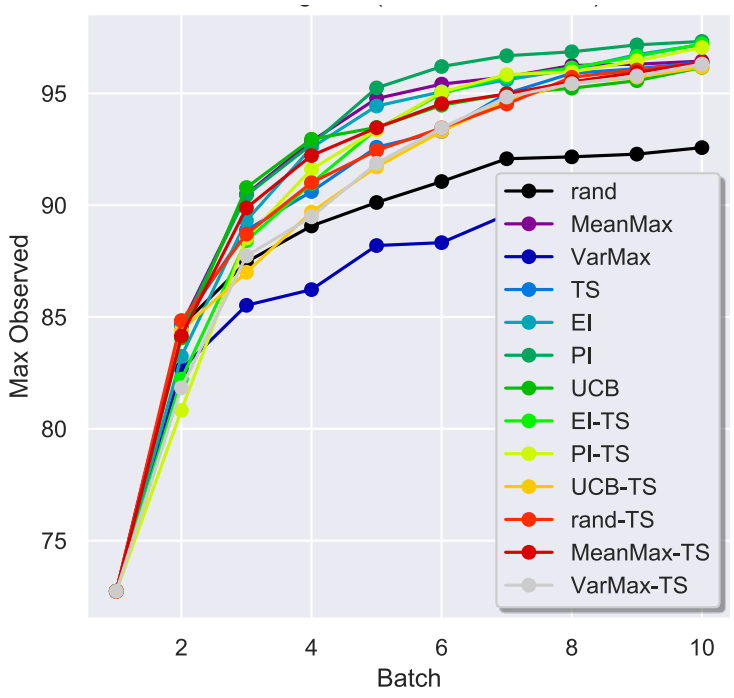



**Figure S36**. Acquisition function performance on development objective **2e**. Average of 30 random initializations with a GP surrogate model, DFT encoding, and batch size = 5.

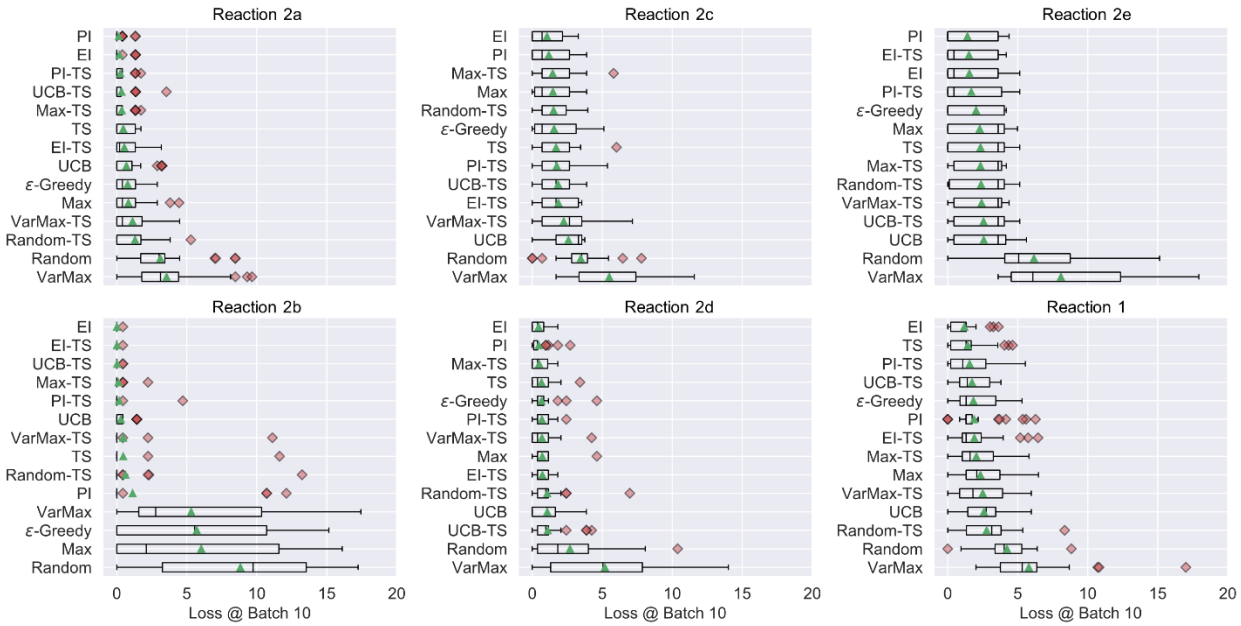



**Figure S37.** Summary of acquisition function performance on development objectives. Box and whisker plots showing performance after 10 batches of experiments using 30 random initializations, a GP surrogate model, DFT encoding, and batch size = 5. Green triangles and red diamonds denote the mean and outliers, respectively. Plots are sorted by average performance.

| | 1 | 2a | 2b | 2c | 2d | 2e |
|---|---|---|---|---|---|---|
| EI | 1 | 0 | 0 | 1 | 0 | 2 |
| EI-TS | 2 | 1 | 0 | 2 | 1 | 2 |
| ε-greedy(ε=0.1) | 2 | 1 | 6 | 2 | 1 | 2 |
| MeanMax-TS | 2 | 0 | 0 | 1 | 0 | 2 |
| MeanMax | 2 | 1 | 6 | 1 | 1 | 2 |
| PI-TS | 2 | 0 | 0 | 2 | 1 | 2 |
| PI | 2 | 0 | 1 | 1 | 0 | 1 |
| rand-TS | 3 | 1 | 1 | 2 | 1 | 2 |
| rand | 4 | 3 | 9 | 3 | 3 | 6 |
| TS | 1 | 0 | 0 | 2 | 1 | 2 |
| UCB-TS | 2 | 0 | 0 | 2 | 1 | 3 |
| UCB | 3 | 1 | 0 | 3 | 1 | 3 |
| VarMax-TS | 3 | 1 | 0 | 2 | 1 | 2 |
| VarMax | 6 | 4 | 5 | 6 | 5 | 8 |

**Table S6.** Mean loss for BO with acquisition functions after 10 batches of 5 experiments using 30 seeded random initializations, a GP surrogate model, and DFT encoding.

| | 1 | 2a | 2b | 2c | 2d | 2e |
|---|---|---|---|---|---|---|
| EI | 0.9 | 0.5 | 0.1 | 1.1 | 0.5 | 1.9 |
| EI-TS | 1.6 | 0.7 | 0.1 | 1.4 | 0.6 | 1.8 |
| ε-greedy(ε=0.1) | 1.5 | 0.8 | 5.8 | 1.5 | 0.9 | 2.0 |
| MeanMax-TS | 1.5 | 0.5 | 0.4 | 1.5 | 0.5 | 1.7 |
| MeanMax | 1.8 | 1.1 | 6.2 | 1.4 | 0.8 | 2.0 |
| PI-TS | 1.5 | 0.5 | 0.8 | 1.4 | 0.6 | 1.9 |
| PI | 1.6 | 0.3 | 3.4 | 1.3 | 0.6 | 1.8 |
| rand-TS | 1.9 | 1.4 | 2.4 | 1.2 | 1.3 | 2.1 |
| rand | 1.7 | 2.2 | 5.2 | 1.7 | 2.9 | 3.8 |
| TS | 1.4 | 0.6 | 2.1 | 1.4 | 0.7 | 1.9 |
| UCB-TS | 1.3 | 0.7 | 0.1 | 1.2 | 1.1 | 1.8 |
| UCB | 1.3 | 1.1 | 0.5 | 1.3 | 1.1 | 1.9 |
| VarMax-TS | 1.9 | 1.3 | 2.0 | 1.9 | 0.9 | 1.8 |
| VarMax | 3.0 | 2.6 | 4.7 | 2.6 | 3.5 | 4.1 |

**Table S7.** Standard deviation for BO with different acquisition functions after 10 batches of 5 experiments using 30 seeded random initializations, a GP surrogate model, and DFT encoding.

| | 1 | 2a | 2b | 2c | 2d | 2e |
|---|---|---|---|---|---|---|
| EI-TS | 0.044 | 0.035 | 1.000 | 0.016 | 0.055 | 0.953 |
| ε-greedy(ε=0.1) | 0.049 | 0.001 | 0.000 | 0.162 | 0.285 | 0.347 |
| MeanMax-TS | 0.012 | 0.207 | 0.196 | 0.242 | 0.773 | 0.100 |
| MeanMax | 0.004 | 0.008 | 0.000 | 0.207 | 0.160 | 0.148 |
| PI-TS | 0.261 | 0.597 | 0.328 | 0.043 | 0.105 | 0.789 |
| PI | 0.047 | 0.737 | 0.083 | 0.683 | 0.926 | 0.769 |
| rand-TS | 0.000 | 0.000 | 0.186 | 0.127 | 0.024 | 0.117 |
| rand | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| TS | 0.427 | 0.058 | 0.263 | 0.061 | 0.216 | 0.120 |
| UCB-TS | 0.069 | 0.375 | 0.562 | 0.010 | 0.008 | 0.043 |
| UCB | 0.000 | 0.027 | 0.008 | 0.000 | 0.008 | 0.046 |
| VarMax-TS | 0.002 | 0.001 | 0.245 | 0.005 | 0.217 | 0.077 |
| VarMax | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |

**Table S8.** Welch's t-test p-values for the null hypothesis of equal average performance to EI for BO with different acquisition functions after 10 batches of 5 experiments using 30 seeded random initializations, a GP surrogate model, and DFT encoding.

| | 1 | 2a | 2b | 2c | 2d | 2e |
|---|---|---|---|---|---|---|
| EI | 4 | 1 | 0 | 3 | 2 | 5 |
| EI-TS | 6 | 3 | 0 | 4 | 2 | 4 |
| ε-greedy(ε=0.1) | 5 | 3 | 15 | 5 | 5 | 4 |
| MeanMax-TS | 6 | 2 | 2 | 6 | 2 | 4 |
| MeanMax | 6 | 4 | 16 | 4 | 5 | 5 |
| PI-TS | 6 | 2 | 5 | 5 | 2 | 5 |
| PI | 6 | 1 | 12 | 4 | 3 | 4 |
| rand-TS | 8 | 5 | 13 | 4 | 7 | 5 |
| rand | 9 | 8 | 17 | 8 | 10 | 15 |
| TS | 5 | 2 | 12 | 6 | 3 | 5 |
| UCB-TS | 4 | 4 | 0 | 4 | 4 | 5 |
| UCB | 6 | 3 | 1 | 4 | 4 | 6 |
| VarMax-TS | 6 | 4 | 11 | 7 | 4 | 4 |
| VarMax | 17 | 10 | 17 | 12 | 14 | 18 |

**Table S9.** Worst-case loss for BO with different acquisition functions after 10 batches of 5 experiments using 30 seeded random initializations, a GP surrogate model, and DFT encoding.

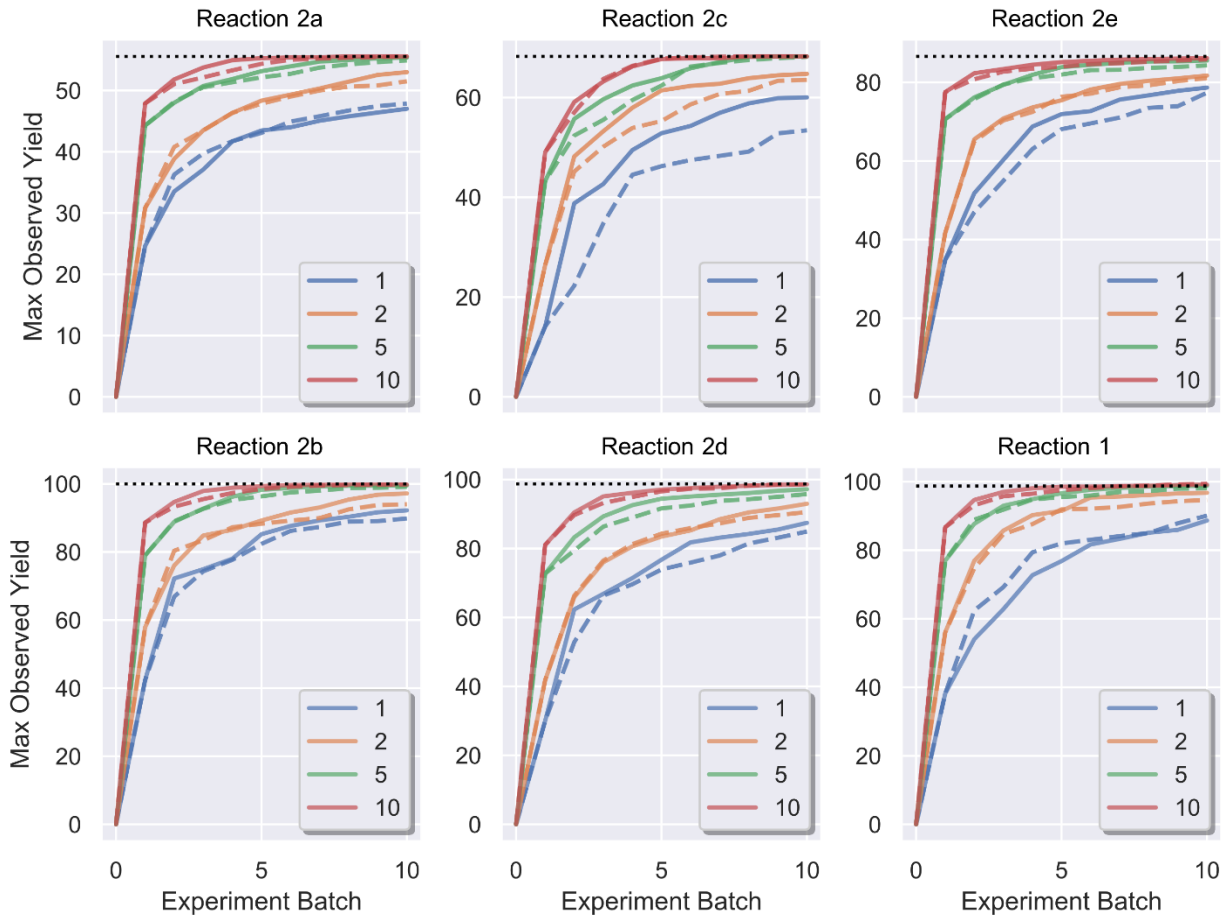



**Figure S38.** Summary of parallel EI (solid) and TS (dashed) performance in time with different batch sizes. Increasing batch size increases convergence rate in time. Average of 30 random initializations.

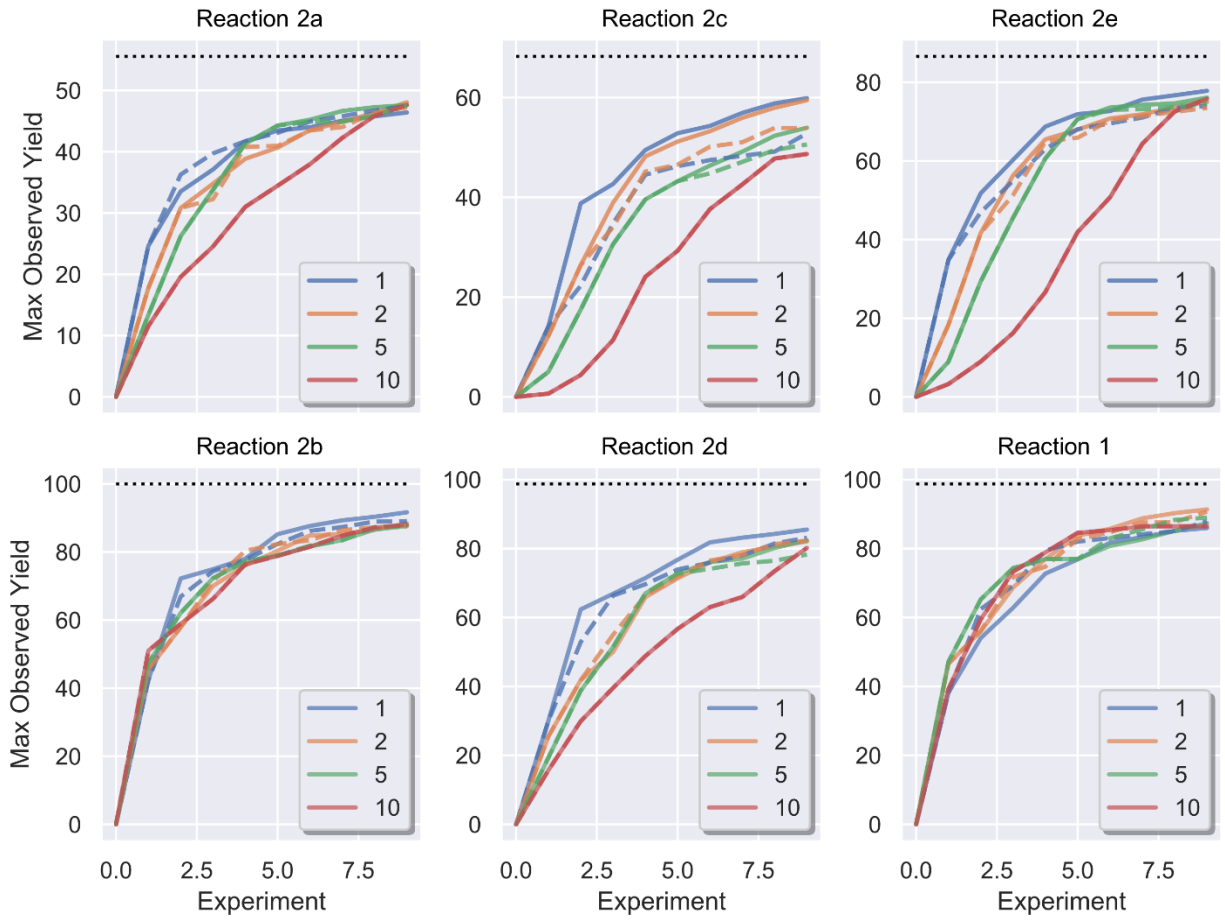



**Figure S39.** Summary of parallel EI (solid) and TS (dashed) performance in experiments with different batch sizes. Increasing batch size decreases convergence rate in experiments. Average of 30 random initializations.

| | 1 | 2a | 2b | 2c | 2d | 2e |
|---|---|---|---|---|---|---|
| BS=1 p-value | 0.259 | 0.046 | 0.346 | 0.112 | 0.445 | 0.805 |
| BS=1 $\mu_{Loss}$ | 2 | 0 | 0 | 2 | 0 | 2 |
| BS=1 σ | 1.3 | 0.1 | 1.9 | 2.0 | 0.4 | 1.9 |
| BS=1 ξ | 6 | 0 | 11 | 11 | 1 | 5 |
| BS=2 p-value | 0.312 | 0.474 | 0.326 | 0.084 | 0.343 | 0.939 |
| BS=2 $\mu_{Loss}$ | 1 | 0 | 0 | 2 | 0 | 2 |
| BS=2 σ | 0.6 | 0.3 | 0.0 | 1.3 | 0.3 | 1.9 |
| BS=2 ξ | 4 | 1 | 0 | 4 | 1 | 4 |
| BS=5 p-value | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| BS=5 $\mu_{Loss}$ | 1 | 0 | 0 | 1 | 0 | 2 |
| BS=5 σ | 0.9 | 0.5 | 0.1 | 1.1 | 0.5 | 1.9 |
| BS=5 ξ | 4 | 1 | 0 | 3 | 2 | 5 |
| BS=10 p-value | 0.219 | 0.369 | 0.187 | 0.187 | 0.540 | 0.611 |
| BS=10 $\mu_{Loss}$ | 2 | 0 | 0 | 2 | 1 | 2 |
| BS=10 σ | 1.4 | 0.5 | 0.4 | 1.5 | 0.8 | 1.9 |
| BS=10 ξ | 4 | 1 | 2 | 5 | 4 | 4 |

**Table S10.** Parallel acquisition statistics for BO with EI as an acquisition function after 50 total experiments. Statistics are for 30 seeded random initializations, a GP surrogate model, and DFT encoding: (p-value) Welch's t-test p-values for the null hypothesis of equal average performance to BS=5, ($\mu_{Loss}$) mean loss, (σ) standard deviation, (ξ) worst-case loss.

| | **1** | **2a** | **2b** | **2c** | **2d** | **2e** |
|---|---|---|---|---|---|---|
| BS=1 p-value | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| BS=1 $\mu_{Loss}$ | 10 | 8 | 7 | 9 | 9 | 13 |
| BS=1 $\sigma$ | 9.0 | 5.3 | 7.3 | 5.9 | 7.2 | 10.0 |
| BS=1 $\xi$ | 44 | 19 | 21 | 27 | 30 | 37 |
| BS=2 p-value | 0.000 | 0.000 | 0.003 | 0.000 | 0.000 | 0.000 |
| BS=2 $\mu_{Loss}$ | 3 | 3 | 3 | 5 | 3 | 6 |
| BS=2 $\sigma$ | 1.8 | 2.9 | 5.7 | 2.9 | 3.1 | 5.0 |
| BS=2 $\xi$ | 8 | 13 | 16 | 14 | 11 | 21 |
| BS=5 p-value | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| BS=5 $\mu_{Loss}$ | 1 | 0 | 0 | 1 | 0 | 2 |
| BS=5 $\sigma$ | 0.9 | 0.5 | 0.1 | 1.1 | 0.5 | 1.9 |
| BS=5 $\xi$ | 4 | 1 | 0 | 3 | 2 | 5 |
| BS=10 p-value | 0.349 | 0.031 | 0.326 | 0.085 | 0.005 | 0.001 |
| BS=10 $\mu_{Loss}$ | 1 | 0 | 0 | 1 | 0 | 0 |
| BS=10 $\sigma$ | 1.1 | 0.0 | 0.0 | 0.8 | 0.2 | 0.6 |
| BS=10 $\xi$ | 4 | 0 | 0 | 3 | 0 | 4 |

**Table S11.** Parallel acquisition statistics for BO with EI as an acquisition function after 10 batches of experiments. Statistics are for 30 seeded random initializations, a GP surrogate model, and DFT encoding: (p-value) Welch's t-test p-values for the null hypothesis of equal average performance to BS=5, ($\mu_{Loss}$) mean loss, ($\sigma$) standard deviation, ($\xi$) worst-case loss.

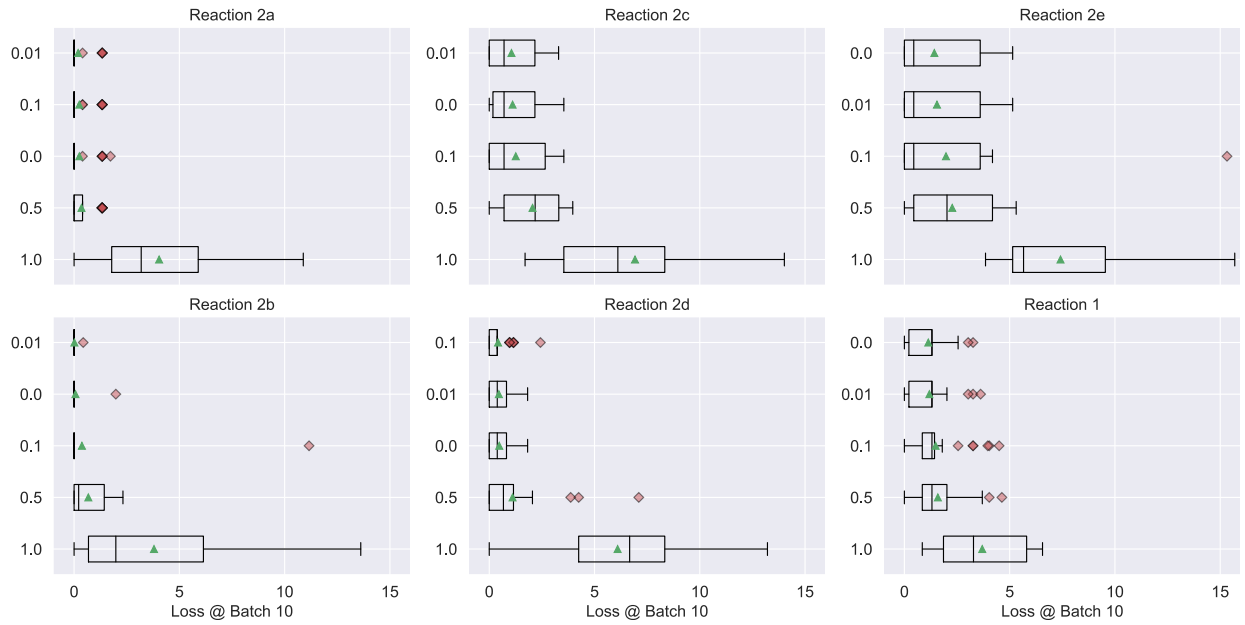



**Figure S40.** Exploring the exploration parameter for EI. Box and whisker plots showing performance after 10 batches of experiments using 30 random initializations, a Gaussian Process surrogate model, DFT encoding, and batch size = 5. Green triangles and red diamonds denote the mean and outliers, respectively. Plots are sorted by average performance.

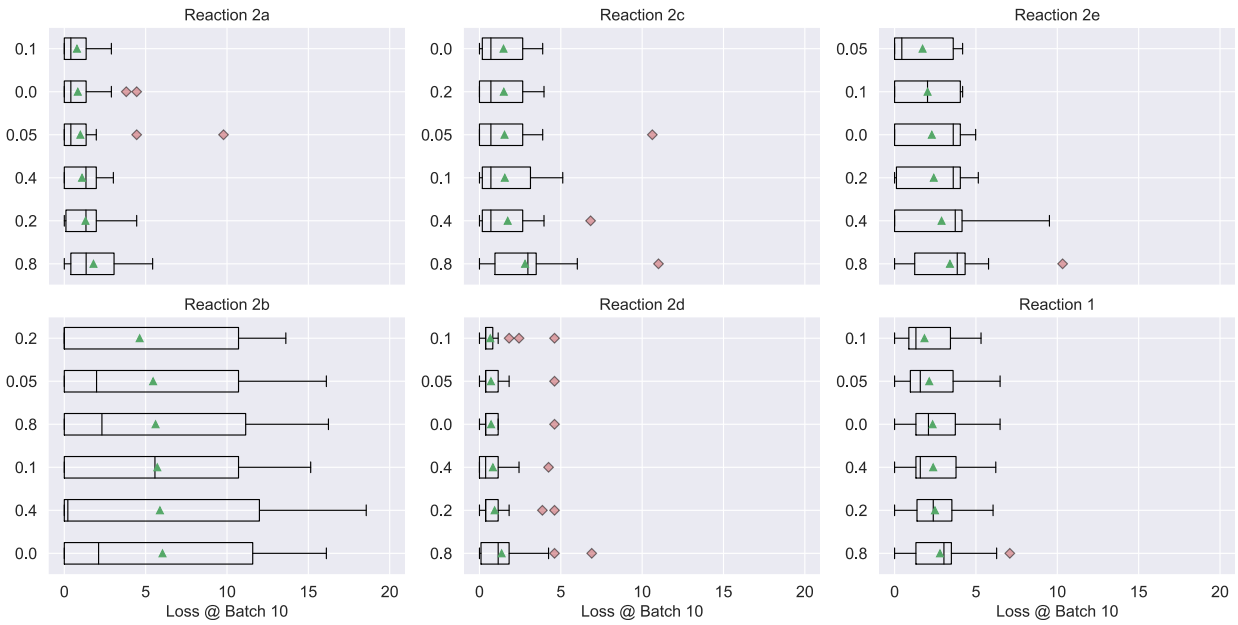



**Figure S41.** Selecting ε for ε-greedy experiments. Box and whisker plots showing performance after 10 batches of experiments using 30 random initializations, a Gaussian Process surrogate model, DFT encoding, and batch size = 5. Green triangles and red diamonds denote the mean and outliers, respectively. Plots are sorted by average performance.

## VII. Test Reaction

### Experiment Design

Background. With our BO framework tuned for reaction optimization, we next sought to test its performance in a new reaction space. We chose to investigate the direct arylation of imidazole, exemplified by reaction 3 (Figure S42). Importantly, the direct arylation of this imidazole is a key step in the commercial synthesis of the JAK2 inhibitor BMS-911543 (Figure 1).[26,27] Our approach was to leverage HTE to generate a search space and then use simulations to investigate BO performance under the most promising configuration identified via the development data (DFT encoding, optimized GP surrogate model, and parallel EI as an acquisition function). Thus, our first step was to design the search space over which we would optimize.

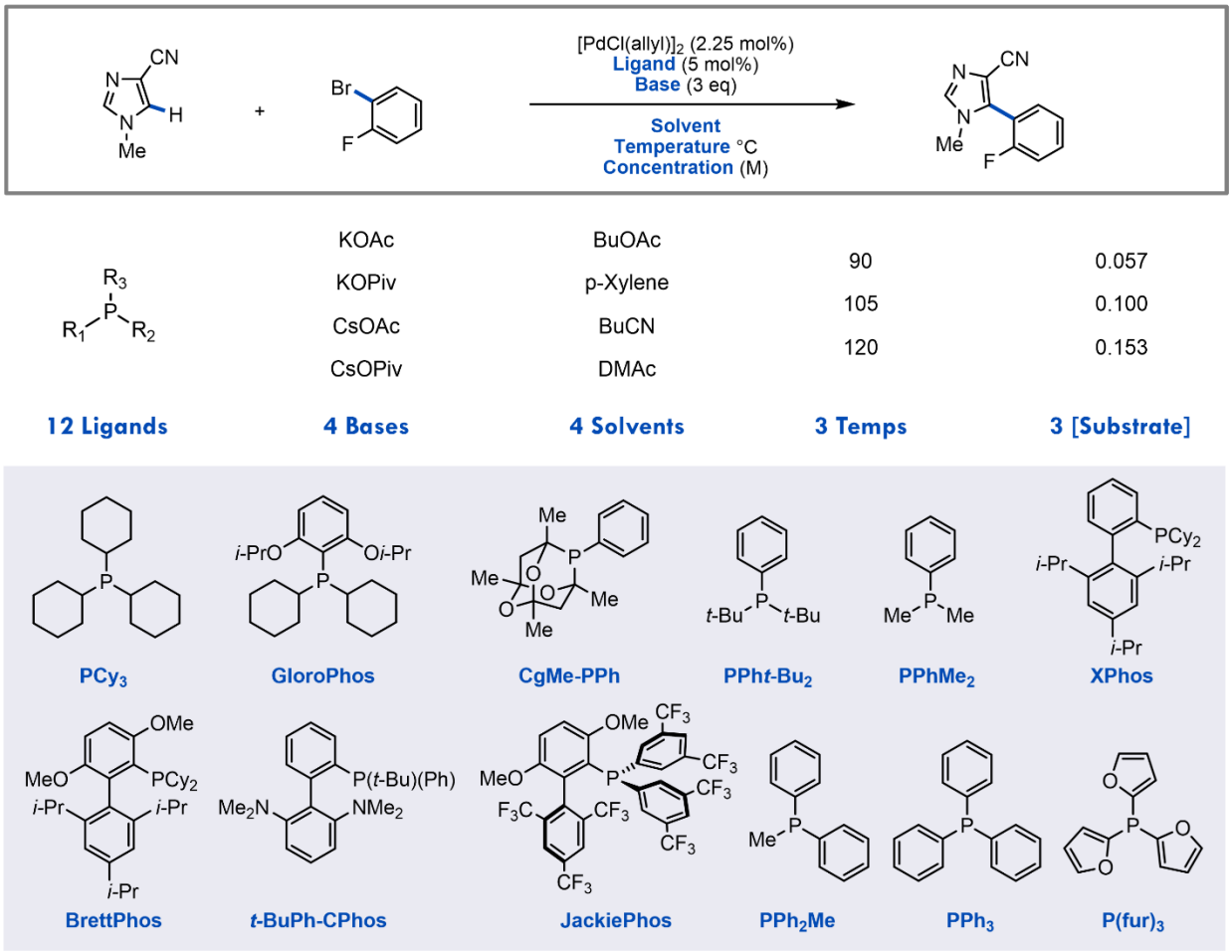



**Figure S42.** Test reaction **3** including the selected search space.

Search space selection. We advocate a hybrid approach to selecting candidates, wherein we first consider the larger set of plausible configurations, then quantify similarities between potential reaction components via unsupervised learning and select those which ensure a satisfactory distribution across the larger search space. Pd-catalyzed direct arylation of imidazoles had already been investigated at BMS due to the transformation's importance in the synthesis of BMS-911543. Over the course of these studies the following parameters were found to significantly impact the yield of desired product: (1) ligand, (2) base, (3) solvent, (4) temperature, (5) concentration of substrate, and (6) Pd-precatalyst. For consistency, we decided to carry out experiments with only a single precatalyst. In particular, in preliminary HTE experiments, we found that allylpalladium(II) chloride dimer gave the most consistent results over other Pd precatalysts. Thus, we selected reaction parameters 1–5 as the dimensions which would define the reaction space. Each of these parameters have numerous possible choices. Conservatively we estimated that on the order of 100 ligands, 10 bases, 10 solvents, 10 temperatures, and 10 concentrations would reasonably capture the reaction space defined by availability (in the BMS library) and amenability to HTE (e.g. no low boiling solvents). This would give 1 million reactions in the conservative case. Obviously, this number of experiments is not tractable. Thus, our next step was to prune the reaction space to a reasonable number of experiments. We relied chemical intuition and previous studies to select a subset of 4 bases, 4 solvents, 3 temperatures, and 3 concentrations of substrate. Next, we reduced the candidate ligands to a subset of phosphines available in the BMS HTE library (70 ligands). This reduced the search space to 10080 possible reactions. Still too many. Thus, we were tasked with choosing a representative subset of ligands to include in the study which spanned the properties of the larger set. To do this elected to (1) compute chemical descriptors for the full set of ligands, (2) cluster the ligands into similar groups, and (3) select ligands from each cluster.

We began by computing Mordred fingerprints for each ligand. Next, we utilized *k*-means clustering with *k* selected by optimizing the silhouette width to partition the ligands into sets in similar areas of chemical space (Figure S43). We identified 9 distinct groups of ligands. Finally, we selected ligands from each relevant cluster which span the space and satisfy our prior understanding of the reaction. For example, the accepted mechanism for this direct arylation suggests the presence of an open coordination site on the metal center is important for achieving reactivity.[28,29] Indeed, while a bidentate ligand was employed in the commercial transformation, it

was found that the formation of a hemilabile phosphine mono-oxide was critical to achieve reactivity. Thus, as a simplifying assumption we excluded bidentate ligands from the search space (pink and light green clusters in figure S43).

Overall, we selected a search space consisting of 12 ligands, 4 bases, 4 solvents, 3 temperatures, and 3 concentrations (Figure S42). Finally, for each of the selected chemical components we carried out conformational analysis (where appropriate), quantum mechanical modeling, and DFT encoding using *auto-QChem*. The DFT output files and resulting DFT encodings can be found on GitHub in the examples folder.

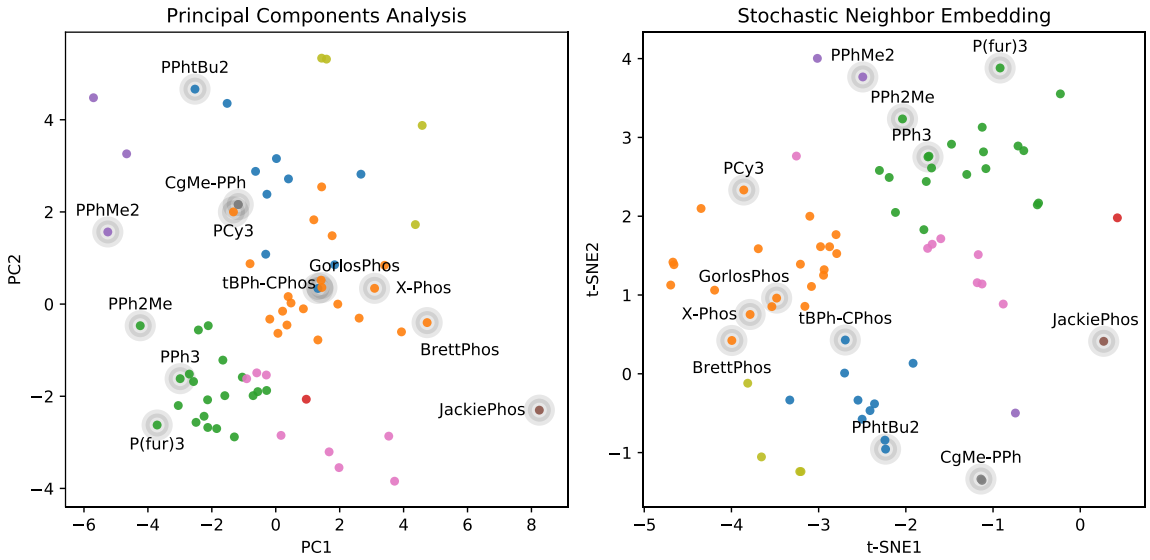



**Figure S43**. Pruning the search space via unsupervised learning. Embeddings of BMS HTE ligand library. Ligands clustered according to cheminformatics fingerprints via *k*-means with *k* selected by optimizing silhouette width. Screening ligands were selected from each cluster, excluding those which contained bidentate ligands.

## High-Throughput Experimentation

Materials & methods. Standard glovebox techniques were employed for handling air-sensitive reagents. NMR spectra were recorded on a 400 or 600 MHz spectrometer. UHPLCMS analysis was conducted on a Water Acquity with QDA detector and processed using Empower software.

All reagents and materials were acquired from commercial suppliers. Data are presented as follows: chemical shift (ppm), multiplicity (s = singlet, d = doublet, t = triplet, m = multiplet, br = broad), coupling constant J (Hz), and integration. HRMS samples were run on the Thermo LTQ-Orbitrap with Acquity Classic inlet.

Bulk Ligand Plate Preparation. In a glovebox, 8 mL vials containing 25 umol ligand were dissolved in 1,2-dichloroethane (2.5 mL) and stirred for 5 min. A 100 uL aliquot of each of the resultant ligand solutions was dispensed to the desired location in the 96 well plate. The solvent was removed *in vacuo* using a Genevac centrifugal evaporator inside the glovebox. Ligand plates were sealed and stored in the glovebox until time of use.

Bulk Base Plate Preparation. Bases were dispensed to each 96 well plate using Unchained Labs Powder Protégé. The bases were stored open in a glovebox for no less than three days prior to use to remove trace amounts of water and then sealed and stored in the glovebox until time of use.

Potassium Acetate (4.4 mg, 0.045 mmol)/well
Potassium Pivalate (6.3 mg, 0.045 mmol)/well
Cesium Acetate (8.6 mg, 0.045 mmol)/well
Cesium Pivalate (10.5 mg, 0.045 mmol)/well

Reaction Execution. In a glovebox, 0.5 M solutions of 1-bromo-2-fluorobenzene (100.8 mg, 0.58 mmol) in each butyronitrile, *n*-butyl acetate, *N*,*N*-dimethylacetamide, and p-xylene (1.15 mL) were prepared and dispensed to the designated ligand vials (40 uL, 0.02 mmol). Solutions of $[Pd(allyl)Cl]_2$ (12.5 mg, 0.034 mmol) in each butyronitrile, *n*-butyl acetate, *N*,*N*-dimethylacetamide, and p-xylene (0.76 mL) were prepared (0.045 M) and dispensed to the designated ligand vials (10 uL, 0.45 umol). The resultant reaction mixtures were sealed and stirred on a shaker block in the glovebox for no less than 30 min.

For 0.05 M reactions: 0.13 M solutions of 1-methyl-1H-imidazole-4-carbonitrile (123.4 mg, 1.15 mmol) containing 4,4'-di-*tert*-butylbiphenyl (15.3 mg, 0.057 mmol) in each butyronitrile, *n*-butyl acetate, *N*,*N*-dimethylacetamide, and p-xylene (8.60 mL) were prepared and dispensed to the entire

plate (300 uL, 0.04 mmol). The reaction mixtures were stirred for 2 min in the glovebox and 263 uL from each well was transferred to the base plate using a multichannel pipettor.

For 0.1 M reactions: 0.27 M solutions of 1-methyl-1H-imidazole-4-carbonitrile (123.4 mg, 1.15 mmol) containing 4,4'-di-*tert*-butylbiphenyl (15.3 mg, 0.057 mmol) in each butyronitrile, *n*-butyl acetate, *N*,*N*-dimethylacetamide, and p-xylene (4.30 mL) were prepared and dispensed to the entire plate (150 uL, 0.04 mmol). The reaction mixtures were stirred for 2 min in the glovebox and 150 uL from each well was transferred to the base plate using a multichannel pipettor.

For 0.15 M reactions: 0.50 M solutions of 1-methyl-1H-imidazole-4-carbonitrile (123.4 mg, 1.15 mmol) containing 4,4'-di-*tert*-butylbiphenyl (15.3 mg, 0.057 mmol) in each butyronitrile, *n*-butyl acetate, *N*,*N*-dimethylacetamide, and p-xylene (2.30 mL) were prepared and dispensed to the entire plate (80 uL, 0.04 mmol). The reaction mixtures were stirred for 2 min in the glovebox and 98 uL from each well was transferred to the base plate using a multichannel pipettor.

The reactions were then sealed and stirred at the desired temperature (90, 105 or 120 °C) for 24 h in the glovebox and subsequently cooled to 23 °C. The plate was removed from the glovebox, opened, and diluted to a 900 uL total volume with *N*,*N*-dimethylacetamide. The plate was stirred for 5 min and a 75 uL sample was taken and filtered into an HPLC analysis plate. The filter was rinsed with 400 uL acetonitrile/water (4:1) solution and analyzed by UHPLCMS

Calibration Curve: A solution of 5-(2-fluorophenyl)-1-methyl-1H-imidazole-4-carbonitrile (181.2 mg, 0.9 mmol) in 5.40 mL *N*,*N*-dimethylacetamide was prepared. A serial dilution from this solution was performed into vials containing 4,4'-di-*tert*-butylbiphenyl (0.91 mg, 3.42 umol) to generate solutions that contain 10%, 20%, 40%, 60%, 80%, and 100% of the original 5-(2-fluorophenyl)-1-methyl-1H-imidazole-4-carbonitrile vs the consistent amount of 4,4'-di-*tert*-butylbiphenyl internal standard. A 75 uL sample of each was taken and filtered into an UHPLCMS analysis plate. The filter was rinsed with 400 uL acetonitrile/water (4:1) solution and analyzed by UHPLCMS.

UHPLCMS Method:

Solvent A: Water with 5% acetonitrile and 0.05% TFA

Solvent B: acetonitrile with 5% water and 0.05% TFA

Gradient: 95% A/B to 5% A/B over 2.5 min, hold 1.0 min at 95% B

Stop Time: 3.5 min

Flow Rate: 0.8 mL/min

Wavelength1: 220

Wavelength2: 254

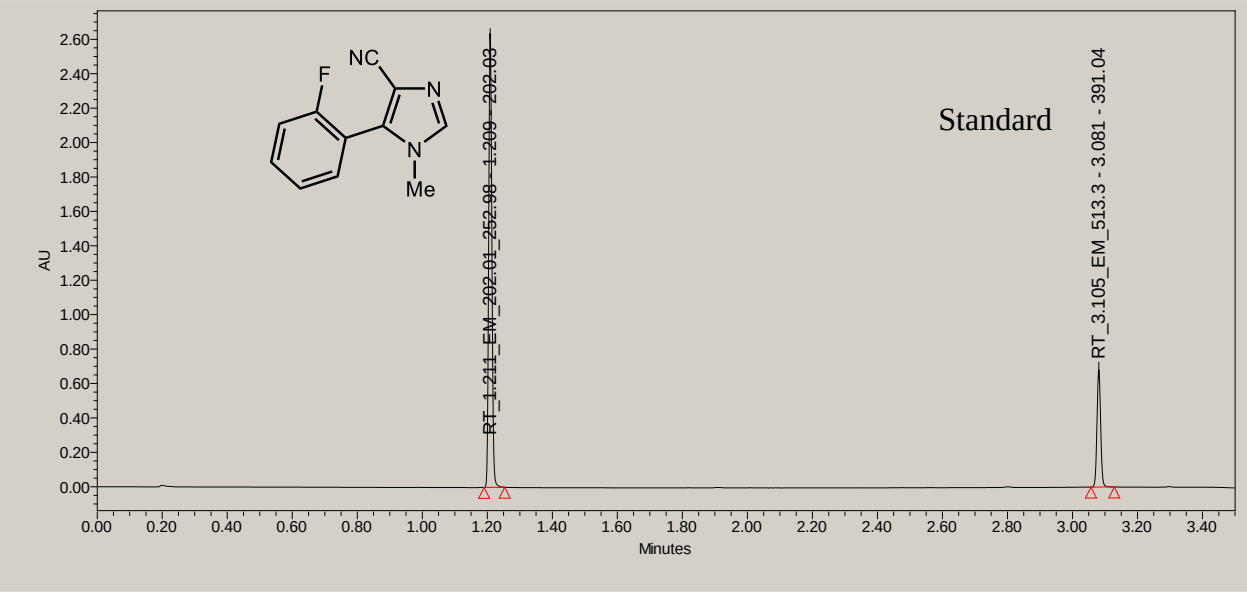
Column: Agilent Poroshell C18 2.7 um 2.1x50mm

Oven Temperature: 40

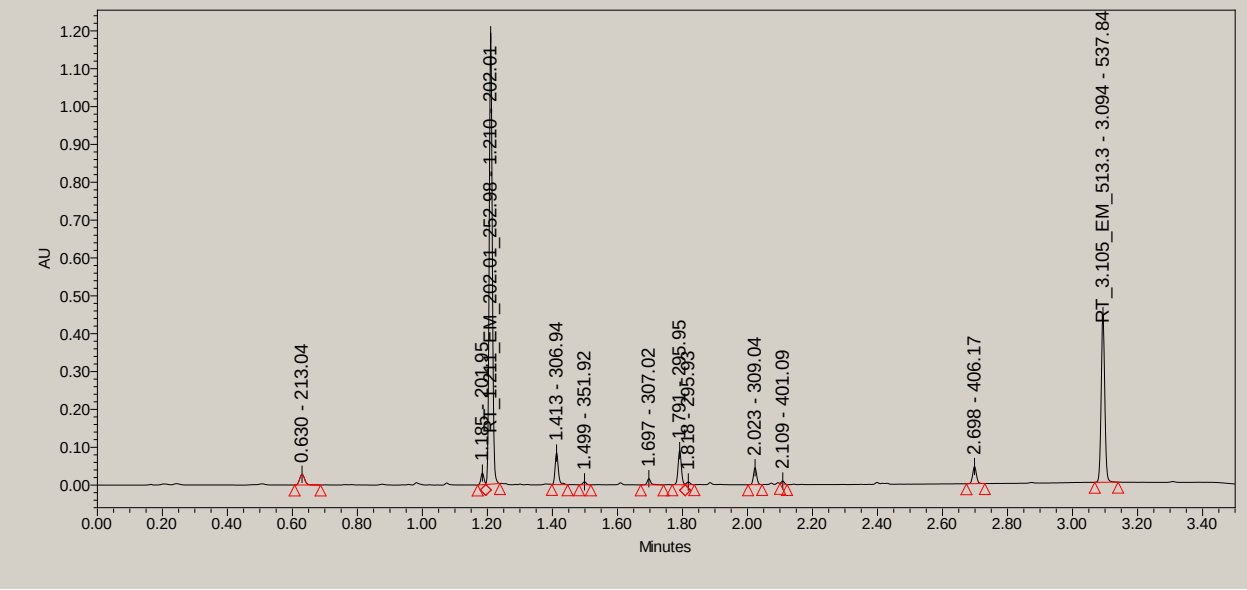
Data Processing. Data processing was completed in R 3.4.4 software installed on Ubuntu 16.04 using the tidyverse 1.2.1 package. The experimental results (Experimental_Data) and calibration results (Calibration_Data) from the UHPLCMS instrument were imported as .csv files and merged with the experimental design. The relative yields (RelYield_PDT) for the experimental and calibration data were calculated by dividing the area percent of the desired product by the area percent of the internal standard for each entry. The yields were calculated by fitting a linear model on the Calibration_Data and applying to the Experimental_Data using the function below.

```
Model <- function(Calibration_Data, Experimental_Data) {

Model <- lm(Yield ~ RelYield_PDT, Calibration_Data)
Model_coeffs <- coefficients(Model)

Experimental_Data $Yield <- Model_coeffs[1] + Model_coeffs[2]* Experimental_Data
$RelYield_PDT

Experimental_Data $Yield <- round(Experimental_Data $Yield, digits = 2)

Experimental_Data
}
```

## UHPLC-MS calibration sample: 100% yield




## UHPLC-MS reaction sample



**5-(2-fluorophenyl)-1-methyl-1H-imidazole-4-carbonitrile.** In a glovebox a 20 mL vial containing 1-bromo-2-fluorobenzene (875 mg, 5 mmol) was added a solution of $PCy_3$ $HBF_4$ (92 mg, 0.25 mmol) in *N*,*N*-dimethylacetamide (5 mL). The solution stirred for 5 min and was treated with a solution of $[Pd(allyl)Cl]_2$ (41 mg, 0.11 mmol) in *N*,*N*-dimethylacetamide (5 mL). The resulting solution stirred for 30 min at 23 °C. To the reaction mixture was added a solution of 1-methyl-1H-imidazole-4-carbonitrile (1.07 g, 10.0 mmol) in *N*,*N*-dimethylacetamide (5 mL). The resulting solution stirred for 1 min and was then decanted into a 40 mL vial containing potassium pivalate (2.10 g, 15.0 mmol) and rinsed with *N*,*N*-dimethylacetamide (5 mL). The reaction mixture was heated to 140 °C for 48 h, cooled to 23 C, filtered through diatomaceous earth, and rinsed with *N*,*N*-dimethylacetamide (10 mL). The solvent was removed *in vacuo* and the crude solids were purified by reverse phase preparative chromatography (Waters XBridge C18 19 x 250 mm, 5 micron; "A" 95/5 Water/acetonitrile + 10mM ammonium acetate, "B" 5/95 Water/acetonitrile + 10mM ammonium acetate; 20% "B" over 11.5 min at 40 mL/min; 220 nm detection, injection volume 0.75 mL of ~ 59 mg/mL in DMF) and concentration of the isolated fractions to afford the desired product (767 mg, 3.81 mmol) in 76% yield.

**$^1$H NMR (600 MHz, DMSO-$d_6$):** $\delta$ 8.03 (s), 7.65 (m), 7.59 (td, $J$ = 7.5, 1.8 Hz), 7.46 (m), 7.42 (td, $J$ = 7.5, 1.1 Hz), 3.58 (s).

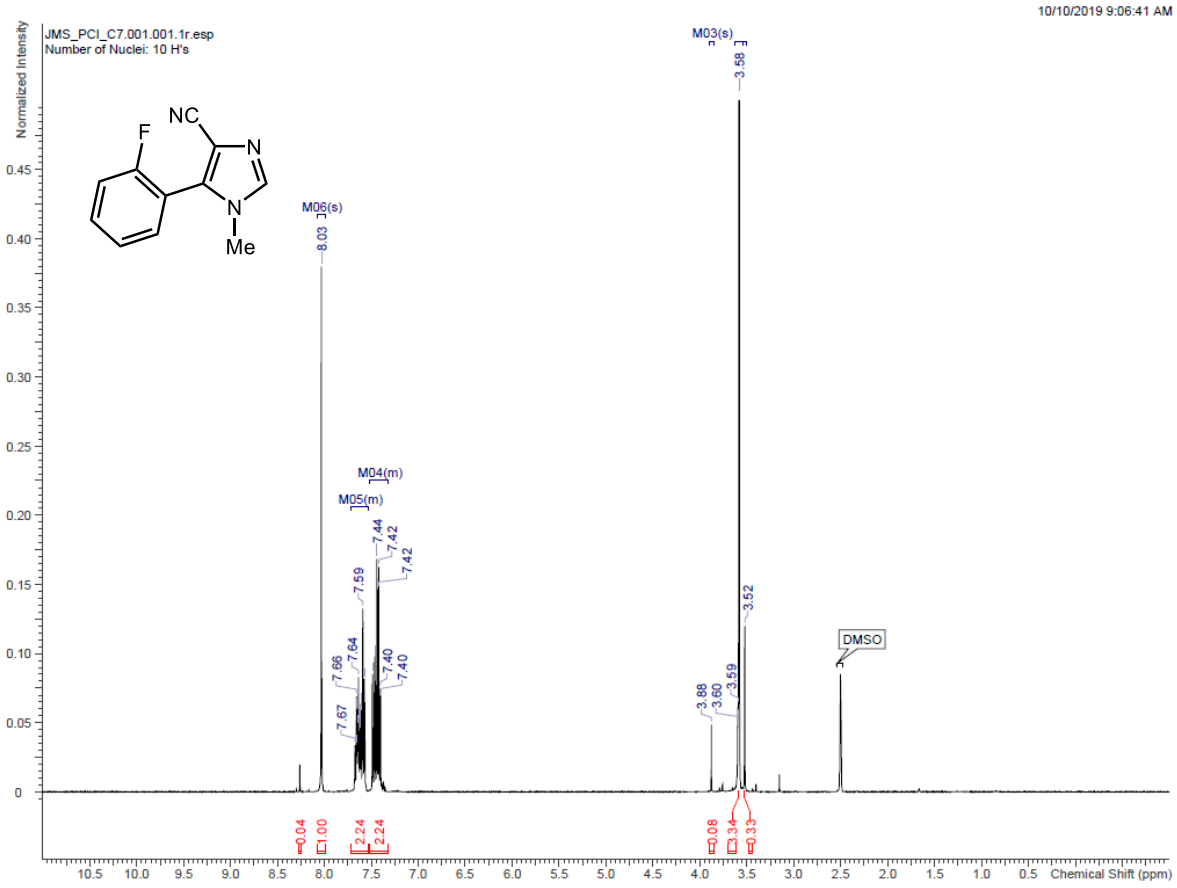
**$^{13}$C NMR (150 MHz, DMSO-$d_6$):** $\delta$ 159.3 (d, $J$ = 248.1 Hz), 141.6, 135.6, 133.1 (d, $J$ = 8.5 Hz), 132.1 (d, $J$ = 1.6 Hz), 125.5 (d, $J$ = 3.6 Hz), 116.6 (d, $J$ = 21.3 Hz), 115.5, 114.0 (d, $J$ = 15.0 Hz), 112.6, 32.8.

**$^{19}$F NMR (376 MHz, DMSO-$d_6$):** d -113.04 (m).

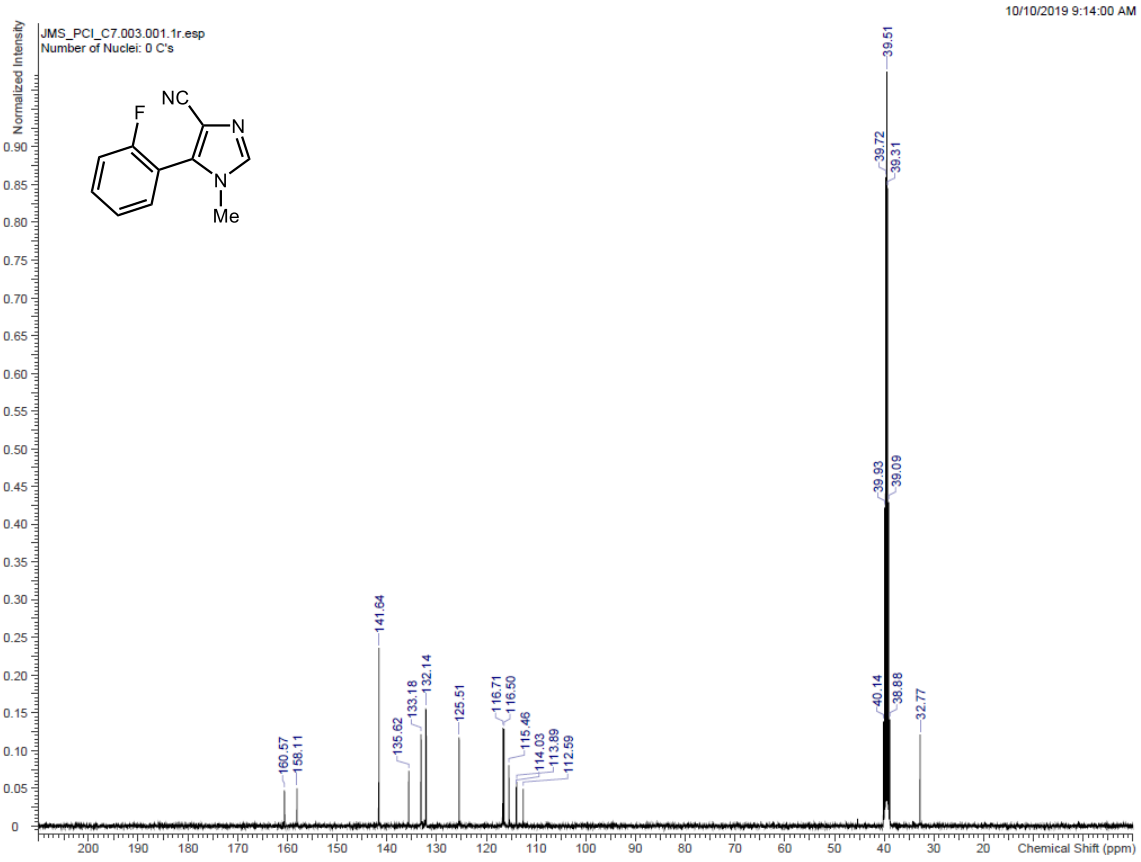
**IR (ATR, cm$^{-1}$):** 3115, 2225, 1576, 1503, 1475, 1449, 1421, 1377, 1359, 1307, 1277, 1254, 1215, 1189, 1152, 1107, 1051, 1029, 1004, 949, 873, 818, 763, 727.

**HRMS (ESI):** Calculated for $[C_{11}H_8FN_3+H]^+$ 201.0702, Found 202.0781.

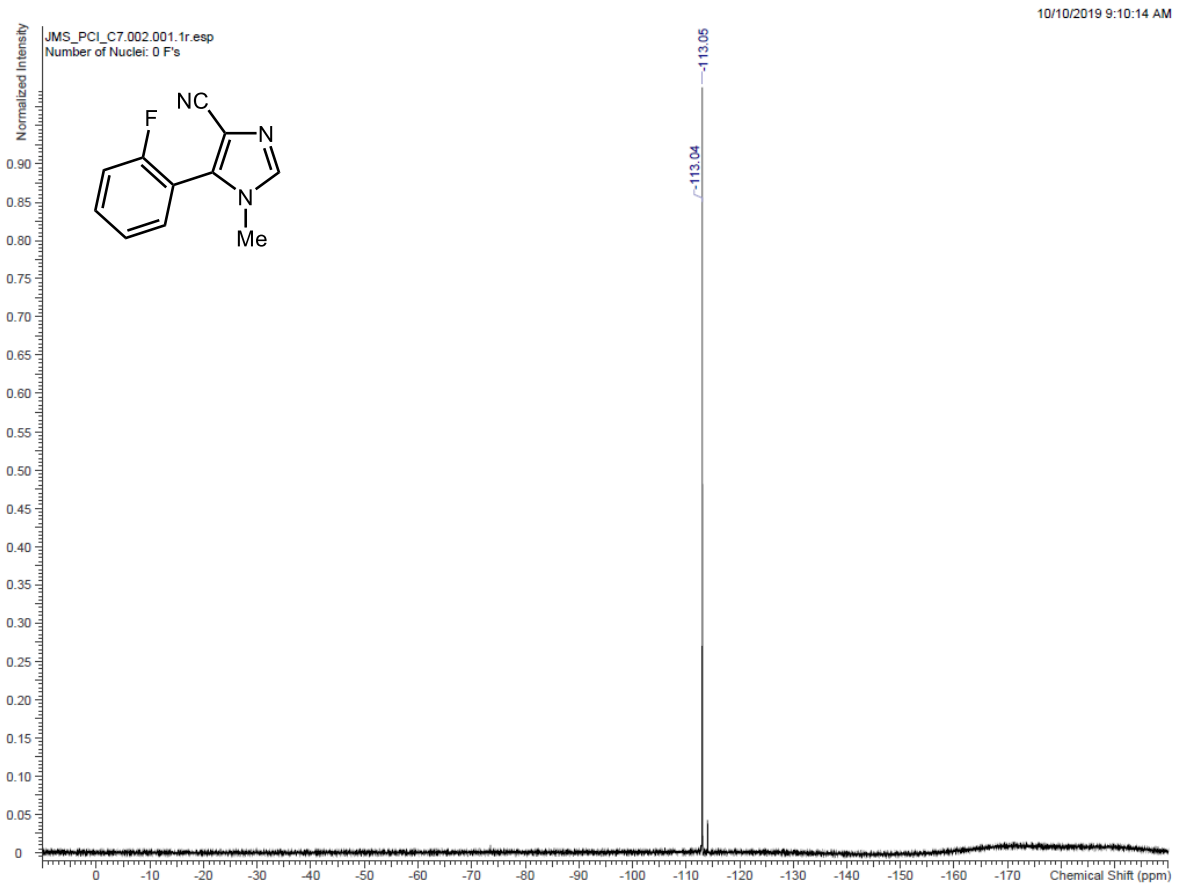
## $^{1}$H NMR (600 MHz, DMSO-$d_6$)



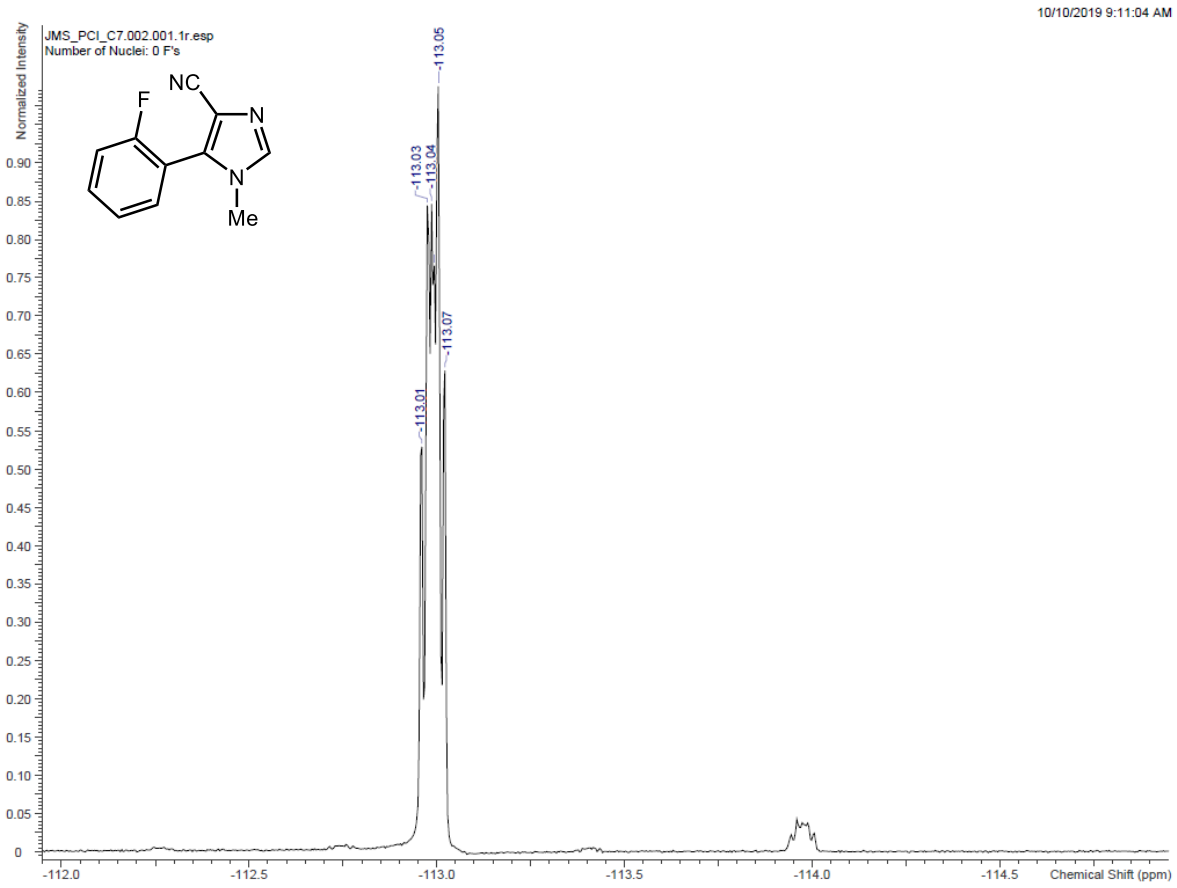
## $^{13}$C NMR (150 MHz, DMSO-$d_6$)



## $^{19}$F NMR (376 MHz, DMSO-$d_6$)



## $^{19}$F NMR (376 MHz, DMSO-$d_6$)



HTE Summary. Processed HTE data is available on GitHub under experiments. In brief, the observed yields for the experimental space were distributed between 0 and >99% with the majority of reaction outcomes resulting in little to no desired product (Figure S44).

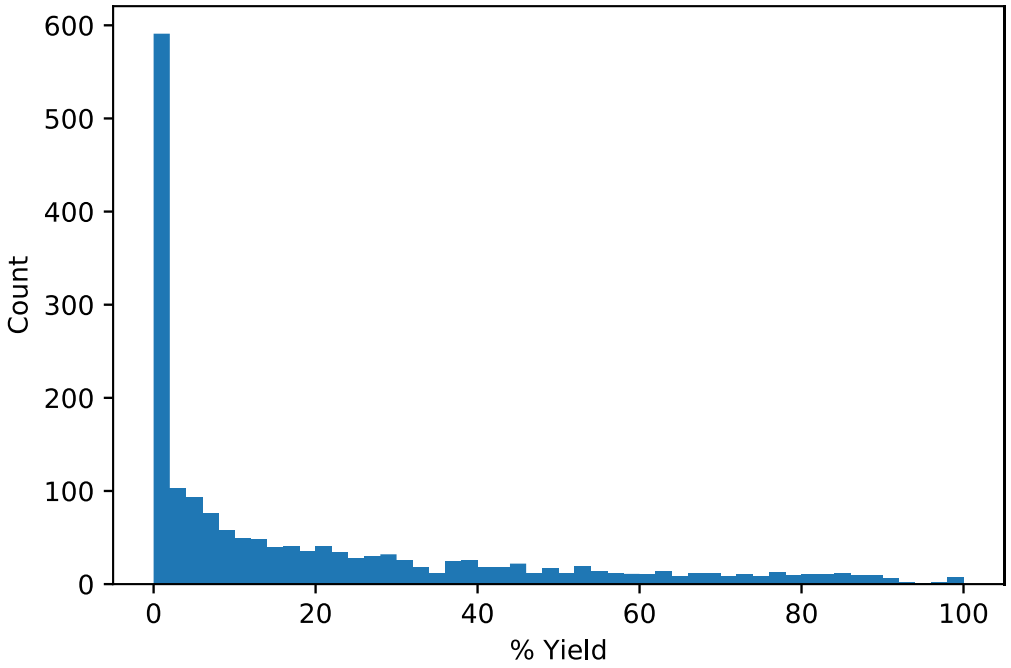



**Figure S44**. Distribution of outcomes in direct arylation reaction **3**.

Out of sample prediction. In this work we emphasized the use of structure-based reaction encodings because they enable the selection of components based on molecular similarity, should facilitate out of sample prediction, and support mechanistic inference. When selecting the ligands for this work we utilized unsupervised learning to group ligands based on computed similarity and then selected from amongst the groups to ensure a satisfactory distribution of ligands across the larger space (*vide supra*). While not the focus of this work, we were curious to see if models trained with the resulting data set would be effective in predicting ligands which were not in the training data (out of sample prediction). Thus, *after completion of BO and human testing* we collected HTE data for 2 additional ligands to be used as an out of sample test set (Figure S45). Here we present preliminary modeling results comparing GP and RF models (using parameters optimized with reaction **1** and **2a–e**) and different encodings in out of sample prediction.

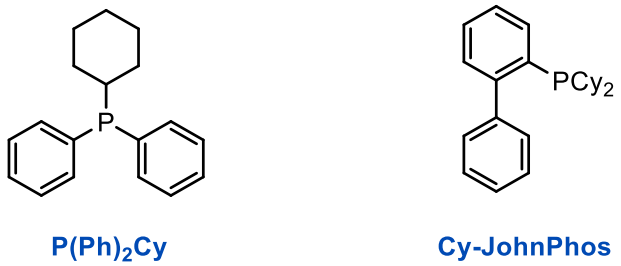



**Figure S45.** Extended ligand set for testing model out of sample prediction performance.

We selected $P(Ph)_2Cy$ and Cy-JohnPhos as ligands to include in the test set and carried out HTE experiments utilizing the majority of the combinatorial space defined by our initially selected 4 bases, 4 solvents, 3 temperatures, and 3 concentrations (Figure S42, Figure S46). Thus, the test set is only out of sample in ligand. Then, with test data in hand, we trained GP and RF models using OHE, Mordred, and DFT encoded data from the initial 1728 experiments. Next, for each trained model we independently evaluated performance in predicting reaction outcomes for $P(Ph)_2Cy$ (Figure S47) and Cy-JohnPhos (Figure S48). For $P(Ph)_2Cy$ it was observed that the RF model using Mordred encoding gave the best predictions ($R^2$ predicted versus observed = 0.88). In general, for this ligand, models with chemical encodings tended to perform better. In contrast, for Cy-JohnPhos we found that all models performed quite poorly in out of sample prediction. This poor performance was largely due to over prediction. Importantly, while the model's predictions were poor, the highest yielding conditions were still amongst the top predicted outcomes. Thus, in practice these models would still provide useful information to users.

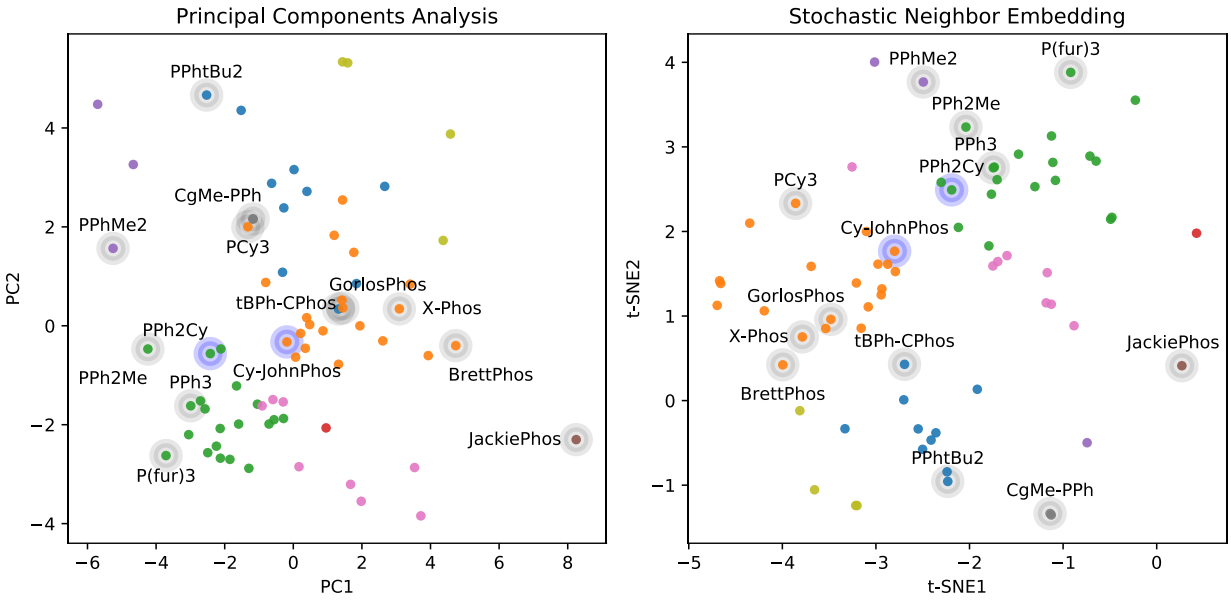



**Figure S46**. Embeddings of BMS HTE ligand library. Ligands clustered according to cheminformatics fingerprints via *k*-means with *k* selected by silhouette width. Initial ligands are highlighted in gray and out of sample ligands are highlighted in blue.

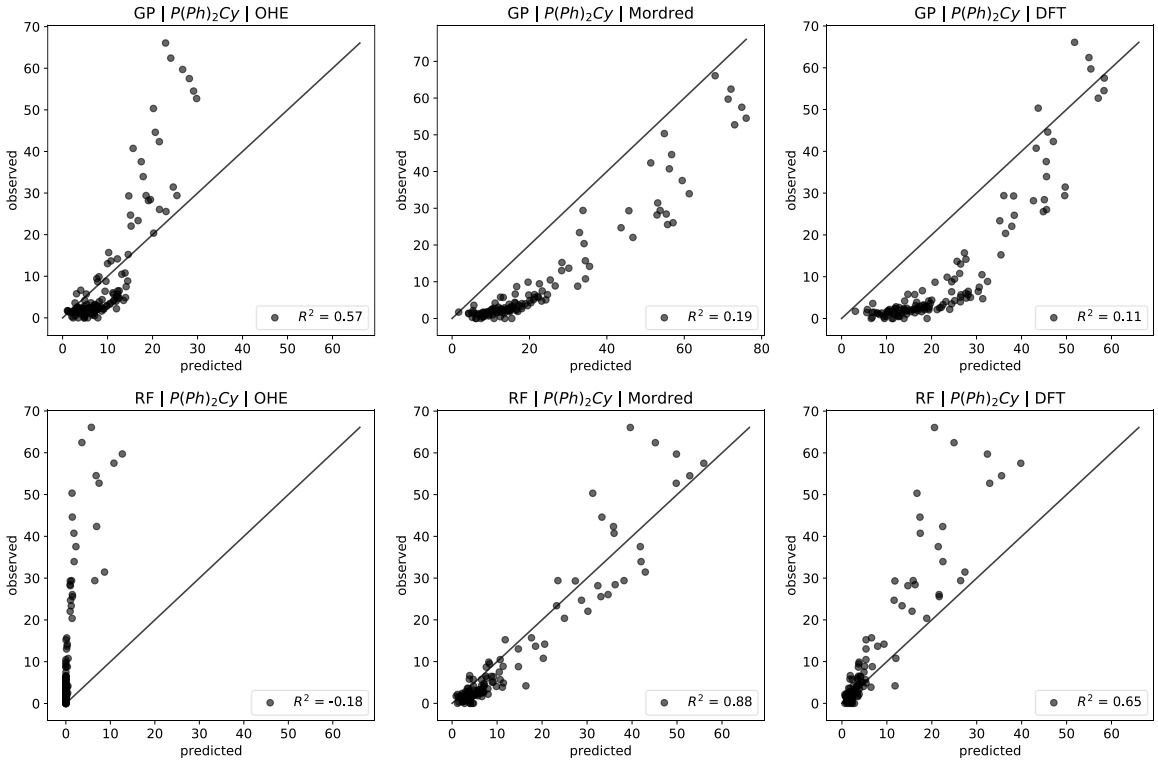



**Figure S47**. Model performance in predicting $P(Ph)_2Cy$ reaction outcomes.

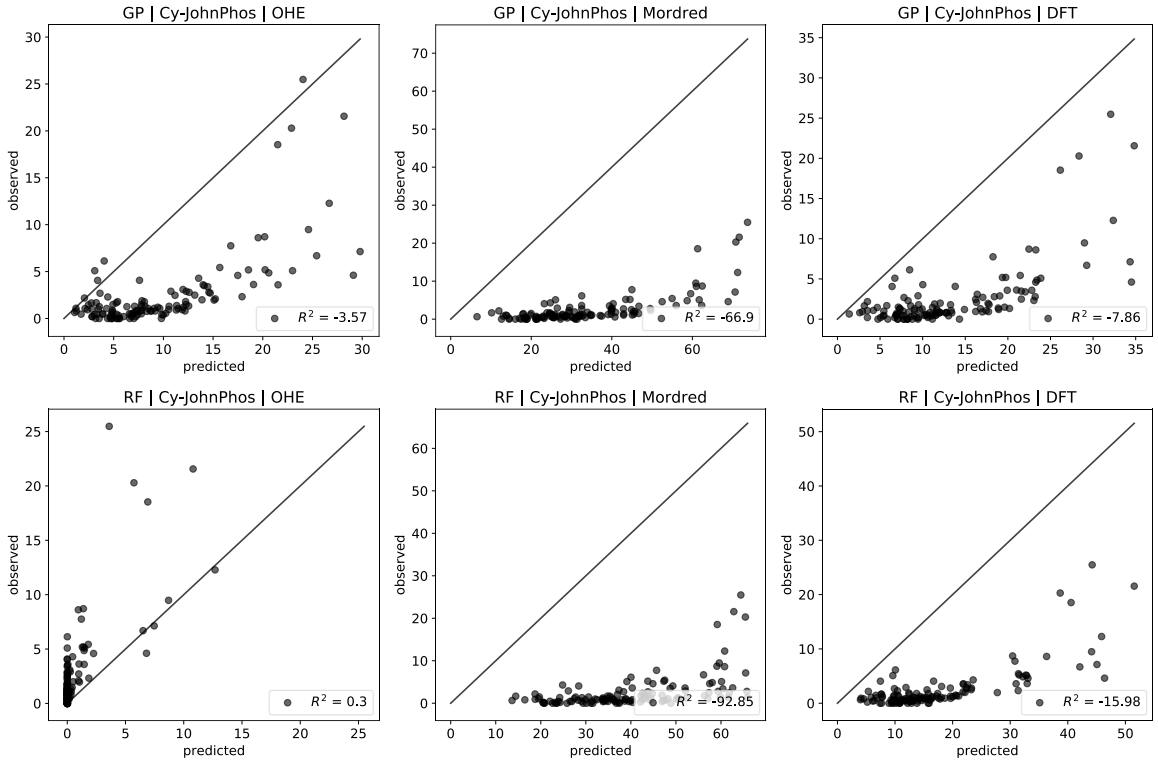



**Figure S48**. Model performance in predicting Cy-JohnPhos reaction outcomes.

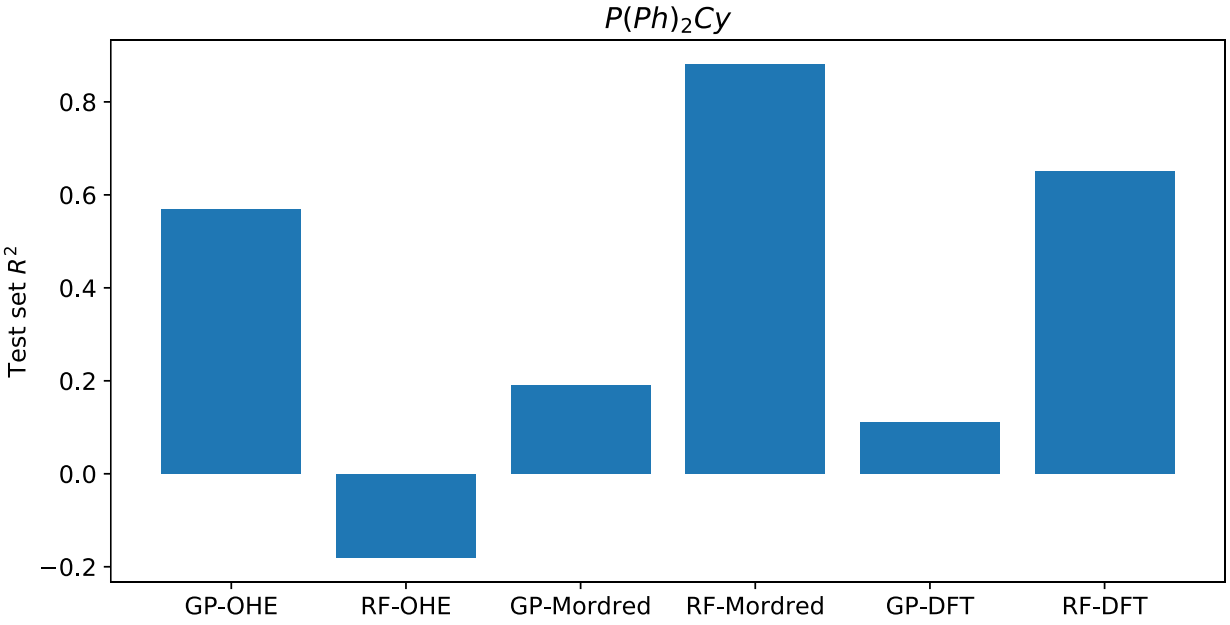



**Figure S49**. Summary of model performance in predicting $P(Ph)_2Cy$ reaction outcomes.

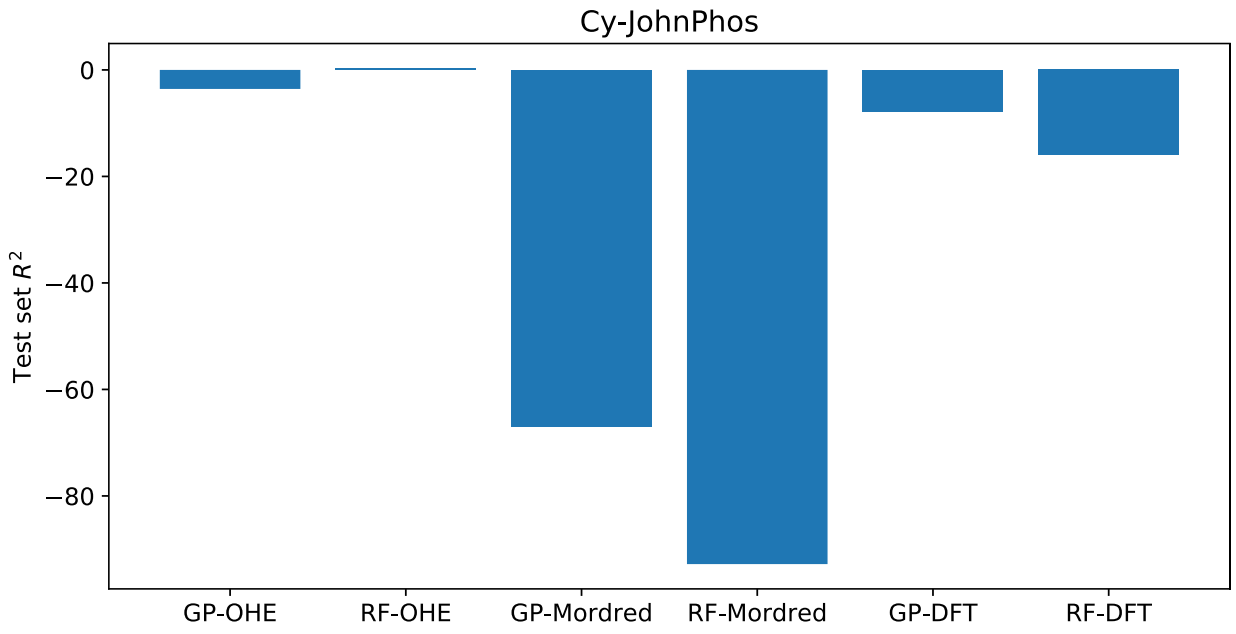



**Figure S50**. Summary of model performance in predicting Cy-JohnPhos reaction outcomes.

Bayesian optimization performance. On the basis of data for rection **1** and **2a–e** we established that BO performance was most promising using DFT reaction representations, a GP surrogate model, and expected improvement with batches selected via the Kriging believer algorithm as an acquisition function. In addition, using all 6 development reactions and multi-objective self-optimization we identified optimal model priors and hyperparameters (*vide supra*). Our experiments with reactions **1** and **2a–e** suggested that optimization was feasible with only a small fraction of the search space (typically well below 100 experiments). Importantly, in a real application, it is not typically feasible to screen parameters like temperature using a single 96 well plates for HTE. Thus, to fully explore temperature for this search space using fully populated HTE plates would require a minimum of 288 reactions. Given the importance of parallel acquisition functions and the number of candidate experiments in the reaction space, we elected to carry out tests with a batch size of 5. This gives a reasonable model for optimization at the bench. In addition, we imagined selecting 5 parallel experiments would work well in human benchmarking experiments (*vide infra*). To minimize contamination of the test set we elected to carry out a minimal number of data experiments. That is, we exclusively carried out BO simulations with randomly selected initial conditions using the encoding and optimizer configuration identified via the development data (Figure S51). In these experiments we found that the optimizer typically

converged to conditions which maximize yield of test reaction **3** withing 10 batches of 5 experiments (< 3% of experiments in the search space), irrespective of the 5 randomly selected reactions used for initialization. Importantly, in contrast to human experts which may be more biased towards ligands utilized in related transformations, the optimizer rapidly identified one of three optimal conditions which included the less conventional ligand CgMe-PPh (Figure S52).

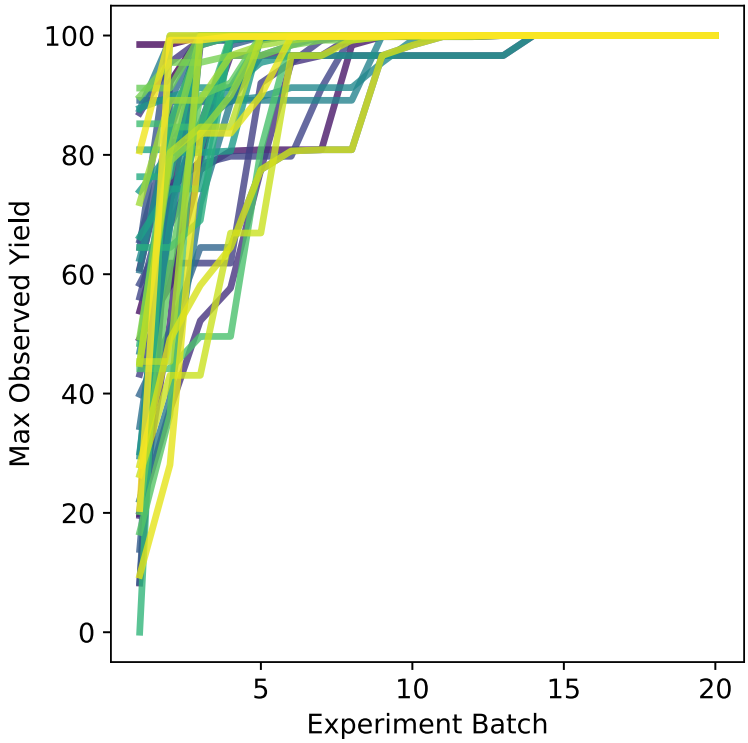
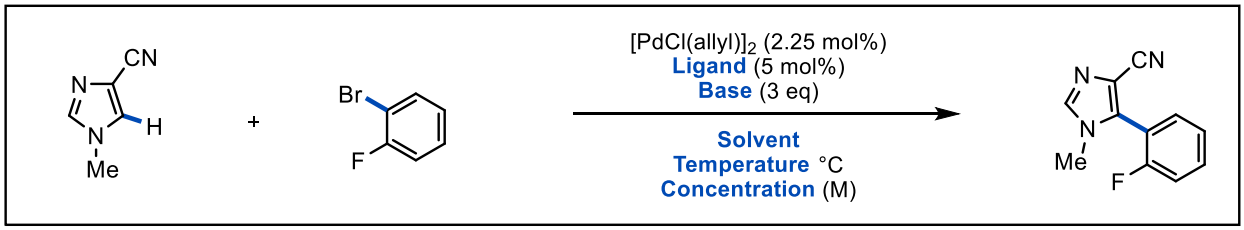



**Figure S51.** Bayesian optimization of direct arylation reaction **3**. BO performance over 50 random initializations with a GP surrogate model, DFT encoding, EI as an acquisition function, and batch size = 5.



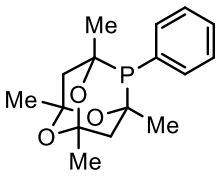
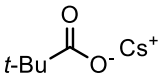
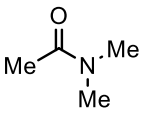
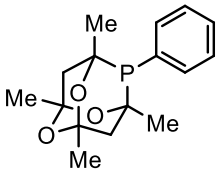
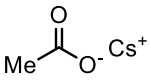
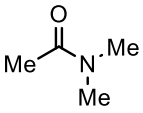
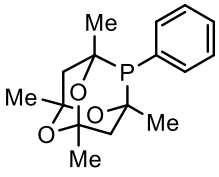
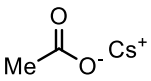
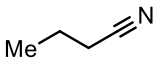

| Ligand | Base | Solvent | Temp. (°C) | Concentration (M) | Yield |
|---|---|---|---|---|---|
|  |  |  | 105 | 0.153 | >99% |
|  |  |  | 105 | 0.153 | >99% |
|  |  |  | 120 | 0.153 | >99% |

**Figure S52.** Top 3 reaction conditions for direct arylation reaction **3**. These reactions gave quantitative yield of the desired arylation product. The ligand CgMe-PPh was uniformly important for achieving optimal yield.

## VIII. Experts Versus Machine Learning

Background. The aim of the reaction optimization game was to probe the decisions made by automated machine learning systems (BO) and chemists of different backgrounds and levels of experience (e.g. process chemists, medicinal chemists, graduate students, and chemistry faculty). In the game, chemists were armed with only their personal knowledge and data they gain by running experiments. By comparison the machine learning algorithm had no knowledge prior to running experiments. While the game was intended to simulate reaction optimization on a fixed experimental budget, the data was real. Each experiment you "run" returned the result of the corresponding experiment in the lab!

Game rules. *The following rules were given to participants in the game.* Your boss has tasked you with preparing 5-(2-fluorophenyl)-1-methyl-1H-imidazole-4-carbonitrile and you have identified palladium-catalyzed direct arylation as a promising strategy. Your objective is to optimize the yield of this reaction by varying the following parameters: base, ligand, solvent, concentration, and temperature. The reaction scheme and identity of each component are summarized above. Your bench has room to run 5 experiments per "day". Each round you must select which experiments to run in the Run Experiments tab. After specifying the components and conditions for each reaction you can "run" the experiments by hitting the Run experiments button. The results will then be added to a summary table under the Experiment Results tab. Your boss is giving you "1 month" to find optimal conditions and you can run 1 batch of 5 experiments "per weekday" (graduate students you are allowed to take the weekend off). This means you will have up to 20 batches of reactions for a total of 100 experiments to find the best experimental conditions (around 6% of the experimental space). Please make sure to complete all the trials (5 experiments per submission and up to 100 experiments) in a single sessions on the web page!!! The app will finalize your results if you leave the page. You can submit your responses as quickly as you want. If you believe that you have optimized the reaction, you may stop the trials early. Play fairly. You can use only your personal knowledge about the reaction and the information you gain from each of the screens to make your decisions. No use of online resources, books, etc. No crowd sourcing. No sharing of "experimental" results between those participating. Finally, this is a statistical study

so please do not play the game more than once. If you would like to try again after your official play through, use the same username and append rerun so that it can be excluded from the analysis.

Note: while participants were asked to not share results no attempt was made to ensure that they were not communicating.

Web application. We implemented the reaction optimization game as a web application so that we could share a link with participants to collect data on human decision making in the optimization of reaction **3**. The application was developed in R/shiny and hosted on an AWS EC2 instances. The basic components of the game web application were: (1) an information page which gave a preface, background, and the rules for the game (Figure S53). (2) A user input tab which would enable players to select and submit the experiments they wanted to "run" (Figure S54). In this section we also collected basic demographic data about the player. (3) An experiment results tab which provides a sortable index of results from experiments selected by the user (Figure S55). (4) A leader board which would be populated after a player completed the game (Figure S56). Notice that we limited players to submitting 5 experiments at a time in order to make the results directly comparable to BO simulations. For source code, instructions on setting up the shiny server on AWS, and a results summary see the GitHub site.[4]

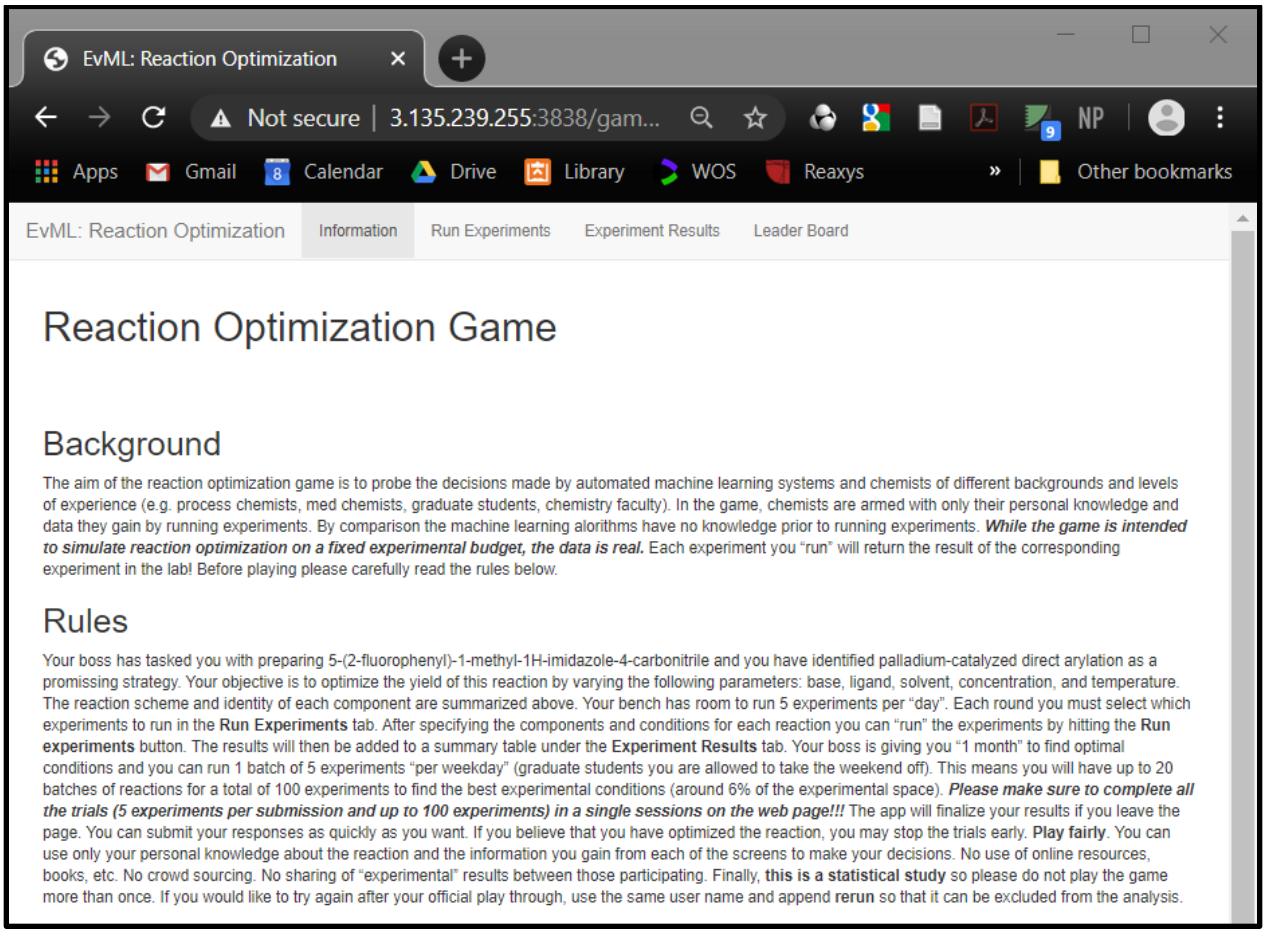



**Figure S53**. Web application for the reaction optimization game. Screen shot of part of the Information tab which gives background information and rules for the game.

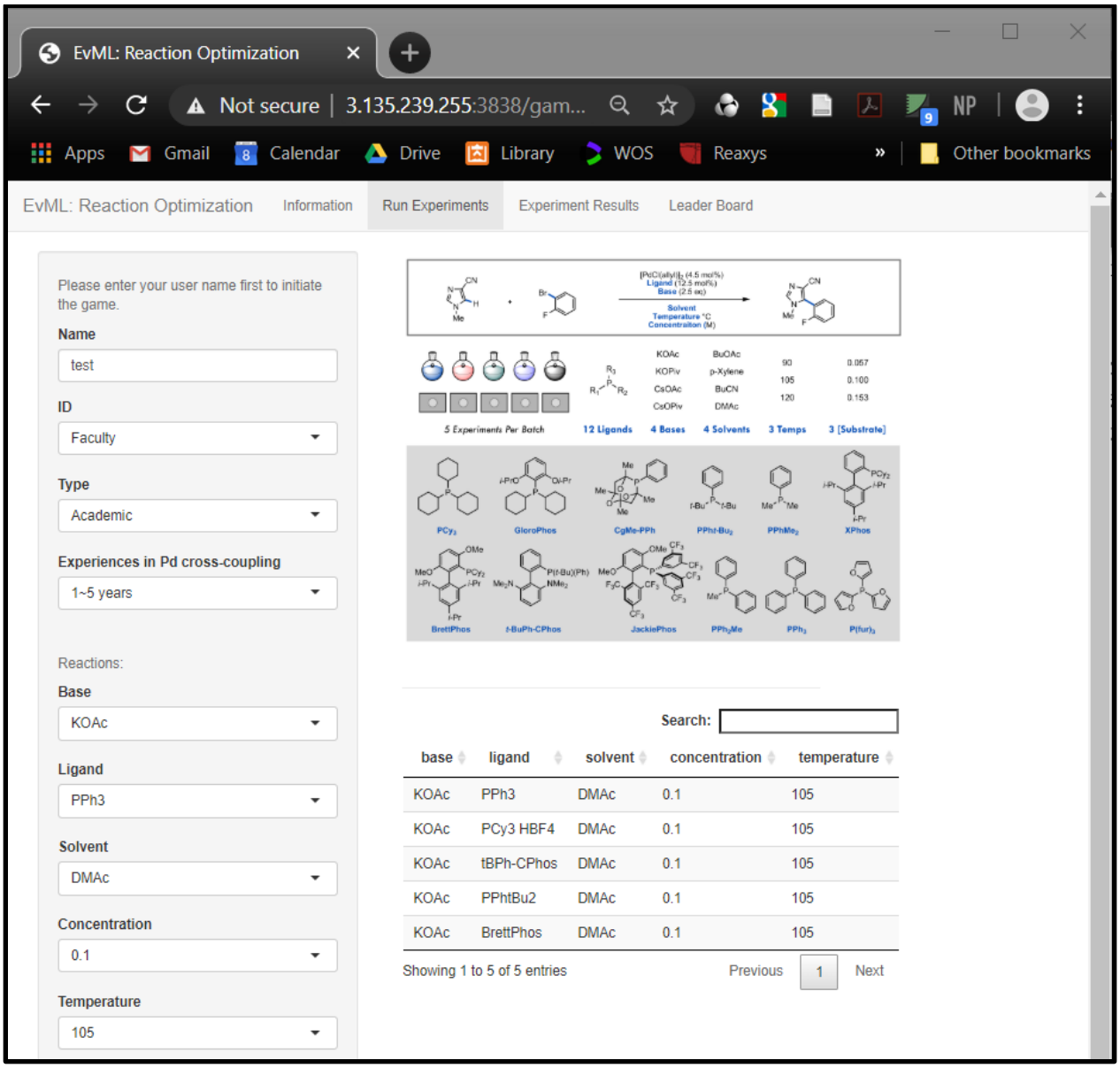



**Figure S54**. Web application for the reaction optimization game. Screen shot of Run Experiments tab which allows players to make choices about which experiments to run next. This section also enabled the collection of basic demographic data about the player.

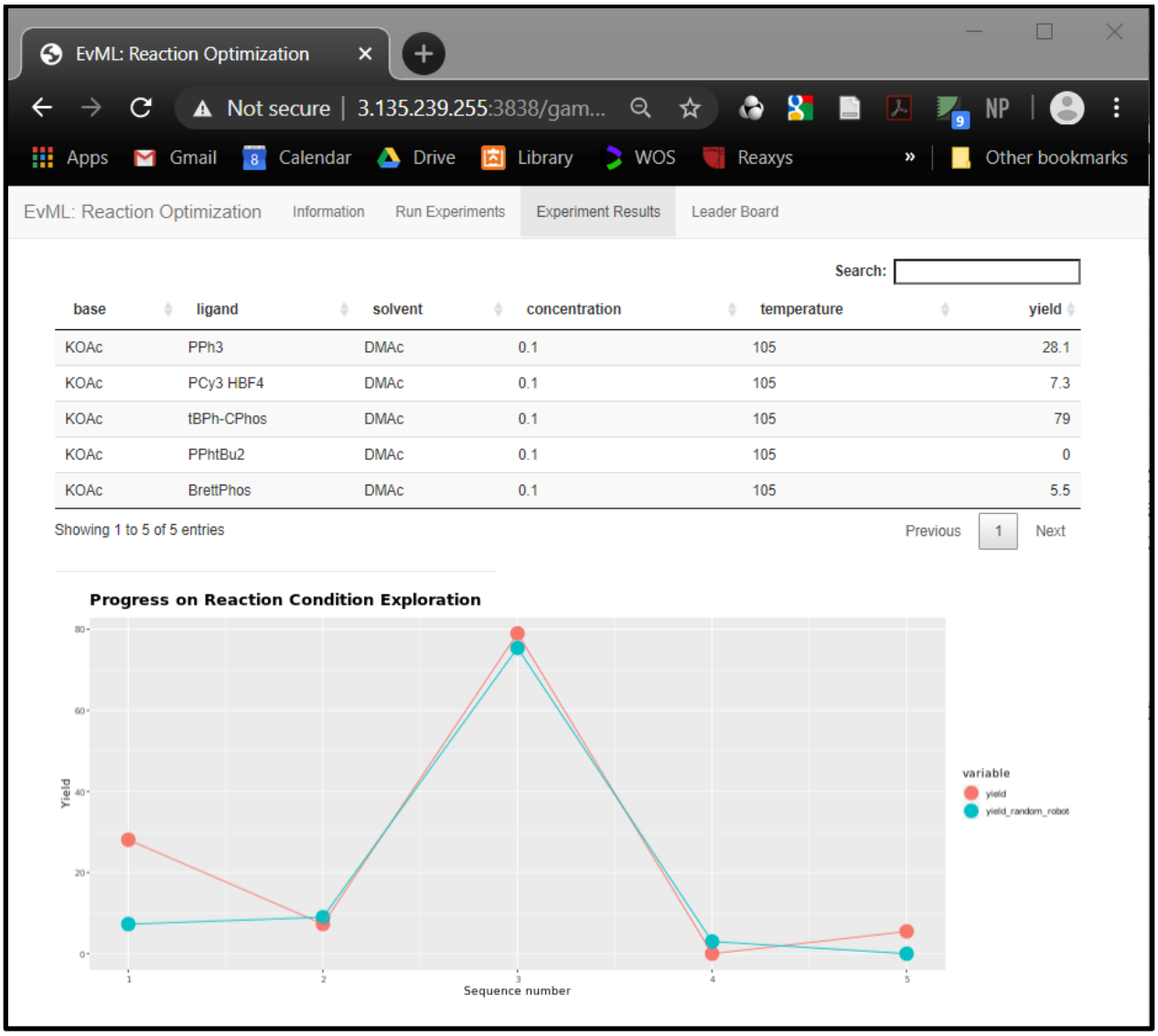



**Figure S55**. Web application for the reaction optimization game. Screen shot of Experimental Results tab which returns results for selected experiments to players.

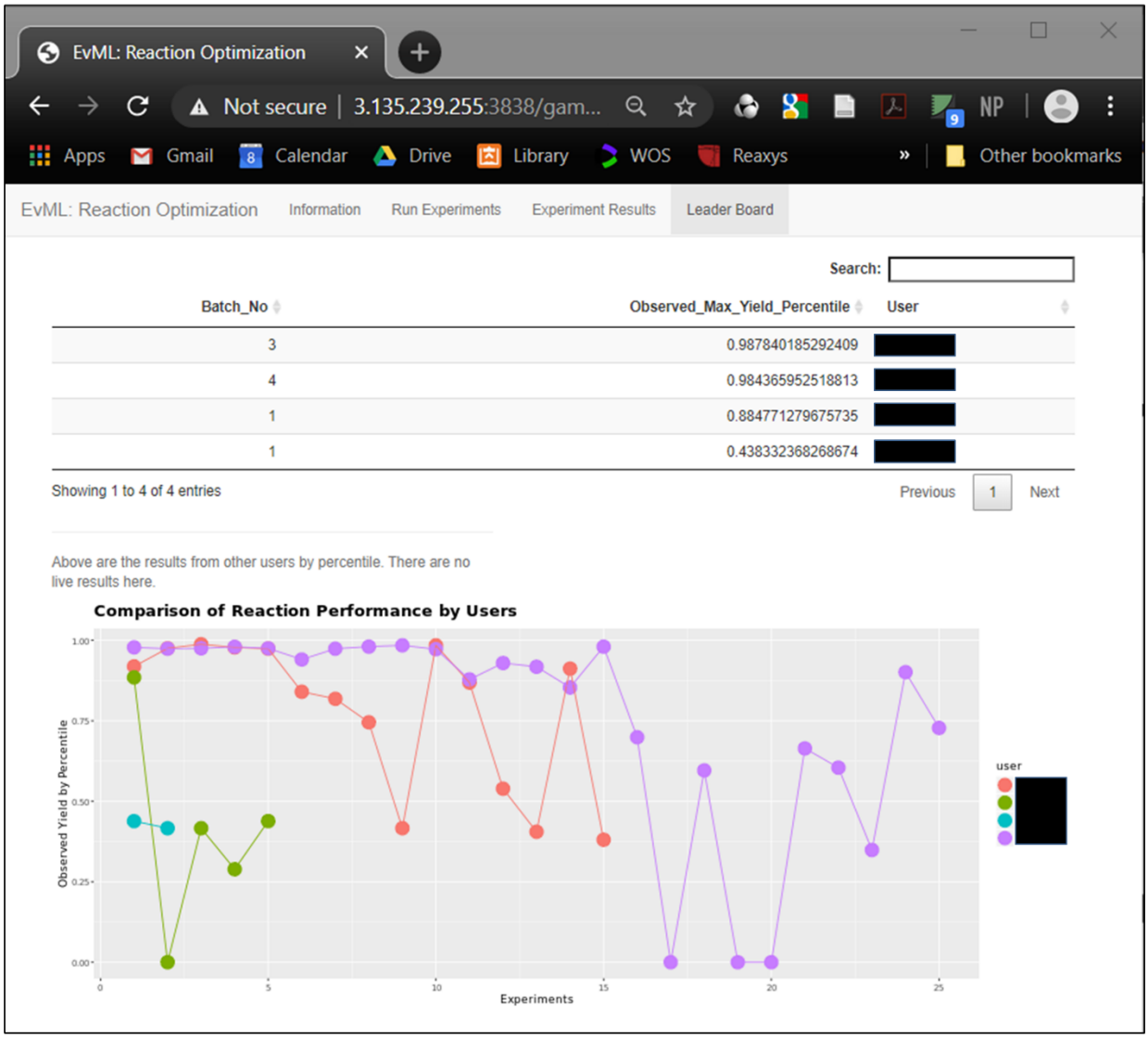



**Figure S56**. Web application for the reaction optimization game. Screen shot of leader Board tab which gives a summary of player results as a percentile. This component was designed to give the game an element of competition.

Participant data. In total, 50 participants in academia and industry played the game. Industrial players included engineers, process chemists, and medicinal chemists from Bristol-Myers Squibb. Academic players included graduate students, postdoctoral researchers, and faculty from chemistry departments at Princeton University, University of Utah, Colorado State University, and University of California, Berkeley. A summary table containing data and tabulated decisions for each player can be found on the *EvML* GitHub site.

Summary of results. The convergence behavior of all human participants can be found in figure S57. We found that 28/50 and 32/50 human participants achieved a yield that was within 1% and 5% of the maximum yield, respectively. In contrast, 100% of randomly initialized BO runs achieved the optimal yield. This is in part due to some human participants not utilizing all 20 of their possible experiment batches (we deal with this issue below).

Bounding human performance. The objective of this experiment was to statistically compare the optimization performance of BO and human experts. Each player had up to 20 batches of reactions for a total of 100 experiments to find the best experimental conditions. However, most participants played fewer than 20 rounds of experiments, for example because they believed they had achieved a globally optimal solution. Thus, in addition to comparing the raw optimization paths we sought to put hard bounds on average human performance in the data set. If we assume that players who stopped early would not have achieved any higher yielding conditions had they continued playing, we get a hard lower bound on human performance (Figure S58 and Figure S59). Conversely, if we assume optimistically that had players continued, they would have achieved 100% yield in their next batch of experiments, we get a hard upper bound. The average lower and upper bound, compared to BO performance, is displayed in figure S60. Each of these bounds is unrealistic and the actual average performance of course would fall somewhere in between. However, notice that the raw averaged data (where we only average the available data for each batch of experiments and the number of points in batch $i$ is not necessarily the same as batch $i + 1$) closely follows the lower bound up until batch 11. In addition, the upper bound closely follows the average path of the optimizer.

Hypothesis testing. Comparing the bounds on average human convergence to average BO convergence suggests that BO offers improved performance in optimization. However, we wanted to assess their performance statistically. To do this, at each step in the optimization we conducted a hypothesis test under the null hypothesis that the central tendency of human and machine performance was identical. Here, we remove the assumption of equal variance between the two population samples and utilize an unpaired Welch's t-test with Satterthwait degrees of freedom.[30] For example, figure S61 shows the distribution of max observed yields for batch 1 of human and BO decisions. A Welch's t-test with $H_0$: $\mu_{Human} = \mu_{BO}$ gives $t = 2.99$ and $p = 0.0035$. Thus, we reject the null hypothesis and conclude the initial choice distributions do not have an identical average. Further, we infer from their averages that humans tended to make better initial decisions than random selection. Now, by carrying out this analysis at each step in the optimization for the null hypotheses $\mu_{Human} = \mu_{BO}$, $\mu_{Lower} = \mu_{BO}$, and $\mu_{Upper} = \mu_{BO}$ we can statistically compare convergence behavior of each case. In figure S62 we plot the p-values for each case. A p-value < 0.05 indicates that we can reject the null hypothesis. That is, the performance of humans and machines are statistically different. For the raw data and lower bound, we found that on average after the 5th batch of experiments the optimizers performance is better than that of the humans. In contrast, for the upper bound we found that there is not a statistically significant difference between the two average curves. Thus, we conclude that in the optimization of reaction 3, Bayesian reaction optimization on average outperforms human experts, tracing the unrealistic absolute upper bound of the games recorded data. This observation, combined with the reliable convergence of the optimizer in the face of random initial reaction choices, suggests that Bayesian reaction optimization has the potential to enable chemists to make better, more consistent, data-driven decisions about which experiments to run in the course of routine reaction optimizations.

Initializing BO with human decisions. In the example in figure S61 we inferred that humans tended to make better initial decisions than random selection. Thus, in practice one would expect that initializing BO using human choices could provide some improvement in performance. Accordingly, after completing our initial studies, we carried out a second set of simulations in which the optimizer, with an otherwise identical configuration, was seeded with the initial human decisions (Figure S63). Figure S64 shows a comparison between the two initialization strategies. Interestingly, while initialization with human decisions did provide an initial average

improvement, the intermediate and long-term convergence behavior remained largely unchanged. In fact, based on statistical analysis (*vide supra*) the only steps in which their average maximum observed yield deviated ($p$-value < 0.05) were in batches 1, 8, and 9 (Figure S65).

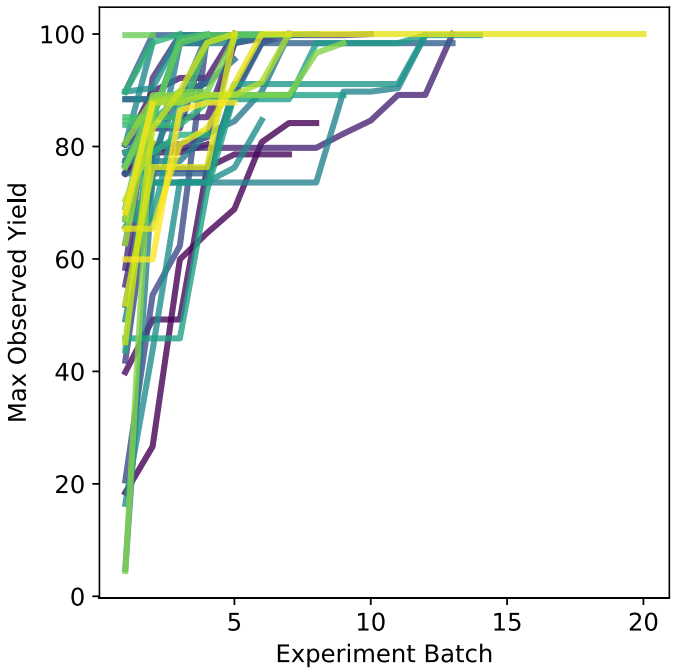



**Figure S57.** Convergence for each player. Plot showing convergence toward optimal reaction conditions for each human participant.

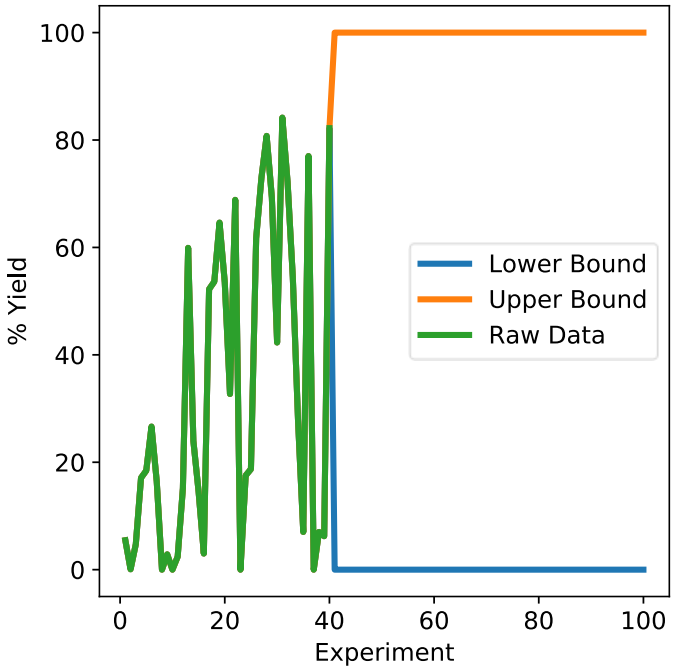



**Figure S58.** Bounding human performance. Example of individual humans' optimization path with bounds. This corresponds to the same player as figure S62.

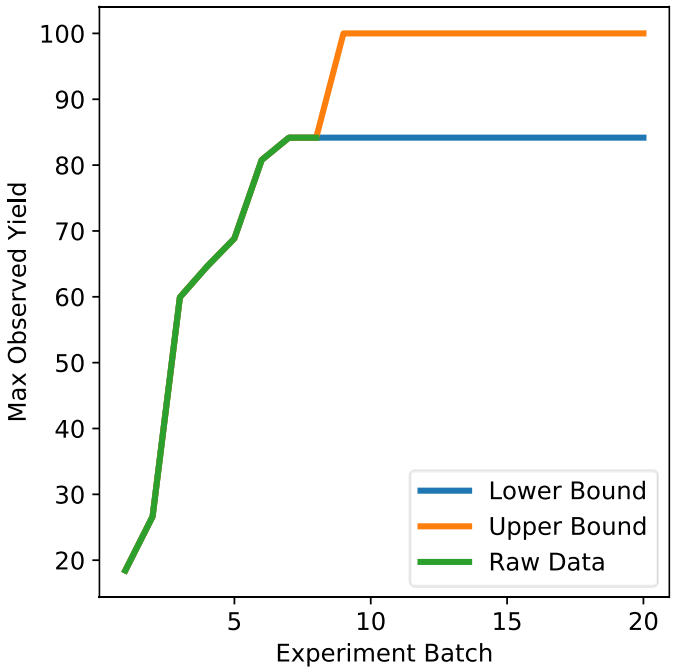



**Figure S59.** Bounding human performance. Example of individual humans' convergence with bounds. This corresponds to the same player as figure S61.

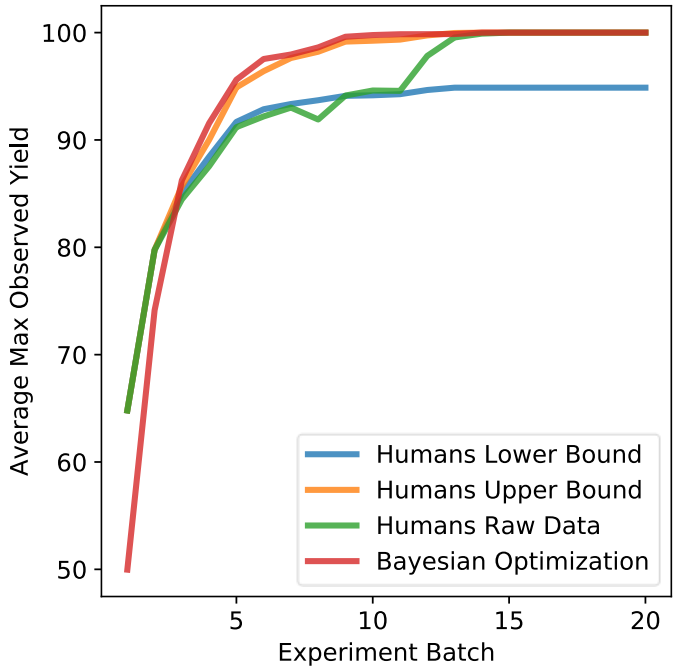



**Figure S60.** Human versus BO performance. Summary of average human and average BO performance.

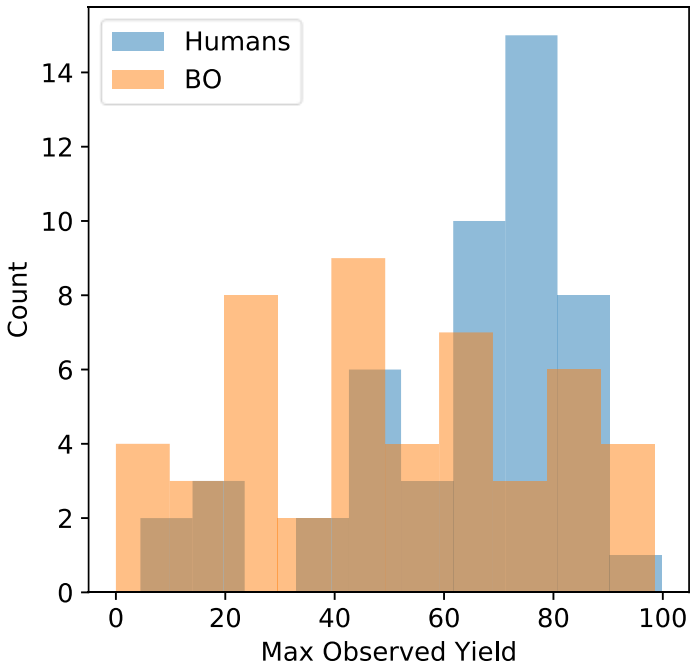



**Figure S61.** Hypothesis testing. Distribution of max observed yields for batch 1 of human and BO decisions. A Welch's t-test with $H_0$: $\mu_{Human} = \mu_{BO}$ gives $t = 2.99$ and $p = 0.0035$.

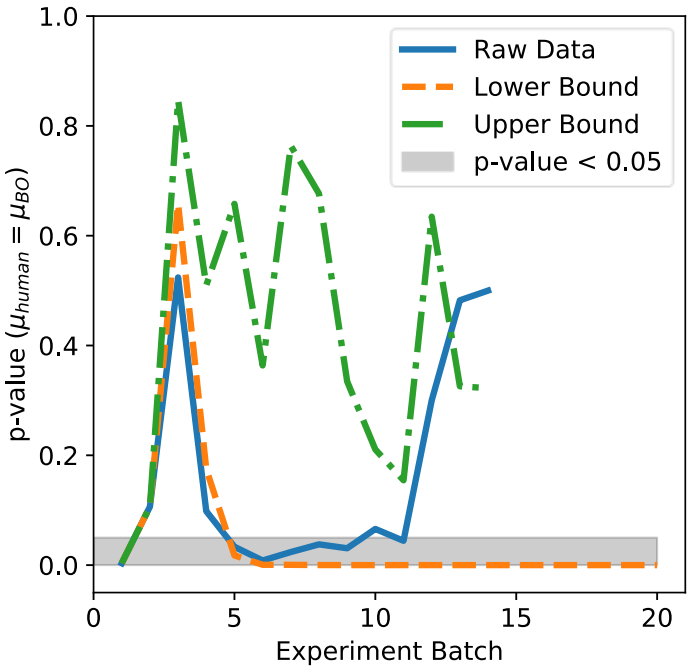



**Figure S62.** Hypothesis testing summary. Central tendency equivalence of max observed yields for human and BO decisions were tested using Welch's t-test with $H_0$: $\mu_{Human} = \mu_{BO}$.

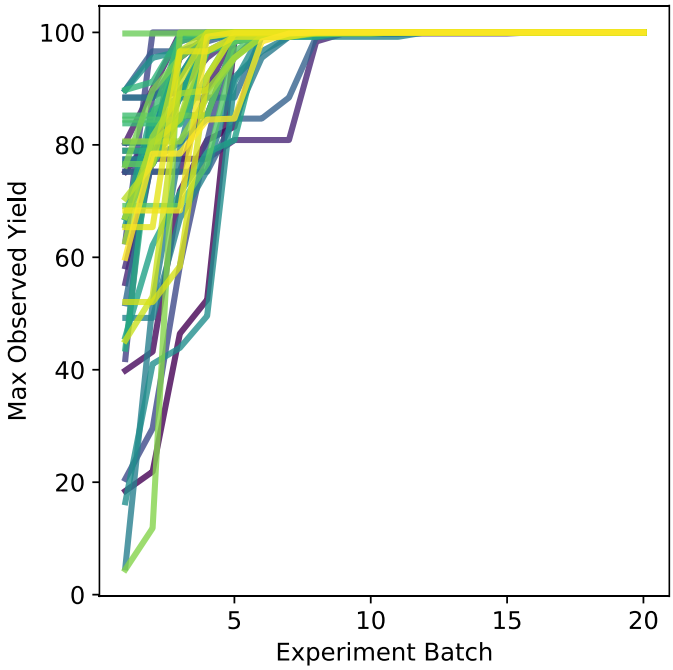



**Figure S63.** Bayesian optimization of direct arylation reaction **3**. BO performance when initialized with human choices with a GP surrogate model, EI as an acquisition function, and batch size = 5.

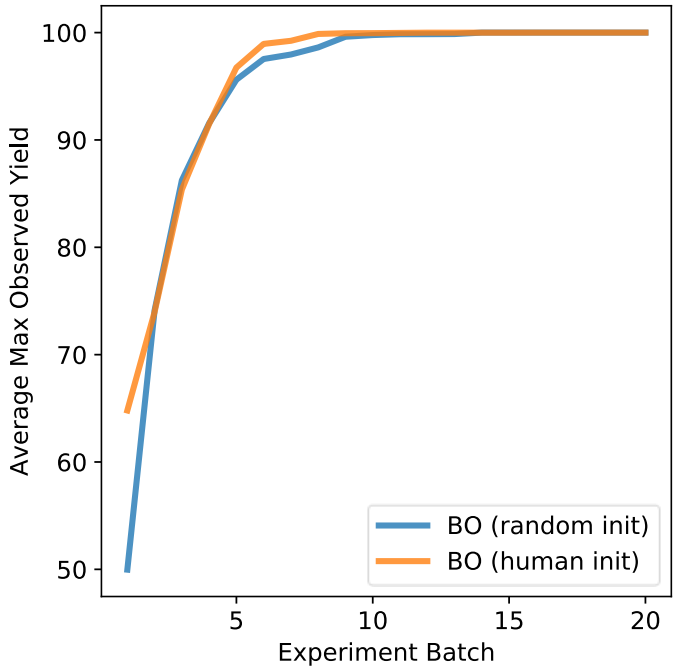



**Figure S64.** Initializing BO with human decisions. Comparison of average performance for BO initialized with random choices and human decisions.

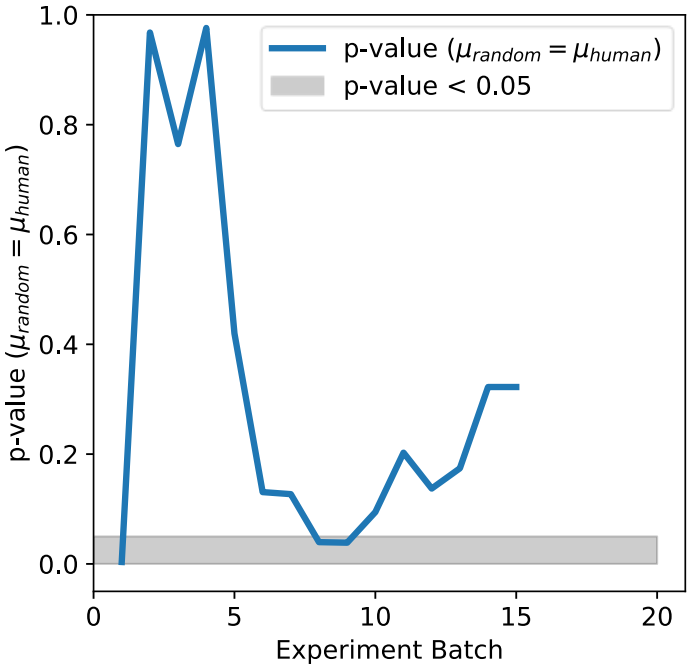
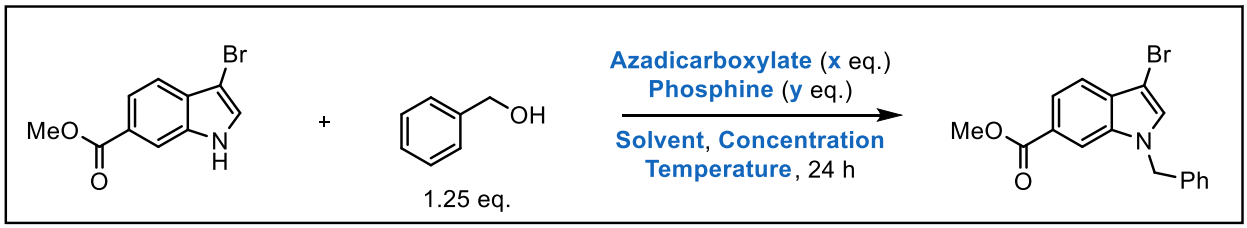


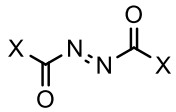

**Figure S65.** Hypothesis testing summary BO with and without human initialization. Central tendency equivalence was tested using Welch's t-test with $H_0$: $\mu_{Human} = \mu_{Random}$.

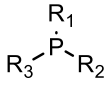
# IX. Mitsunobu Reaction Optimization

## Optimization Details

The reaction space, defined by the combinatorial set of experiment given by the selected components (Figure S66), was selected based on experience at BMS. For each of the selected chemical components we carried out conformational analysis (where appropriate), quantum mechanical modeling, and generated DFT encoding using *auto-QChem*. The DFT output files and resulting DFT encodings can be found on GitHub in the examples folder. Then the reaction space was encoded using DFT computed descriptors for categorical components (azadicarboxylate, phosphine, and solvent) and a grid of numerical values for continuous parameters (concentration, azadicarboxylate eq., phosphine eq., and temperature). Reaction optimization was the carried out by iteratively selecting experiments using EDBO. A copy of the Jupyter notebook used for optimization and collected experimental data can be found on GitHub.




| 6 Azadicarboxylates | 12 Phosphines | 5 Solvents | Conc. (M) | x | y | T (C) |
|---|---|---|---|---|---|---|
| | | PhMe | | 1.1 | 1.1 | 5 |
| | | THF | 0.05 | 1.3 | 1.3 | 15 |
| | | EtOAc | 0.10 | 1.5 | 1.5 | 25 |
| | | MeCN | 0.15 | 1.7 | 1.7 | 35 |
| | | DMF | 0.20 | 1.9 | 1.9 | 45 |

**Figure S66.** Mitsunobu reaction **4** optimization parameters and reaction space.

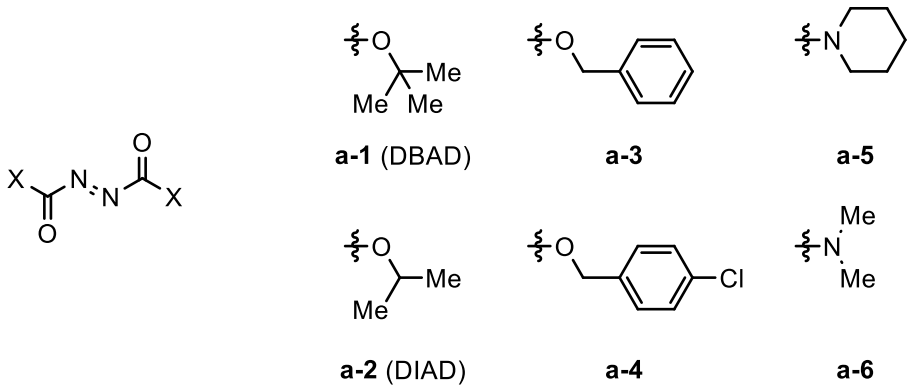



**Figure S67.** Azadicarboxylates included in Mitsunobu reaction space.

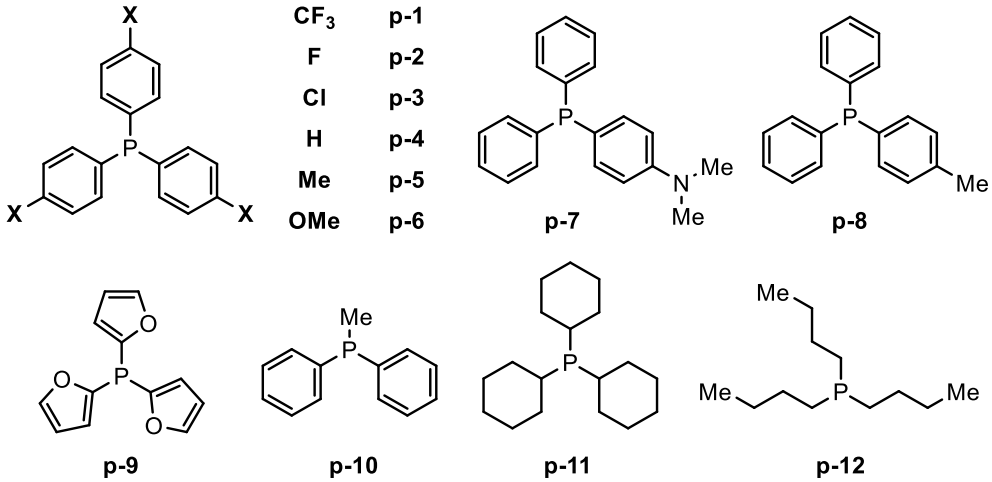



**Figure S68.** Phosphines included in Mitsunobu reaction space.

## Experiments

Materials & methods. Standard glovebox techniques were employed for handling air-sensitive reagents. NMR spectra were recorded on a 400 or 600 MHz spectrometer. UHPLCMS analysis was conducted on a Water Acquity with QDA detector and processed using Empower software. All reagents and materials were acquired from commercial suppliers. Data are presented as follows: chemical shift (ppm), multiplicity (s = singlet, d = doublet, t = triplet, m = multiplet, br = broad), coupling constant J (Hz), and integration. HRMS samples were run on the Thermo LTQ-Orbitrap with Acquity Classic inlet.

Preparation of UPLCMS Standards. To each of six vials was dispensed a 2.66 mg/mL solution of 4,4'-di-*t*-butyl-1,1'-biphenyl in THF (1.0 mL). The solvent was removed by centrifugal evaporation. To these vials was added methyl 1-benzyl-3-bromo-1H-indole-6-carboxylate (34.4 mg, 27.6 mg, 20.7 mg, 13.8 mg, 6.9 mg, and 3.5 mg) and each treated with 4.0 mL DMF to generate representative samples of 100, 80, 60, 40, 20 and 10% yields, respectively. A 50 uL sample of each of these solutions was diluted into 1.5 mL 4:1 MeCN/water for UPLCMS analysis.

Bulk Reaction Preparation. A 50 mL volumetric flask was treated with methyl 3-bromo-1H-indole-6-carboxylate (2.54 g, 10 mmol) and 4,4'-di-*t*-butyl-1,1'-biphenyl (266.4 mg, 1 mmol) then diluted to 50 mL with THF. The resulting solution was suspended in an ultrasonic bath for 5 min and 1.0 mL aliquots were distributed to 50 separate vials. The solvent was removed by centrifugal evaporation *in vacuo*.

Reaction Execution. On the benchtop each prescribed Mitsunobu reagent (DIAD, DBAD, DBzAD, DCAD, ADDP, TMAD) was dispensed into 4.0 mL vials, capped, and transferred to the glovebox. In the glovebox, vials with pre-plated methyl 3-bromo-1H-indole-6-carboxylate/4,4'-di-*t*-butyl-1,1'-biphenyl were treated with solid phosphine reagents. Both the methyl 3-bromo-1H-indole-6-carboxylate/4,4'-di-*t*-butyl-1,1'-biphenyl/phosphine vials and the Mitsunobu reagent vials were treated with the prescribed amount of solvent and equilibrated to the desired temperature. If necessary, methyl 3-bromo-1H-indole-6-carboxylate/4,4'-di-*t*-butyl-1,1'-biphenyl vials were treated with liquid phosphine ($Bu_3P$ or $PPh_2Me$). The solution of Mitsunobu reagent was transferred to the reaction vial containing methyl 3-bromo-1H-indole-6-carboxylate/4,4'-di-*t*-butyl-1,1'-biphenyl/phosphine. The reactions were sealed and transferred to the benchtop where they stirred for 22 h. The reactions were diluted to 4.0 mL total volume with DMF pre-equilibrated to the reaction temperature for each vial, stirred for 5 min at the target temperature, and a 25 uL sample of each reaction was diluted into 1.5 mL 4:1 MeCN/water for UPLCMS analysis.

UHPLCMS Method:
Solvent A: Water with 5% acetonitrile and 0.05% TFA
Solvent B: acetonitrile with 5% water and 0.05% TFA
Gradient: 95% A/B to 5% A/B over 4.0 min, hold 1.0 min at 95% B

Stop Time: 5.0 min

Flow Rate: 0.5 mL/min

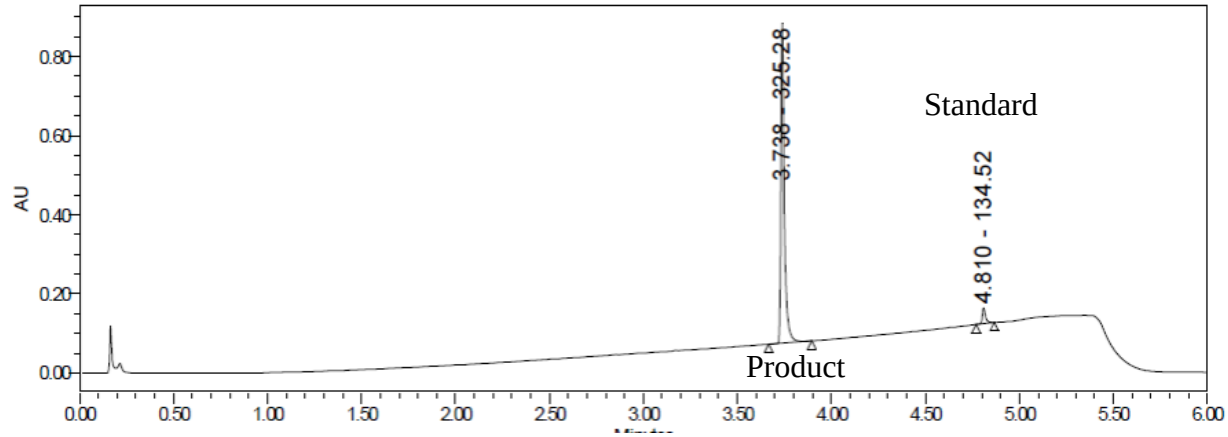
PDA Detection: Extracted 254 nm for analysis

Column: Agilent Poroshell C18 2.7 um 2.1x50mm

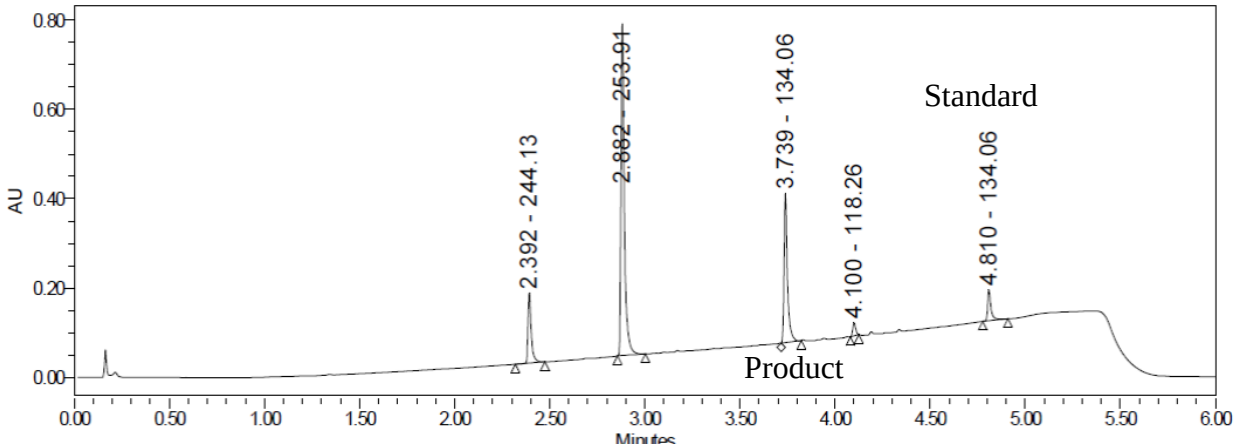
Oven Temperature: 40

Reaction Reproducibility. Two reactions were set up simultaneously according to the reaction execution protocol on 0.2 mmol scale containing (DIAD, 1.1 equiv; $PPh_3$, 1.1 equiv, THF, 0.1 M, at 25 C). The reactions were found to produce solution yields of 59% and 60% using the UPLCMS calibration vs internal standard.

**UHPLC-MS calibration sample**




**UHPLC-MS experiment sample**



**Methyl 1-benzyl-3-bromo-1H-indole-6-carboxylate.** To a 500 mL round-bottom flask in a glovebox under inert atmosphere was added methyl 3-bromo-1H-indole-6-carboxylate (1.27 g, 5 mmol), THF (50 mL), and diphenylmethylphosphine (1.14 mL, 6.25 mmol). The flask was transferred to the benchtop and treated with (E)-diazene-1,2-diylbis(piperidin-1-ylmethanone) (1.58 g, 6.25 mmol). The reaction stirred at 23 C for 3 h then quenched by the addition of water (25 mL) and diluted with ethyl acetate (250 mL). The aqueous phase was removed, and the desired organic phase was washed twice with water (25 mL). The isolated organic phase was concentrated to orange oil. Purification by silica gel chromatography (120 g ISCO Redi-sep gold; 5-20% ethyl acetate/heptanes gradient) afforded the title compound (1.36 g, 3.95 mmol) in 79% isolated yield as white solids.

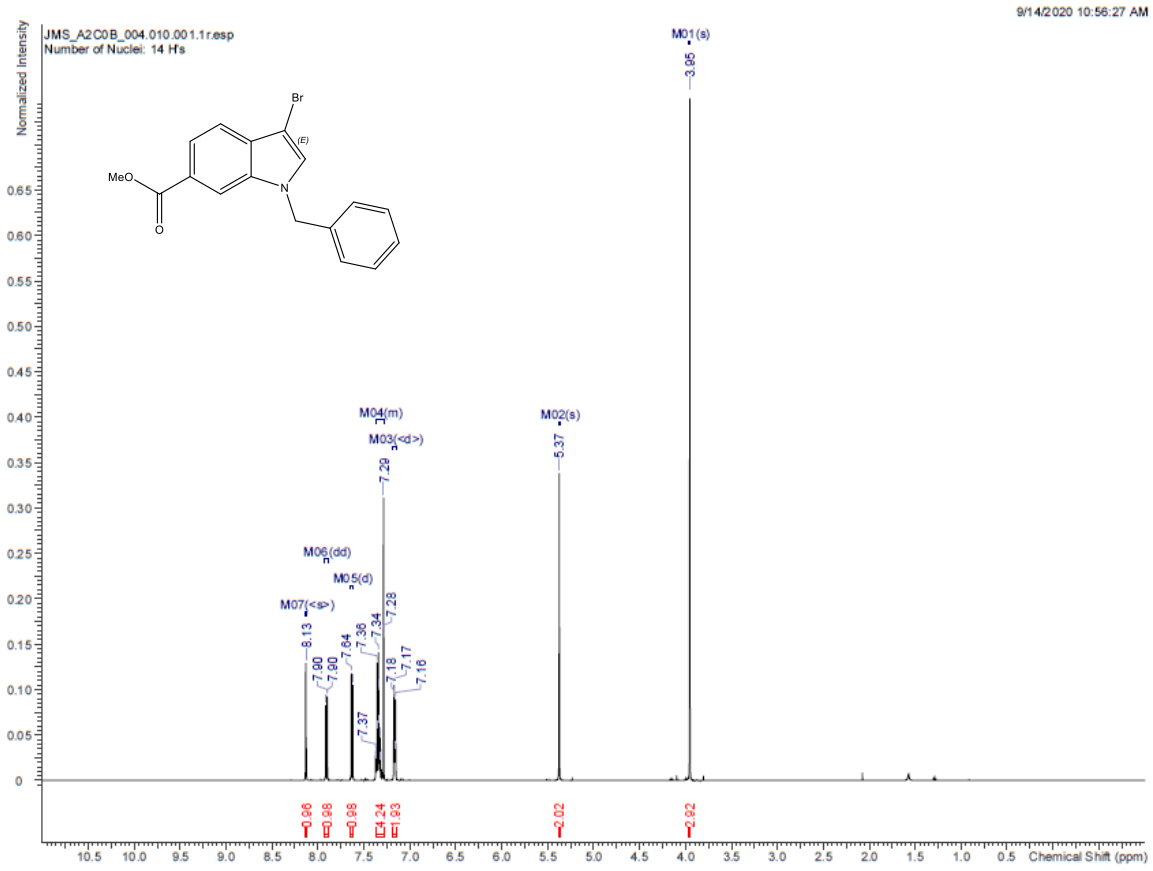
**$^{1}$H NMR (600 MHz, $CDCl_3$):** $\delta$ 8.13 (s, 1H), 7.91 (dd, *J*=8.4, 1.4 Hz, 1H), 7.63 (d, *J*=8.4 Hz, 1H), 7.37 - 7.28 (m, 4H), 7.17 (d, *J*=6.9Hz, 2H), 5.37 (s, 2H), 3.95 (s, 3H).

**$^{13}$C NMR (126 MHz, $CDCl_3$):** $\delta$ 167.6, 136.2, 135.3, 130.8, 130.1, 128.9, 128.1, 127.0, 124.6, 121.3, 119.1, 112.2, 90.6, 52.0, 50.3.

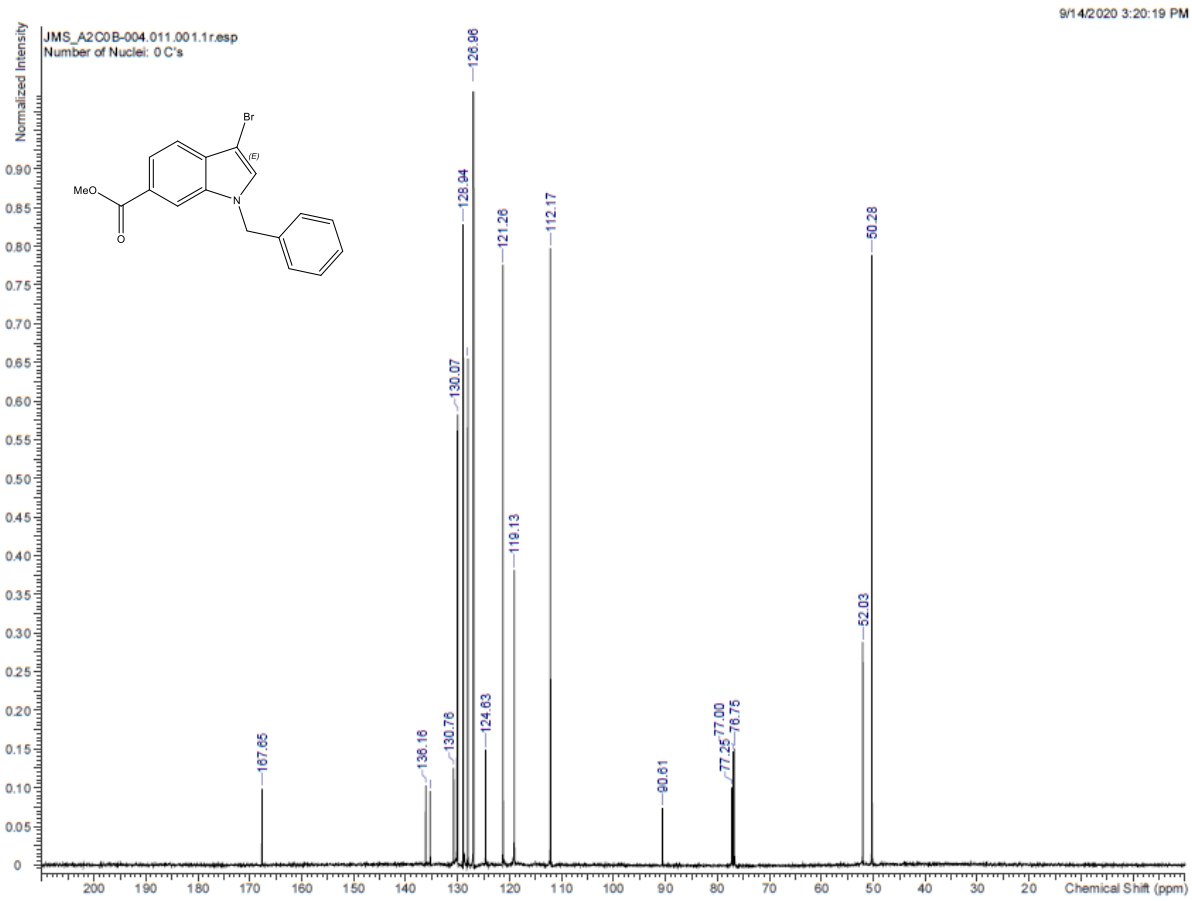
**IR (thin film)$\nu^{-1}$:** 2950.8, 1704.6, 1610.2, 1496.3, 1485.2, 1446.7, 1431.7, 1399.6, 1347.7, 1293.8, 1239.2, 1204.3, 1186.7, 1166.7, 1099.7, 1080.0, 1056.2, 1028.0, 975.3, 872.5, 851.1, 814.6, 788.7, 763.2, 738.0, 694.6.

**HRMS (ESI):** calcd for $[C_{17}H_{14}BrNO_2+H]^+$ 344.0281, found 344.0276.

## $^{1}$H NMR (600 MHz, CDCl3)



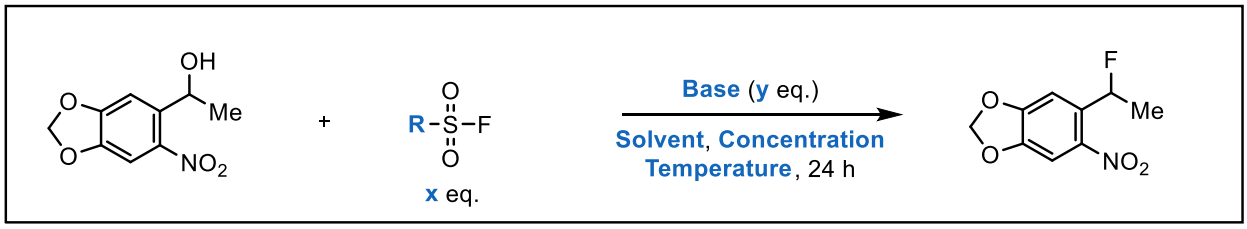
## $^{13}$C NMR (126 MHz, $CDCl_3$)



## X. Deoxyfluorination Reaction Optimization

### Optimization Details

The reaction space, defined by the combinatorial set of experiment given by the selected components (Figure S69), was selected based on experience in the Doyle lab. For each of the selected chemical components we carried out conformational analysis (where appropriate), quantum mechanical modeling, and generated DFT encoding using *auto-QChem*. The DFT output files and resulting DFT encodings can be found on GitHub in the examples folder. Then the reaction space was encoded using DFT computed descriptors for categorical components (sulfonyl fluoride, base, and solvent) and a grid of numerical values for continuous parameters (concentration, sulfunylfluoride eq., base eq., and temperature). Reaction optimization was the carried out by iteratively selecting experiments using EDBO. A copy of the experimental data and Jupyter notebook used for optimization can be found on GitHub.





| | | | | |
|---|---|---|---|---|
| PhMe | 0.1 | 1.1 | 1.1 | 20 |
| THF | 0.2 | 1.3 | 1.3 | 30 |
| $PhCF_3$ | 0.3 | 1.5 | 1.5 | 40 |
| DMSO | 0.4 | 1.7 | 1.7 | 50 |
| DMF | 0.5 | 1.9 | 1.9 | 60 |
| 10 Sulfonylfluorides | 10 Bases | 5 Solvents | Conc. (M) | x | y | T (C) |

**Figure S69.** Deoxyfluorination reaction **5** optimization parameters and reaction space.

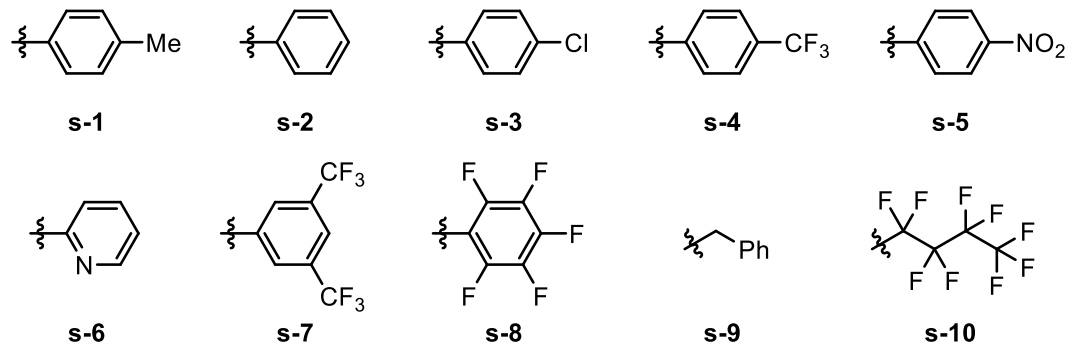



**Figure S70.** Sulfonyl fluorides in the deoxyfluorination reaction space.

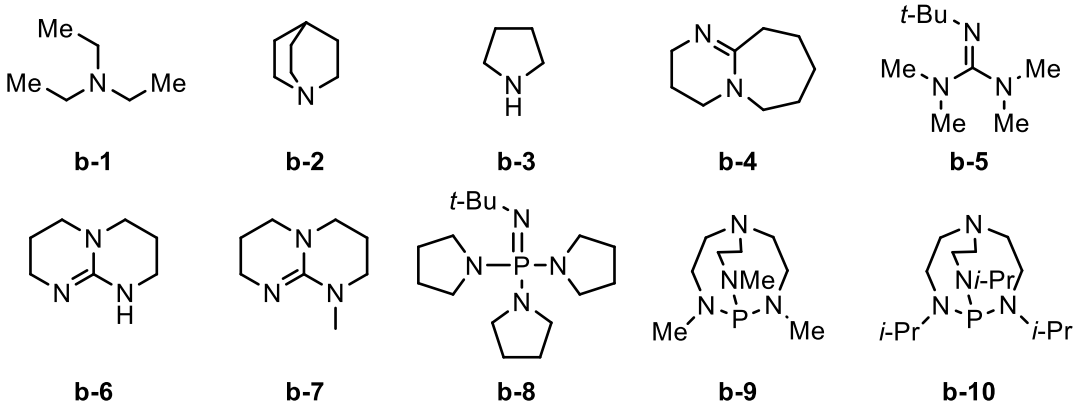



**Figure S71.** Bases in the deoxyfluorination reaction space.

## Experiments

Materials & methods. 4-Phenyl-2-butanol was obtained from TCI. 1-Fluoronaphthalene was acquired from Oakwood. 7-Methyl-1,5,7-triazabicyclo[4.4.0]dec-5-ene (MTBD) was purchased from TCI. 2-(tert-butyl)-1,1,3,3-tetramethylguanidine, N-tert-butyl-1,1,1-tri(pyrrolidin-1-yl)-l5-phosphanimine, 2,8,9-trimethyl-2,5,8,9-tetraaza-1-phosphabicyclo[3.3.3]undecane, 2,8,9-triisopropyl-2,5,8,9-tetraaza-1-phosphabicyclo[3.3.3]undecane, Quinuclidine, triethylamine, pyrrolidine, 1,8-Diazabicyclo[5.4.0]undec-7-ene, 1,3,4,6,7,8-hexahydro-2H-pyrimido[1,2-a]pyrimidine were purchased from Aldrich. Phenylmethanesulfonyl fluoride was purhased from BACHEM. 4-methylbenzenesulfonyl fluoride, 1,1,2,2,3,3,4,4,4-nonafluorobutane-1-sulfonyl fluoride, pyridine-2-sulfonyl fluoride, benzenesulfonyl fluoride were purchased from aldrich. Anhydrous aldrich solvent bottles were used for solvents except for toluene, DMSO, and THF, which were dispensed from a dry solvent systems. Organic solutions were concentrated under

reduced pressure using a rotary evaporator (23-30 °C, <50 torr). Automated column chromatography was performed us-ing silica gel cartridges (40-53 μm, 60 Å) on a Biotage SP4. Proton nuclear magnetic resonance (1H NMR) spectra were recorded on a Bruker 500 AVANCE spectrometer (500 MHz), a Bruker NB 300 spectrometer (300 MHz), or a Bruker Avance III HD NanoBay (400 MHz) spectrometer. Deuterium nuclear magnetic reso-nance (2H NMR) spectra were recorded on a Bruker 500 AVANCE spectrometer (77 MHz). Carbon nuclear magnetic resonance ($^{13}$C NMR) spectra were recorded on a Bruker 500 AVANCE spectrometer (126 MHz). Fluorine nuclear magnetic resonance ($^{19}$F NMR) spectra were recorded on a Bruker NB 300 spectrometer (282 MHz). Chemical shifts for protons are reported in parts per million downfield from tetramethylsilane and are referenced to residual protium in the NMR solvent ($CHCl_3$ = δ 7.26 ppm). Chemical shifts for carbon are reported in parts per million downfield from tetramethylsilane and are referenced to the carbon resonanc-es of the solvent residual peak ($CDCl_3$ = δ 77.16 ppm). NMR data are represented as follows: chemical shift (δ ppm), multiplicity (s = singlet, d = doublet, t = triplet, q = quartet, p = pentet, m = multiplet), coupling constant in Hertz (Hz), integration. Reversed phase liquid chromatog-raphy/mass spectrometry (LC/MS) was performed on an Agilent 1260 Infinity analytical LC and Agilent 6120 Quadrupole LC/MS system, using electrospray ionization/atmospheric-pressure chemical ionization (ESI/APCI), and UV detection at 254 and 280 nm. High-resolution mass spectra were obtained on an Agilent 6220 LC/MS using electrospray ioniza-tion time-of-flight (ESI-TOF) or Agilent 7200 gas chromatography/mass spectrometry using electron impact time-of-flight (EI-TOF). Gas chromatography was performed on an Agilent 7890A series instrument equipped with a split-mode capillary injection system and flame ioni-zation detectors. Fourier transform infrared (FT-IR) spectra were recorded on a Perkin-Elmer Spectrum 100 and are reported in terms of frequency of absorption ($cm^{-1}$).

Reaction Execution. An oven-dried 0.5 dram vial (VWR® glass vials, 66011-020) equipped with a PTFE-coated stir bar (VWR® Micro stir bars, 2 x 7 mm, 58948-976) was charged with benzylic alcohol (0.1 mmol, 21.12 mg) followed by sulfonyl fluoride (125 μL), additional solvent, and base (125 μL). The vials were capped and stirred at 600 rpm at the temperature specified by BO for 24 hours. All reactions were carried out on the bench without exclusion of oxygen. Stock solutions were made such that two aliquots could be dispenesed if needed. For reactions at 0.5 M

concentration, no stock solutions were made and the reagents were individually weighed out on the bench. Whenever stock solutions of base or sulfonyl fluoride were insoluble after rigours shaking or sonication, the reagents were added directly by weighing out instead. Solubility was particulary problematic for 1,5,7-Triazabicyclo[4.4.0]dec-5-ene (TBD) in non-polar solvents like toluene. After 24 of stirring, reactions were stopped and yield measured via $^{19}F$ NMR versus 1-fluoronapthalene as an external standard.

Reaction Reproducibility. Duplicates of two control reactions were set up simultaneously according to the reaction execution protocol on 0.1 mmol scale. The results are presented in figure S72.

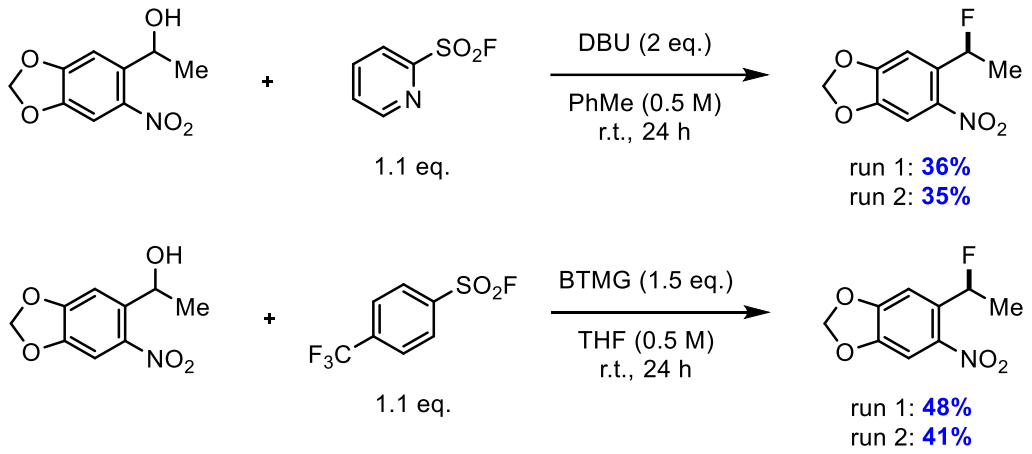



**Figure S72.** Control reactions and reproducability for deoxyfluorination reaction **5**.

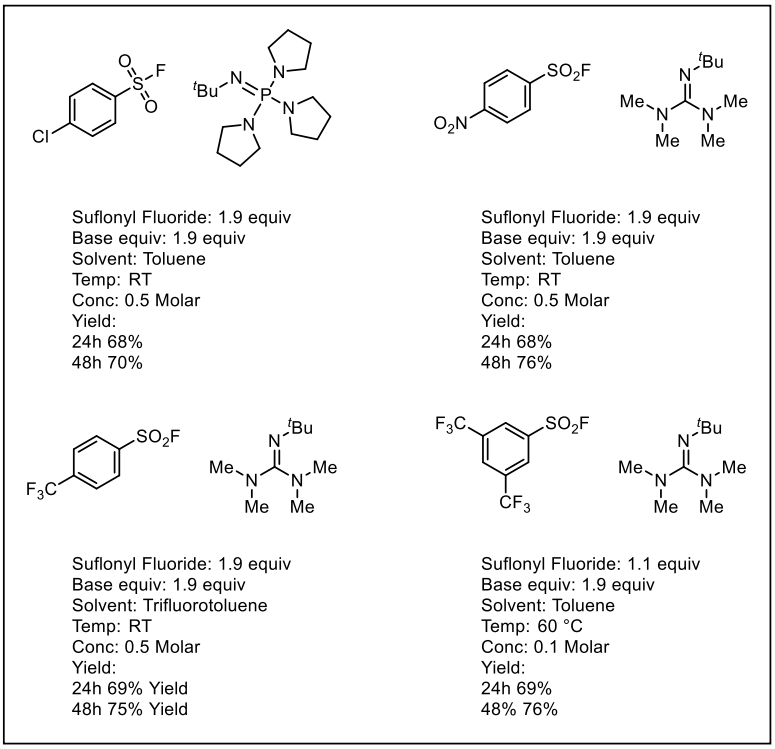
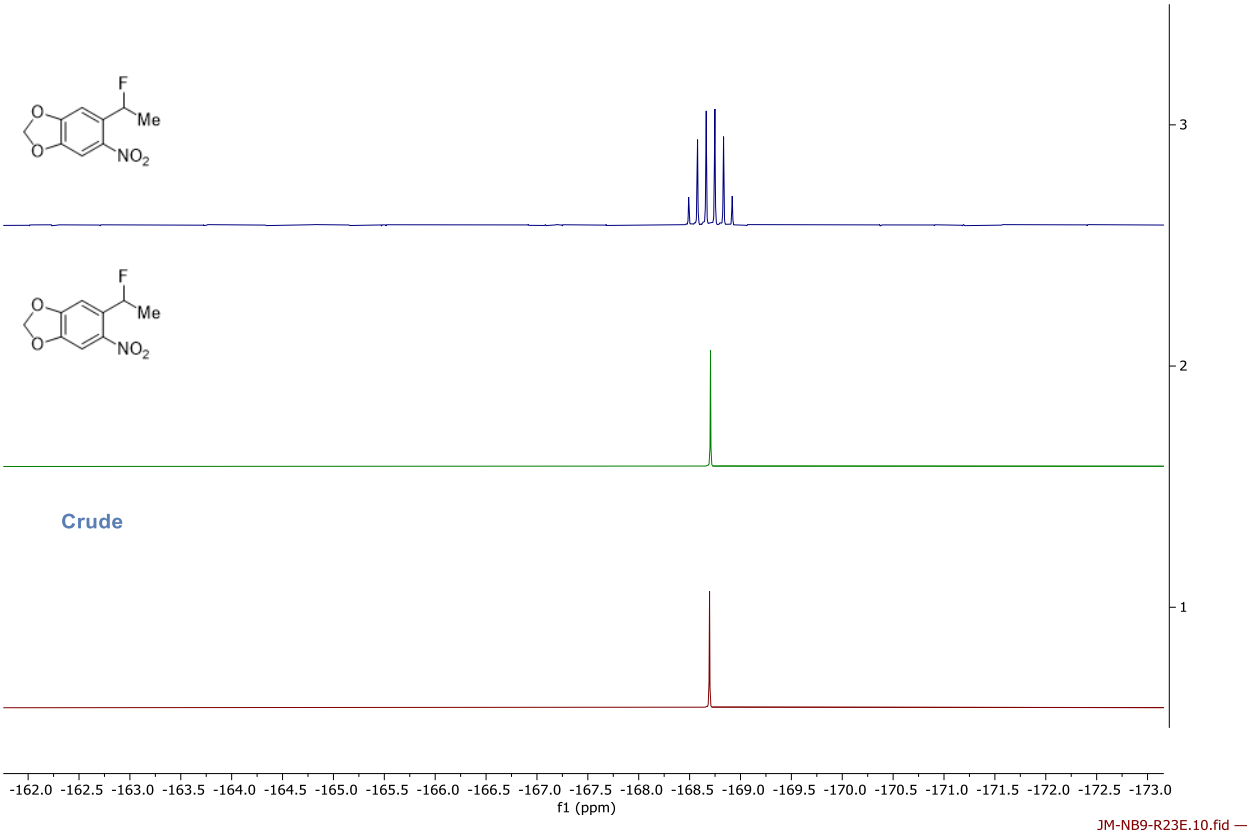
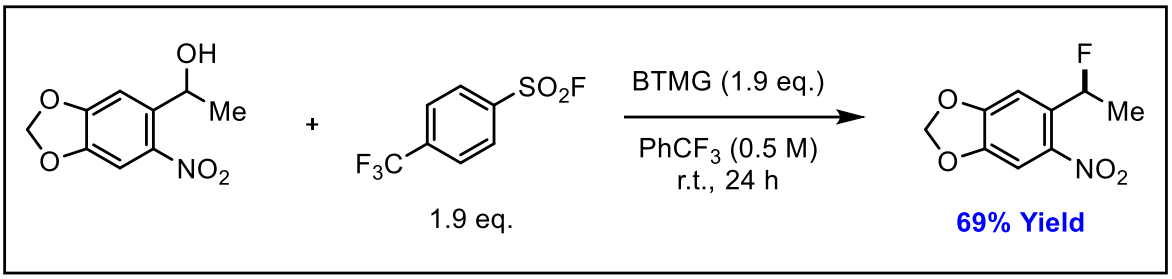


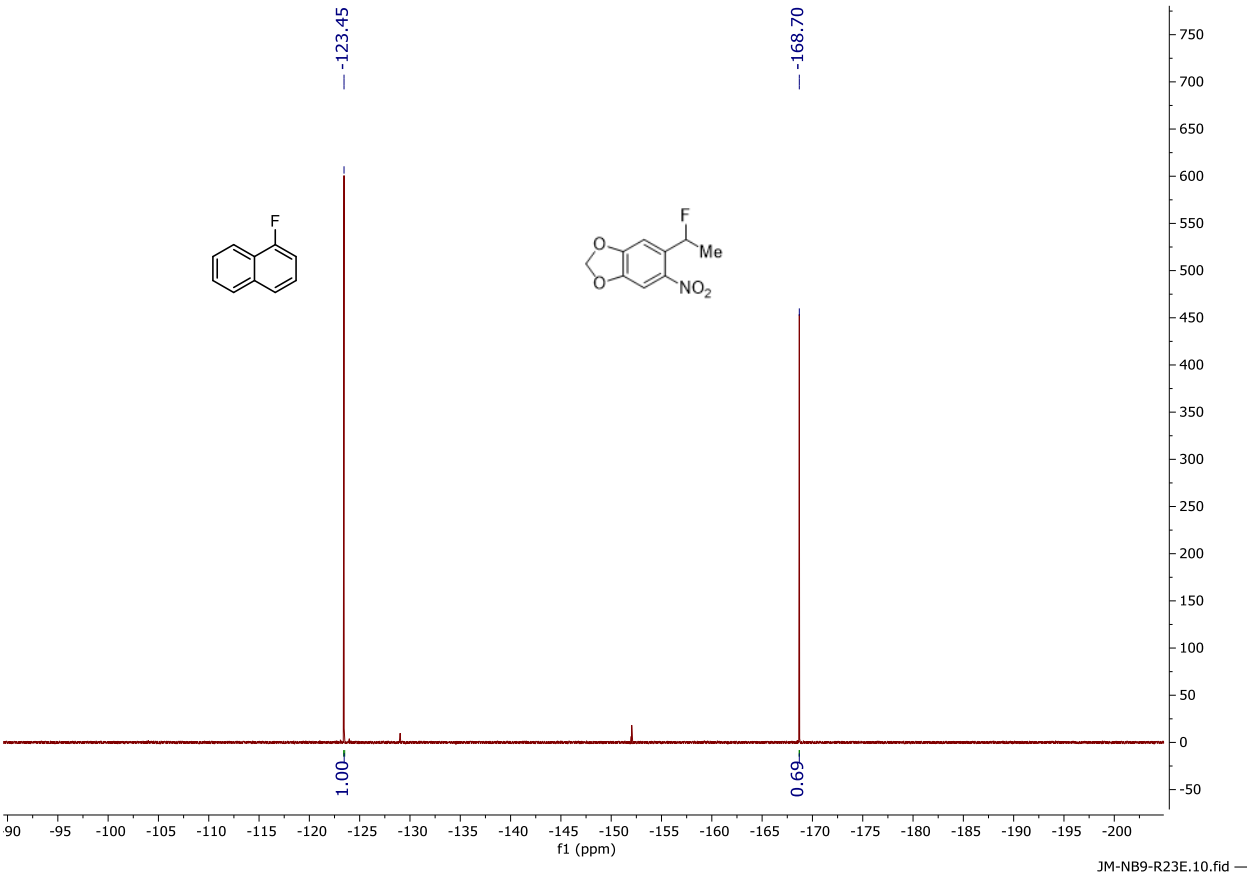

**Figure S73.** Time point measurements for the top four optimal conditons that were found after 10 rounds of optimization.

**$^{19}$F NMR Peak Assignment:** Overlay of proton coupled (top) and decoupled (middle, bottom) $^{19}$F NMR spectra for authentic product (top, middle) and crude reaction mixture (bottom).



## $^{19}$F NMR Assay Example



**5-(1-fluoroethyl)-6-nitrobenzo[d][1,3]dioxole.** was isolated from a crude reaction in screening. The title compound was isolated using automated column chromatography eluting with Ether:Hexanes (10 g column; 0% 5 CV, 0-10% 10 CV, 10% 4 CV). A fraction was collected for characterization.

**$^{1}$H NMR (500 MHz, $CDCl_3$):** δ 7.56 (apps, 1H), 7.16 (apps, 1H), 6.23 (dq, J = 48.3, 6.2 Hz, 1H), 6.14 (d, J = 4.4 Hz, 2H), 1.68 (dd, J = 24.1, 6.1 Hz, 3H).

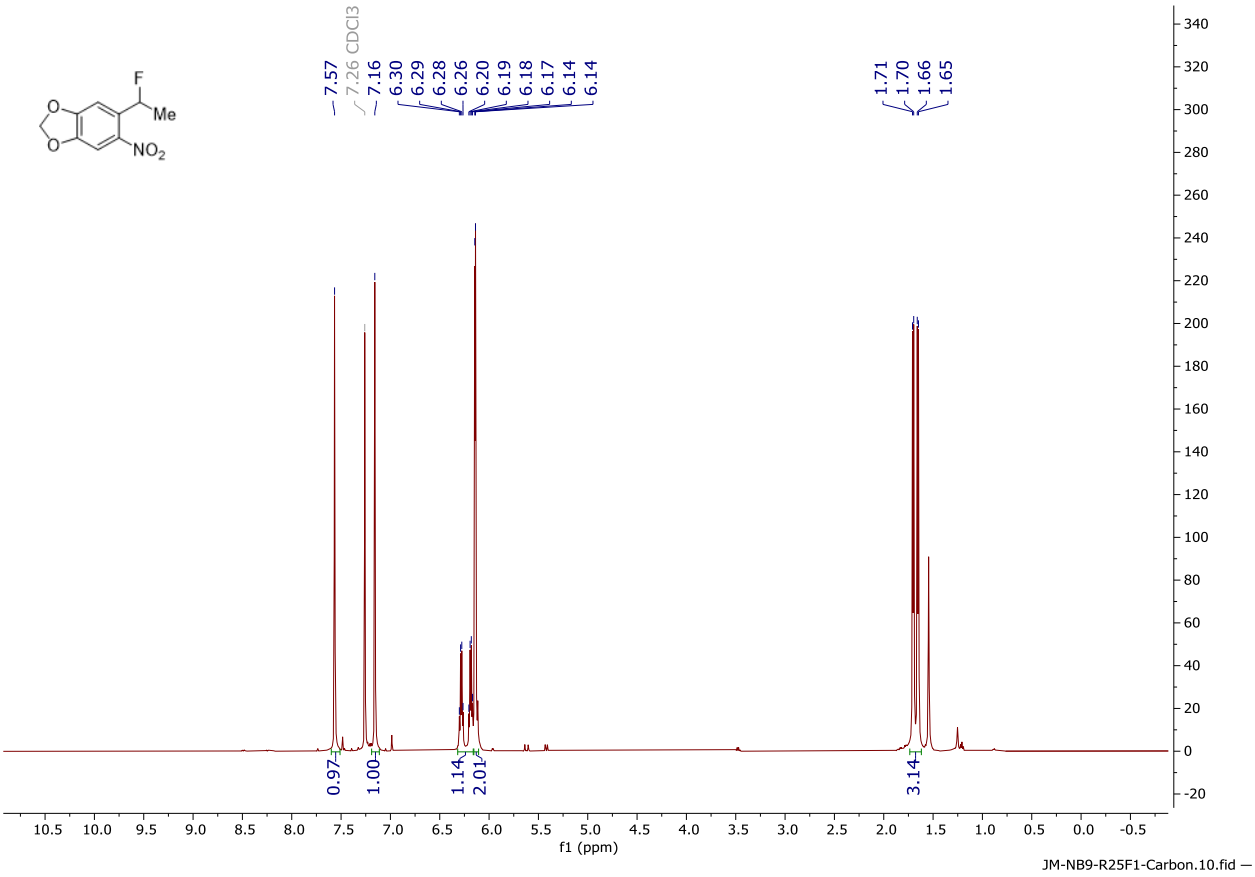
**$^{13}$C NMR (126 MHz, $CDCl_3$):** δ 152.97 (d, J = 2.1 Hz), 147.59 (d, J = 1.3 Hz), 140.46, 136.30 (d, J = 21.6 Hz), 105.80 (d, J = 16.7 Hz), 105.33, 103.29, 87.74 (d, J = 170.0 Hz), 23.14 (d, J = 25.2 Hz).

**$^{19}$F NMR (282 MHz, $CDCl_3$):** δ -168.71 (dq, J = 48.3, 24.2 Hz).

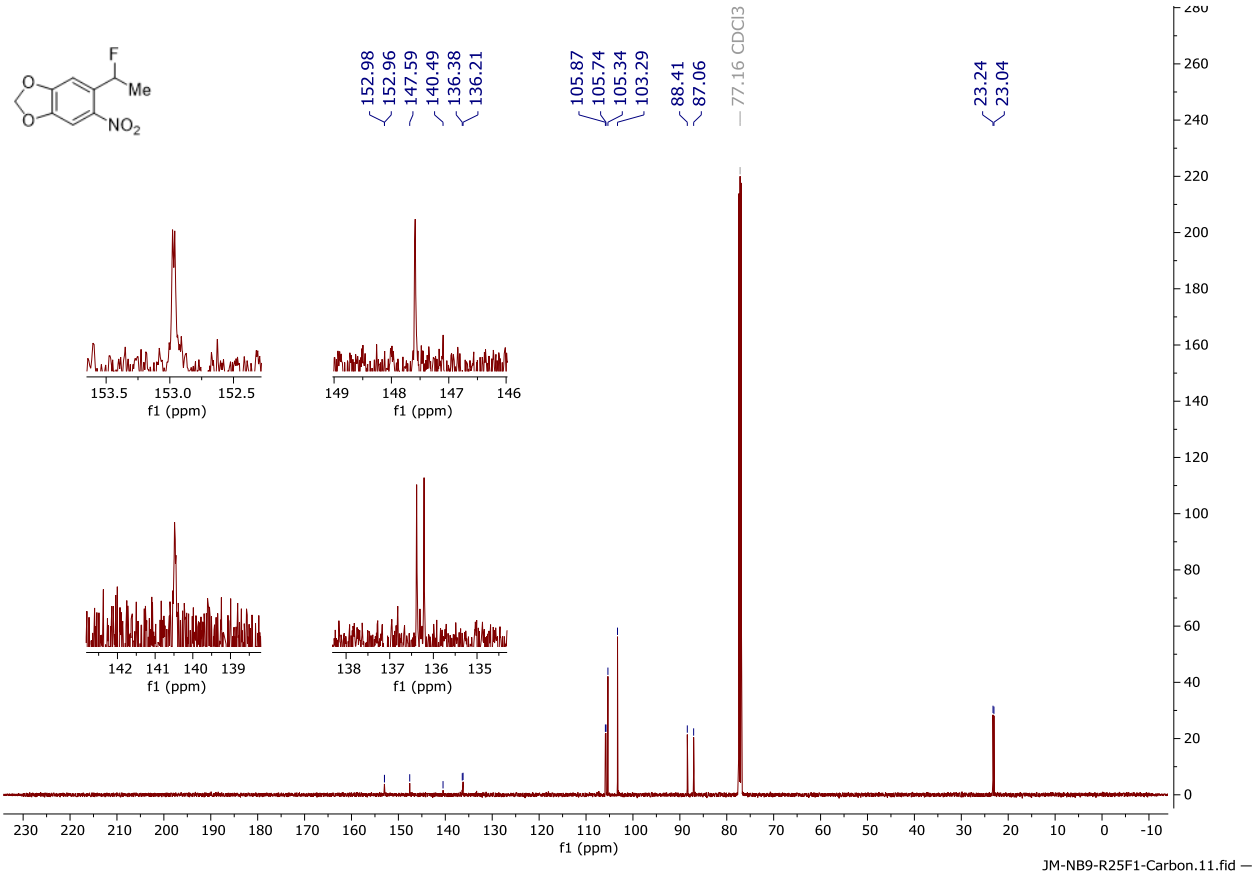
**HRMS:** (EI-TOF) calculated for $C_9H_8FNO_4$+: 213.0432, found 213.0426.

**FTIR (ATR cm$^{-1}$):** 1618, 1519, 1503, 1480, 1449, 1423, 1374, 1332, 1255, 1141, 1067, 1030, 984, 927, 878, 818, 765, 727, 661, 613, 530.

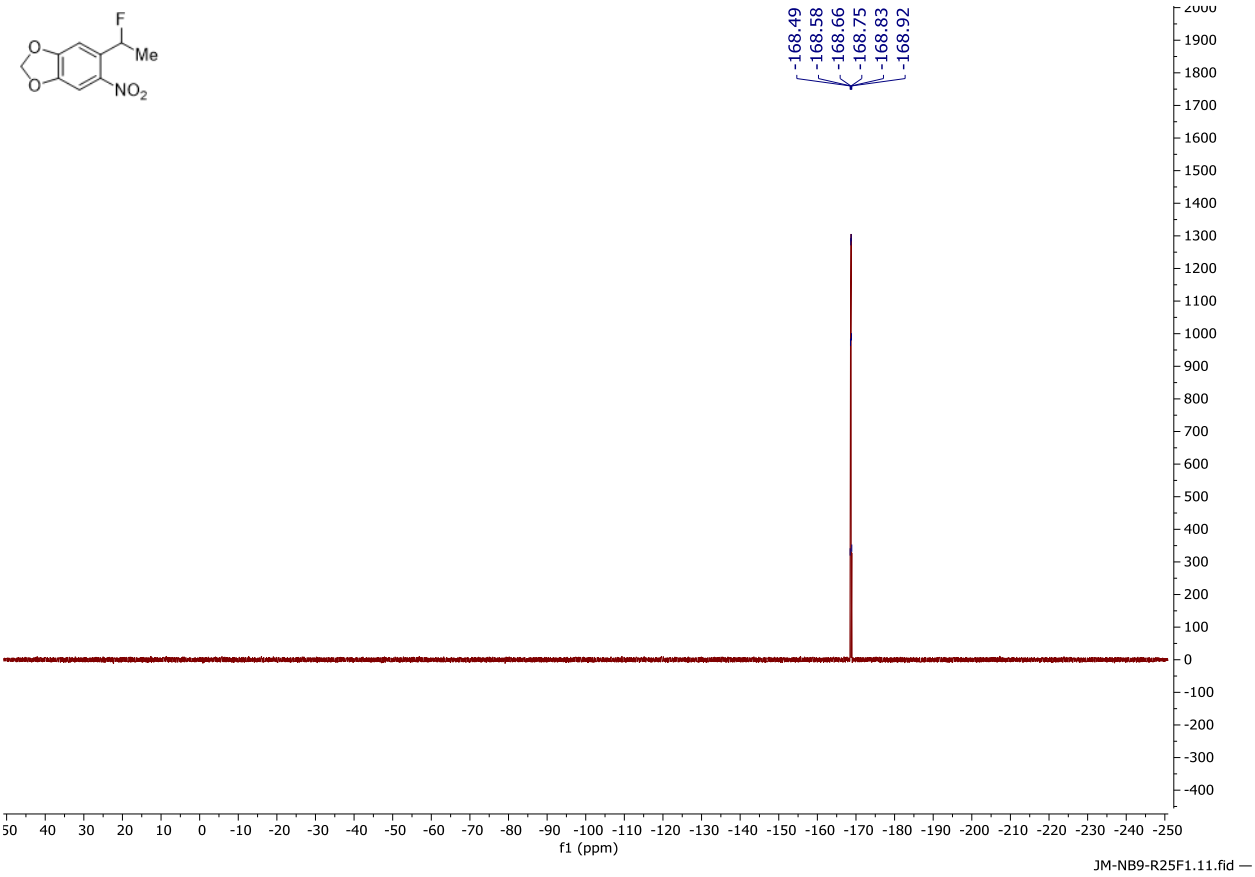
**$^{1}$H NMR (500 MHz, $CDCl_3$)**



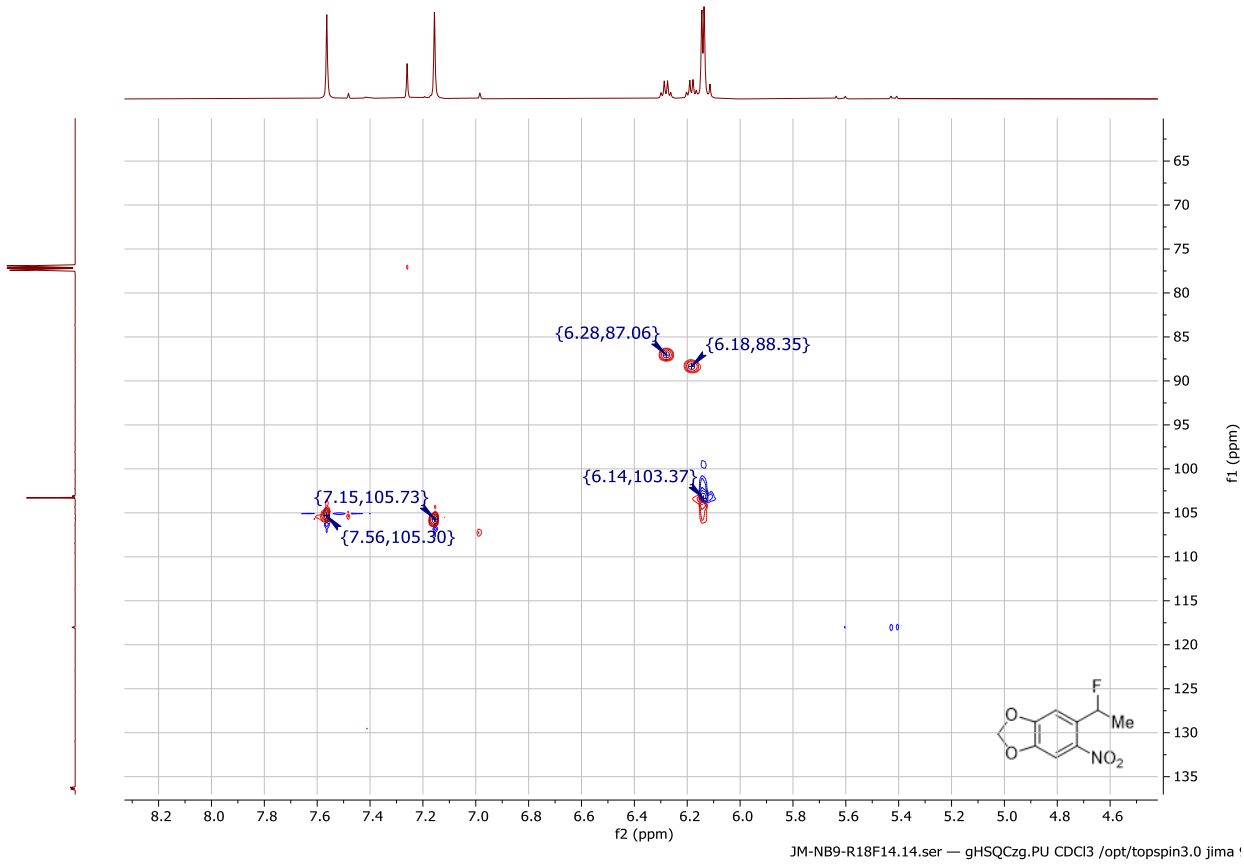
## $^{13}$C NMR (126 MHz, $CDCl_3$)



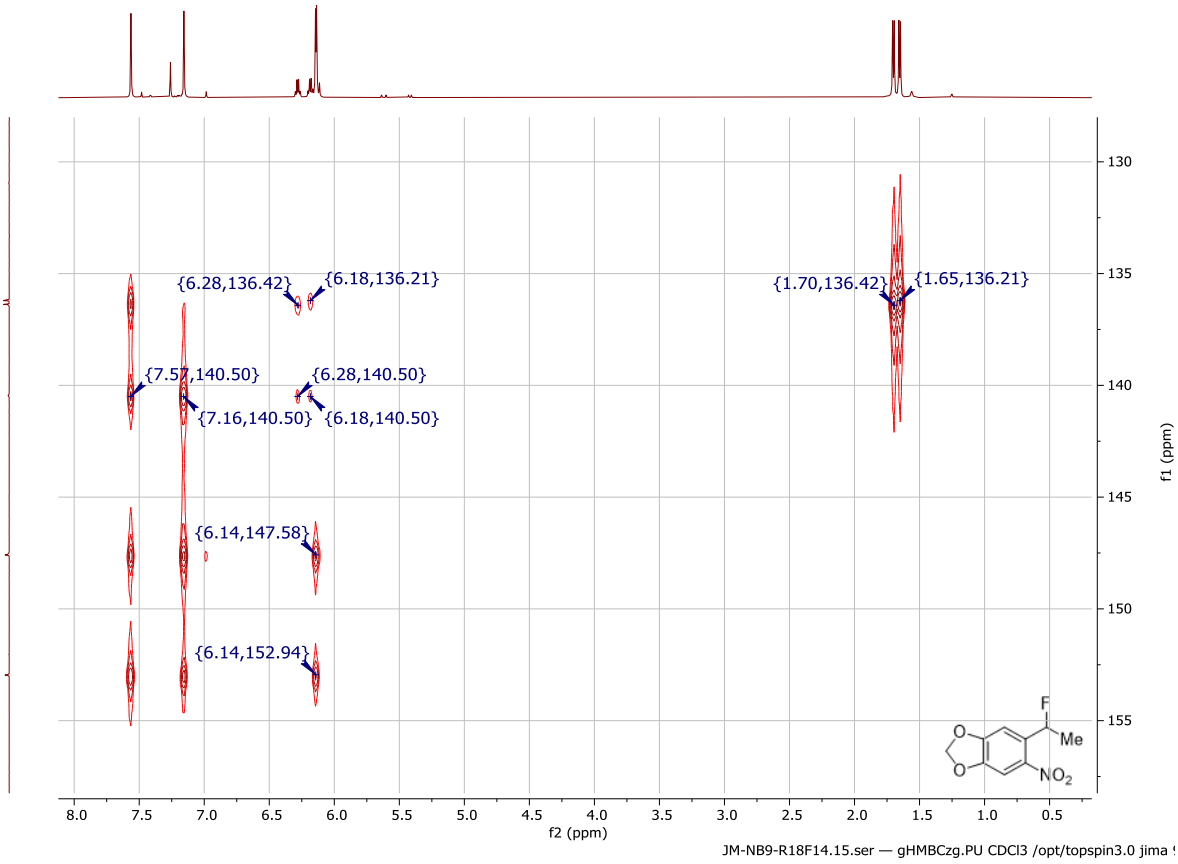
## $^{19}F$ NMR (282 MHz, $CDCl_3$)



## HSQC (500 MHz, $CDCl_3$)



## HMBC (500 MHz, $CDCl_3$)

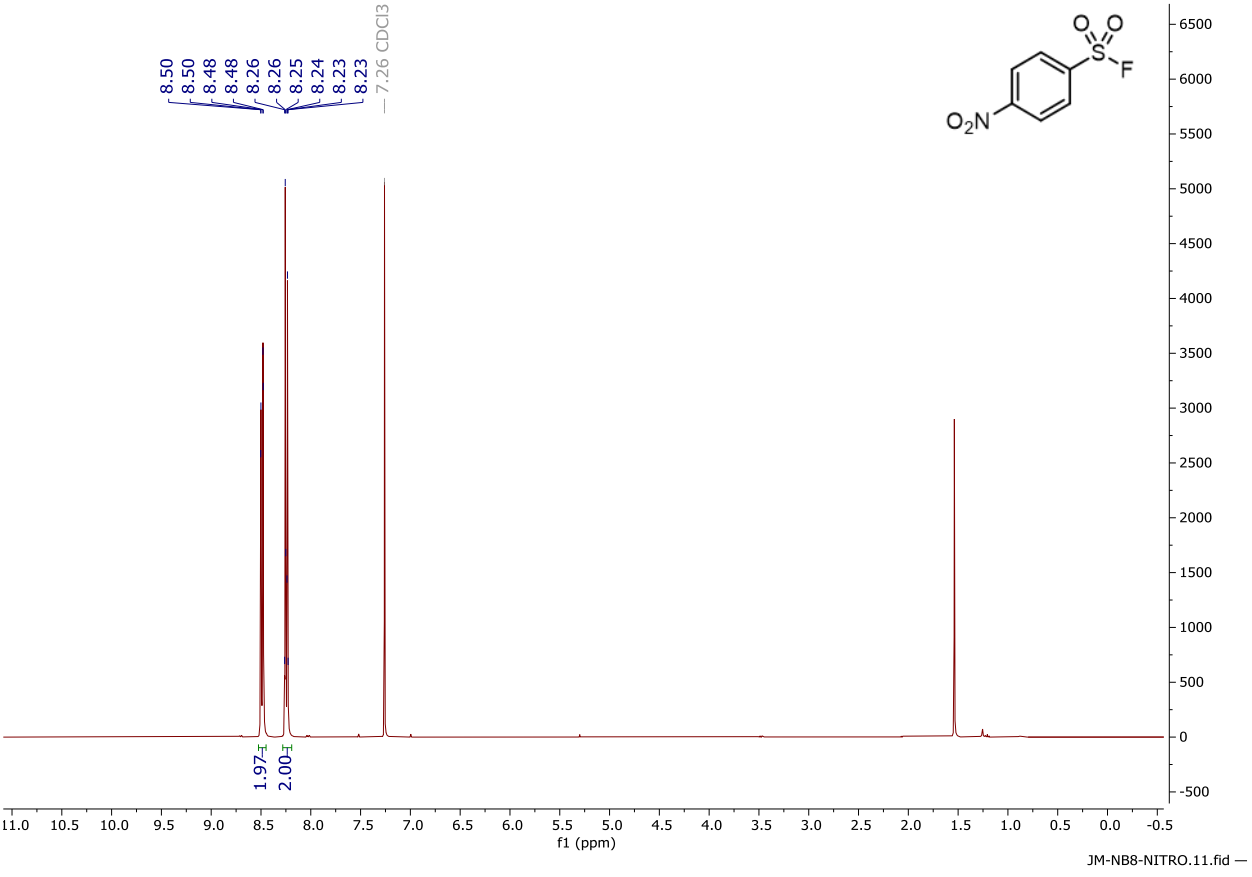


**4-nitrobenzenesulfonyl fluoride.** According to the published[31] procedure, this reaction was run with 4-nitrobenzenesulfonyl chloride (1.67 g, 7.52 mmol), potassium;fluoride;hydrofluoride (2.94 g, 37.62 mmol) and 15 mL of 3:1 water:acetonitrile. The mixture was stirred vigorously for one hour at room temperature. The mixture was diluted with 16 mL brine and extracted once with 40 mL ethyl acetate. The organic extract was filtered through 6 g of silica on a fritted filter, and subsequently rinsed with 10 mL ethyl acetate. The filtrate was concentrated to afford 4-nitrobenzenesulfonyl fluoride (1.44 g, 7.02 mmol, 93.33% yield) as a light brown solid. Compound characterization was in agreement with published procedure.

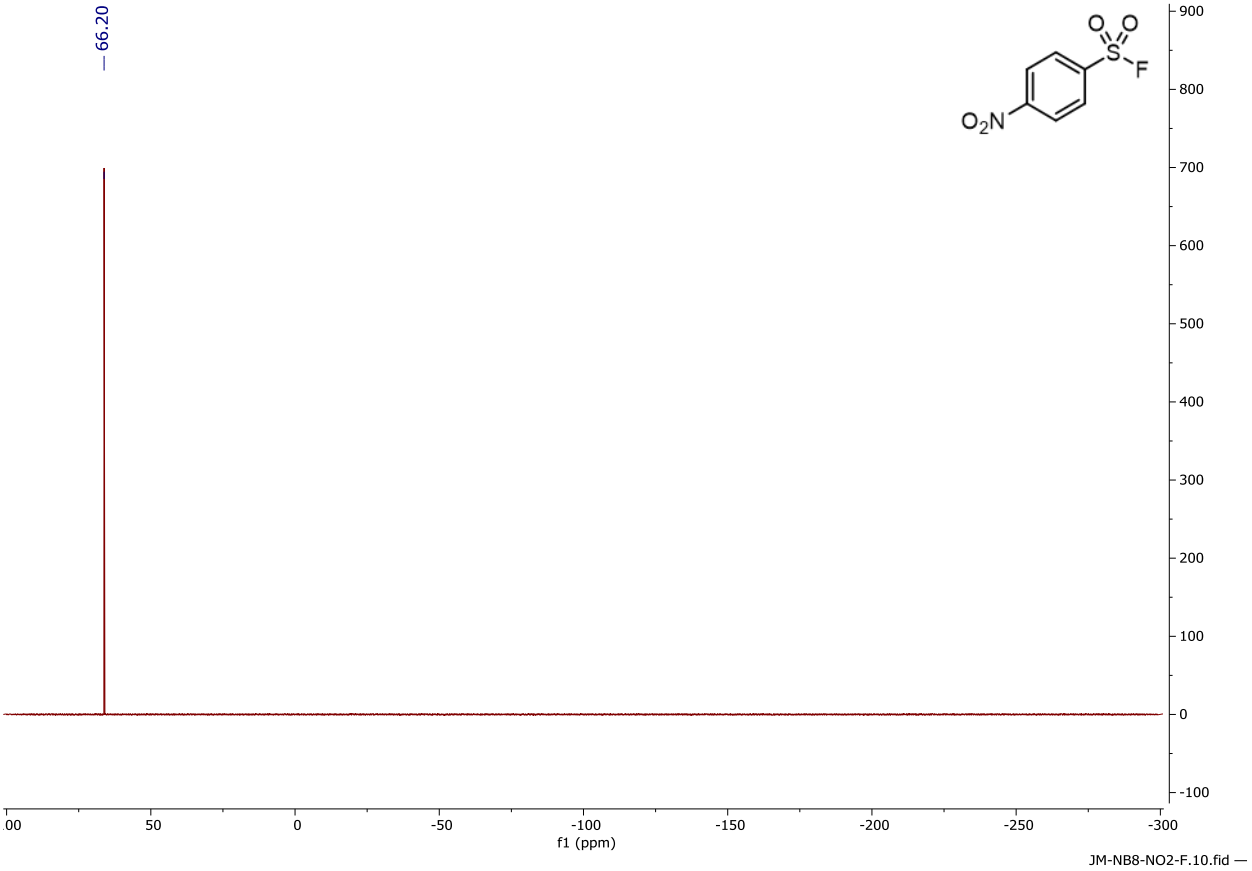
**$^{1}$H NMR (400 MHz, $CDCl_3$):** δ 8.55 – 8.44 (m, 2H), 8.30 – 8.15 (m, 2H).

**$^{19}$F NMR (282 MHz, $CDCl_3$):** δ 66.20 (s).

**$^{1}$H NMR (500 MHz, $CDCl_3$)**



## $^{19}F$ NMR (282 MHz, $CDCl_3$)

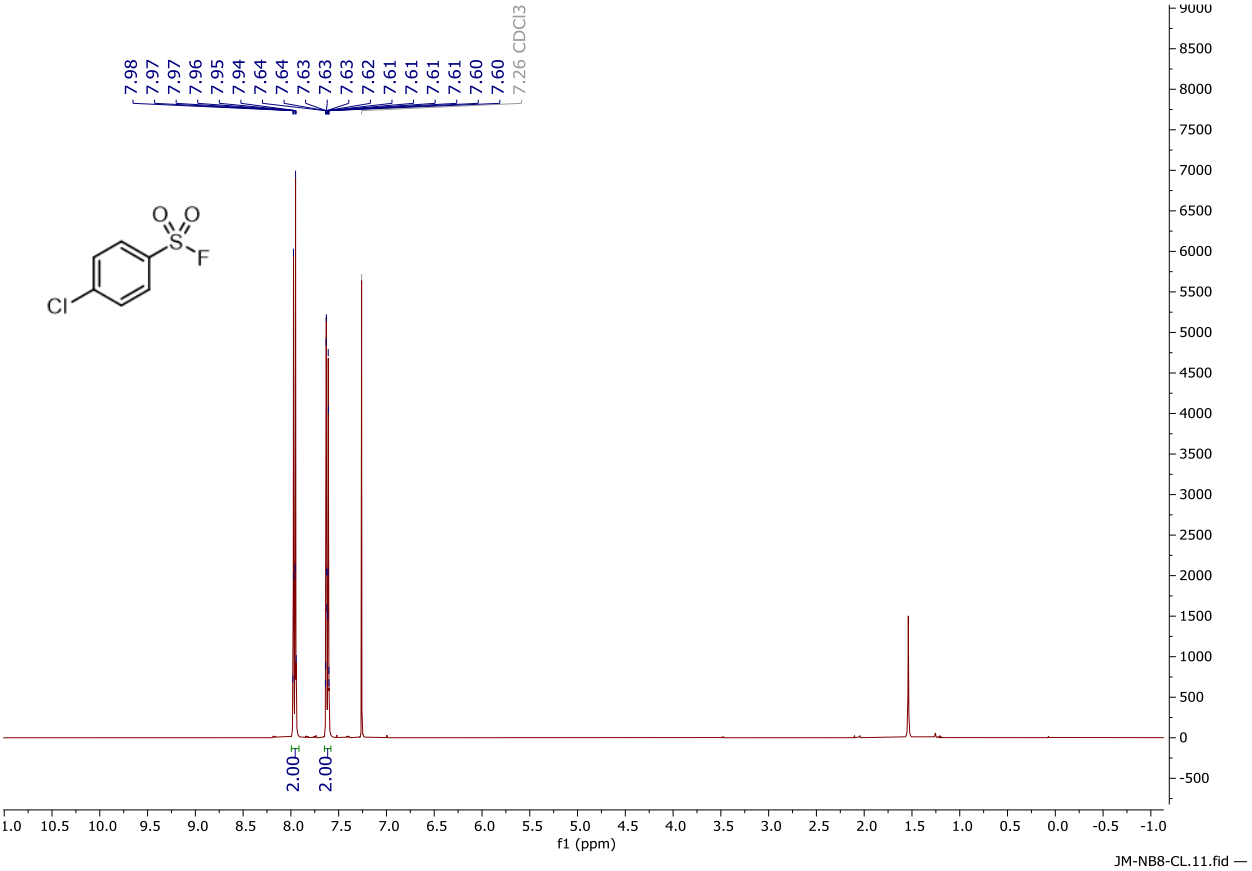


**4-chlorobenzenesulfonyl fluoride.** According to the published[31] procedure, this reaction was run with 4-chlorobenzenesulfonyl chloride (1.59 g, 7.52 mmol), potassium;fluoride;hydrofluoride (2.94 g, 37.62 mmol) and 15 mL of 3:1 water:acetonitrile. The mixture was stirred vigorously for one hour at room temperature. The mixture was diluted with 16 mL brine and extracted once with 40 mL ethyl acetate. The organic extract was filtered through 6 g of silica on a fritted filter, and subsequently rinsed with 10 mL ethyl acetate. The filtrate was concentrated to afford 4-chlorobenzenesulfonyl fluoride (1.34 g, 6.87 mmol, 91.28% yield) as a white solid. Compound characterization was in agreement with published procedure.

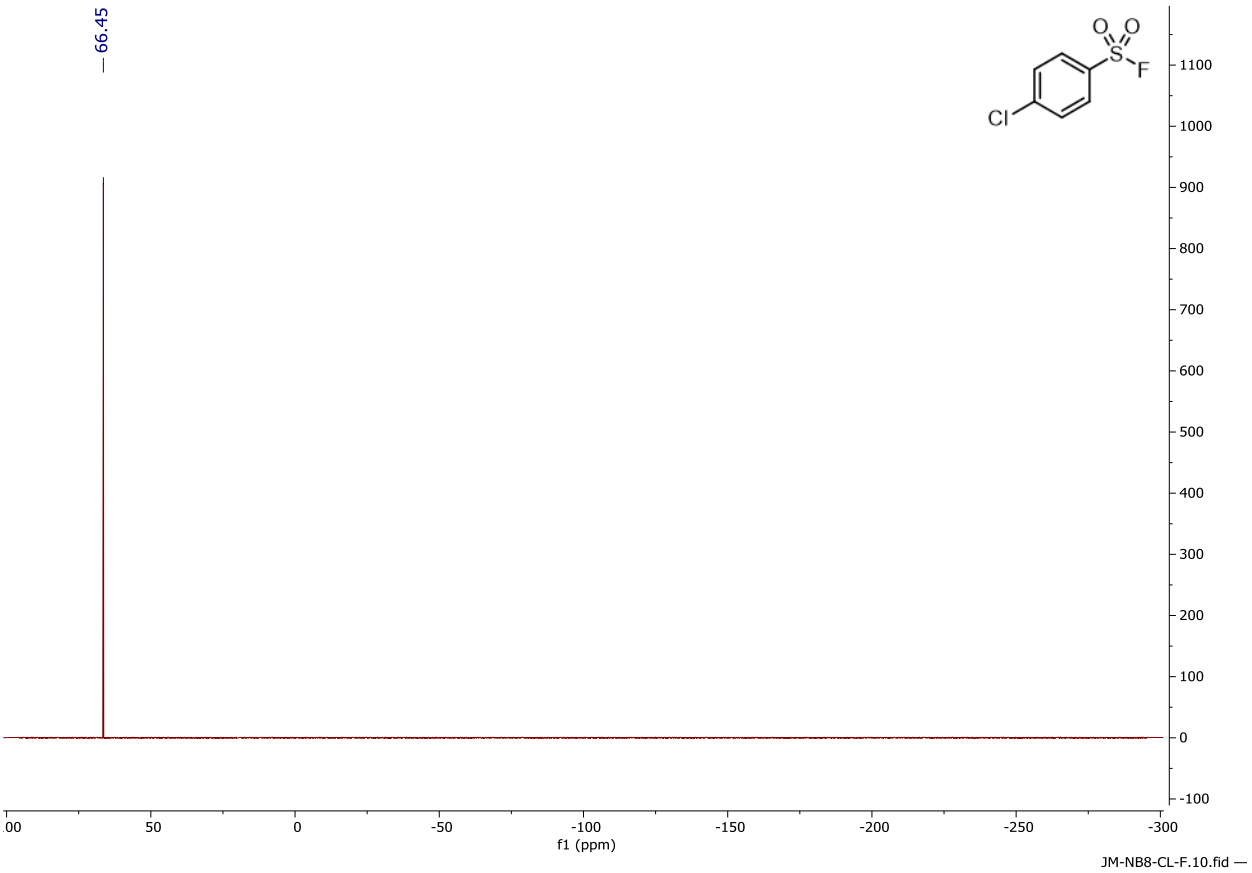
**$^{1}$H NMR (400 MHz, $CDCl_3$):** δ 8.05 – 7.92 (m, 2H), 7.67 – 7.58 (m, 2H).

**$^{19}$F NMR (282 MHz, $CDCl_3$):** δ 66.45 (s).

**$^{1}$H NMR (500 MHz, $CDCl_3$)**



## $^{19}F$ NMR (282 MHz, $CDCl_3$)

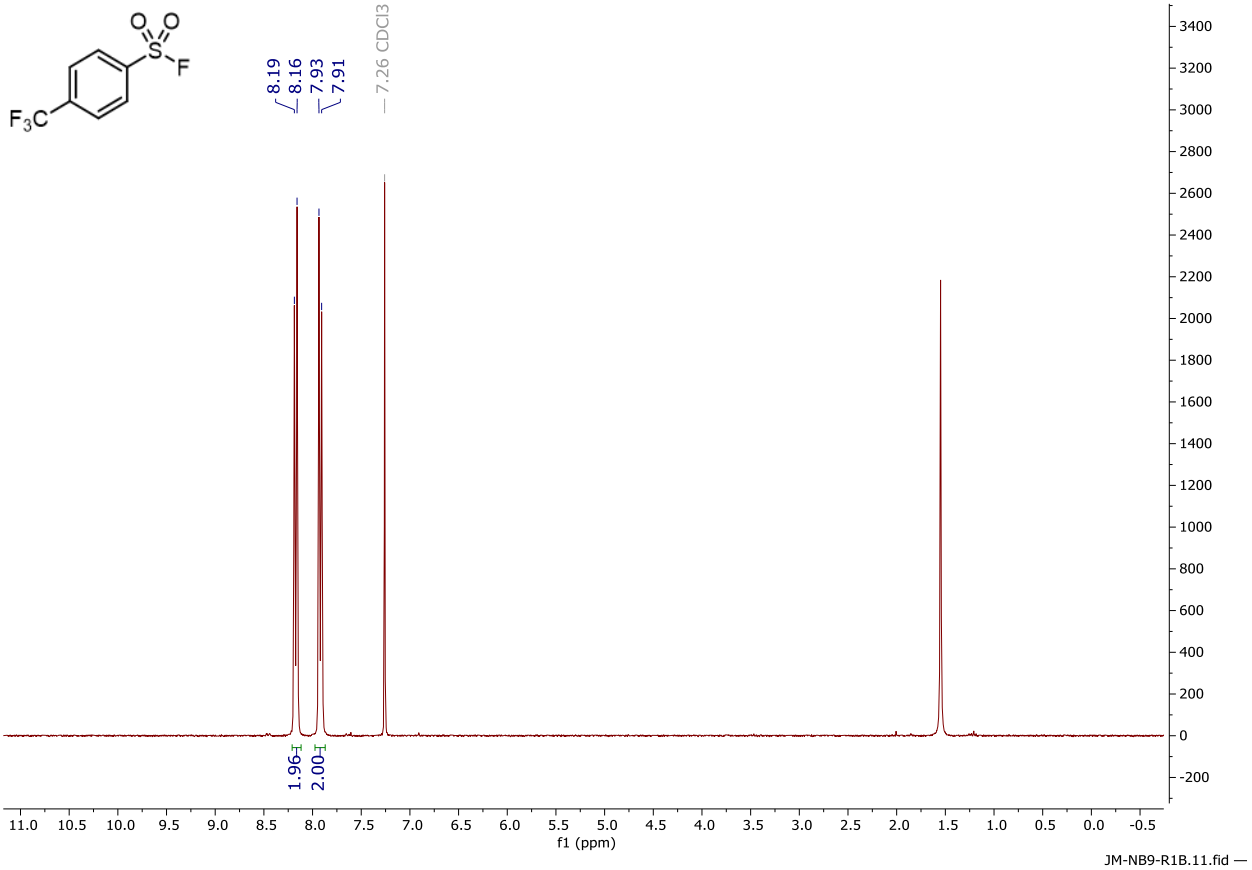


**4-(trifluoromethyl)benzenesulfonyl fluoride.** According to the published[31] procedure, this reaction was run with 4-(trifluoromethyl)benzenesulfonyl chloride (1.65 g, 6.75 mmol, 1.08 mL), potassium;fluoride;hydrofluoride (5.80 g, 74.28 mmol) and 40 mL of 1:1 water:acetonitrile. The mixture was stirred vigorously for two hours at room temperature. The mixture was extracted once with 80 mL ether. The organic layer was passed through a short silica plug and concentrated to afford 4-trifluoromethylbenzenesulfonyl fluoride (1.07 g, 4.67 mmol, 69.20% yield) as a white solid. Compound characterization was in agreement with published procedure.

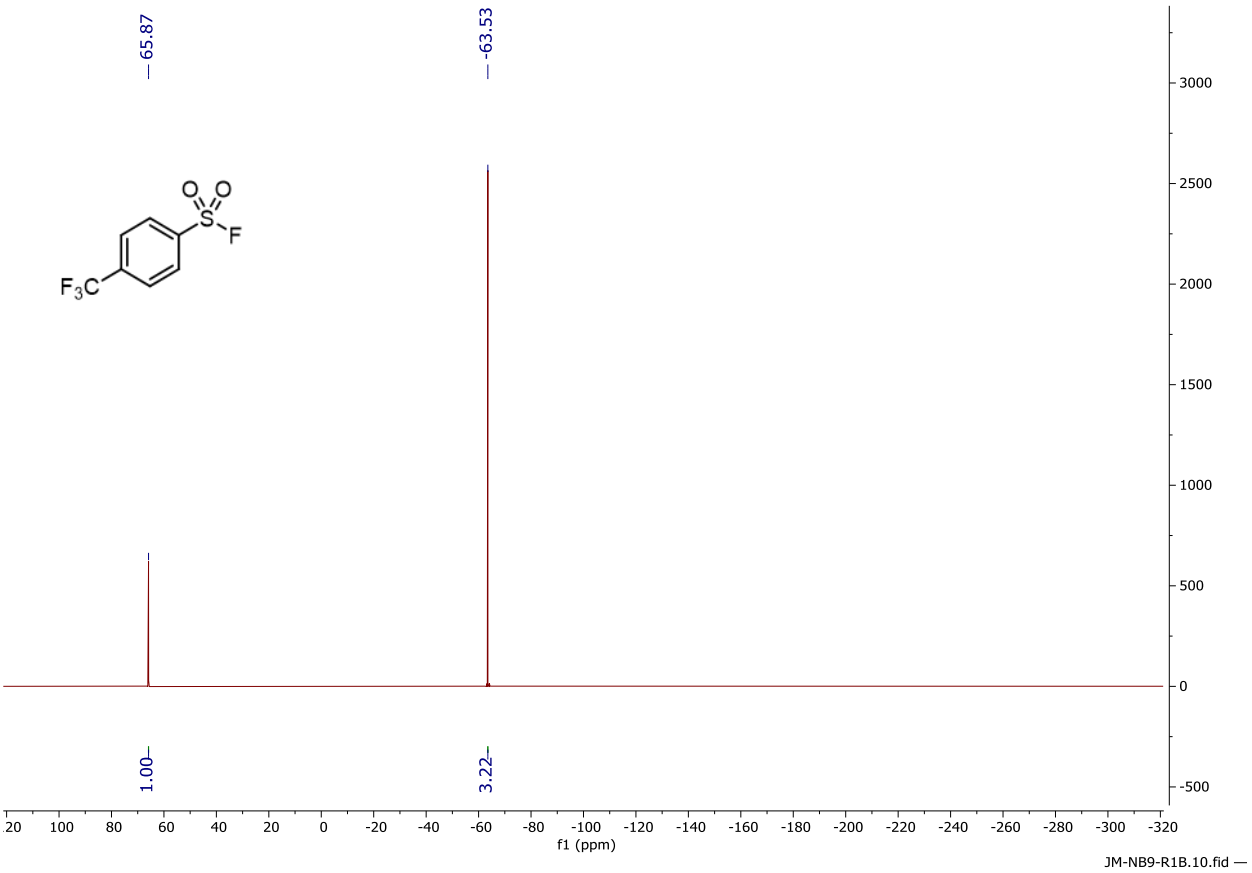
**$^{1}$H NMR (300 MHz, $CDCl_3$):** δ 8.17 (d, J = 8.2 Hz, 2H), 7.92 (d, J = 8.2 Hz, 2H).

**$^{19}$F NMR (282 MHz, $CDCl_3$):** δ 65.87 (s, 1F), -63.53 (s, 3F).

**$^{1}$H NMR (500 MHz, $CDCl_3$)**



## $^{19}F$ NMR (282 MHz, $CDCl_3$)

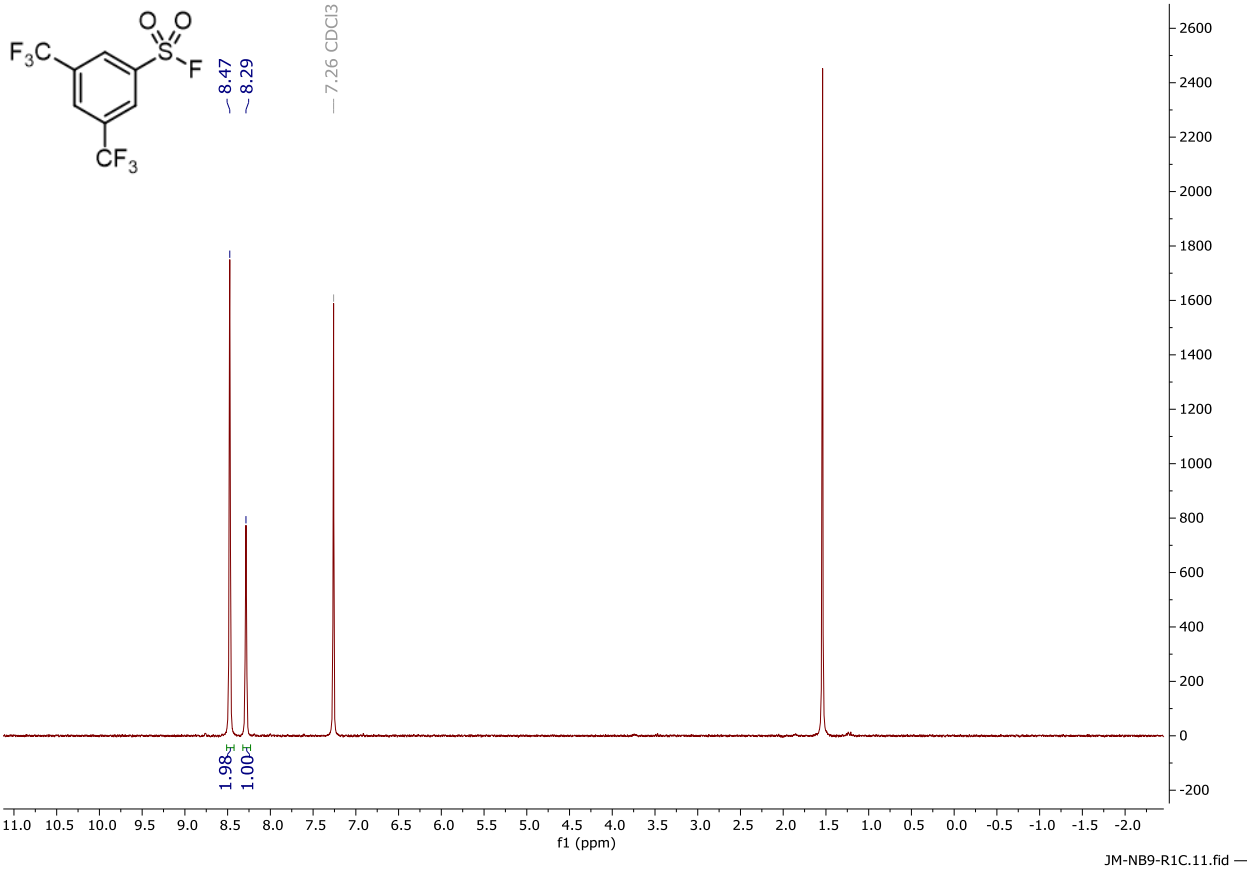


**3,5-bis(trifluoromethyl)benzenesulfonyl fluoride.** According to the published[31] procedure, this reaction was run with 3,5-bis(trifluoromethyl)benzenesulfonyl chloride (2.11 g, 6.75 mmol), potassium;fluoride;hydrofluoride (5.80 g, 74.28 mmol) and 40 mL of 1:1 water:acetonitrile. The mixture was stirred vigorously for two hours at room temperature. The mixture was extracted once with 80 mL ether. The organic layer was concentrated and isolated by automated column chromatography (25 g column, Hexanes 0% 12 CV) to afford 3,5-bis(trifluoromethyl)benzenesulfonyl fluoride (1.09 g, 3.69 mmol, 54.59% yield) as a pale white oil. Compound characterization was in agreement with a previously reported spectrum.[32]

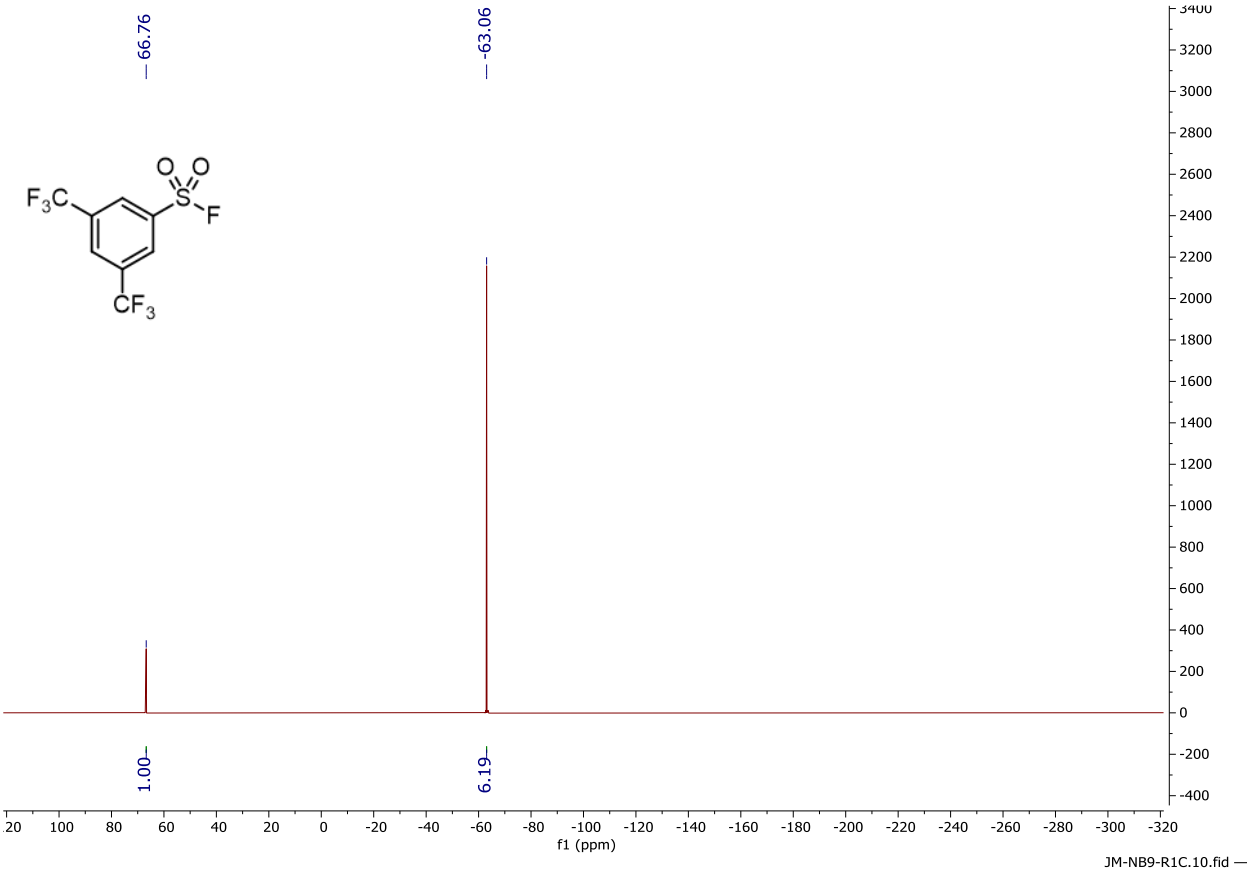
**$^{1}$H NMR (300 MHz, $CDCl_3$):** δ 8.47 (bs, 2H), 8.29 (bs, 1H).

**$^{19}$F NMR (282 MHz, $CDCl_3$):** δ 66.76 (s, 1F) , -63.06 (s, 6F).

**$^{1}$H NMR (500 MHz, $CDCl_3$)**



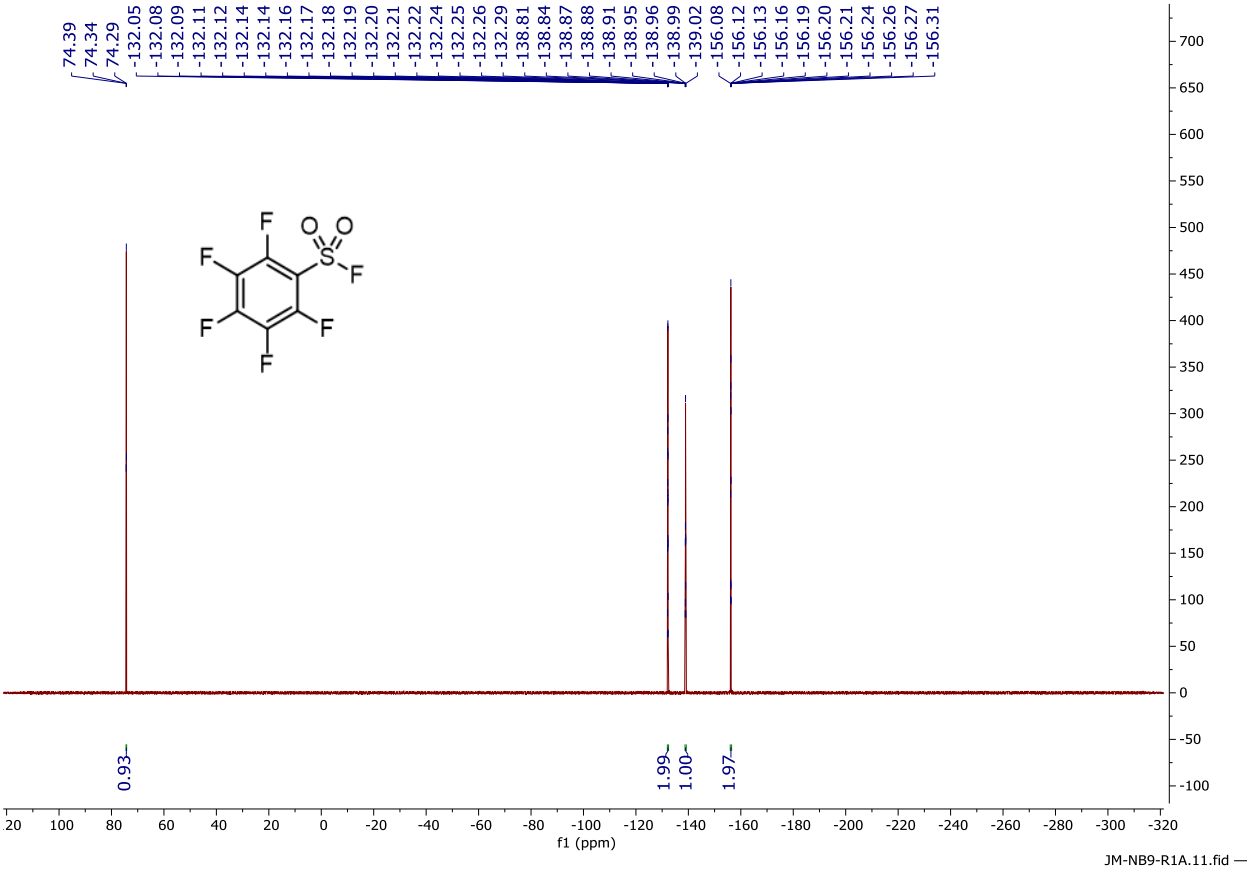
## $^{19}F$ NMR (282 MHz, $CDCl_3$)



**2,3,4,5,6-pentafluorobenzenesulfonyl fluoride.** According to a modified published[31] procedure, this reaction was run with 2,3,4,5,6-pentafluorobenzenesulfonyl chloride (1.80 g, 6.75 mmol), potassium;fluoride;hydrofluoride (5.80 g, 74.28 mmol) and 40 mL of 1:1 water:acetonitrile. The mixture was stirred vigorously for two hours at room temperature. The mixture was extracted once with 80 mL ether. The organic layer was concentrated and isolated by automated column chromatography (25 g column, EtOAc:Hex 0% 4 CV, 0-30% 12 CV, 30% 4 CV) to afford 2,3,4,5,6-pentafluorobenzenesulfonyl fluoride (1.69 g, 5.74 mmol, 85.03% yield) as a white oil. Compound characterization was in agreement with a previously reported spectrum.[33]

**$^{19}$F NMR (282 MHz, $CDCl_3$):** δ 74.34 (t, *J* = 14.6 Hz, 1F), -132.00 – -132.34 (m, 2F), -138.91 (tt, *J* = 21.3, 8.9 Hz, 1F), -155.97 – -156.55 (m, 2F).

**$^{19}$F NMR (282 MHz, $CDCl_3$)**



## XI. Design of Experiments

Optimization experiments. Design of experiments is commonly utilized for the optimization and understanding of continuous process parameters in chemistry. In general, an experimental design will be executed, the results used to build a model, and the model used to predict optimal conditions. However, the process of running DOE for a given system relies heavily on operator intuition and experience. What type of design to use? How to treat categorical variables? How to model responses? These questions require answers for a given application. Here, we benchmarked DOE as an optimization methodology using the following simulation design: (1) run $M$ experiments associated with an experimental design, (2) train a quadratic model using the data, and (3) predict the $N = B - M$ best conditions (where $B$ is the number of experiments in the budget). Then the best observed yield distribution in the experimental data can be compared to Bayesian optimization (BO) using the same number of experiments. We selected a total budget of 50 experiments for direct comparison to BO simulations.

Designs. Since many optimization reaction spaces are built from categorical variables, we need to use designs compatible with this data type. For example, response surface designs (e.g., central composite) require a sense of order and factorial designs require the use of only 2 levels, thus these designs are not readily adaptable to typical synthetic tasks. Hence DOE is typically only used for optimizations over continuous variables). One design compatible with multilevel categorical variables is the Generalized Subset Design (GSD).[34] GSD is a generalization of traditional fractional factorial designs to problems where factors can have more than two levels. By analogy to fractional factorial designs, GSDs enable the reduction of the number of experiments required to fill the design via a user-specified reduction factor. To carry out simulations we utilized the reduction factor which gave $N$ closest to the budget (50) such that $M < 50$. Another type of design compatible with multilevel categorical variables are D-optimal designs.[35] D-optimal designs are computer-aided designs which can be utilized when classical designs fail (e.g., reaction optimization with categorical variables). These designs have been employed effectively towards the optimization of chemical processes.[36] We will employ Federov's exchange algorithm to calculate the approximate design for each reaction.[37] Finally, we also investigated random designs as a point of comparison.

Response surface model. For response surface modeling we utilized a polynomial model since this is typical in DOE as a means to capture interactions. Due to the large number of features associated with such an expansion we will also use automatic relevance determination to fit the weights of the regression model.[38] We used OHE as descriptors for the data sets since most applications do not utilize chemical descriptors.

Implementation. While these designs are compatible with multilevel categorical variables, as the dimensionally of the search space grows, filling either design quickly become intractable experimentally. In addition, one point of experimental variability is that the ordering of the levels is arbitrary. That is, the experiments selected depend upon the order in which the components in each category are listed. This of course is not a desirable feature as different users may list their reagents in a different order for the same reaction space. Thus, to gauge the variance in outcome to the ordering of the categorical variables we carried out simulations by shuffling the ordering of variables in the data. Then with the optimization outcome distributions in hand we can compute some test statistics to evaluate the performance of BO versus DOE in the optimization of the 6 development reactions with the same experimental budget. To do this, we conducted a hypothesis test under the null hypothesis that the central tendency of BO and DOE performance is identical. Here, we remove the assumption of equal variance between the two population samples and utilize an unpaired Welch's t-test with Satterthwait degrees of freedom. As an additional demonstration of the versatility of EDBO, we wrote a model class that is compatible with the framework and then use EDBO methods to carry out the modeling. We used pyDOE2[39] to generate the GSD in python and AlgDesign[40] to generate the D-optimal design in R. The python and R code can be found on the EDBO github page.

GSD Results. For each of the reactions in the development set GSD optimization was able to identify the global maximum in some instances. However, upon the shuffling of levels there was a considerable amount of variance in outcome observed (Table S6). Therefore, while the difference in mean optimization outcome between the GSD and BO are small in some instances, the difference between the worst-case outcomes was significantly larger. For each reaction, the statistics tell us: (1) BO performs better on average (p-values $< 0.05$ in all cases), (2) BO has less

variance in outcome even with randomly selected initial conditions, and (3) BO has a better worst case scenario (closer to the global max yield).

| Reaction | $r$ | $M$ | $N$ | $\mu_{BO} - \mu_{GSD}$ | $\sigma_{BO} - \sigma_{GSD}$ | $min_{BO} - min_{GSD}$ | $t$ | $p$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 19 | 49 | 1 | 3 | -0.8 | 5 | 7.33 | 0.000 |
| 2a | 21 | 44 | 6 | 2 | -1.8 | 8 | 3.77 | 0.001 |
| 2b | 21 | 44 | 6 | 8 | -6.8 | 16 | 4.84 | 0.000 |
| 2c | 21 | 44 | 6 | 2 | -1.9 | 8 | 3.47 | 0.002 |
| 2d | 21 | 44 | 6 | 2 | -1.5 | 6 | 3.78 | 0.001 |
| 2e | 21 | 44 | 6 | 4 | -1.6 | 11 | 4.46 | 0.000 |

**Table S12.** Comparison of generalized subset design optimization results (20 simulations) to BO (30 simulations): (*r*) is the selected GSD reduction factor, (*M*) number of design points, (*N*) number of predicted points, ($\mu_{BO} - \mu_{GSD}$) difference in mean best observed yield, ($\sigma_{BO} - \sigma_{GSD}$) difference in best observed yield standard deviation, ($min_{BO} - min_{GSD}$) difference in lowest achieved yield, (*t*) t-statistic for a Welch's t-test for the null hypothesis of equal means, (*p*) p-value corresponding to *t*.

D-Optimal Results. For each of the reactions in the development set D-Optimal optimization was able to identify the global maximum in some or many instances. However, upon the shuffling of levels there was a considerable amount of variance in outcome observed (Table S7). Therefore, while the difference in mean optimization outcome between BO and D-optimal methods is small and not statistically significant in the case of reaction 2b, the difference between the worst-case outcomes was still large. For each reaction, the statistics tell us: (1) BO performs somewhat better on average (p-values < 0.05 in 5 out of 6 cases), (2) BO has less variance in outcome even with randomly selected initial conditions and more simulations, and (3) BO has a better worst case scenario (close to the global max yield).

| Reaction | $M$ | $N$ | $\mu_{BO} - \mu_{Dopt}$ | $\sigma_{BO} - \sigma_{Dopt}$ | $min_{BO} - min_{Dopt}$ | $t$ | $p$ |
|---|---|---|---|---|---|---|---|
| 1 | 45 | 5 | 3 | -2.4 | 11 | 3.30 | 0.003 |
| 2a | 45 | 5 | 2 | -1.3 | 6 | 4.19 | 0.000 |
| 2b | 45 | 5 | 1 | -2.7 | 10 | 1.95 | 0.066 |
| 2c | 45 | 5 | 1 | -0.7 | 2 | 2.68 | 0.012 |
| 2d | 45 | 5 | 1 | -1.0 | 4 | 3.18 | 0.004 |
| 2e | 45 | 5 | 2 | 0.0 | 1 | 3.94 | 0.000 |

**Table S13.** Comparison of D-Optimal design optimization results (20 simulations) to BO (30 simulations): (*M*) number of design points, (*N*) number of predicted points, ($\mu_{BO} - \mu_{Dopt}$) difference in mean best observed yield, ($\sigma_{BO} - \sigma_{Dopt}$) difference in best observed yield standard deviation, ($min_{BO} - min_{Dopt}$) difference in lowest achieved yield, (*t*) t-statistic for a Welch's t-test for the null hypothesis of equal means, (*p*) p-value corresponding to *t*.

Random Results. When compared to a random design (Table S8), only the D-Optimal design offered a modest improvement in performance statistics.

| Reaction | $M$ | $N$ | $\mu_{BO} - \mu_{Rand}$ | $\sigma_{BO} - \sigma_{Rand}$ | $min_{BO} - min_{Rand}$ | $t$ | $p$ |
|---|---|---|---|---|---|---|---|
| 1 | 45 | 5 | 2 | -1.5 | 6 | 3.23 | 0.004 |
| 2a | 45 | 5 | 1 | -1.2 | 4 | 3.30 | 0.003 |
| 2b | 45 | 5 | 3 | -4.1 | 13 | 3.31 | 0.004 |
| 2c | 45 | 5 | 2 | -1.4 | 8 | 3.01 | 0.006 |
| 2d | 45 | 5 | 2 | -1.5 | 6 | 3.28 | 0.004 |
| 2e | 45 | 5 | 3 | -2.2 | 11 | 3.13 | 0.004 |

**Table S14.** Comparison of random design optimization results (20 simulations) to BO (30 simulations): (*M*) number of design points, (*N*) number of predicted points, ($\mu_{BO} - \mu_{Rand}$) difference in mean best observed yield, ($\sigma_{BO} - \sigma_{Rand}$) difference in best observed yield standard deviation, ($min_{BO} - min_{Rand}$) difference in lowest achieved yield, (*t*) t-statistic for a Welch's t-test for the null hypothesis of equal means, (*p*) p-value corresponding to *t*.

## XII. References


1. Shields, B. *b-shields/auto-QChem*. (2020).
2. *PrincetonUniversity/auto-qchem*. (PrincetonUniversity, 2020).
3. b-shields/edbo. *GitHub* https://github.com/b-shields/edbo.
4. b-shields/EvML. *GitHub* https://github.com/b-shields/EvML.
5. Perera, D. *et al.* A platform for automated nanomole-scale reaction screening and micromole-scale synthesis in flow. *Science* **359**, 429–434 (2018).
6. 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* **360**, 186–190 (2018).
7. Domingos, P. A few useful things to know about machine learning. *Commun. ACM* **55**, 78–87 (2012).
8. Moriwaki, H., Tian, Y.-S., Kawashita, N. & Takagi, T. Mordred: a molecular descriptor calculator. *J. Cheminformatics* **10**, 4 (2018).
9. O'Boyle, N. M. *et al.* Open Babel: An open chemical toolbox. *J. Cheminformatics* **3**, 33 (2011).
10. *Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Keith, T.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski,*

*J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. Gaussian 09, revision D.01; Gaussian, Inc.: Wallingford, CT, 2009.*

11. Falivene, L. *et al.* Towards the online computer-aided design of catalytic pockets. *Nat. Chem.* **11**, 872–879 (2019).

12. Nagpal, A., Jatain, A. & Gaur, D. Review based on data clustering algorithms. in *2013 IEEE Conference on Information Communication Technologies* 298–303 (2013). doi:10.1109/CICT.2013.6558109.

13. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).

14. Gower, J. C. A General Coefficient of Similarity and Some of Its Properties. *Biometrics* **27**, 857–871 (1971).

15. Huang, Z. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Min. Knowl. Discov.* **2**, 283–304 (1998).

16. Biau, G. Analysis of a random forests model. *J. Mach. Learn. Res.* **13**, 1063–1095 (2012).

17. Shahriari, B., Swersky, K., Wang, Z., Adams, R. P. & de Freitas, N. Taking the Human Out of the Loop: A Review of Bayesian Optimization. *Proc. IEEE* **104**, 148–175 (2016).

18. Rasmussen, C. E. & Williams, C. K. I. *Gaussian processes for machine learning*. (MIT Press, 2006).

19. Gardner, J., Pleiss, G., Weinberger, K. Q., Bindel, D. & Wilson, A. G. GPyTorch: Blackbox Matrix-Matrix Gaussian Process Inference with GPU Acceleration. in *Advances in Neural Information Processing Systems 31* (eds. Bengio, S. et al.) 7576–7586 (Curran Associates, Inc., 2018).

20. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *ArXiv14126980 Cs* (2017).

21. Paszke, A. *et al.* PyTorch: An Imperative Style, High-Performance Deep Learning Library. in *Advances in Neural Information Processing Systems 32* (eds. Wallach, H. et al.) 8026–8037 (Curran Associates, Inc., 2019).

22. Jones, D. R., Schonlau, M. & Welch, W. J. Efficient Global Optimization of Expensive Black-Box Functions. *J. Glob. Optim.* **13**, 455–492 (1998).

23. Kushner, H. J. A New Method of Locating the Maximum Point of an Arbitrary Multipeak Curve in the Presence of Noise. *J. Basic Eng.* **86**, 97–106 (1964).

24. Srinivas, N., Krause, A., Kakade, S. & Seeger, M. Gaussian process optimization in the bandit setting: no regret and experimental design. in *Proceedings of the 27th International Conference on International Conference on Machine Learning* 1015–1022 (Omnipress, 2010).

25. Kandasamy, K., Krishnamurthy, A., Schneider, J. & Poczos, B. Parallelised Bayesian Optimisation via Thompson Sampling. in *International Conference on Artificial Intelligence and Statistics* 133–142 (2018).

26. Fox, R. J. *et al.* C–H Arylation in the Formation of a Complex Pyrrolopyridine, the Commercial Synthesis of the Potent JAK2 Inhibitor, BMS-911543. *J. Org. Chem.* **84**, 4661–4669 (2019).

27. Ji, Y. *et al.* Mono-Oxidation of Bidentate Bis-phosphines in Catalyst Activation: Kinetic and Mechanistic Studies of a Pd/Xantphos-Catalyzed C–H Functionalization. *J. Am. Chem. Soc.* **137**, 13272–13281 (2015).

28. Lapointe, D. & Fagnou, K. Overview of the Mechanistic Work on the Concerted Metallation–Deprotonation Pathway. *Chem. Lett.* **39**, 1118–1126 (2010).

29. Gorelsky, S. I., Lapointe, D. & Fagnou, K. Analysis of the Concerted Metalation-Deprotonation Mechanism in Palladium-Catalyzed Direct Arylation Across a Broad Range of Aromatic Substrates. *J. Am. Chem. Soc.* **130**, 10848–10849 (2008).

30. Delacre, M., Lakens, D. & Leys, C. Why Psychologists Should by Default Use Welch's *t*-test Instead of Student's *t*-test. *Int. Rev. Soc. Psychol.* **30**, 92 (2017).

31. 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.* **140**, 5004–5008 (2018).

32. Lee, C., Ball, N. D. & Sammis, G. M. One-pot fluorosulfurylation of Grignard reagents using sulfuryl fluoride. *Chem. Commun.* **55**, 14753–14756 (2019).

33. Laudadio, G. *et al.* Sulfonyl Fluoride Synthesis through Electrochemical Oxidative Coupling of Thiols and Potassium Fluoride. *J. Am. Chem. Soc.* **141**, 11832–11836 (2019).

34. Surowiec, I. *et al.* Generalized Subset Designs in Analytical Chemistry. *Anal. Chem.* **89**, 6491–6497 (2017).

35. 5.5.2.1. D-Optimal designs. https://www.itl.nist.gov/div898/handbook/pri/section5/pri521.htm.

36. Hsieh, H.-W., Coley, C. W., Baumgartner, L. M., Jensen, K. F. & Robinson, R. I. Photoredox Iridium–Nickel Dual-Catalyzed Decarboxylative Arylation Cross-Coupling: From Batch to Continuous Flow via Self-Optimizing Segmented Flow Reactor. *Org. Process Res. Dev.* **22**, 542–550 (2018).

37. optFederov function | R Documentation. https://www.rdocumentation.org/packages/AlgDesign/versions/1.2.0/topics/optFederov.

38. sklearn.linear_model.ARDRegression — scikit-learn 0.23.2 documentation. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.ARDRegression.html.

39. *clicumu/pyDOE2*. (clicumu, 2020).

40. AlgDesign package | R Documentation. https://www.rdocumentation.org/packages/AlgDesign/versions/1.2.0.