Roadmap to Pharmaceutically Relevant Reactivity Models Leveraging High-Throughput Experimentation

Jessica Xu[1‡], Dipannita Kalyani[2*‡], Thomas Struble[2], Spencer Dreher[2], Shane Krska[2], Stephen L. Buchwald[1*], Klavs F. Jensen[1*]

[1]Massachusetts Institute of Technology, Cambridge MA, USA
[2]Department of Discovery Chemistry, Merck & Co., Inc. Kenilworth, NJ 07033, USA

Introduction

The merger of High-Throughput Experimentation (HTE) and data science presents an enormous opportunity to both accelerate and inspire innovations in synthetic chemistry. Contemporary HTE enables the rapid and efficient exploration of diverse chemical spaces and reaction conditions while ensuring consistent data.[1–6] Similarly, developments in machine learning (ML) have enabled the distillation of large and complex data sets into predictive models capable of generalizing patterns in the data.[4,7–13] Despite these advances, efforts to merge HTE with ML remains largely limited to a few reported datasets with limited structural diversity[14–20] and corresponding trained models that do not extrapolate well to substrates beyond the training set.

An important application of HTE is optimization of reaction conditions for single or multiple target compounds.[19,21,22] While identifying reaction conditions using multi-parameter HTE is fruitful, it is time and resource intensive, especially in the context of medicinal chemistry campaigns for which large libraries of target analogues are often required. An efficient and forward–looking approach entails the creation of predictive reactivity models to assess reaction performance *in silico*, which would enable researchers to identify low– or high–yielding reactions thereby focusing time-intensive screening efforts only on targets predicted to be low yielding. Importantly, screening reaction conditions for low yielding targets will result in focused innovations to continually expand the synthetically accessible chemical space. Despite their potential impact, predictive models for drug–discovery relevant transformations that encompass the structural complexity and diversity encountered in pharmaceutical targets remain elusive. To this end, we embarked on a quest to build predictive models for palladium–catalyzed C–N cross–coupling, a workhorse reaction in medicinal chemistry, by leveraging the state–of–the–art HTE and ML methods.

A survey of prior work reveals that efforts to acquire and model reaction outcome of C–N coupling data falls under two distinct categories, employing historical (**Strategy I**), or HTE (**Strategy II**) datasets. These categories have distinct strengths and weaknesses and quite divergent outcomes. Pd–catalyzed C–N coupling data have been extracted from historical reaction sets such as patent databases and Reaxys[8,9,23–26] as well as Electronic Laboratory Notebooks (ELNs)[27]. Yield prediction for historical datasets (**Strategy I**) results in models with relatively poor performance (as evidenced by low coefficient of determination ($R^2 \sim 0.2$)) in part due to significant heterogeneity in the data quality which can be time consuming and often impossible to curate systematically. Additionally, these historical datasets might not contain the state–of–the–art HTE–amenable conditions for library synthesis due to temporal characteristics of the reagent use in these datasets, as adoption of modern reagents often lags behind the current literature.[25,27] It is imperative

‡ Authors contributed equally
* Corresponding authors:
dipannita.kalyani@merck.com (DK), sbuchwal@mit.edu (SLB), kfjensen@mit.edu (KFJ)

that reaction ML models be generated using the most robust set of HTE amenable automation friendly reaction conditions to maximize applicability across a broad range pharmaceutically relevant chemical space. Importantly, automation amenable reaction conditions are particularly important when large libraries are needed at nanomole scale which can only be executed using robotic equipment.

Yield–modeling efforts on the few reported HTE datasets demonstrate the potential advantage of **Strategy II**. Modeling the yield of these datasets (4K C–N couplings[15], or 2K Suzuki–Miyaura couplings in flow[14]) produces predictive models with R^2 or AUROC > 0.9.[11,15,27–34] However, models trained on these datasets demonstrate limited ability to extrapolate beyond the molecules in their training sets, in part due to the minimal structural diversity in the dataset.

Herein, we detail the first ML models for Pd-catalyzed C–N couplings using pharmaceutically relevant structurally diverse large data sets (~ 5000 unique products) generated using nanomole scale compatible chemistry. Careful consideration was given to both the diversity of the data set, and accurate model predictions for substrates bearing features beyond those present in the training set. The structural diversity in the data set was enabled by leveraging the Merck & Co., Inc Building Block Collection (MBBC) which comprises >22000 amines and >7000 aryl halides, theoretically allowing access to hundreds of millions of products. Our initial efforts focused on C–N coupling using secondary amines thereby reducing this virtual chemical space to ~15M products using ~5000 aryl bromides and ~3000 secondary amines (Figure 1).

The sections below address the challenges and outline a workflow to build predictive models for this enormous chemical space through the prudent use of HTE. The generation of the large, pharmaceutically relevant dataset was enabled by the discovery of state–of–the–art nano–chemistry compatible C–N coupling conditions. To the best of our knowledge, this is the first report that addresses the experimental considerations for generating structurally diverse HTE data sets through the careful, and systematic assessment of multiple aspects of data quality and their impact on resulting ML models.

Results

General considerations for the generation of large diverse datasets: As mentioned above, our efforts aimed to generate predictive models for the Pd–catalyzed C–N coupling of secondary amines. In view of the limitations of the existing datasets discussed in the introduction, this goal necessitated the *de novo* generation of diverse, pharmaceutically relevant datasets. Several experimental considerations were important for the generation of requisite data sets, including 1) the identification of appropriate nanomole scale amenable, automation friendly C–N coupling conditions[1] that would enable the generation of large datasets, 2) selection and generation of a structurally diverse molecular dataset mirroring the complexity present in pharmaceuticals, and 3) systematic interrogation of data quality. The subsequent sections detail our workflow, addressing each of these considerations and their resulting impact on the predictive models.

[1] e.g., the use of DMSO, a high boiling point solvent, compatible with nano-well materials and capable of solubilizing reagents

Identification of nanomole scale compatible pharmaceutically relevant reaction conditions: To date, the modelling efforts on C–N couplings generated using nanomole scale (0.1 µmol) chemistry have employed ^{t}BuXPhos Pd G3 catalyst and P_2Et as the base.[17] The performance of these reaction conditions for Pd-catalyzed C–N coupling have been benchmarked for the coupling of Informer halides with piperidine at microscale (2.5 µmol).[35] The Informer halides are a standardized set of 18 complex drug-like halides to assess the relevance of synthetic methods for the functionalization of pharmaceutically-relevant substrates.[36,37] These studies revealed significant opportunity for the identification of more robust reaction conditions since only 5/18 Informer halides were transformed to the desired products with >20% product LCAP (Liquid Chromatography Area Percent). Hence, in a series of preliminary studies, we aimed to identify the catalyst system and base that enabled the broadest scope with respect to the coupling of drug-like aryl halides with secondary amines. LCAP was used to assess the degree of product formation.

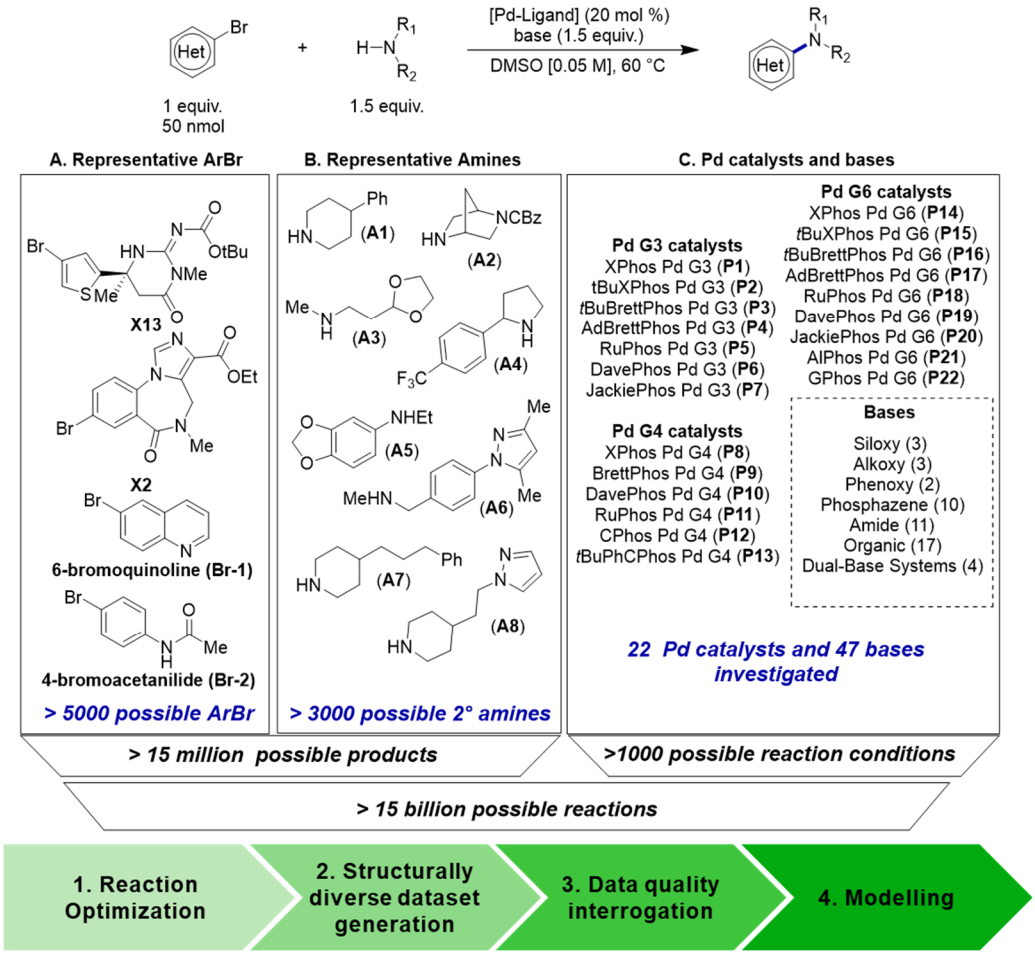


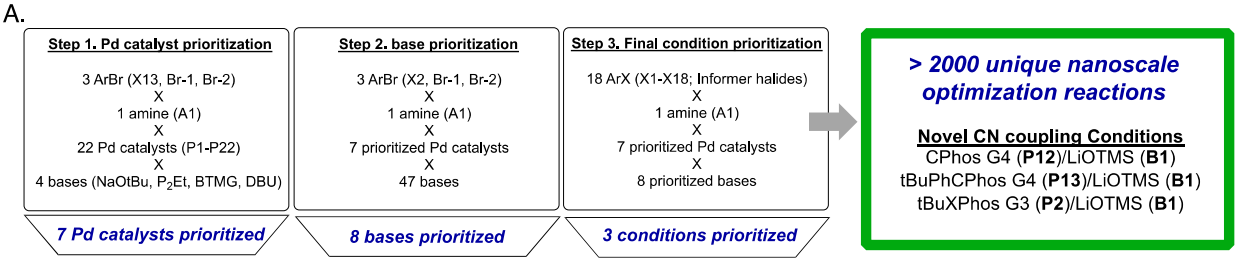
Figure 1: *Dataset Scope and overall workflow*

As shown in Figure 1, we investigated the use of 22 distinct catalysts generated through combinations of 12 biaryl ligands and three different generations of precatalysts (G3, G4 or G6). Additionally, 47 distinct bases were used for these studies. The catalysts and bases were largely chosen to maximize structural diversity and ensure solubility in DMSO. Theoretically, >1000 reagent combinations exist among these catalysts and bases. Together with the >15 million

possible products, the total number of reactions for a full factorial evaluation of all reaction parameters and substrates exceeds 15 billion. As detailed below, our experimental workflow enabled us to identify suitable reaction conditions for the generation of large data sets using a subset of this vast possibility of reaction permutations.

The couplings of three aryl bromides, 6-bromoquinoline (**Br-1**), 4–bromoacetanilide (**Br-2**), and **X13**, with 4-phenyl piperidine (**A1**) were conducted using 22 Pd precatalysts (see Figure 1C for catalyst identities) and four bases. Informer halide **X13** was intentionally included to identify conditions that are amenable to couplings with pharmaceutically relevant halides. This screen led to the prioritization of seven precatalysts for further study (Figure 2A, Step 1). These seven precatalysts were used in combination with 47 bases for the coupling of **X2**, **Br-1** and **Br-2** with 4-phenylpiperidine (**A1**) to identify eight bases that led to the highest product LCAPs (Figure 2A, Step 2). To assess the robustness and generality of the 56 conditions (seven precatalysts × eight bases) resulting from the prioritized precatalysts and bases, the reactions of 4–phenyl piperidine with 18 complex aryl halides (Informer halides, **X1**–**X18**) were assessed (Figure 2A, Step 3). As shown in Figure 2A, three sets of conditions: CPhos Pd G4 (**P12**) with LiOTMS (**B1**) as the base, (^{t}Bu)PhCPhos Pd G4 (**P13**) with LiOTMS (**B1**) as the base, and ^{t}BuXPhos Pd G3, (**P2**) with LiOTMS (**B1**) as the base demonstrated a more general scope than conditions previously used (^{t}BuXPhos Pd G3 precatalyst (**P2**) with P_2Et as base).[17] The success rates and average product LCAPS with the newly identified conditions (Figure 2B, entries 1-3) are the best reported to date for nano-scale amenable Pd-catalyzed C–N couplings. These conditions led to ~3-4-fold increase in the number of Informer halides leading to products with >20% LCAP versus the previous method using catalyst **P2** with P_2Et as the base. Ultimately, we utilized a single set of conditions, **P13** and **B1**, for all subsequent studies.

A.



B.

		X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	X17	X18	Average LCAP	# >20% LCAP
1	CPhos G4/LiOTMS	3	60	55	28	13	74	5	55	24	3	0	68	53	43	41	33	2	6	31	11
2	tBuXPhos G3/LiOTMS	25	11	73	45	4	61	4	33	53	0	0	41	51	32	35	9	0	1	27	10
3	tBuPhCPhos G4 LiOTMS	35	4	17	26	22	42	3	33	38	0	0	79	17	35	41	9	1	0	22	9
4	tBuXPhos G3/P2Et	0	15	59	19	2	20	0	17	0	0	0	29	6	7	22	11	0	4	12	3

Figure 2: (A) Reaction optimization and elucidation of best conditions. (B) Product LCAPS for Figure 2A, Step 3; see SI for structures of Informer Halides.

Selection of aryl halides and amines: Having determined appropriate automation friendly reaction conditions set the stage to generate the requisite large HTE datasets for modeling. The first step toward this goal was to select a diverse array of pharmaceutically relevant aryl halides and amines. Of the aryl halides (>5000) and amines (>3000) that met the functional group (FG)

filtration criteria to eliminate FG's with obvious reaction compatibility issues (e.g., highly reactive electrophiles such as acyl chlorides), 383 amines (**Set A**), and 366 aryl halides (**Set B**), were randomly selected as representative structures from MBBC (Figure 3C). Furthermore, the HTE data used comprises a small fraction (~5000 experiments) of the full factorial of 384 amines × 384 aryl halides space (n^2 space). Generating a fully factorial dataset would require ~ 160K experiments, which while achievable, is prohibitively time and resource intensive with the state-of-the-art HTE data analysis techniques. Considering this practical limitation, the ability to generate predictive models is highly desirable using a fraction of the n^2 space, such as represented by our HTE data.[16]

Generating a large, diverse HTE data set: The generation of the large data sets for modeling was conducted in two parts (**Part I** and **Part II**, Figure 3A). **Part I** involved the coupling of 384 secondary amines (4–phenyl piperidine, the substrate used for reaction optimization + **Set A**) with four aryl bromides. Since aryl halides from the Informer set were used to identify the optimal C–N coupling conditions (Figure 2), three of the four bromides used for this experiment, **X2**, **X3**, and **X13**, were also selected from that set. The fourth halide was the simplest partner aryl bromide, 6-bromoquinoline (**Br-1**).

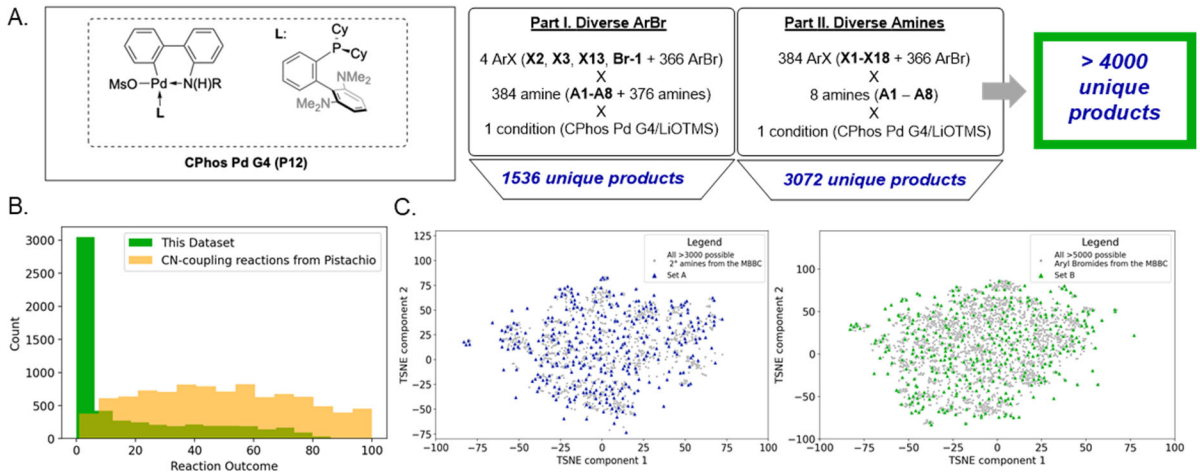


Figure 3: *Summary of constructed dataset.* (A) Workflow for generation of structurally diverse datasets. (B) Yield distributions for the generated dataset versus an isolated set of C–N coupling reactions from the Pistachio patent dataset. Reaction outcome = product % yield for Pistachio dataset and Product LCAP for HTE dataset.[38,39] (C) Aryl halide chemical space visualizations: t-distributed stochastic neighbor embedding (t-SNE) plots of MBBC Pools of suitable amines (left) and aryl halides (right) with chosen molecules marked. The plots visualize 2048–bit, radius–2 Morgan fingerprints of molecules in 2–D space.

Part II entailed the coupling of the 384 aryl halides (**Set B** + 18 Informer halides) with 8 secondary amines (**A1**-**A8**) (for amine structures see Figure 1). The amines for **Part II** were chosen using the results from **Part I** to maximize structural diversity and varied reactivity. Importantly, the chemical space covered by over half of the approved drugs intersects with the structural diversity of the products generated in **Part I** and **Part II,** affirming the relevance of the generated HTE data for applications in drug discovery (Figure S1). Furthermore, the product yield distribution from the HTE dataset is representative of realistic reaction performance with significant failures and successes versus the literature derived datasets that largely report successful reactions (Figure 3B).

Importantly, the inclusion of reaction failures is critical to build predictive models and identify opportunities for innovation to expand the accessible chemical spa

Data Quality Interrogation: As emphasized in the literature, the availability of high quality systematic data is imperative for building reliable and robust predictive models.[40] Hence, careful consideration of potential experimental sources of error was deemed important. Beyond the modeling context, the ability to draw conclusions and extract general principles from data depends on how accurately the data reflects the actual reaction performance.[41] As detailed below, three major sources of experimental error were considered: substrate quality, analytical error, and experimental reproducibility.

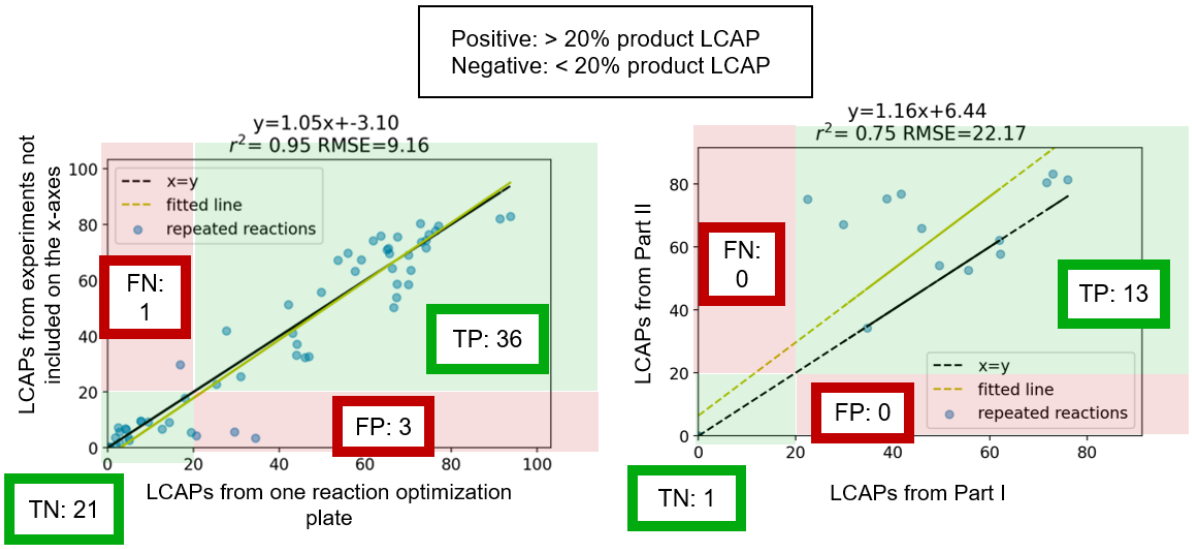


Figure 4: Regression and classification analyses for assessment of reaction reproducibility. Left Plot: reproducibility study using a subset of optimization reactions. Right plot: reproducibility study using a subset of the scope studies. TP = true positive; FP = false positive; TN = true negative; FN = false negative.

Assessment of substrate quality: Due to the large number of substrates involved in this study, it was deemed impractical to individually confirm the identity of each aryl halide and amine. Instead, proxy reactions were performed. For each of the aryl halides used in **Part I** (Figure 3A), a Suzuki–Miyaura coupling reaction was conducted. The presence of Suzuki–Miyaura coupling product was confirmed for 85% of the aryl bromides and taken to be a positive indicator of the identity of the aryl bromide suggesting the potential for the formation of the corresponding C–N coupling products. For a random subset of the amines (~113/384), an independent sample from an external vendor (Enamine) was acquired and the set of previously performed C–N coupling reactions in **Part I** was repeated using the new samples. Based on these studies 73% of products generated for this proxy study were formed in comparable LCAPs using amines from either MBBC or Enamine. The comparable reactivity was taken to be a positive indicator of amine quality obtained from MBBC. The impact of this quality assessment of aryl halides and amines is explored further in the modeling section.

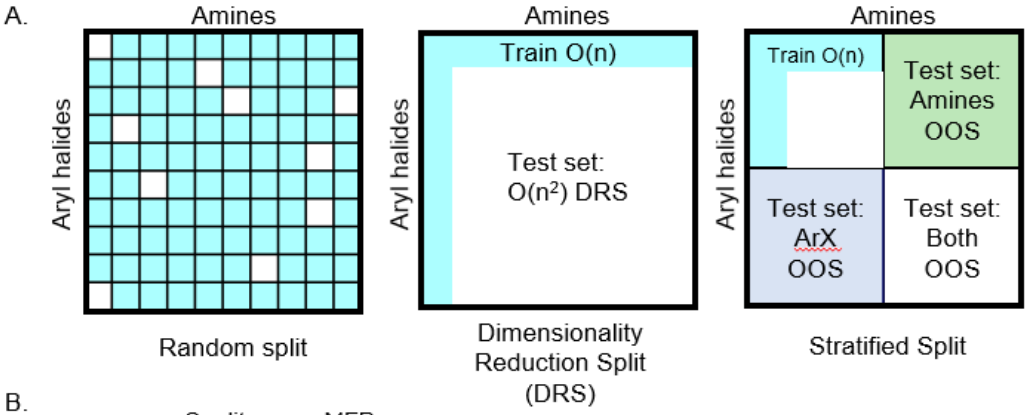
Assessment of UPLC data quality: Occasionally, overlapping peaks in the UPLC chromatograms resulted in the inability to accurately determine the reaction outcome.[42] This was factored into the overall assessment of data quality and analyzed in the modeling section.

Assessment of reaction reproducibility: Variance due to minor differences in reaction setup performed on different days and/or reaction plates was also studied. To assess this variance on each reaction plate, commonly used metrics used for evaluating ML models were calculated to compare reactions that were repeated across different reaction plates using regression and classification analysis. This analysis was done separately for a small subset of the reactions that were used for the optimization (Figure 2) and the scope studies (Figure 3A). Regression analysis entailed graphing reaction outcomes (as product LCAPs) from one reaction plate against the product LCAPs from the same reaction on another reaction plate. As shown in Figure 4, the absolute reproducibility is significantly higher for optimization reactions (left plot, $r^2 = 0.95$) versus the scope studies (right plot, $r^2 = 0.75$). This difference could be in part due to higher control over the reagent quality (pure by ^{1}H NMR analysis) for the optimization studies due to smaller number of substrates. Additionally, the number of repeats per reaction is higher for optimization studies than for the scope studies which can also influence the overall r^2. In contrast to the regression analysis, a binary classification analysis using a product LCAP threshold of 20% (> 20% or less than 20% product LCAP) is highly consistent for all reactions that were repeated in different reaction plates for both optimization and scope studies. This is reflected by the zero or low percentage of false positives/negatives for the scope and optimization studies respectively. These results suggest that the classification analysis is a more robust strategy to account for inherent minor experimental noise than the corresponding regression study. The evaluation of the three sources of experimental error detailed above were employed to determine the effect of data quality on trained models.

Modelling: We next used our structurally diverse datasets to build and evaluate predictive models using both regression and classification analysis. The overall goal of predictive models is to predict reaction failures and successes with high fidelity. Literature reports suggest that models trained on random splits of HTE data generally perform well within the modeled datasets,[11,15,27–31] an observation that is hypothesized to be a result of hidden patterns in the dataset and bias in its construction.[43] However, extending models to unseen structures is often difficult and limited by narrow substrate scopes.[44] In contrast, the generation of the large diverse dataset using ~400 aryl halides and amines allowed us to arrange the train/validation/test split in various ways to systematically interrogate the out–of–scope (OOS) extrapolative performance of the models using regression and classification models. In addition to the commonly used random splits (Figure 4A, left), two different paradigms for dividing the data were defined to probe the ability of models to extrapolate to new molecules. The first splitting strategy, the "dimensionality reduction split" (DRS), is designed to reduce the dimensionality of the space from squared order ($O(n^2)$) to linear order ($O(n)$) (Figure 5A, middle). In this scenario, the training set contains the reactions of a small set of amines against all 384 aryl halides, and vice versa, ensuring that at least one reaction of each substrate (amine and aryl halide) in the dataset is present in the training data. The test set probes the ability of the model to predict in the n^2 space. Specifically, the model predicts the reaction performance of products for which both the aryl halide and the amine components have been seen by the model but not in the combination present in the test set (See SI for details). Notably, if the modeling using a DRS is successful, this would imply that predictions for 6×10^6 aryl halide and

amine combinations could be achieved by ~2 orders of magnitude smaller training sets (~10^4 reactions) significantly reducing the experimental demands. The importance of dimensionality reduction to reduce experimental demands has also been previously stated in the context of assessing the scope and limitations of four different HTE amenable C–N coupling methods.[15]

A.



B.

Model Type	Quality Level	MFP Length	Random	DRS[1]	ArX OOS	Amine OOS	Both OOS[1]
Ridge	1	256	0.63	0.60	0.12	0.20	0.02
	1	512	0.66	0.63	0.23	0.24	0.06
	1	1024	0.66	0.63	**0.28**	0.25	0.08
Random Forest	1	256	0.62	0.52	0.18	**0.50**	**0.14**
	1	512	0.62	0.51	0.17	0.51	0.09
	1	1024	0.62	0.54	0.18	0.51	0.09
MLP	1	1024	**0.70**	**0.68**	0.27	0.32	0.11

C.

Random Forest	1	1024	0.62	0.54	0.18	0.51	0.09
	2	1024	0.66	0.56	0.18	0.56	0.06
	3	1024	0.65	0.54	0.18	0.55	0.07
	4	1024	0.66	0.59	0.19	0.52	0.04
	5	1024	0.68	0.57	0.19	0.52	0.03
	6	1024	0.67	0.45	0.18	0.46	0.06
	7	1024	0.62	0.45	0.26	0.35	0.10
	8	1024	0.63	0.50	0.24	0.36	0.11

Figure 5: (A) *Splitting strategies.* Hypothetical training data in light blue. O(n) = linear order, i.e., few amines against many halides and few halides against many amines. O(n^2) = squared order, i.e., products generated using a full factorial combination of all halides and amines. The HTE data generated in this manuscript has linear order O(n). (B) *Regression Model Outcomes.* R^2s for a selected set of regression models are shown across five (Random, DRS, ArX OOS, Amine OOS, and Both OOS) splits. The results are reported as a *N*–fold cross validation, where *N*=10 when possible. Top: Model performances (R^2) for different Morgan Fingerprint (MFP) lengths. (C) The impact of data quality on a random forest model for all splits. The explicit construction of the quality levels is discussed in the SI. [1]The results are reported as an 8–fold cross validation, see SI for complete details.

The second modelling strategy is based on a stratified split, where the training, validation, and test sets are designed to have separate sets of amines and aryl bromides. Figure 5A depicts this space as a square divided into four quadrants (rightmost square). In this scenario, the training set is

similar to the DRS split but fulfills the requisite criteria for the stratified splits (Amine OOS, ArX OOS and both OOS). Specifically, the training data is contained in one quadrant with an associated set of aryl halides and amines and there are three distinct options for the test set. The two quadrants that share one edge with the training data (Amine OOS and ArX OOS) are considered partially out of scope, as the model will have one coupling partner in its training data (e.g., ArX for Amine OOS), but not the other. For example, if a model can be trained to be predictive in the Amine OOS region, then training data for a fully predictive model would not need to cover all amines and a reduced number would need to be experimentally evaluated. This is especially important considering that specific molecules can be expensive, unavailable, or otherwise have properties that result in analytical barriers to facile data collection. The final quadrant (Both OOS) contains the datapoints where neither coupling partner has been seen by the model, which evaluates the model's ability to fully extrapolate to unseen amine *and* aryl bromide chemical space.

Figure 5B reports modeling results for the random, DRS and the stratified split strategies described above. We first pursued regression approaches to predict product LCAP directly as a continuous, numerical value. Models investigated included linear regression derivatives (Ridge), decision tree-based models (Random Forest), and multilayer perceptrons (MLPs). Two different classes of featurization were investigated, Morgan fingerprints and quantum-mechanical (QM) features. The QM features were calculated using a trained ML model.[45] A full description of the calculation of QM features and results for the one–hot encoded baseline are provided in the SI. In general, MLP and random forest models afford comparable results. Interestingly, however, regression models for Amine OOS split are superior with random forest methods. Consistent with previous work[11,15,27–31], random splitting of the data yields better models than any OOS scaffold splitting (Amine OOS, ArX OOS, Both OOS, Figure 4B, top). The performance of all three OOS cases for the stratified split was significantly inferior to the DRS strategy indicating that the model has a limited ability to extrapolate beyond molecules in its training set. In addition, models trained on the DRS reach R^2s that are identical or better than the random splitting, which supports the hypothesis that DRS is an appropriate dataset generation strategy.

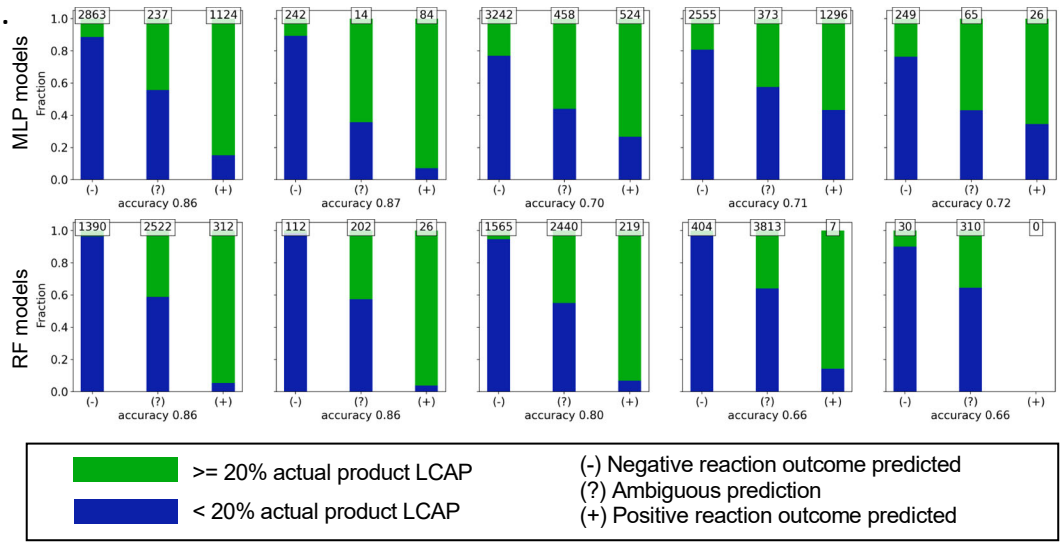
We also explored the effect of quality, or implicitly, noise, on the dataset (Figure 5C). As described above three sources of experimental noise were systematically evaluated: substrate quality, UPLC/MS data quality, and reaction reproducibility. Data quality levels were manually assigned from 1 (lowest) to 8 (highest) as described in the SI. At every quality level, reaction outcomes that exceeded a threshold for experimental noise were systematically removed. For example, the lowest quality dataset (quality 1) consisted of all generated data whereas the highest quality dataset (quality 8) consisted of only reactions that were devoid of any identified experimental noise. The resulting analysis showed comparable R^2 regardless of the data quality suggesting modest impact of the experimental noise on model training. This is consistent with a previous study of a HTE dataset of C–N coupling reactions that demonstrated minimal impact on model training by introduction of simulated noise to experimental data.[12,40] These results suggest that models using HTE data sets are largely unimpacted by minor experimental noise, thereby obviating the necessity for time intensive assessment of data quality for future efforts.[2] Instead, the time and resources should be expended on enhancing the diversity of the dataset.

[2] Studies of noise in historical reaction sets suggest that de–noising historical data does have a significant impact on model outcomes.[40,47]

A.

Model Type	Product LCAP Threshold	Fingerprint	Baseline Positive Rate	random precision @5	random precision @10%	DRS precision @5	DRS precision @10%	ArX OOS precision @5	ArX OOS precision @10%	Amine OOS precision @5	Amine OOS precision @10%	Both OOS precision @5	Both OOS precision @10%
Pytorch MLP Model	0%	MFP	54%	94(4)%	91(1)%	77(6)%	75(6)%	88(4)%	85(2)%	80(4)%	83(2)%	80(7)%	78(9)%
	10%		40%	94(2)%	90(1)%	87(4)%	95(3)%	92(4)%	86(3)%	60(9)%	71(5)%	62(5)%	41(7)%
	20%		33%	**96(2)%**	**95(1)%**	90(3)%	**100(0)%**	**92(5)%**	**88(4)%**	62(9)%	70(8)%	50(10)%	58(11)%
	0%	MFP + QM Fingerprints	54%	90(4)%	93(1)%	87(4)%	90(6)%	86(5)%	86(3)%	82(8)%	81(3)%	65(7)%	65(7)%
	10%		40%	90(4)%	93(1)%	85(5)%	91(5)%	90(4)%	85(3)%	70(9)%	74(5)%	70(7)%	70(9)%
	20%		33%	94(4)%	94(1)%	92(3)%	91(5)%	86(6)%	84(5)%	64(9)%	67(7)%	57(6)%	58(5)%
Random Forest Model	0%	MFP + QM Fingerprints	54%	98(1)%	96(0)%	95(3)%	100(0)%	94(2)%	82(3)%	88(4)%	85(2)%	75(5)%	68(7)%
	10%		40%	94(2)%	94(1)%	92(3)%	100(0)%	80(8)%	77(5)%	88(7)%	85(5)%	52(9)%	58(7)%
	20%		33%	88(5)%	90(2)%	**100(0)%**	**100(0)%**	80(7)%	74(4)%	**90(5)%**	**84(5)%**	57(8)%	58(11)%

B.



C.

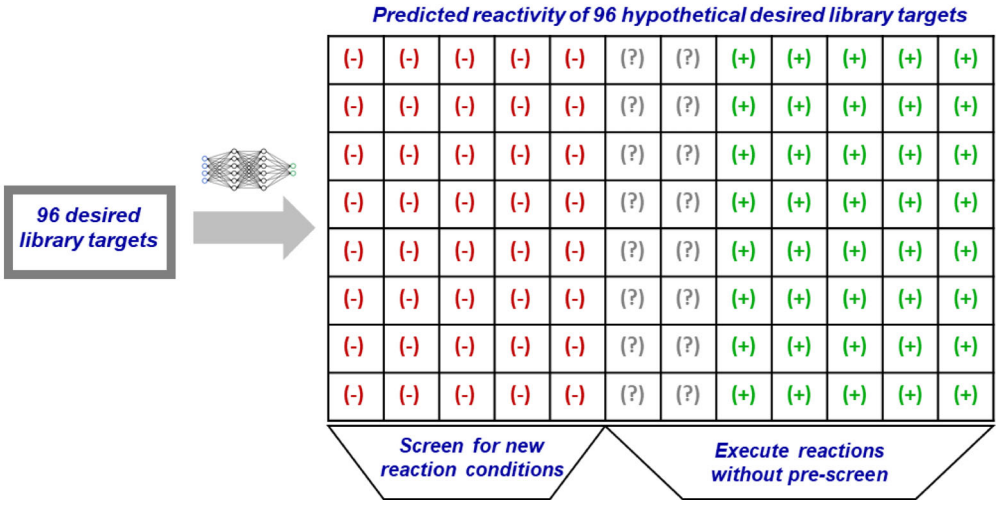


Figure 6: *Classification Results*. (A) precision@5, and precision@10% are given for two different classifiers for a number of different parameters. Results are shown for two input types (MFP and MFP + QM). The baseline positive rate in the test set of specific splits can vary 1-2%. Precision@5 and Precision@10% metrics are given in the table. The values are given as averages across the N-fold validation, with the margin of error in parentheses (standard deviation/ √# samples). The color gradient represents the difference between each value and its relevant baseline positive rate. (B) Accuracies for high confidence (highest class > 0.9, denoted as (+) or lowest class >= 0.9, denoted as (-)) are graphed in a bar chart across two splits and 5 models. The remaining model predictions are considered ambiguous because the confidence is less than 0.9 for highest or lowest class (denoted as (?)). Highest

class: >20% product LCAP and lowest class: <20% product LCAP. The numbers on the bars are the absolute number of products. (C) Illustration of a hypothetical propsective application of ML models. The symbols (+), (-) and (?) have the same definitions as for Figure 6B.

Categorical Modeling of Reaction Space: As illustrated in Figure 4 above, binary classification using a product LCAP threshold is more robust to experimental noise than regression analysis. Hence, we next modelled the dataset by treating the LCAP as a categorical variable split into two groups according to a product LCAP threshold value (0, 10% or 20% product LCAP, Figure 6A). Importantly, the resulting binary classification task enables the identification of sufficiently productive reactions (> 10% or 20% product LCAP) for medicinal chemistry campaigns.[46] Each model was evaluated based on its ability to recommend hits (reactions with a product LCAP of >= 0, 10 or 20%). The corresponding metric is precision@N, which is calculated by using the model to recommend N reactions it predicts to be hits and then evaluating the percentage of those reactions that are classified correctly in the desired product LCAP bin. For example, the precision@10% for the random split is 95% using the MLP classifier for 20% product LCAP threshold implicating that 95% of the top 10% recommended hits are actually formed with >20% product LCAP. Table 2 shows the precision@N performances (precision@5 or precision@10%) of multi-layer perceptron (MLP) and random forest models for all splitting strategies (See SI for other metrics). In addition, the baseline positive rate for all LCAP thresholds is also shown. The baseline positive rate reflects the percentage of data points in the test set that exceed the chosen LCAP threshold. While the MLP and random forest classifiers, in general afford comparable results, the model predictivity for Amine OOS split is in general higher using the random forest models. Moreover, supplementing the Morgan fingerprint (MFP with 1024 input length) with a QM fingerprint does not yield significant enhancement of precision@N in most cases. The highest calculated precision@10% for >20% product LCAP threshold exceeds the baseline PR by ~3–fold for random, DRS, ArX OOS and Amine OOS splits (numbers in bold, Figure 6A). This precision@10%:PR of ~3 could enhance the percentage of recommended hits (product LCAP > 20%) by up to 200% which is the maximum achievable enrichment. Excitingly, even when both the aryl halide and the amine are not part of the training set (Both OOS), the MLP classifier has a precision@10% accuracy exceeding the baseline by a factor of 1.75 for >20% product LCAP threshold thereby exemplifying the models' ability to extrapolate beyond the molecules contained in the training set. For completely new substrates (Both OOS), this improvement in the precision@10% accuracy could increase the percent of successful C–N couplings by ~75%.

While the metrics detailed in Figure 6A above can be used to enrich for successful reactions, they provide no information on the overall predictivity of the models. Specifically, the precision@N metric does not provide information on whether a given C–N coupling reaction will yield the product above the desired LCAP threshold. As such, an important metric for predicting the library success rates is model accuracy which is defined as the percentage of accurate predictions for the test set products being in the desired LCAP bin. To this end, Figure 6B shows the model accuracy for all splits using the MLP and Random Forest classifiers. Consistent with the regression modelling results detailed above, the accuracy of product LCAP classification is higher for random split and DRS versus the OOS splits. Using MLP models, 81–96% of the products in the test set can be binned above (+) or below 20% LCAP (-) with high confidence (>0.9) depending on the splitting strategy. The accuracy of this binning increases with increasing model accuracy. In general, random forest models are significantly inferior than the corresponding MLP models for

binning products as most products fall within the ambiguous category (**?**). The data depicted in Figure 6B is particularly important for prospective applications for which reaction condition screening efforts will largely be devoted for products predicted to form with <20% LCAP (Figure 6C).

Conclusion and Outlook. To summarize, this manuscript details the first ML models for C–N couplings using large structurally diverse pharmaceutically relevant datasets generated using nanomole scale HTE. The dataset generation was enabled by the identification and use of novel nanomole scale compatible automation friendly C–N coupling reaction conditions. The structural diversity in the dataset was achieved by leveraging MBBC. The large dataset enabled the systematic evaluation of model performance using five different data splitting strategies. These five splits were carefully designed to evaluate the model's ability to extrapolate beyond the substrates in the training set. Regression analysis generally led to low to modest R^2's which can be likely improved using active learning strategies. Classification models were also built with a lens toward application to medicinal chemistry campaigns. The precision@N accuracy exceeded the baseline PR by 25-67% depending on the splitting strategy. These results would manifest as significant enrichment of successful C–N couplings using the hits recommended by the models. In addition, the accuracy of the best models for each of the five splits ranged between 70-87% suggesting excellent overall predictivity of the models even for completely unseen substrates. Furthermore, the models enable reasonable to excellent binning of the products below or above 20% LCAP. In a prospective application, this would enable *in silico* identification of reaction performance with high fidelity thereby focusing the time–consuming screening efforts on low–yielding targets. Importantly, systematic investigation of the data quality showed that minor experimental noise does not significantly influence the model performances. Hence, the time intensive data quality evaluation studies might not be required in future efforts. Ultimately, this manuscript lays the groundwork for building predictive models for structurally diverse HTE derived datasets. Such efforts have enormous potential to accelerate and enhance resource efficiency of drug discovery efforts. Furthermore, the ML models will serve as a mechanism to identify synthetic chemistry innovation opportunities by predicting the dark space of chemical reactions.

Acknowledgements: JX and KFJ thank the Machine Learning for Pharmaceutical Discovery and Synthesis consortium for financial support. SLB is supported by National Institutes of Health under award number R35-GM122483. The authors also acknowledge Timothy Nowak and Debopreeti Mukherjee for assisting with the acquisition of analytical data for the HTE experiments. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Health.

References:

(1) Strieth-Kalthoff, F.; Sandfort, F.; Segler, M. H. S.; Glorius, F. Machine Learning the Ropes: Principles, Applications and Directions in Synthetic Chemistry. *Chem. Soc. Rev.* **2020**, *49* (17), 6154–6168. https://doi.org/10.1039/c9cs00786e.

(2) Shi, Y.; Prieto, P. L.; Zepel, T.; Grunert, S.; Hein, J. E. Automated Experimentation Powers

Data Science in Chemistry. *Acc. Chem. Res.* **2021**, *54* (3), 546–555. https://doi.org/10.1021/acs.accounts.0c00736.

(3) Shim, E.; Kammeraad, J. A.; Xu, Z.; Tewari, A.; Cernak, T.; Zimmerman, P. M. Predicting Reaction Conditions from Limited Data through Active Transfer Learning. *Chem. Sci.* **2022**, *13* (22), 6655–6668. https://doi.org/10.1039/d1sc06932b.

(4) Fu, Z.; Li, X.; Wang, Z.; Li, Z.; Liu, X.; Wu, X.; Zhao, J.; Ding, X.; Wan, X.; Zhong, F.; Wang, D.; Luo, X.; Chen, K.; Liu, H.; Wang, J.; Jiang, H.; Zheng, M.; Jiang, H.; Zheng, M. Optimizing Chemical Reaction Conditions Using Deep Learning: A Case Study for the Suzuki-Miyaura Cross-Coupling Reaction. *Org. Chem. Front.* **2020**, *7* (16), 2269–2277. https://doi.org/10.1039/d0qo00544d.

(5) Kite, S.; Hattori, T.; Murakami, Y. Estimation of Catalytic Performance by Neural Network - Product Distribution in Oxidative Dehydrogenation of Ethylbenzene. *Appl. Catal. A, Gen.* **1994**, *114* (2). https://doi.org/10.1016/0926-860X(94)80169-X.

(6) Reid, J. P.; Sigman, M. S. Holistic Prediction of Enantioselectivity in Asymmetric Catalysis. *Nature* **2019**, *571* (7765), 343–348. https://doi.org/10.1038/s41586-019-1384-z.

(7) de Almeida, A. F.; Moreira, R.; Rodrigues, T. Synthetic Organic Chemistry Driven by Artificial Intelligence. *Nat. Rev. Chem.* **2019**, *3* (10), 589–604. https://doi.org/10.1038/s41570-019-0124-0.

(8) Jiang, S.; Zhang, Z.; Zhao, H.; Li, J.; Yang, Y.; Lu, B. L.; Xia, N. When SMILES Smiles, Practicality Judgment and Yield Prediction of Chemical Reaction via Deep Chemical Language Processing. *IEEE Access* **2021**, *9*, 85071–85083. https://doi.org/10.1109/ACCESS.2021.3083838.

(9) Lu, J.; Zhang, Y. Unified Deep Learning Model for Multitask Reaction Predictions with Explanation. *J. Chem. Inf. Model.* **2022**, *62* (6), 1376–1387. https://doi.org/10.1021/acs.jcim.1c01467.

(10) Sandfort, F.; Strieth-Kalthoff, F.; Kühnemund, M.; Beecks, C.; Glorius, F. A Structure-Based Platform for Predicting Chemical Reactivity. *Chem* **2020**, *6* (6), 1379–1390. https://doi.org/10.1016/j.chempr.2020.02.017.

(11) Haywood, A. L.; Redshaw, J.; Hanson-Heine, M. W. D.; Taylor, A.; Brown, A.; Mason, A. M.; Gärtner, T.; Hirst, J. D. Kernel Methods for Predicting Yields of Chemical Reactions. *J. Chem. Inf. Model.* **2021**. https://doi.org/10.1021/acs.jcim.1c00699.

(12) Kwon, Y.; Lee, D.; Choi, Y. S.; Kang, S. Uncertainty-Aware Prediction of Chemical Reaction Yields with Graph Neural Networks. *J. Cheminform.* **2022**, *14* (1), 1–10. https://doi.org/10.1186/s13321-021-00579-z.

(13) Schwaller, P.; Laino, T. Data - Driven Learning Systems for Chemical Reaction Prediction :

An Analysis of Recent Approaches. In *ACS Symposium Series*; American Chemical Society, 2019. https://doi.org/10.1021/bk-2019-1326.ch004.

(14) Perera, D.; Tucker, J. W.; Brahmbhatt, S.; Helal, C. J.; Chong, A.; Farrell, W.; Richardson, P.; Sach, N. W. A Platform for Automated Nanomole-Scale Reaction Screening and Micromole-Scale Synthesis in Flow. *Science (80-.).* **2018**, *359* (6374), 429–434. https://doi.org/10.1126/science.aap9112.

(15) 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 (80-.).* **2018**, *360* (6385), 186–190. https://doi.org/10.1126/science.aar5169.

(16) Lin, S.; Dikler, S.; Blincoe, W. D.; Ferguson, R. D.; Sheridan, R. P.; Peng, Z.; Conway, D. V.; Zawatzky, K.; Wang, H.; Cernak, T.; Davies, I. W.; DiRocco, D. A.; Sheng, H.; Welch, C. J.; Dreher, S. D. Mapping the Dark Space of Chemical Reactions with Extended Nanomole Synthesis and MALDI-TOF MS. *Science (80-.).* **2018**, *361* (6402). https://doi.org/10.1126/science.aar6236.

(17) Santanilla, A. B.; Regalado, E. L.; Pereira, T.; Shevlin, M.; Bateman, K.; Campeau, L.; Schneeweis, J.; Berritt, S.; Shi, Z.; Nantermet, P.; Liu, Y.; Helmy, R.; Welch, C. J.; Vachal, P.; Davies, I. W.; Cernak, T.; Dreher, S. D. Nanomole-Scale High-Throughput Chemistry for the Synthesis of Complex Molecules. *Science (80-.).* **2015**, *347* (6217), 443–448. https://doi.org/10.1126/science.1259859.

(18) Gesmundo, N. J.; Sauvagnat, B.; Curran, P. J.; Richards, M. P.; Andrews, C. L.; Dandliker, P. J.; Cernak, T. Nanoscale Synthesis and Affinity Ranking. *Nature* **2018**, *557* (7704), 228–232. https://doi.org/10.1038/s41586-018-0056-8.

(19) Torres, J. A. G.; Lau, S. H.; Anchuri, P.; Stevens, J. M.; Tabora, J. E.; Li, J.; Borovika, A.; Adams, R. P.; Doyle, A. G. A Multi-Objective Active Learning Platform and Web App for Reac- Tion Optimization. *ChemRxiv. Cambridge Cambridge Open Engag.* https://doi.org/10.26434/chemrxiv-2022-cljcp.

(20) Schleinitz, J.; Langevin, M.; Smail, Y.; Wehnert, B.; Grimaud, L.; Vuilleumier, R. Machine Learning Yield Prediction from NiCOlit, a Small-Size Literature Data Set of Nickel Catalyzed C − O Couplings. *J. Am. Chem. Soc.* **2022**, *14*, 14722–14730. https://doi.org/10.1021/jacs.2c05302.

(21) Dombrowski, A. W.; Aguirre, A. L.; Shrestha, A.; Sarris, K. A.; Wang, Y. The Chosen Few: Parallel Library Reaction Methodologies for Drug Discovery. *J. Org. Chem.* **2021**, acs.joc.1c01427. https://doi.org/10.1021/ACS.JOC.1C01427.

(22) Shevlin, M. Practical High-Throughput Experimentation for Chemists. *ACS Med. Chem. Lett.* **2017**, *8* (6), 601–607. https://doi.org/10.1021/acsmedchemlett.7b00165.

(23) Ryou, S.; Maser, M. R.; Cui, A. Y.; Delano, T. J.; Yue, Y.; Reisman, S. E. Graph Neural

Networks for the Prediction of Substrate-Specific Organic Reaction Conditions. *arXiv* **2020**, No. 2018. https://doi.org/https://doi.org/10.48550/arXiv.2007.04275.

(24) Fitzner, M.; Wuitschik, G.; Koller, R. J.; Adam, J.-M.; Schindler, T.; Reymond, J.-L. What Can Reaction Databases Teach Us about Buchwald–Hartwig Cross-Couplings? *Chem. Sci.* **2020**, *11*, 13085–13093. https://doi.org/10.1039/d0sc04074f.

(25) Beker, W.; Roszak, R.; Wołos, A.; Angello, N. H.; Rathore, V.; Burke, M. D.; Grzybowski, B. A. Machine Learning May Sometimes Simply Capture Literature Popularity Trends: A Case Study of Heterocyclic Suzuki–Miyaura Coupling. *J. Am. Chem. Soc.* **2022**, *144* (11), 4819–4827. https://doi.org/10.1021/jacs.1c12005.

(26) Maser, M. R.; Cui, A. Y.; Ryou, S.; Delano, T. J.; Yue, Y.; Reisman, S. E. Multilabel Classification Models for the Prediction of Cross-Coupling Reaction Conditions. **2021**. https://doi.org/10.1021/acs.jcim.0c01234.

(27) Saebi, M.; Nan, B.; Herr, J.; Wahlers, J.; Wiest, O.; Chawla, N.; Guo, Z.; Zurański, A.; Kogej, T.; Norrby, P.-O.; Doyle, A. On the Use of Real-World Datasets for Reaction Yield Prediction. **2021**. https://doi.org/10.33774/CHEMRXIV-2021-2X06R-V3.

(28) Sato, A.; Miyao, T.; Funatsu, K. Prediction of Reaction Yield for Buchwald-Hartwig Cross-Coupling Reactions Using Deep Learning. *Mol. Inform.* **2021**, *40*, 1–12. https://doi.org/10.1002/minf.202100156.

(29) Probst, D.; Schwaller, P.; Reymond, J.-L. Reaction Classification and Yield Prediction Using the Differential Reaction Fingerprint DRFP. **2021**. https://doi.org/10.33774/CHEMRXIV-2021-MC870.

(30) Żurański, A. M.; Alvarado, J. I. M.; Shields, B. J.; Doyle, A. G. Predicting Reaction Yields via Supervised Learning. *Acc. Chem. Res.* **2021**, *54* (8), 1856–1865. https://doi.org/10.1021/ACS.ACCOUNTS.0C00770.

(31) Schwaller, P.; Vaucher, A. C.; Laino, T.; Reymond, J. L. Prediction of Chemical Reaction Yields Using Deep Learning. *Mach. Learn. Sci. Technol.* **2021**, *2* (1). https://doi.org/10.1088/2632-2153/abc81d.

(32) Schwaller, P.; Vaucher, A. C.; Laino, T.; Reymond, J. Data Augmentation Strategies to Improve Reaction Yield Predictions and Estimate Uncertainty. *Mach. Learn. Mol. Work. NeurIPS 2020.* **2020**, No. 16, 1–6.

(33) Dong, J.; Peng, L.; Yang, X.; Zhang, Z.; Zhang, P. XGBoost-Based Intelligence Yield Prediction and Reaction Factors Analysis of Amination Reaction. *J. Comput. Chem.* **2022**, *43* (4), 289–302. https://doi.org/10.1002/jcc.26791.

(34) Viet Johansson, S.; Gummesson Svensson, H.; Bjerrum, E.; Schliep, A.; Haghir Chehreghani, M.; Tyrchan, C.; Engkvist, O. Using Active Learning to Develop Machine

Learning Models for Reaction Yield Prediction. *Mol. Inform.* **2022**, *2200043*, 1–16. https://doi.org/10.1002/minf.202200043.

(35) Santanilla, A. B.; Regalado, E. L.; Pereira, T.; Shevlin, M.; Bateman, K.; Campeau, L.; Schneeweis, J.; Berritt, S.; Shi, Z.; Nantermet, P.; Liu, Y.; Helmy, R.; Welch, C. J.; Vachal, P.; Davies, I. W.; Cernak, T.; Dreher, S. D. Nanomole-Scale High-Throughput Chemistry for the Synthesis of Complex Molecules. *Science (80-.).* **2015**, *347* (6217), 443–448. https://doi.org/10.1126/science.1259203.

(36) Kutchukian, P. S.; Dropinski, J. F.; Dykstra, K. D.; Li, B.; DiRocco, D. A.; Streckfuss, E. C.; Campeau, L.-C.; Cernak, T.; Vachal, P.; Davies, I. W.; Krska, S. W.; Dreher, S. D. Chemistry Informer Libraries: A Chemoinformatics Enabled Approach to Evaluate and Advance Synthetic Methods. *Chem. Sci.* **2016**, *7* (4), 2604–2613. https://doi.org/10.1039/C5SC04751J.

(37) Dreher, S. D.; Krska, S. W. Chemistry Informer Libraries: Conception, Early Experience, and Role in the Future of Cheminformatics. *Acc. Chem. Res.* **2021**, *54* (7), 1586–1596. https://doi.org/10.1021/ACS.ACCOUNTS.0C00760.

(38) Pistachio https://www.nextmovesoftware.com/pistachio.html (accessed Jul 15, 2022).

(39) Mayfield, J.; Lagerstedt, I.; Sayle, R. Pistachio: Fantastic Reactions and How to Use Them. In *NIH Virtual Workshop on Reaction Informatics*; 2021.

(40) Strieth-Kalthoff, F.; Sandfort, F.; Kühnemund, M.; Schäfer, F. R.; Kuchen, H.; Glorius, F. Machine Learning for Chemical Reactivity: The Importance of Failed Experiments. *Angew. Chemie - Int. Ed.* **2022**. https://doi.org/10.1002/anie.202204647.

(41) Bender, A.; Schneider, N.; Segler, M.; Patrick Walters, W.; Engkvist, O.; Rodrigues, T. Evaluation Guidelines for Machine Learning Tools in the Chemical Sciences. *Nat. Rev. Chem.* **2022**, *6* (6), 428–442. https://doi.org/10.1038/s41570-022-00391-9.

(42) Grainger, R.; Whibley, S. A Perspective on the Analytical Challenges Encountered in High-Throughput Experimentation. *Org. Process Res. Dev.* **2021**, *25* (3), 354–364. https://doi.org/10.1021/acs.oprd.0c00463.

(43) Chuang, K. V.; Keiser, M. J. Comment on "Predicting Reaction Performance in C–N Cross-Coupling Using Machine Learning." *Science (80-.).* **2018**, *362* (6416), 1–3. https://doi.org/10.1126/science.aat8603.

(44) Zahrt, A. F.; Henle, J. J.; Denmark, S. E. Cautionary Guidelines for Machine Learning Studies with Combinatorial Datasets. *ACS Comb. Sci.* **2020**, *22* (11), 586–591. https://doi.org/10.1021/ACSCOMBSCI.0C00118.

(45) Guan, Y.; Coley, C. W.; Wu, H.; Ranasinghe, D.; Heid, E.; Struble, T. J.; Pattanaik, L.; Green, W. H.; Jensen, K. F. Regio-Selectivity Prediction with a Machine-Learned Reaction

Representation and on-the-Fly Quantum Mechanical Descriptors. *Chem. Sci.* **2021**, *12* (6), 2198–2208. https://doi.org/10.1039/d0sc04823b.

(46) Qi, N.; Wismer, M. K.; Conway, D. V.; Krska, S. W.; Dreher, S. D.; Lin, S. Development of a High Intensity Parallel Photoreactor for High Throughput Screening. *React. Chem. Eng.* **2022**, *7* (2), 354–360. https://doi.org/10.1039/d1re00317h.

(47) Toniato, A.; Schwaller, P.; Cardinale, A.; Laino, T. Unassisted Noise Reduction of Chemical Reaction Data Sets. 1–30. https://doi.org/10.26434/chemrxiv.12395120.v2.