A Multi-Objective Active Learning Platform and Web App for Reaction Optimization

Jose A. Garrido Torres[†], Sii Hong Lau[†], Pranay Anchuri[†], Jason M. Stevens, Jose E. Tabora, Jun Li, Alina Borovika, Ryan P. Adams, and Abigail G. Doyle*

Department of Chemistry, Princeton University, Princeton, New Jersey 08544, United States

Department of Chemistry & Biochemistry, University of California, Los Angeles, California 90095, United States

Chemical Process Development, Bristol Myers Squibb, New Brunswick, New Jersey 08901, United States

Center of Information Technology Policy, Princeton University, Princeton, New Jersey 08544, United States

Department of Computer Science, Princeton University, Princeton, New Jersey 08544, United States

Supporting Information Placeholder

ABSTRACT: We report the development of an open-source Experimental Design via Bayesian Optimization platform for multi-objective reaction optimization. Using high-throughput experimentation (HTE) and virtual screening datasets containing high-dimensional continuous and discrete variables, we optimized the performance of the platform by fine-tuning the algorithm components such as reaction encodings, surrogate model parameters and initialization techniques. Having established the framework, we applied the optimizer to real-word test scenarios for the simultaneous optimization of reaction yield and enantioselectivity in a Ni/photoredox-catalyzed enantioselective cross-electrophile coupling of styrene oxide with two different aryl iodide substrates. Starting with no previous experimental data, the Bayesian optimizer identified reaction conditions that surpassed the previously human-driven optimization campaigns within 15 and 24 experiments, for each substrate, among 1,728 possible configurations available in each optimization. To make the platform more accessible to non-experts, we developed a Graphical User Interface (GUI) that can be accessed online through a web-based application and incorporated features such as conditions modification on-the-fly and data visualization. This web-application does not require software installation, removing any programming barrier to use the platform, which enables chemists to integrate Bayesian optimization routines into their everyday laboratory practices.

INTRODUCTION

Reaction optimization is essential to synthetic chemistry. Typically, an optimization campaign requires the exploration of reaction conditions consisting of multiple categorical and continuous reaction variables, such as catalyst, additive, solvent, temperature, etc. In a synthetic chemistry laboratory, a common optimization strategy involves searching the literature for similar reactions to select components that are anticipated to give a higher chance of success, testing one factor/variable at a time (OFAT or OVAT) to isolate the effect of a single component, and studying the structure-activity relationship to predict better conditions. This approach has served chemists well for reaction optimization, but it neglects interactions between variables which are essential in searching for the global optimum.

Another viable strategy to determine the optimal conditions is to evaluate all possible combinations of the search space. For example, recent advances in high-throughput experimentation (HTE) have allowed chemists to rapidly screen up to thousands of reactions in parallel.[1,2] However, the number of possible reaction condition configurations scales exponentially as reaction variables vary from tens to thousands of components. As a result, given limited time and material resources, evaluating the entire condition space is often inefficient from an economic and environmental standpoint.

The simultaneous improvement of multiple reaction objectives adds another layer of complexity to the existing multidimensional challenge in reaction optimization.[3] In fact, many optimization problems in chemistry, both in academia and the chemical industry, require simultaneous optimization of two or more reaction objectives.[4] Examples of these objectives are yield, selectivity (regio-, site-, enantio-, chemo-), cost, environmental sustainability, and properties of products. An example of a multi-objective optimization in chemistry is shown in Figure 1A.[5] In many cases, there is no single solution to multi-objective optimizations such as this one. Instead, locating a set of non-dominated optimal conditions, or the Pareto front, requires balancing the trade-offs in the objectives.[6] In other words, the improvement of one objective is sometimes only possible at the expense of other objectives, which makes the identification of global maxima in a condition search space much more challenging.

In the past decade, data science and machine learning methods have been applied to address numerous

challenges in synthetic chemistry, such as multi-step synthetic planning,[7–9] prediction of reaction outcomes,[10,11] automated synthesis,[12–14] and drug design and discovery.[15,16] There have also been important advances in applying machine learning methods to reaction optimization,[17,18] building off of data science tools such as partial or full factorial design of experiments (DOE).[19–21,22] Recently, our group developed EDBO (Experimental Design via Bayesian Optimization), a platform for Bayesian reaction optimization for chemical synthesis (Figure 1B).[18] Bayesian optimization (BO) is a global optimization algorithm that can interpolate response surfaces by evaluating only a small subset of total possible combinations, thus minimizing requirements to generate a large number of experimental observations.[23,24]

However, EDBO can only perform single-objective optimization and limited effort thus far has been reported for the application of active learning strategies like BO to the simultaneous optimization of multiple objectives in synthetic chemistry. [25–27] Aspuru-Guzik and coworkers developed Chimera[28] and Gryffin,[29] packages for multi-objective optimization that combine the concepts of *a priori* scalarizing with lexicographic approaches. The same group, in collaboration with Hein, Sigman, and Merck, later demonstrated its utility in an autonomous process optimization of a stereoselective Suzuki-Miyaura coupling.[30] The group of Jensen and Jamison also applied multi-objective BO to a computer-proposed multistep synthesis of the small molecule sonidegib on an automated robotic flow platform.[31] However, these tool are less accessible to non-experts and lack valuable functionality such as the ability to visualize output predictions and modify condition space during the course of an optimization campaign. Recently, the Vlacho group developed NEXTorch,[32] a toolkit that implements BO routines through PyTorch.[33] However, its application in multi-objective optimization was only demonstrated using a search space consisting of continuous variables.

These important advances notwithstanding, for these tools to be integrated with the current synthetic chemistry practices it is essential to develop machine learning surrogate models that are not only tuned, validated, and tested on synthetic experimental chemistry data, but also provide improved accessibility and functionality tailored to reaction optimization. For example, enhancements related to augmentation of the condition space on-the-fly (adding or removing reaction condition configurations), data

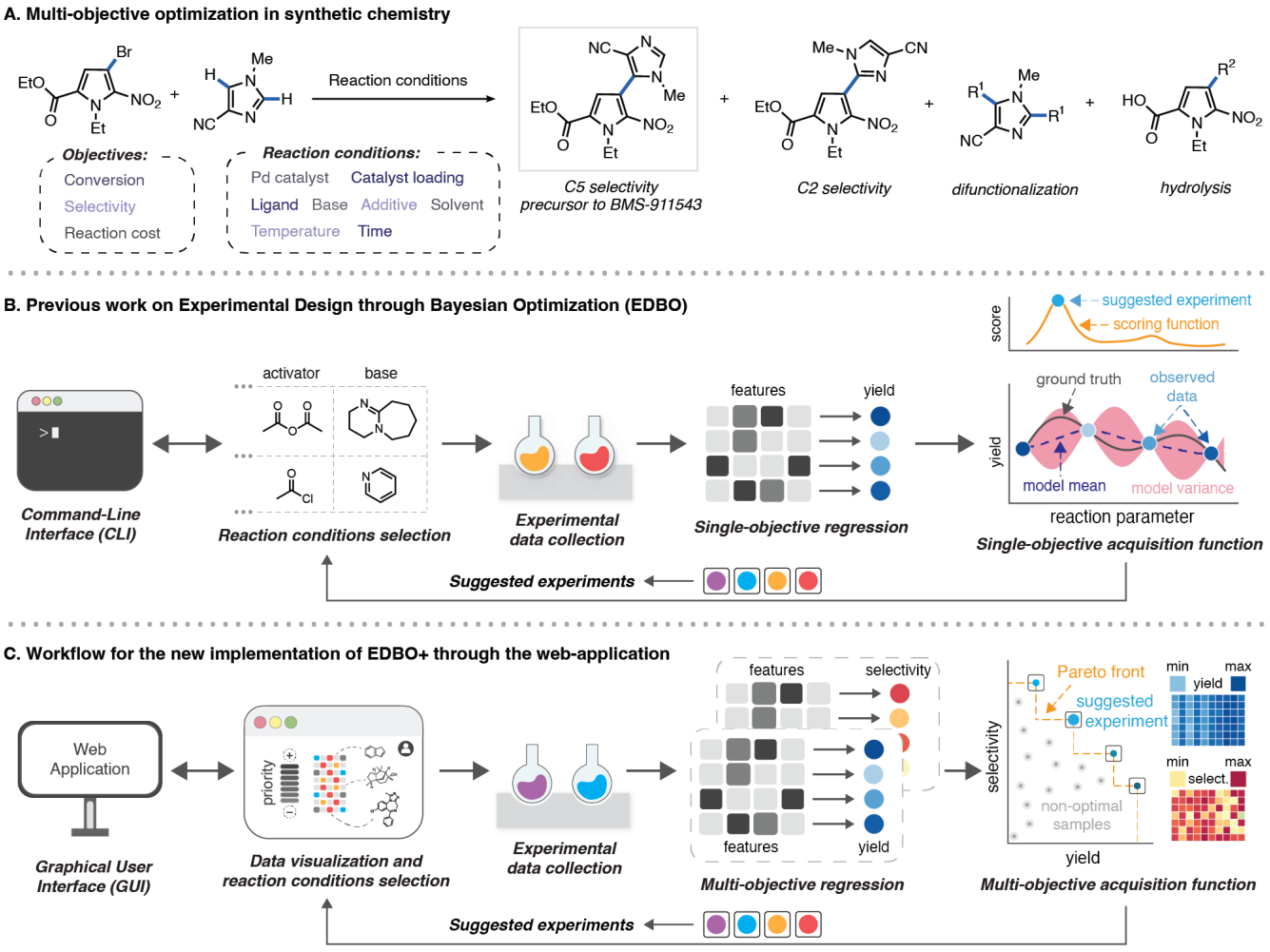


Figure 1. (A) Example of a multi-objective optimization problem in chemistry. R[1] = pyrrole fragment, R[2] = imidazole fragment, or Br,[5] (B) Previous workflow: single-objective Experimental Design via Bayesian Optimization (EDBO). (C) Current workflow: multi-objective reaction optimization framework using EDBO+ through its web application.

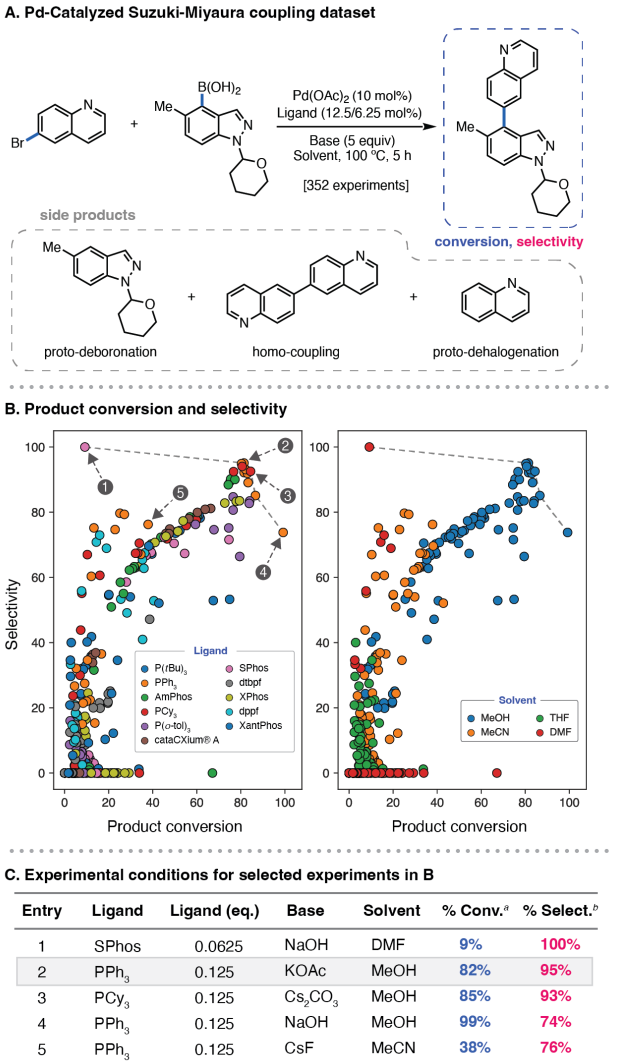
visualization and access to the predictive estimates of the surrogate models can enable the adoption of Bayesian tools in chemistry. Furthermore, the requirement of prior coding knowledge is a major obstacle for most synthetic chemists to apply BO in their day-to-day laboratory activities.

Herein, we report EDBO+, an open-source multi-objective active-learning platform based on Bayesian theory and its accompanying web application (https://www.edbowebapp.com/) (Figure 1C). Several features have been incorporated into EDBO+ including the ability to modify the reaction conditions space during an optimization campaign and the inclusion of visualizations of model predictions and uncertainties. The online platform can be accessed through a web browser, removing a requirement for any software installation, which would allow users with limited programming experience to adopt single- and multi-objective BO. In this work, we use HTE and virtual screening datasets to optimize the performance of EDBO+ by fine-tuning the algorithm components such as reaction encodings, surrogate model parameters and initialization techniques. We then apply EDBO+ to a real-word test case – a Ni/photoredox-catalyzed enantioselective cross-electrophile coupling of styrene oxides with two different aryl iodide substrates.

RESULT AND DISCUSSIONS

General Workflow. The general workflow for EDBO+ begins with input from the synthetic chemist on identifying (a) the reaction conditions space (e.g., catalysts, temperatures and concentrations) that will be explored in the optimization campaign, (b) the featurization for categorical variables (i.e., mathematical representation of the reaction components), (c) the objectives and accompanying thresholds to be optimized, and (d) the number of experiments to be evaluated in parallel per round (batch size). This initial search space can be modified at any stage of the optimization (expanding or reducing the number of components to consider). Once these are defined, the algorithm will suggest an initial set of experimental conditions (following an initialization method, see Optimizer Development section). After completing the suggested experiments in the laboratory, the chemist introduces the outputs of these experiments (e.g., yields and selectivities) back into the platform. EDBO+ builds a regression model using the experimental data and predicts the target objectives for all the remaining untested conditions included in the reaction condition space. Next, an acquisition function ranks the untested conditions based on model predictions and recommends the next set of conditions for experimental evaluation to close the active-learning cycle. Iterations of the active learning cycle will increase the accuracy of the regression predictions by providing the algorithm with more experimental observations, ultimately improving the predictions of the surrogate model. This workflow can be executed through either a command-line interface or a web-based application for single- and multi- objective optimizations.

Optimizer Development. To optimize the performance of EDBO+ (e.g., initialization methods, featurization techniques, and acquisition function), we selected two high-dimensional screening datasets: (a) Pd-catalyzed Suzuki-Miyaura coupling,[34] and (b) Pd-catalyzed C–H arylation[18] as ground truth. The condition space for these two datasets consists of a combination of continuous (e.g., temperature and concentration) and categorical variables (e.g., solvent, base and ligand). The Pd-catalyzed Suzuki-Miyaura cross coupling[35,36] dataset involves the reaction of an indazole-containing boronic acid and 6-bromoquinoline, in which the objectives are to maximize the conversion and selectivity simultaneously (Figure 2A).[34] Heteroaromatic biaryls are attractive scaffolds due to their prevalence in bioactive molecules[37,38] but their preparation via cross coupling is often accompanied by homocoupling, protodeboronation, and protodehalogenation, as captured in the selectivity objective.[39–41] This dataset consists of 352 datapoints, including 11 ligands, 4 solvents, and 8 bases.



C. Experimental conditions for selected experiments in B

Entry	Ligand	Ligand (eq.)	Base	Solvent	% Conv.[a]	% Select.[b]
1	SPhos	0.0625	NaOH	DMF	9%	100%
2	PPh_3	0.125	KOAc	MeOH	82%	95%
3	PCy_3	0.125	Cs_2CO_3	MeOH	85%	93%
4	PPh_3	0.125	NaOH	MeOH	99%	74%
5	PPh_3	0.125	CsF	MeCN	38%	76%

Figure 2. Overview of the Pd-catalyzed Suzuki-Miyaura coupling dataset. (A) Schematic representation of the reaction and its components along with the desired and side products. [a]conversion = (total product)/(total product + remaining starting material)*100%, [b]selectivity = (desired product)/(total products)*100% (B) Ground truth scatter plots for the two objectives in this reaction (product conversion and selectivity) color-coded by (left) ligand and (right) solvent. The dashed gray lines show the connections for the set of 'non-inferior' solutions in the objective space (Pareto optimal solutions). (C) Experimental conditions for labeled experiments in B.

The second HTE dataset consists of 1,728 total conditions (12 ligands, 4 solvents, 4 bases, 3 temperatures, and 3 concentrations) for the Pd-catalyzed C–H arylation of N1-methyl-1H-imidazole-4-carbonitrile and 1-bromo-2-fluorobenzene (see Ref. 18). In this case, we set the optimization goal to be finding reaction conditions that maximize reaction yield while minimizing the overall cost of the reaction. To extend the range of applicability, we also tested the performance of EDBO+ against a virtual-experimentation dataset built for nucleophilic substitution reactions[42] which exclusively contains continuous variables (see SI).

Using all three datasets, we found that optimal optimization performance can be achieved using a Gaussian process surrogate model and q-Expected HyperVolume Improvement (q-EHVI) as the acquisition function (See SI).[43] q-EHVI has been shown to maximize hypervolume of predicted experimental outputs with respect to the Pareto, and is intrinsically formulated to be efficient for batch sampling. Independent of the featurization methods used, q-EHVI is found to be optimal when compared to other common acquisition functions such as upper confidence bound (UCB) and ε-greedy (see SI). It requires fewer experiments to find the optimal values and achieves the highest rate of hypervolume expansion at the end of the optimization campaign. The hypervolume indicator is one of the most used set-quality indicators in multi-objective optimization problems since it allows evaluation of the performance of optimizers by considering the diversity, spread and proximity of the collected experimental values to the Pareto front.

Next, we compared the performance of EDBO+ for the Suzuki-Miyaura dataset using different featurization methods: (a) One-Hot Encoding (OHE) which creates a new variable for each categorical feature, (b) quantum mechanics-based features from Density Functional Theory (DFT) calculations, and (c) chemical informatics-based features using Mordred featurization[44]. To visualize the distribution of the objective values for this reaction, we color-coded the datapoints in Figure 2B according to the two categorical variables in this dataset: ligands (left panel) and solvents (right panel). Interestingly, we observe that no single ligand dominates the Pareto front (see Figure 2B, C). From an algorithm design standpoint, this allows us to test the performance of EDBO+ on data that can be represented either as discrete or continuous depending on the featurization. On the other hand, methanol (MeOH) appeared to populate the Pareto front as the optimal solvent for this transformation.

For each of the three featurization methods, we completed five optimization campaigns starting from different initial experimental conditions. First, we analyzed the distribution of conversion and selectivity values at each step of the optimization campaigns (Fig. 3A). The left panels in Fig. 3A show the evolution of the objective values in each of the five optimization runs and the right panels indicate the density of the objective values after completing these campaigns (after 30 experiments). The density plots obtained using the random sampling (Fig. 3A, in green) show that, in absence of a predictive model, there is a high probability of finding low yield and selectivity values in this dataset. In contrast, the probability of obtaining optimal conditions (with higher yield and selectivity) is increased when using EDBO+ and DFT featurization (see blue density plots in Fig. 3A). We observe this trend for all three featurization methods and in all three datasets (see SI).

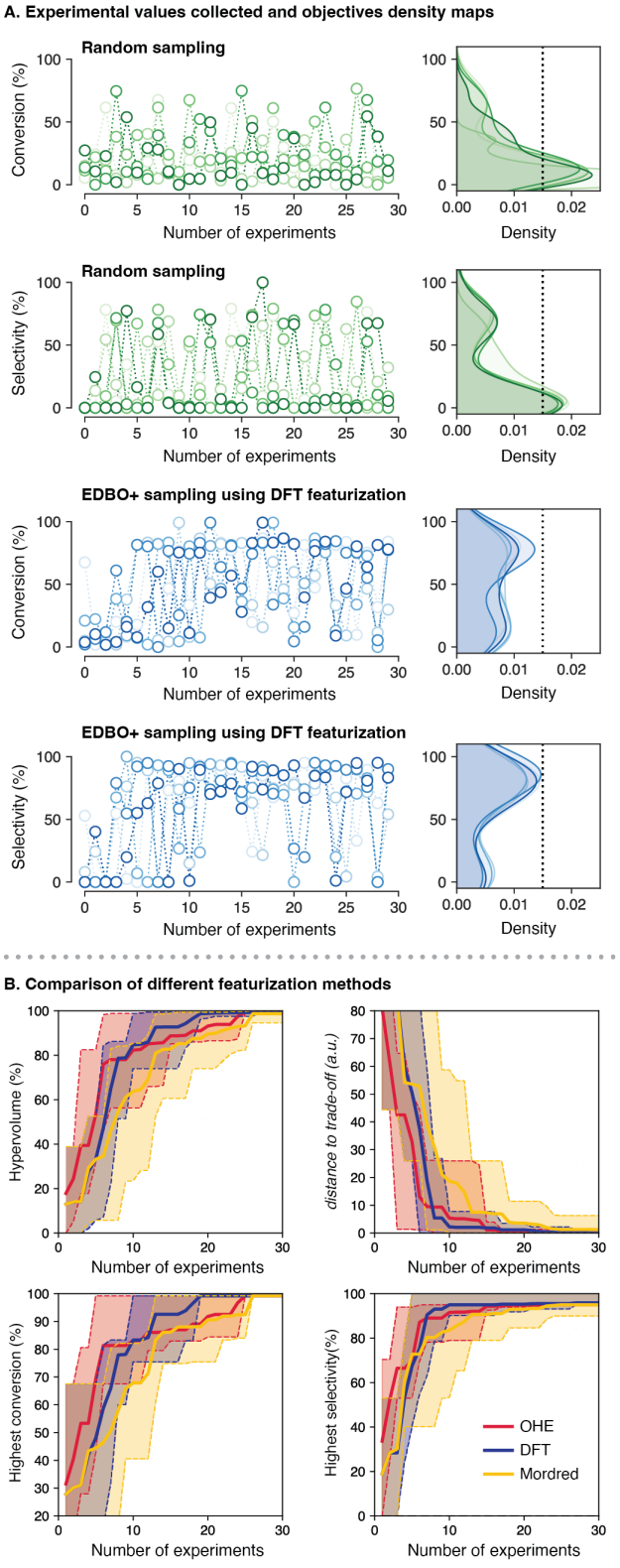


Figure 3. Optimizer performance as a function of the featurization method. (A) Conversion and selectivity values at each step of the optimization campaigns when using DFT featurization (in blue) and random sampling (in green) are shown in the left panels while the right panels show their corresponding distribution of conversion and selectivity over the 30 experiments collected for each run. Different color shades are used to distinguish the five different optimization campaigns. (B) Normalized hypervolume, minimum distance to trade-off experimental values, highest conversion and selectivity as a function of the collected experimental values, averaged over 5 runs with seeded initialization. The solid lines

indicate the average, and the shaded areas represent the upper and lower values at each stage of the optimization campaign.

In order to obtain a deeper understanding of the algorithm's performance when using different featurization methods, we measured the hypervolume covered by the collected experimental values at each optimization step (Figure 3B). In addition, we tracked the minimum distance from any collected experimental output to the high-tradeoff experimental value (in the knee region of the Pareto front, see Ref. 45) as well as the maximum values for conversion and selectivity collected at each step of the optimization. We found that DFT-encoded features provide slightly improved performance over other featurization methods, suggesting experimental conditions with optimal conversion and selectivity values (above 90%) in earlier stages compared to the optimizations using OHE and Mordred featurization. This is consistent with the single-objective optimization results previously obtained with EDBO.[18] We also note that the DFT featurization displays the lowest variance (difference between the upper and lower bounds at each step, highlighted by the shaded regions in Fig. 3B) showing its robustness against the selection of the initial experiments.[46]

Another important consideration in the success of an optimization is the choice of the initial conditions to start the optimization campaign. We illustrate the impact of the initialization method using the Pd-catalyzed C–H arylation dataset (see Ref. 18). The values for yield and cost for this HTE dataset are presented in Figure 4A. We tested the performance of the algorithm when the optimization campaigns are initialized using the Centroidal Voronoi Tessellation (CVT), Latin Hypercube sampling (LHS) and random sampling methods. We assessed the performance of the different methods and batch sizes using the dominated hypervolume metric (Figure 4B). On average, the LHS and CVT methods display a higher rate of hypervolume expansion than the random sampling method. In particular, the highest hypervolume value and lowest Mean Absolute Error (MAE) are achieved when using the CVT method and a batch size of three experiments per round (Figure 4B, C).

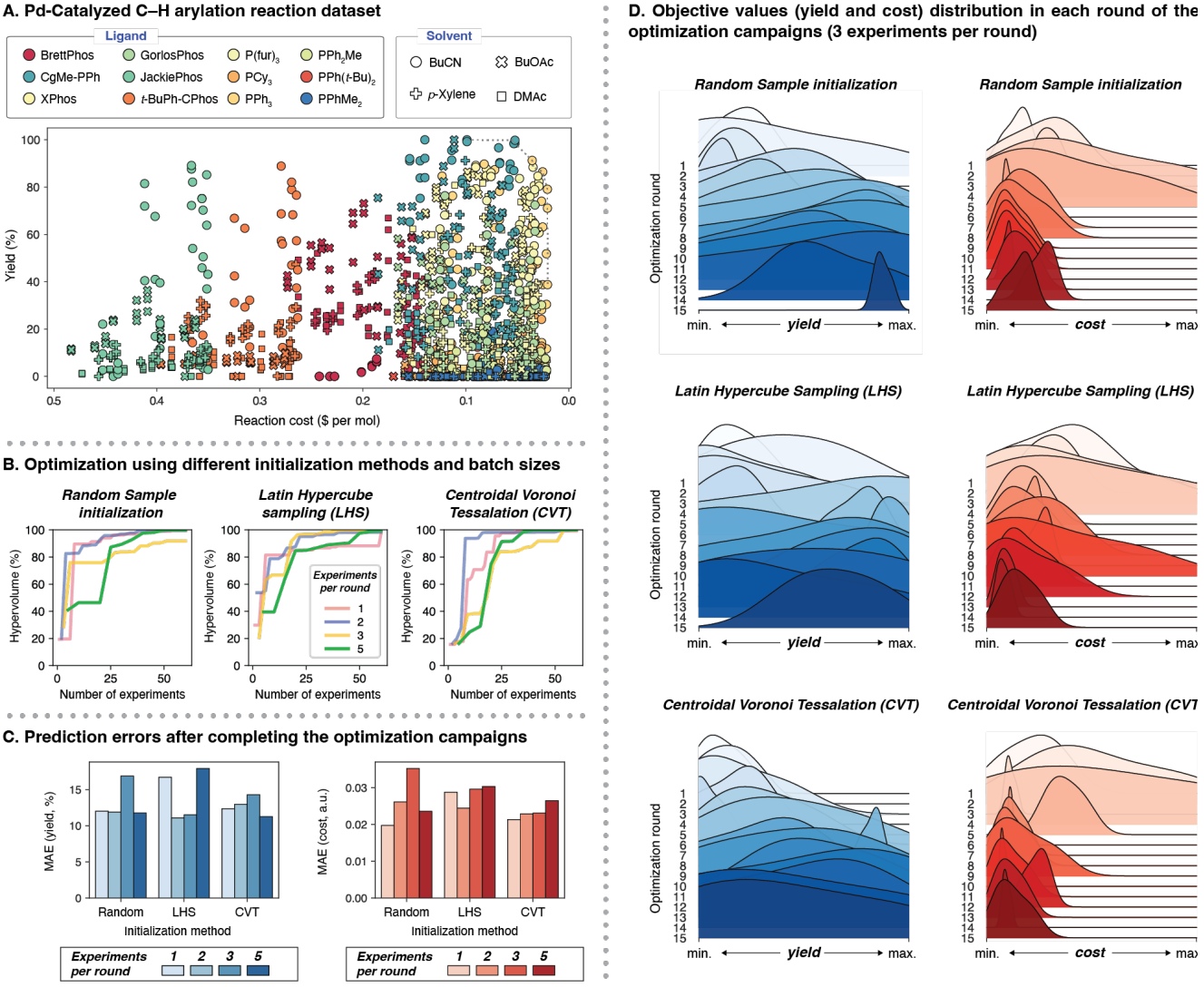
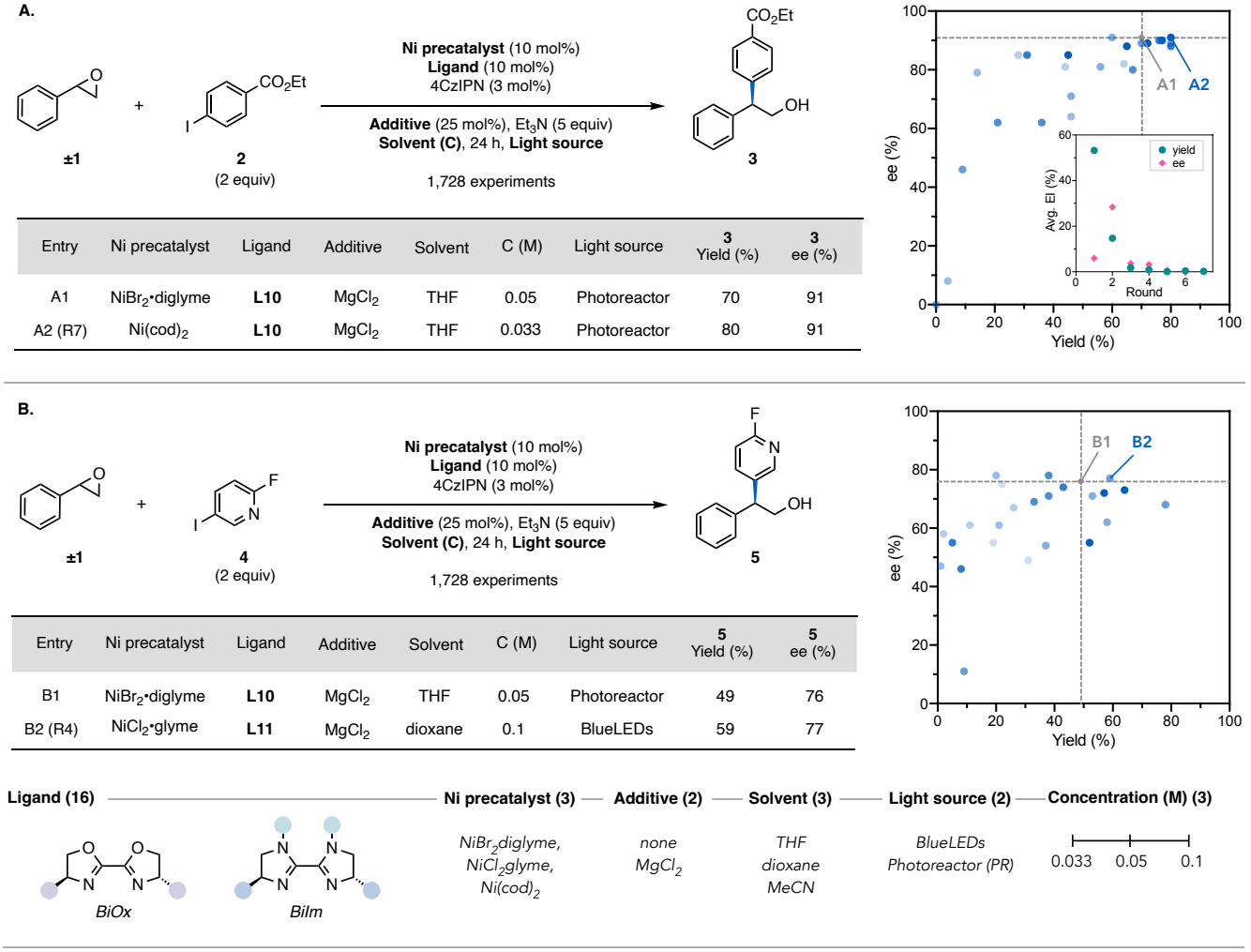


Figure 4. Model performance for the Pd-catalyzed C–H arylation dataset. (A) Overview of the objectives (yield and cost) values, the dashed lines highlight the Pareto front. The different ligands are color-coded while and different symbols are used to distinguish between solvents. (B) Hypervolume covered by the experimental values collected at each stage of the optimization campaign when using different initialization methods and batch sizes. (C) Mean Absolute Error (MAE) for the different initialization methods and batch sizes. (D) the distribution of yield (in blue) and cost (in red) values at each optimization step when initializing the optimizations using the different sampling methods.

In Figure 4D, we show the distribution of the yield and cost values of the experimental conditions for the different initialization methods using three experiments per round. A similar sampling pattern is found for all initialization methods: (1) an exploratory phase in the first rounds of the optimizations, collecting a wide range of objective values, followed by (2) exploitation behavior, with a narrow distribution of objective values closer to the optimal regions (see Figure 4D). This indicates that the algorithm can suggest optimal values starting from a variety of initial experiments, showing that the combination of the qEHVI acquisition function with the GPR hyperparameters provide a good balance between exploration and exploitation.

Application of EDBO+. Having established an optimized framework for EDBO+ on the HTE datasets, we sought to apply EDBO+ to a real-world test case for the simultaneous optimization of multiple objectives. Recently, our lab developed an enantioselective cross-electrophile coupling of styrene oxides and aryl iodides via the merger of nickel and photoredox catalysis.[47] This transformation generates enantioenriched 2,2-diarylalcohols which could be readily derivatized into chiral 1,1-diarylalkanes, an important medicinally relevant motif found in pharmaceuticals such as tolterodine, sertraline, and podophyllotoxins.[48–50] This reaction presented an ideal test case of EDBO+ for the optimization of both yield and enantioselectivity simultaneously as a yield-ee tradeoff presented a hurdle in our previous optimization campaign. In fact, the tradeoff between yield and stereoselectivity has been a longstanding challenge in enantioselective reactions, yet the two objectives must be optimized concertedly. In this study, we selected two examples to evaluate: the first example involves the model substrate, styrene oxide **1** and 4-iodobenzoate **2**, and the second is with a challenging heteroaryl iodide, 2-fluoro-5-iodopyridine **4**, from the scope studies. The reaction conditions space that we selected comprised 3 nickel precatalysts, 16 bioxazoline and biimidazoline ligands, 2 additives, 3 solvents, 3 concentrations, and 2 light source to give a total space of 1,728 possible configurations.



Entry	Ni precatalyst	Ligand	Additive	Solvent	C (M)	Light source	3 Yield (%)	3 ee (%)
A1	$NiBr_2$•diglyme	**L10**	$MgCl_2$	THF	0.05	Photoreactor	70	91
A2 (R7)	$Ni(cod)_2$	**L10**	$MgCl_2$	THF	0.033	Photoreactor	80	91

Entry	Ni precatalyst	Ligand	Additive	Solvent	C (M)	Light source	5 Yield (%)	5 ee (%)
B1	$NiBr_2$•diglyme	**L10**	$MgCl_2$	THF	0.05	Photoreactor	49	76
B2 (R4)	$NiCl_2$•glyme	**L11**	$MgCl_2$	dioxane	0.1	BlueLEDs	59	77

Figure 5. Applications of EDBO+: Ni/photoredox-catalyzed enantioselective cross-electrophile coupling of styrene oxides and aryl iodides. DFT featurization for ligand and OHE for other variables, CVT initialization and three experiments per round. Gray spots show datapoints collected using previously optimized condition, and the shades of the blue spots show the progress of the optimization (darker spots represents datapoints collected later in the campaign). The inset plot in A shows the average expected improvement values for yield and ee at each round of the optimization.

We carried out multi-objective Bayesian optimization using DFT encoded features for the ligands, running three experiments in parallel per batch, with initial experiments selected using CVT initialization. The optimizer surpassed the benchmark result within 7 rounds of optimization (24 reactions), affording an improved yield of 80% at the same enantioselectivity (91% ee, Figure 5A). In comparison, the previously reported conditions for the synthesis of **3** were identified via a one-factor-at-a-time (OFAT) method and afforded 63% yield and 91% ee after roughly 500 experiments. However, it is important to note that this comparison between the number of experiments to obtain the optimal result does not take into consideration that the optimal ligand **L10** was not available during the earliest phases of our human-driven optimization campaign. Nevertheless, this example showcases the potential of EDBO+ to identify conditions close to or at the Pareto front and outperformed the previously human-drive optimization campaign by evaluating only a small subset of the total possible configurations.

In reaction discovery, the optimal conditions identified for the model substrate are often applied to a broad range of substrates to evaluate the generality of the method. However, the optimal conditions for one substrate do not always translate to more complex or different variants. In our previous study, the conditions optimized for the model reaction to generate **3** afforded 47% yield and 75% ee for the coupling between styrene oxide **1** and pyridyl iodide **4**.[49] Without pretraining EDBO+ with prior experimental data, we optimized the reaction of **2** within the same conditions space. We found that within 4 rounds of optimization (15 reactions), EDBO+ identified conditions that afforded higher yield and enantioselectivity (59% yield, 77% ee, Figure 5B). These conditions are unique in that they feature a different ligand (biimidazolines **L10** and **L11** feature the same isopropyl substituents but vary in the aniline moiety), solvent, nickel precatalyst, solvent, concentration, and light source when compared to the previously optimized condition. This presented a case where Bayesian optimization learned about interactions between variables that would not typically be identified in a OFAT optimization campaign.

Optimizer features and user interface. Given the potential utility of this multi-objective optimization tool for reaction development efforts, we wanted to make the algorithm more accessible to practicing synthetic chemists. To this end, we developed EDBOApp (www.edbowebapp.com), a web application supported by a cloud-computing platform. No prior programming or coding experience is required to use the web application.

We also incorporated a number of functions into the workflow to make EDBO+ amenable to human-in-the-loop intervention and decision-making. First, the ability to modify the condition space during an optimization campaign allows users to alter the search space by either adding or removing reaction components or dimensions. Second, we added a data visualization tool that shows the objective predictions and uncertainties across all conditions throughout the optimization. This function enables chemists to track the expected improvement (EI) of the target objectives at any stage of the optimization and informs when to terminate the optimization campaign. For instance, the small average EI of yield and ee (~1%) toward the end of the optimization for the Ni/photoredox coupling with styrene oxide **1** and aryl iodide **2** indicates significant diminishing return to performing additional experiments (See Figure 5a inset).

To improve the functionality and adaptability of the framework, we also incorporated the ability to select different batch sizes based on constraints in experimental set up and accessibility of material resources. Thresholds can be applied to the objectives to prioritize one reaction objective over the others or to focus on specific regions of the pareto front. Finally, previous experimental data can be imported into EDBO+ to pretrain the surrogate model, giving the user a head start in the optimization process. These features, available in the EDBO+ package via command-line or graphic user interface, are intended to provide flexibility as each individual or process has distinct requirements.

CONCLUSIONS

We report the development of EDBO+, an open-source multi-objective optimization platform and an accompanying web application that allows chemists to apply Bayesian optimization methods into everyday synthetic chemistry practices. The framework relies on building a surrogate machine learning model by combining the predictive estimates with acquisition functions that balance the exploration/exploitation trade-off of single- and multi-objective optimizations. EDBO+ was tested on a selection of datasets that include both categorical and continuous reaction dimensions to identify surrogate model configurations that could be broadly applicable to optimization problems in synthetic chemistry. In a real-world test case of a Ni/photoredox-catalyzed enantioselective cross-electrophile coupling of styrene oxides with aryl iodides, the optimizer identified conditions that surpassed the originally reported conditions within 15 and 24 experiments (for two different aryl iodide substrates) among a total of 1,728 possible conditions. Further investigations will focus on exploring the use of recommender systems for the expansion of the reaction condition space and its application in autonomous process optimization.

ASSOCIATED CONTENT

Code availability and implementation.
The command-line interface of EDBO+ used to create and optimize reaction conditions presented in this work, along with the scripts to analyze the performance of the optimizer, are available in the following GitHub repository: https://github.com/doyle-lab-ucla/edboplus.

Web application.
We developed EDBOWebApp, a web application that makes the Bayesian optimizer more accessible to users with limited knowledge of programming languages. This web application can be accessed in https://www.edbowebapp.com/ through a web-browser. The back-end is supported by a cloud-computing platform to perform the computations required for creating and optimizing reaction conditions. This makes our routines accessible through any device that

enables web browsing without having to install any software or packages.

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website.

Experimental details, optimization studies and characterization data (PDF)
R Source Code (TXT)

AUTHOR INFORMATION

Corresponding Author

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

Authors

Jose A. Garrido Torres – *Department of Computer Science, Princeton University, Princeton, New Jersey 08544, United States*; orcid.org/0000-0002-1727-0862

Sii Hong Lau – *Department of Chemistry, Princeton University, Princeton, New Jersey 08544, United States; Department of Chemistry & Biochemistry, University of California, Los Angeles, California 90095, United States.*

Pranay Anchuri – *Center for Information Technology Policy, Princeton University, Princeton, New Jersey 08544, United States; orcid.org/0000-0003-4377-5036*

Jason M. Stevens – *Chemical Process Development, Bristol Myers Squibb, 556 Morris Ave, Summit, NJ 07901, United States; orcid.org/0000-0003-1671-1539*

Jose E. Tabora – *Chemical Process Development, Bristol Myers Squibb, New Brunswick, New Jersey 08903, United States.*

Jun Li – *Chemical Process Development, Bristol Myers Squibb, 1 Squibb Drive, New Brunswick, NJ 08903, United States; orcid.org/0000-0002-0594-7143*

Alina Borovika – *Chemical Process Development, Bristol Myers Squibb, 1 Squibb Drive, New Brunswick, NJ 08903, United States.*

Ryan P. Adams – *Department of Computer Science, Princeton University, Princeton, New Jersey 08544, United States*; orcid.org/0000-0002-5704-6654

Author Contributions

†These authors contributed equally.

Funding

The authors declare no competing financial interest.

ACKNOWLEDGMENT

This work was supported financially by NSF through the Center for Computer Assisted Synthesis C-CAS (CHE-1925607), Bristol-Myers Squibb through the Princeton Catalysis Initiative, and the Dreyfus Program for Machine Learning in the Chemical Sciences and Engineering. J.A.G.T. and P.A. acknowledge support from the Schmidt DataX Fund at Princeton University made possible through a major gift from the Schmidt Futures Foundation. J.A.G.T and A.G.D. acknowledge support of Azure cloud computing credits at Princeton University made possible through a gift from the Microsoft Corporation. The authors thank Neal Sach (Pfizer) and Paul Richardson (Pfizer) for kindly providing the Pd-catalyzed Suzuki-Miyaura HTE dataset included in this work.

REFERENCES

(1) Cernak, T.; Gesmundo, N. J.; Dykstra, K.; Yu, Y.; Wu, Z.; Shi, Z.-C.; Vachal, P.; Sperbeck, D.; He, S.; Murphy, B. A.; Sonatore, L.; Williams, S.; Madeira, M.; Verras, A.; Reiter, M.; Lee, C. H.; Cuff, J.; Sherer, E. C.; Kuethe, J.; Goble, S.; Perrotto, N.; Pinto, S.; Shen, D.-M.; Nargund, R.; Balkovec, J.; Devita, R. J.; Dreher, S. D. Microscale High-Throughput Experimentation as an Enabling Technology in Drug Discovery: Application in the Discovery of (Piperidinyl)Pyridinyl-1H-Benzimidazole Diacylglycerol Acyltransferase 1 Inhibitors. *J. Med. Chem.* **2017**, *60*, 3594–3605.

(2) Häse, F.; Roch, L. M.; Aspuru-Guzik, A. Chimera: Enabling Hierarchy Based Multi-Objective Optimization for Self-Driving Laboratories. *Chem. Sci.* **2018**, *9*, 7642–7655.

(3) Mariette, A.; Rahul, K. *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*; Springer Nature. **2015**.

(4) Rangaiah, G. P.; Feng, Z.; Hoadley, A. F. Multi-Objective Optimization Applications in Chemical Process Engineering: Tutorial and Review. *Processes* **2020**, *8*, 508.

(5) Fox, R. J.; Cuniere, N. L.; Bakrania, L.; Wei, C.; Strotman, N. A.; Hay, M.; Fanfair, D.; Regens, C.; Beutner, G. L.; Lawler, M.; Lobben, P.; Soumeillant, M. C.; Cohen, B.; Zhu, K.; Skliar, D.; Rosner, T.; Markwalter, C. E.; Hsiao, Y.; Tran, K.; Eastgate, M. D. C–H Arylation in the Formation of a Complex Pyrrolopyridine, the Commercial Synthesis of the Potent JAK2 Inhibitor, BMS-911543. *J. Org. Chem.* **2019**, *84*, 4661–4669.

(6) Alessio, I.; Philippe, N. *Multi-Criteria Decision Analysis: Methods and Software*; John Wiley & Sons. **2013**.

(7) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI. *Nature* **2018**, *555*, 604–610.

(8) Coley, C. W.; Green, W. H.; Jensen, K. F. Machine Learning in Computer-Aided Synthesis Planning. *Acc. Chem. Res.* **2018**, *51*, 1281–1289.

(9) Szymkuć, S.; Gajewska, E. P.; Klucznik, T.; Molga, K.; Dittwald, P.; Startek, M.; Bajczyk, M.; Grzybowski, B. A. Computer-Assisted Synthetic Planning: The End of the Beginning. *Angew. Chem. Int. Ed.* **2016**, *55*, 5904–5937.

(10) Wei, J. N.; Duvenaud, D.; Aspuru-Guzik, A. Neural Networks for the Prediction of Organic Chemistry Reactions. *ACS Cent. Sci.* **2016**, *2*, 725–732.

(11) Coley, C. W.; Barzilay, R.; Jaakkola, T. S.; Green, W. H.; Jensen, K. F. Prediction of Organic Reaction Outcomes Using Machine Learning. *ACS Cent. Sci.* **2017**, *3*, 434–443.

(12) Steiner, S.; Wolf, J.; Glatzel, S.; Andreou, A.; Granda, J. M.; Keenan, G.; Hinkley, T.; Aragon-Camarasa, G.; Kitson, P. J.; Angelone, D.; Cronin, L. Organic Synthesis in a Modular Robotic System Driven by a Chemical Programming Language. *Science* **2019**, *363*.

(13) Kitson, P. J.; Marie, G.; Francoia, J.-P.; Zalesskiy, S. S.; Sigerson, R. C.; Mathieson, J. S.; Cronin, L. Digitization of Multistep Organic Synthesis in Reactionware for On-Demand Pharmaceuticals. *Science* **2018**, *359*, 314–319.

(14) Coley, C. W.; ThomasIII, D. A.; Lummiss, J. A. M.; Jaworski, J. N.; Breen, C. P.; Schultz, V.; Hart, T.; Fishman, J. S.; Rogers, L.; Gao, H.; Hicklin, R. W.; Plehiers, P. P.; Byington, J.; Piotti, J. S.; Green, W. H.; Hart, A. J.; Jamison, T. F.; Jensen, K. F. A Robotic Platform for Flow Synthesis of Organic Compounds Informed by AI Planning. *Science* **2019**, *365*.

(15) Ma, J.; Sheridan, R. P.; Liaw, A.; Dahl, G. E.; Svetnik, V. Deep Neural Nets as a Method for Quantitative Structure–Activity Relationships. *J. Chem. Inf. Model* **2015**, *55*, 263–274.

(16) Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The Rise of Deep Learning in Drug Discovery. *Drug Discov. Today* **2018**, *23*, 1241–1250.

(17) Zhou, Z.; Li, X.; Zare, R. N. Optimizing Chemical Reactions with Deep Reinforcement Learning. *ACS Cent. Sci.* **2017**, *3*, 1337–1344.

(18) Shields, B. J.; Stevens, J.; Li, J.; Parasram, M.; Damani, F.; Alvarado, J. I. M.; Janey, J. M.; Adams, R. P.; Doyle, A. G. Bayesian Reaction Optimization as a Tool for Chemical Synthesis. *Nature* **2021**, *590*, 89–96.

(19) Weissman, S. A.; Anderson, N. G. Design of Experiments (DoE) and Process Optimization. A Review of Recent Publications. *Org. Process Res. Dev.* **2015**, *19*, 1605–1633.

(20) Lee, R. Statistical Design of Experiments for Screening and Optimization. *Chem. Ing. Tech* **2019**, *91*, 191–200.

(21) Murray, P. M.; Bellany, F.; Benhamou, L.; Bučar, D.-K.; Tabor, A. B.; Sheppard, T. D. The Application of Design of Experiments (DoE) Reaction Optimisation and Solvent Selection in the Development of New Synthetic Chemistry. *Org. Biomol. Chem.* **2015**, *14*, 2373–2384.

(22) Rolf, C.; E, C., Johan. *Design and Optimization in Organic Synthesis*; Elsevier. **2005**.

(23) Snoek, J.; Larochelle, H.; Adams, R. P. Practical Bayesian Optimization of Machine Learning Algorithms. In *Advances in Neural Information Processing Systems* **2012**, 2951–2959.

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

(25) Jumbam, D. N.; Skilton, R. A.; Parrott, A. J.; Bourne, R. A.; Poliakoff, M. The Effect of Self-Optimisation Targets on the Methylation of Alcohols Using Dimethyl Carbonate in Supercritical CO2. *J. Flow. Chem.* **2012**, *2*, 24–27.

(26) McMullen, J. P.; Jensen, K. F. An Automated Microfluidic System for Online Optimization in Chemical Synthesis. *Org, Process Res, Dev,* **2010**, *14*, 1169–1176.

(27) Krishnadasan, S.; Brown, R. J. C.; deMello, A. J.; deMello, J. C. Intelligent Routes to the Controlled Synthesis of Nanoparticles. *Lab on a Chip* **2007**, *7*, 1434–1441.

(28) Häse, F.; Roch, L. M.; Aspuru-Guzik, A. Chimera: Enabling Hierarchy Based Multi-Objective Optimization for Self-Driving Laboratories. *Chem. Sci.* **2018**, *9*, 7642–7655.

(29) Häse, F.; Aldeghi, M.; Hickman, R. J.; Roch, L. M.; Aspuru-Guzik, A. Gryffin: An Algorithm for Bayesian Optimization of Categorical Variables Informed by Expert Knowledge. *Appl. Phys. Rev.* **2021**, *8*, 031406.

(30) Christensen, M.; Yunker, L. P. E.; Adedeji, F.; Häse, F.; Roch, L. M.; Gensch, T.; Gomes, G. dos P.; Zepel, T.; Sigman, M. S.; Aspuru-Guzik, A.; Hein, J. E. Data-Science Driven Autonomous Process Optimization. *Commun. Chem.* **2021**, *4*, 112.

(31) Nambiar, A. M. K.; Breen, C. P.; Hart, T.; Kulesza, T.; Jamison, T. F.; Jensen, K. F. Bayesian Optimization of Computer-Proposed Multistep Synthetic Routes on an Automated Robotic Flow Platform. *ACS Cent. Sci.* **2022**, *8*, 825–836.

(32) Wang, Y.; Chen, T.-Y.; Vlachos, D. G. NEXTorch: A Design and Bayesian Optimization Toolkit for Chemical Sciences and Engineering. *J. Chem. Inf. Model* **2021**, *61*, 5312–5319.

(33) Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Köpf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; Chintala, S. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Arxiv* **2019**.

(34) 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* **2018**, *359*, 429–434.

(35) Akira, S. Cross-Coupling Reactions Of Organoboranes: An Easy Way To Construct C–C Bonds (Nobel Lecture). *Angew. Chem. Int. Ed.* **2011**, *50*, 6722–6737.

(36) Blackmore, D. Suzuki–Miyaura Coupling. in *Synthetic Methods in Drug Discovery: Volume 1*, **2015**.

(37) Vitaku, E.; Smith, D. T.; Njardarson, J. T. Analysis of the Structural Diversity, Substitution Patterns, and Frequency of Nitrogen Heterocycles among U.S. FDA Approved Pharmaceuticals. *J. Med. Chem.* **2014**, *57*, 10257–10274.

(38) Heravi, M.; Zadsirjan, V. Prescribed Drugs Containing Nitrogen Heterocycles: An Overview. *RSC Adv*. **2020**, *10*, 44247–44311.

(39) Billingsley, K. L.; Buchwald, S. L. A General and Efficient Method for the Suzuki–Miyaura Coupling of 2-Pyridyl Nucleophiles. *Angew. Chem. Int. Ed.* **2008**, *47*, 4695–4698.

(40) Cox, P. A.; Reid, M.; Leach, A. G.; Campbell, A. D.; King, E. J.; Lloyd-Jones. Base-Catalyzed Aryl-B(OH)2 Protodeboronation Revisited: From Concerted Proton Transfer to Liberation of a Transient Aryl Anion. *J. Am. Chem. Soc.* **2017**, *139*, 13156–13165.

(41) Cox, P. A.; Leach, A. G.; Campbell, A. D.; Lloyd-Jones. Protodeboronation of Heteroaromatic, Vinyl, and Cyclopropyl Boronic Acids: PH–Rate Profiles, Autocatalysis, and Disproportionation. *J. Am. Chem. Soc.* **2016**, *138*, 9145–9157.

(42) Domagalski, N. R.; Mack, B. C.; Tabora, J. E. Analysis of Design of Experiments with Dynamic Responses. *Org. Process Res. Dev.* **2015**, *19*, 1667–1682.

(43) Daulton, S.; Balandat, M.; Bakshy, E. Differentiable Expected Hypervolume Improvement for Parallel Multi-Objective Bayesian Optimization. *Arxiv* **2020**.

(44) Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; Takagi, T. Mordred: A Molecular Descriptor Calculator. *J. Cheminformatics* **2018**, *10*, 4.

(45) Rachmawati, L.; Srinivasan, D. Multiobjective Evolutionary Algorithm with Controllable Focus on the Knees of the Pareto Front. *IEEE T. Evolut. Comput.* **2009**, *13*, 810–824.

(46) Pomberger, M.; Pedrina McCarthy, A. A.; Khan, A.; Sung, S.; Taylor, C. J.; Gaunt, M. J.; Colwell, L.; Walz, D.; Lapkin, A. A. The Effect of Chemical Representation on Active Machine Learning towards Closed-Loop Optimization. *React. Chem. Eng.* **2022**, *7*, 1368–1379.

(47) Lau, S. H.; Borden, M. A.; Steiman, T. J.; Wang, L. S.; Parasram, M.; Doyle, A. G. Ni/Photoredox-Catalyzed Enantioselective Cross-Electrophile Coupling of Styrene Oxides with Aryl Iodides. *J. Am. Chem. Soc.* **2021**, *143*, 15873–15881.

(48) Hills, C. J.; Winter, S. A.; Balfour, J. A. Tolterodine. *Drugs* **1998**, *55*, 813–820.

(49) McRae, A. L.; Brady, K. T. Review of Sertraline and Its Clinical Applications in Psychiatric Disorders. *Expert Opin. Pharmaco.* **2005**, *2*, 883–892.

(50) Ameen, D.; Snape, T. J. Chiral 1,1-Diaryl Compounds as Important Pharmacophores. *Medchemcomm* **2013**, *4*, 893–907.

TOC graphic: