IF OPTIMIZING FOR GENERAL PARAMETERS IN CHEMISTRY IS USEFUL, WHY IS IT HARDLY DONE?

Anonymous authors
Paper under double-blind review

ABSTRACT

General parameters are highly desirable in the natural sciences — e.g., reaction conditions that enable high yields across a range of related transformations. This has a significant practical impact since those general parameters can be transfered to related tasks without the need for laborious and time-intensive re-optimization. While Bayesian optimization (BO) is widely applied to find optimal parameter sets for specific tasks, it has remained underused in experiment planning towards such general optima. In this work, we consider the real-world problem of condition optimization for chemical reactions to study whether performing generality-oriented BO can accelerate the identification of general optima, and whether these optima also translate to unseen examples. This is achieved through a careful formulation of the problem as an optimization over curried functions, as well as systematic benchmarking of generality-oriented strategies for optimization tasks on real-world experimental data. Empirically, we find that for generality-oriented optimization, simple optimization strategies that decouple parameter and task selection perform comparably to more complex ones, and that effective optimization is merely determined by an effective exploration of both parameter and task space.

1 INTRODUCTION

Identifying parameters that deliver satisfactory performance on a wide set of tasks, which we refer to as *general parameters*, is crucial for numerous real-world challenges. Examples are the identification of sensor settings that allow the sensor to measure accurately in different environments (Güntner et al., 2019), or the design of footwear that provides good performance for a range of people on different undergrounds (Promjun & Sahachaisaeree, 2012). A prominent example comes from the domain of chemical synthesis, where finding reaction conditions under which different starting materials can be reliably converted into the corresponding products, remains a critical challenge (Wagen et al., 2022; Prieto Kullmer et al., 2022; Rein et al., 2023; Betinol et al., 2023; Rana et al., 2024; Schmid et al., 2024). Such general conditions are of particular interest, *e.g.*, in the pharmaceutical industry, where thousands of reactions are carried out regularly, and optimizing each reaction is unfeasible (Wagen et al., 2022). While Bayesian Optimization (BO) is increasingly adopted within reaction optimization (Clayton et al., 2019; Shields et al., 2021; Guo et al., 2023; Tom et al., 2024), the vast majority of cases neglects generality considerations (Figure 1, left-hand side.)

This lack of consideration can be attributed to the fact that directly observing the generality of selected conditions is associated with largely increased experimental costs, as experimental evaluations on multiple substrates are required. Attempts at reducing the required number of experiments inevitably increase the complexity of the decision-making process. Thus, the usage of generality-oriented optimization in laboratories is hindered in the absence of appropriate decision-making algorithms. Here, generality-oriented optimization turns into a *partial monitoring scenario*, in which each condition can only be evaluated on a subset of all possible substrates. As a consequence, any iterative experiment planning algorithm needs to recommend both the condition and the substrate for the next experimental evaluation (Figure 1, right-hand side). Experimentally measuring the outcome of the recommended experiment corresponds to a partial observation of the generality objective, which needs to be taken into account when recommending the next experiment.

In the past two years, isolated studies have targeted the identification of general reaction conditions through variations of BO (Angello et al., 2022) and multi-armed bandit optimization (Wang

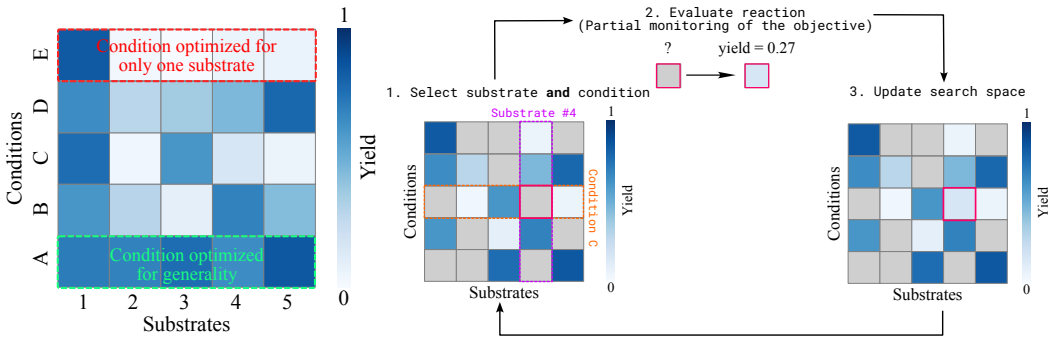


Figure 1: *Left:* While conditions can be optimized to maximize the reaction outcome for only one substrate (red), generality-optimized conditions provide a satisfactory reaction outcome for multiple substrates. *Right:* Optimization loop for generality-oriented optimization under partial monitoring.

et al., 2024). Concurrently, different algorithms have been proposed to optimize similarly structured problems, such as BO with expensive integrands (BOEI; Xie et al., 2012; Toscano-Palmerin & Frazier, 2018) and distributionally robust BO (DRBO; Bogunovic et al., 2018; Kirschner et al., 2020a). Despite these advances, generality-oriented optimizations are still not commonly performed in real-world experiments (see Section 2.2.4).This likely arises from the fact that the applicability and limitations of these algorithms are yet to be understood, which is crucial for their effective integration into real-world laboratory workflows (Tom et al., 2024).

For these reasons, we herein perform a systematic benchmark study into generality-oriented optimization. To obtain a problem flexibility required for real-world applications (Betinol et al., 2023), we formulate generality-oriented optimization as an optimization problem over curried functions. Further, we perform systematic benchmarks on various real-world chemical reaction optimization tasks.Specifically for the latter, we (i) confirm the expectation that optimization over multiple substrates leads to more general optima, and (ii) demonstrate that finding these optima effectively can be achieved through a highly explorative acquisition of the next conditions to evaluate.

In summary, our contributions are four-fold:

- Formulation of generality-oriented optimization as an optimization problem over a curried function.
- Expansion and adaptation of established reaction optimization benchmark tasks, improving their utility as benchmarks for generality-oriented BO.
- Evaluation of different optimization algorithms for identifying general optima.
- *CurryBO* as an open-source extension to *BoTorch* (Balandat et al., 2020) for generality-oriented optimization problems (noa).

2 Foundations of Generality-Oriented Bayesian Optimization

To formalize the generality-oriented optimization problem, we provide a principled outline by considering it as an extension of established global optimization approaches over curried functions. For clarity, we also discuss its distinction to different variations of global optimization, including multiobjective, multifidelity, and mixed-variable optimization.

2.1 Global Optimization

Global black-box optimization is concerned with finding the optimum of an unknown objective function $f(\mathbf{x})$:

$$\hat{\mathbf{x}} = \underset{\mathbf{x}\in\mathcal{X}}{\operatorname{argmax}}\ f(\mathbf{x}) \tag{1}$$

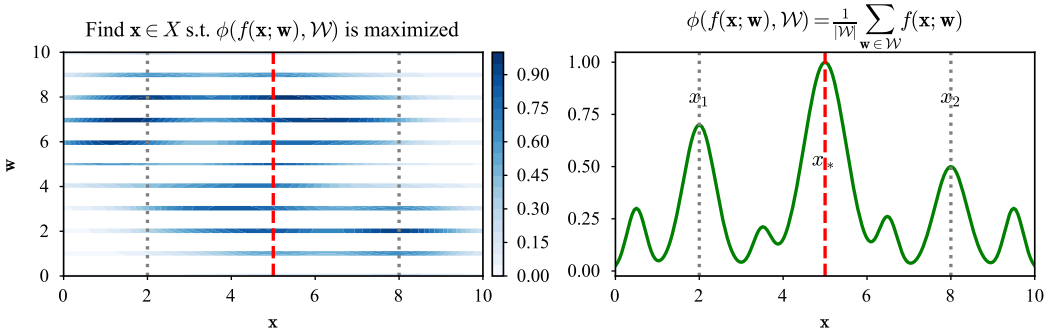


Figure 2: Conceptual overview of the generality-oriented optimization problem. *Left:* The function values across the joint space $\mathcal{X} \times \mathcal{W}$. *Right:* Mean aggregation applied to the function family $f(\mathbf{x};\mathbf{w})$, that is obtained via currying of the joint space $\mathcal{X} \times \mathcal{W}$. The quantity $\phi\big(f(\mathbf{x};\mathbf{w}),\mathcal{W}\big)$ constitutes the partially observable objective function, of which $\hat{\mathbf{x}} = \arg\max_{\mathbf{x}\in\mathcal{X}} \phi(\mathbf{x})$ is the optimum that should be identified.

Suppose $f(\mathbf{x})$ is a function that (a) is not analytically tractable, (b) is very expensive to evaluate, and (c) can only be evaluated without obtaining gradient information. In this scenario, BO has emerged as a ubiquitous approach for finding the global optimum $\hat{\mathbf{x}} \in \mathcal{X}$ in a sample-efficient manner (Garnett, 2023). The working principle of BO involves a probabilistic surrogate model $g(\mathbf{x})$ to approximate $f(\mathbf{x})$, which can be used to compute a predictive posterior distribution over g under all previous observations $\mathcal{D} = \{(\mathbf{x}_i, f(\mathbf{x}_i)\}_{i=1}^k$. The most prominent choice for $p(g(\mathbf{x}) \mid \mathcal{D})$ are Gaussian processes (GPs; Rasmussen & Williams, 2006), with various types of Bayesian neural networks becoming increasingly popular in the past decade (Hernández-Lobato et al., 2017; Kristiadi et al., 2023; Li et al., 2024; Kristiadi et al., 2024). Based on the predictive posterior, an acquisition function α over the input space $\mathcal{X}$ is used to decide at which $\mathbf{x}_{\text{next}} \in \mathcal{X}$ the objective function should be evaluated next. Key to the success of BO is the implicit exploitation–exploration tradeoff in α, which makes use of the posterior distribution $p(g(x) \mid \mathcal{D})$ (Močkus, 1975). Common choices of α are Upper Confidence Bound (UCB; Kaelbling, 1994a;b; Agrawal, 1995), Expected Improvement (EI; Jones et al., 1998), Knowledge Gradient (Gupta & Miescke, 1994; Frazier et al., 2008; 2009) or Thompson Sampling (TS; Thompson, 1933). The hereby selected $\mathbf{x}_{\text{next}}$ is evaluated experimentally, resulting in $f(\mathbf{x}_{\text{next}})$, and the described procedure is repeated until a satisfactory outcome is observed, or the experimentation budget is exhausted.

2.2 Global Optimization for Generality

2.2.1 Problem Formulation

Extending the global optimization framework, we consider a black-box function $f : \mathcal{X} \times \mathcal{W} \to \mathbb{R}$ in joint space $\mathcal{X} \times \mathcal{W}$, where $\mathbf{x} \in \mathcal{X}$ can be continuous, discrete or mixed-variable and $\mathcal{W} = \{\mathbf{w}_i\}_{i=1}^n$ is a discrete parameter space of size n (see Figure 2). Each evaluation of f is expensive and does not provide gradient information. In the example of reaction condition optimization, $\mathbf{x}$ are conditions from the condition space $\mathcal{X}$, e.g. the temperature, and $\mathbf{w} \in \mathcal{W}$ the substrates (starting materials of a reaction) that are considered for generality-oriented optimization. Let curry be a currying operator on the second argument, i.e., $\text{curry}(f) : \mathcal{W} \to \big(\mathcal{X} \to \mathbb{R}\big)$. Then, for some $\mathbf{w} \in \mathcal{W}$, evaluating $\text{curry}(f)(\mathbf{w})$ yields a new function $f(\,\cdot\,;\mathbf{w}) : \mathcal{X} \to \mathbb{R}$, where $f(\mathbf{x};\mathbf{w}) = f(\mathbf{x},\mathbf{w})$. Importantly, these $f(\,\cdot\,;\mathbf{w}) : \mathcal{X} \to \mathbb{R}$ correspond to functions that can be evaluated experimentally (i.e. a reaction for a specific substrate as a function of conditions), even though evaluations are expensive. This allows us to describe all observable functions through an n-sized set $\mathcal{F} = \{f(\,\cdot\,;\mathbf{w}_i) : \mathcal{X} \to \mathbb{R}\}_{i=1}^n$. In the context of reaction condition optimization $\mathcal{F}$ consists of all functions that describe the reaction outcome for each substrate. Evaluation of a specific $f(\mathbf{x}_{\text{obs}};\mathbf{w}_{\text{obs}})$ then corresponds to measuring the reaction outcome of a substrate (described by $\mathbf{w}_{\text{obs}}$) under specific reaction conditions $\mathbf{x}_{\text{obs}}$.

In generality-oriented optimization, the goal is to identify the optimum $\hat{\mathbf{x}} \in \mathcal{X}$ that is generally optimal across $\mathcal{W}$, meaning $\hat{\mathbf{x}}$ maximizes a user-defined generality metric over all $\mathbf{w} \in \mathcal{W}$ (see

Algorithm 1 Generality-oriented Bayesian optimization

Input:

Set of observable functions $\mathcal{F} = \{f(\cdot; \mathbf{w}_i) : \mathcal{X} \to \mathbb{R}\}_{i=1}^{n}$
Initial dataset $\mathcal{D}_k = \left\{\mathbf{x}_j, \mathbf{w}_j, f(\mathbf{x}_j; \mathbf{w}_j)\right\}_{j=1}^{k}$
Aggregation function $\phi\big(f(\mathbf{x}; \mathbf{w}), \mathcal{W}\big)$
Surrogate model $g(\mathbf{x}, \mathbf{w})$ and acquisition policy A
Budget K

while $k \leq K$ **do**
  Compute posterior distribution $p\big(g_k(\mathbf{x}, \mathbf{w}) \mid \mathcal{D}_k\big)$
  Acquire $\mathbf{x}_{k+1}, \mathbf{w}_{k+1} = A\Big(p\big(g_k(\mathbf{x}, \mathbf{w}) \mid \mathcal{D}_k\big), \phi\big(f(\mathbf{x}; \mathbf{w}), \mathcal{W}\big)\Big)$
  Observe $f(\mathbf{x}_{k+1}; \mathbf{w}_{k+1})$
  Update $\mathcal{D}_{k+1} = \mathcal{D}_k \bigcup \left\{(\mathbf{x_{k+1}}, \mathbf{w}_{k+1}, f(\mathbf{x}_{k+1}\mathbf{w}_{k+1}))\right\}$
  $k = k + 1$
end while
return $\hat{\mathbf{x}} = \underset{\mathbf{x} \in \mathcal{X}}{\operatorname{argmax}}\ \mathbb{E}\Big[p\big(\phi(\mathbf{x}) \mid \mathcal{D}_K\big) \mid \mathbf{x}\Big]$

Figure 2 for illustration). We refer to this generality metric as the *aggregation function* ϕ:

$$\hat{\mathbf{x}} = \underset{\mathbf{x} \in \mathcal{X}}{\operatorname{argmax}}\ \phi(\mathbf{x}) := \underset{\mathbf{x} \in \mathcal{X}}{\operatorname{argmax}}\ \phi\big(f(\mathbf{x}; \mathbf{w}), \mathcal{W}\big) \tag{2}$$

In the reaction optimization example, this corresponds to conditions (e.g. reaction temperature) that give e.g. the highest average yield over all considered substrates. In this scenario, the choice of ϕ is the mean $\phi(f(\mathbf{x}; \mathbf{w}), \mathcal{W}) = {}^{1}/_{|\mathcal{W}|} \sum_{\mathbf{w} \in \mathcal{W}} f(\mathbf{x}; \mathbf{w})$. An alternative choice of ϕ could be the number of function values $\{f(\mathbf{x}; \mathbf{w}_i)\}_{i=1}^{n}$ above a user-defined threshold (Betinol et al., 2023). Further practically relevant aggregation functions are described in Appendix A.1.1.

While equation 2 appears like a standard global optimization problem over $\mathcal{X}$, evaluating $\phi(\mathbf{x})$ itself is intractable due to the aggregation over $\mathcal{W}$. Indeed, to evaluate $\phi(\mathbf{x})$ on a single $\mathbf{x}$, one must perform n-many expensive function evaluations to first obtain $\{f(\mathbf{x}; \mathbf{w}_i)\}_{i=1}^{n}$. Due to this intractability, ideally, the number of such function evaluations is minimized. Thus, this setting differs from the conventional global optimization problem, due to its *partial observation* nature: One can only compute $\phi(\mathbf{x})$ via a subset of observations $\{f(\mathbf{x}; \mathbf{w}_j)\}_{j=1}^{m}$ where $m < n$.

To maximize sample efficiency, an optimizer should always recommend a new pair $(\mathbf{x}_{\text{next}}, \mathbf{w}_{\text{next}})$ to evaluate next — in other words: $\phi(\mathbf{x}_{\text{next}})$ is only observed partially via a single evaluation of f, i.e., $m = 1$. Treating this in the conventional framework of BO, we can build a probabilistic surrogate model $g(\mathbf{x}_i; \mathbf{w}_i)$ from all k available observations $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{w}_i, f(\mathbf{x}_i; \mathbf{w}_i)\}_{i=1}^{k}$, referred to as $p\big(g_k(\mathbf{x}, \mathbf{w}) \mid \mathcal{D}\big)$. From the posterior distribution over g, a posterior distribution over ϕ can be estimated for any functional form of ϕ via Monte-Carlo integration (see Appendix A.1.2 for further details; Balandat et al., 2020).

Unlike the conventional BO case, we now need a specific acquisition policy A to decide which $\mathbf{x} \in \mathcal{X}$ *and* $\mathbf{w} \in \mathcal{W}$ the aggregated objective function $\phi(\mathbf{x})$ should be partially evaluated. Note that A plays an important role since it must respect the partial observability constraint. That is, it must also propose *a single* $\mathbf{w}$ at each BO step such that the general (over *all* $\mathbf{w}_i$'s) optimum $\hat{\mathbf{x}}$ is obtained in as few steps as possible. Given the pair $(\mathbf{x}_{k+1}, \mathbf{w}_{k+1})$, the aggregated objective $\phi(\mathbf{x}_{k+1})$ is partially observed, $\mathcal{D}$ is updated, and the discussed steps are repeated until the experimentation budget is exhausted. Eventually, owing to the partial monitoring scenario (Rustichini, 1999; Lattimore & Szepesvári, 2019; 2020), the final optimum after a budget of k experiments, $\hat{\mathbf{x}}_k$, is returned as the $\mathbf{x} \in \mathcal{X}$ that maximizes the mean of the predictive posterior of ϕ. A summary of this is provided in Algorithm 1.

2.2.2 Acquisition Strategies to Select $\mathbf{x}_{\text{next}}$ and $\mathbf{w}_{\text{next}}$

As outlined above, the efficiency of generality-oriented optimization depends on the selection of $\mathbf{x}_{\text{next}}$ and $\mathbf{w}_{\text{next}}$. Given a posterior distribution $p\big(g_k(\mathbf{x}, \mathbf{w}) \mid \mathcal{D}\big)$, an aggregation function

$\phi\big(f(\mathbf{x};\mathbf{w}),\mathcal{W}\big)$, any acquisition policy should determine $\mathbf{x}_{\text{next}}$ and $\mathbf{w}_{\text{next}}$, which formally requires optimization over $\mathcal{X}\times\mathcal{W}$. Assuming weak coupling between $\mathcal{X}$ and $\mathcal{W}$, we can formulate a sequential acquisition policy, as outlined in Algorithm 2. First, $\mathbf{x}_{\text{next}}$ is acquired by optimizing an $\mathbf{x}$-specific acquisition function α_x over the posterior over the aggregation function. Second, a $\mathbf{w}$-specific acquisition α_w is optimized over the posterior distribution at $\mathbf{x}_{\text{next}}$. Notably, in this setting, established one-step-lookahead acquisition functions can be used for both α_x and α_w.

Algorithm 2 Sequential Acquisition Strategy

Input:
posterior distribution $p\big(g_k(\mathbf{x},\mathbf{w}) \mid \mathcal{D}\big)$
aggregation function $\phi\big(f(\mathbf{x};\mathbf{w}),\mathcal{W}\big)$
acquisition function α_x
acquisition function α_w

1: compute posterior distribution $p\big(\phi(\mathbf{x}) \mid \mathcal{D}\big) = p\Big(\phi\big(g_k(\mathbf{x},\mathbf{w}),\mathcal{W}\big) \mid \mathcal{D}\Big)$
2: acquire $\mathbf{x}_{\text{next}} = \underset{\mathbf{x}\in\mathcal{X}}{\text{argmax}}\ \alpha_x\Big(p\big(\phi(\mathbf{x}) \mid \mathcal{D}\big)\Big)$
3: acquire $\mathbf{w}_{\text{next}} = \underset{\mathbf{w}\in\mathcal{W}}{\text{argmax}}\ \alpha_w\Big(p\big(g_k(\mathbf{x}_{\text{next}},\mathbf{w}) \mid \mathcal{D}\big)\Big)$
4: **return** $\mathbf{x}_{\text{next}}, \mathbf{w}_{\text{next}}$

However, the decoupling of $\mathcal{X}$ and $\mathcal{W}$ is a strong simplification, and identifying $\mathbf{x}_{\text{next}}$ and $\mathbf{w}_{\text{next}}$ formally requires a joint optimization over $\mathcal{X}\times\mathcal{W}$ (Algorithm 3). Such a joint optimization necessitates a two-step lookahead acquisition function α'

$$\alpha'(\mathbf{x}_{k+1},\mathbf{w}_{k+1}) = \alpha\left[\underset{\mathbf{x}\in\mathcal{X}}{\text{argmax}}\ \alpha_{\text{final}}\left(p\Big(\phi(\mathbf{x}) \mid \mathcal{D}^*_{k+1}\Big)\right)\right] \tag{3}$$

where α is a classical one-step lookahead acquisition function, which is evaluated at $\mathbf{x}_{k+2}\in\mathcal{X}$ which maximizes the final acquisition function α_{final} (in our case: the posterior mean) over a fantasy posterior distribution $p\Big(\phi(\mathbf{x}) \mid \mathcal{D}^*_{k+1}\Big)$. This distribution is obtained by conditioning the existing posterior on a new fantasy observation at $(\mathbf{x}_{k+1},\mathbf{w}_{k+1})$. An implementation of equation Equation (3) using Monte-Carlo integration is given in Algorithm 4.

$\mathbf{x}_{\text{next}}$ and $\mathbf{w}_{\text{next}}$ are then acquired by optimizing α' in the joint input space $\mathcal{X}\times\mathcal{W}$.

Algorithm 3 Joint Acquisition Strategy

Input:
posterior distribution $p\big(g_k(\mathbf{x},\mathbf{w}) \mid \mathcal{D}\big)$
aggregation function $\phi\big(f(x;w),\mathcal{W}\big)$
two-step lookahead acquisition function α'

1: compute posterior distribution $p\big(\phi(\mathbf{x}) \mid \mathcal{D}\big) = p\Big(\phi\big(g_k(\mathbf{x},\mathbf{w}),\mathcal{W}\big) \mid \mathcal{D}\Big)$
2: acquire $\mathbf{x}_{\text{next}},\mathbf{w}_{\text{next}} = \underset{x,w\in\mathcal{X}\times\mathcal{W}}{\text{argmax}}\ \alpha'\Big(p\big(\phi(\mathbf{x}) \mid \mathcal{D}\big)\Big)$
3: **return** $\mathbf{x}_{\text{next}}, \mathbf{w}_{\text{next}}$

2.2.3 Distinction from Existing Variants of the BO Formalism

Despite seeming similarities with *multiobjective*, *multifidelity*, and *mixed-variable* optimization, the generality-oriented approach describes a distinctly different scenario:

- In contrast to *multiobjective* optimization, here, we consider a single optimization objective, i.e. $\phi(\mathbf{x})$. However, this objective can only be partially observed. Whereas the overall

optimization problem aims to identify $\hat{\mathbf{x}} \in \mathcal{X}$, finding the next recommended observation requires a joint optimization over $\mathcal{X}$ and $\mathcal{W}$.

- In contrast to *multifidelity* BO, the functions parameterized by $\mathbf{w} \in \mathcal{W}$ do not correspond to the same objective with different fidelities. Rather, they are independent functions which all contribute equally to the objective function $\phi(\mathbf{x})$.

- Unlike *mixed-variable* BO (Daxberger et al., 2020), the goal of generality-oriented BO is not to find $(\mathbf{x}, \mathbf{w})$ that maximizes the objective in the *joint* space. Rather, the goal is to find the set optimum $\hat{\mathbf{x}}$ that maximizes $\phi\big(f(\mathbf{x}; \mathbf{w}), \mathcal{W}\big)$ over $f(\mathbf{x}; \mathbf{w})$. In the case of ϕ being a sum, this bears resemblance to maximizing the *marginal* over $\mathbf{x}$ (see Figure 2). Moreover, $\mathcal{X}$ can be continuous or discrete, thus, $\mathcal{X} \times \mathcal{W}$ can be a fully-discrete space.

2.2.4 Related Works

Similarly structured problems have been previously described, mostly for specific formulations of the aggregation function ϕ. Most prominently, if ϕ contains a sum over all $f(\,\cdot\,; \mathbf{w}_i)$ with $\mathbf{w}_i \in \mathcal{W}$, this problem has been referred to as optimization of integrated response functions (Williams et al., 2000), optimizing an average over multiple tasks (Swersky et al., 2013), or optimization with expensive integrands (Toscano-Palmerin & Frazier, 2018). The latter work proposes a BO approach, including a joint acquisition over $\mathcal{X} \times \mathcal{W}$ with the goal of maximizing the value of information. In the framework discussed above, this corresponds to a joint optimization of a two-step lookahead expected improvement, and is included in our benchmark experiments as Joint 2LA-EI. The scenario in which ϕ corresponds to the min operation, i.e. the objective is $\min_{\mathbf{w} \in \mathcal{W}} f(\mathbf{x}; \mathbf{w})$, has been discussed as distributionally robust BO (Bogunovic et al., 2018; Kirschner et al., 2020a; Nguyen et al., 2020; Husain et al., 2023). While these works provide advanced algorithmic solutions for the respective optimization scenarios, our goal was to benchmark the applicability of such algorithms in real-life settings. Therefore, the formulation as optimization over curried functions provides a flexible framework that covers aggregation functions of arbitrary functional form, and the implementation of *CurryBO* allows for rapid integration with the *BoTorch* ecosystem.

In the chemical synthesis, the concept of "reaction generality" has been discussed on multiple occasions, given its enormous importance for accelerating molecular discovery (Wagen et al., 2022; Prieto Kullmer et al., 2022; Rein et al., 2023; Betinol et al., 2023; Rana et al., 2024; Gallarati et al., 2024; Schmid et al., 2024). The first example of actual generality-oriented optimization in chemistry has been reported by Angello et al. (2022), who describe a modification of BO, sequentially acquiring $\mathbf{x}_{\text{next}}$ via $\alpha_x = \text{PI}$ (Probability of Improvement) and $\mathbf{w}_{\text{next}}$ via $\alpha_w = \text{PV}$ (Posterior Variance). The authors demonstrate its applicability in automated experiments on Suzuki–Miyaura cross couplings. A similar algorithm as described in their work is evaluated herein as the Seq 1LA-UCB-var strategy. Following an alternative strategy, Wang et al. (2024) recently formulated generality-oriented optimization as a multi-armed bandit problem, where each arm corresponds to a possible reaction condition. While their algorithm has been successful in campaigns with few possible reaction conditions, the necessity of sampling all conditions at the outset of a campaign renders its application impractical for a high number of discrete conditions or even continuous variables. The algorithm described in their work is evaluated herein as the Bandit strategy.

Despite these recent advances, the applicability and limitations of these algorithmic approaches in real-life settings have remained unclear. Thus, our work provides a systematic benchmark over different generality-oriented optimization strategies, at the example of generality-oriented reaction optimization in chemistry.

Due to the partial monitoring nature of generality-oriented optimization, we want to highlight work that has been conducted on the partial monitoring case for bandits (Rustichini, 1999; Lattimore & Szepesvári, 2019; 2020). However, to the best of our knowledge, works in this field has mostly dealt with an information-theoretic approach towards optimally scaling algorithms. We refer the readers to select publications (Lattimore & Szepesvari, 2019; Kirschner et al., 2020b; Lattimore & Gyorgy, 2021; Lattimore, 2022). Comprehensive benchmark of different strategies in the early stages of an optimization has not been applied to generality-optimization for chemical benchmark tasks.

Table 1: Nomenclature and description of the benchmarked acquisition strategies and acquisition functions in the main text. Further acquisition functions are described in Table 2.

Acquisition Strategy	Acquisition Function
SEQ 1LA: Sequential acquisition of $\mathbf{x}_{\text{next}}$ and $\mathbf{w}_{\text{next}}$, each using a one-step lookahead acquisition function. The final $\hat{\mathbf{x}}$ is selected greedily.	UCB: Upper confidence bound ($\beta = 0.5$).
SEQ 2LA: Sequential acquisition of $\mathbf{x}_{\text{next}}$ and $\mathbf{w}_{\text{next}}$, each using a two-step lookahead acquisition function. The final $\hat{\mathbf{x}}$ is selected greedily.	UCBE: Upper confidence bound ($\beta = 5$).
JOINT 2LA: Joint acquisition of $\mathbf{x}_{\text{next}}$ and $\mathbf{w}_{\text{next}}$ using a two-step lookahead acquisition function. The final $\hat{\mathbf{x}}$ is selected greedily.	EI: Expected Improvement.
BANDIT: Multi-armed bandit algorithm as implemented by Wang et al. (2024).	PV: Posterior Variance.
RANDOM: Random selection of the final $\hat{\mathbf{x}}$.	RA: Random acquisition.

3 METHODS

3.1 EXPERIMENTAL BENCHMARK PROBLEMS

In our benchmarks, we consider four real-world chemical reaction problems stemming from high-throughput experimentation (HTE; Zahrt et al., 2019; Buitrago Santanilla et al., 2015; Nielsen et al., 2018; Stevens et al., 2022; Wang et al., 2024). Each problem evaluates the optimization of a chemically relevant reaction outcome (such as enantioselectivity $\Delta\Delta G^{\ddagger}$, yield, or starting material conversion), and contains an experimental dataset of substrates, conditions and measured outcomes.

Extensive analysis of the benchmark problems can be found in Appendix A.2. At this stage, it should be noted that, while widely used as such, the problems have not been designed as benchmarks for reaction condition optimization. To mitigate the well-known bias of HTE datasets towards high-outcome experiments (Strieth-Kalthoff et al., 2022; Beker et al., 2022), we additionally augment the search space to incorporate larger domains of low-outcome results using a chemically sensible expansion workflow (see Appendix A.2.2 for further details).

3.2 OPTIMIZATION ALGORITHMS

Using the benchmark problems outlined above, we perform systematic evaluations of multiple methods for the identification of general optima. In the main text, we discuss the acquisition strategies and functions for recommending the next data point $(\mathbf{x}_{\text{next}}, \mathbf{w}_{\text{next}})$ as shown in Table 1. We name each experiment according to the acquisition strategy used, followed by specifications of the used acquisition functions α_x and α_w or α for sequential and joint acquisitions, respectively. As an example, a sequential two-step lookahead acquisition strategy with an Upper Confidence Bound as α_x and Posterior Variance as α_w, is referred to as SEQ 2LA-UCB-PV. Each strategy is evaluated under two different generality definitions: the *mean* and the *number-above-threshold* aggregation (threshold aggregation) functions described in Section 2.2.1 (see Appendix A.1.1 for further details).

In all BO experiments, we used a GP surrogate, as provided in *BoTorch* (Balandat et al., 2020), with the Tanimoto kernel from *Gauche* (Griffiths et al., 2023). Molecules were represented using Morgan Fingerprints (Morgan, 1965) with 1024 bits and a radius of 2, generated using RDKit (Landrum, 2023). For each experiment, we provide statistics over 30 independent runs, each performed over different substrates and initial conditions. Further baseline experiments are discussed in Appendix A.4. To ensure cross-task comparability, we calculate the GAP as a normalized, problem-independent optimization metric (GAP = $(y_i - y_0)/(y^* - y_0)$, where y_i is the true generality of the recommendation at experiment k and y^* is the true global optimum; Jiang et al., 2020).

4 RESULTS AND DISCUSSION

To assess the utility of generality-oriented optimization, it is necessary to validate the transferability of these general optima to unseen spaces. Therefore, we commence our analysis by systematically investigating all benchmark tasks using exhaustive grid search. This analysis reveals that, with an

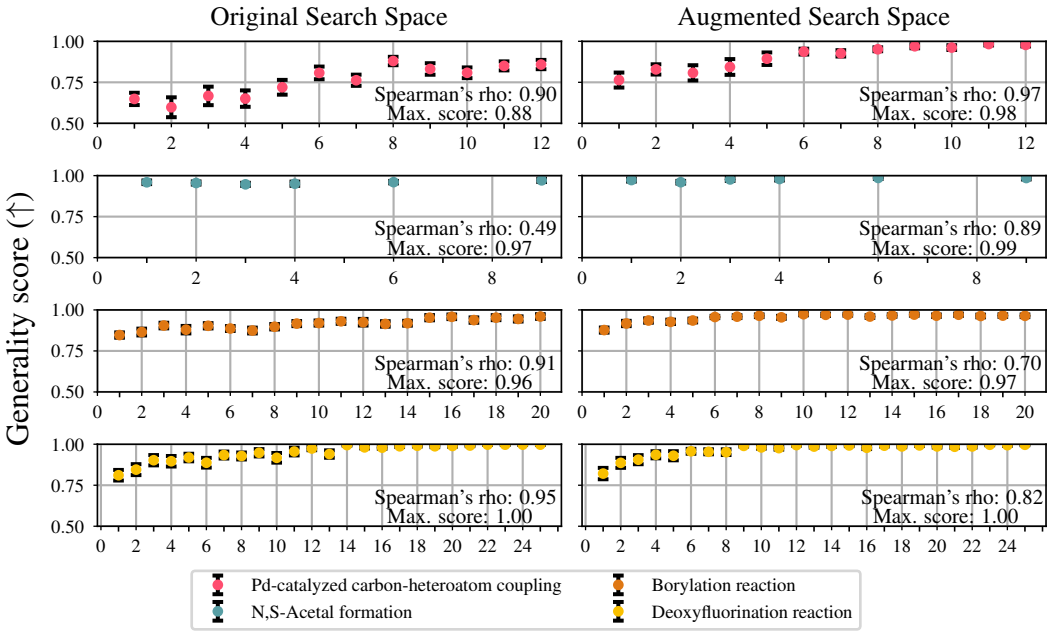


Figure 3: Normalized test-set generality score as determined by exhaustive grid search for the four benchmarks on the original (left) and augmented (right) problems for the mean aggregation. Average and standard error are taken from thirty different train/test substrates splits.

increasing number of substrates in $\mathcal{W}_{\text{train}}$ considered during optimization, the transferability of the found optima to a held-out test set $\mathcal{W}_{\text{test}}$ increases (Figure 3, left), as evidenced by Spearman's $\rho > 0$. While this finding is arguably unsurprising, and merely confirms a common assumption in the field (Wagen et al., 2022), it indicates possible caveats concerning the use of the non-augmented problems as benchmarks for generality-oriented optimization: Even with larger sizes of $\mathcal{W}_{\text{train}}$, the found optima did not consistently lead to optimal outcomes on the corresponding test sets. In contrast, we find that on the augmented benchmark tasks, which are more reflective of experimental reality, transferability of the identified optima to a held-out $\mathcal{W}_{\text{test}}$ is significantly improved. Notably, these observations are not limited to the definition of generality as the average over all $\mathbf{w} \in \mathcal{W}$, but remain valid for further aggregation functions on a majority of benchmarks (see Appendix A.6.1). These findings underline that – especially in "needle in a haystack scenarios" – generality-oriented optimization is indeed necessary for finding transferable optima. Most importantly, such scenarios apply to real-world reaction optimization, where for most reactions, the majority of possible conditions do not lead to observable product quantities. This re-emphasizes the need for benchmark problems that reflect experimental reality.

Having established the utility of generality-oriented optimization, we set out to perform a systematic benchmark of how to identify those optima using iterative optimization under partial objective monitoring. In the first step, we evaluate those approaches that have been developed in the context of reaction optimization (Angello et al., 2022; Wang et al., 2024) on two practically relevant aggregation functions, the mean and threshold aggregation (Appendix A.1.1). As a summary, Figure 4 shows the optimization trajectories of these different algorithms averaged across all augmented benchmark problems.. Overall, we find that the BO-based SEQ 1LA-UCB-PV acquisition strategy, as outlined by Angello et al. (2022), shows faster optimization performance compared to other algorithms used in the chemical domain. In particular, it significantly outperforms the BANDIT algorithm proposed by Wang et al. (2024), which can be attributed to the necessity of evaluating each $\mathbf{w} \in \mathcal{W}_{\text{train}}$ at the outset of each campaign, tying up a notable share of the experimental budget Assuredly, both proposed methods readily outperform the two random baselines RANDOM and SEQ 1LA-RA-RA.

Inspired by these observations, we perform a deeper investigation into the BO approaches formalized in Section 2.2. Initially, different options of the sequential strategy of acquiring $\mathbf{x}_{\text{next}}$ and $\mathbf{w}_{\text{next}}$ are

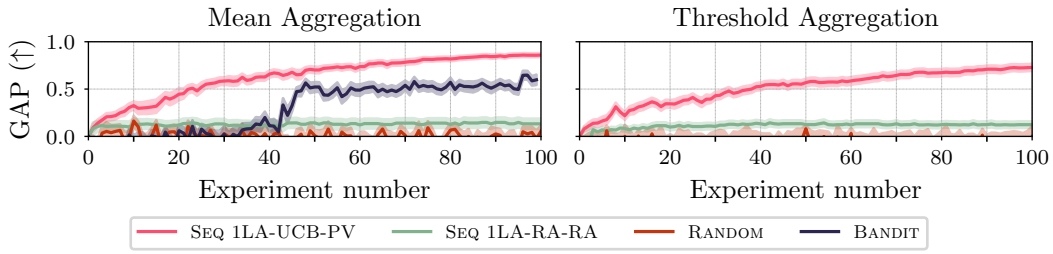


Figure 4: Optimization trajectories of different algorithms for generality-oriented optimization previously reported in the chemical domain. The trajectories are averaged over all augmented benchmark problems. Note that the BANDIT algorithm is incompatible with the threshold aggregation function.

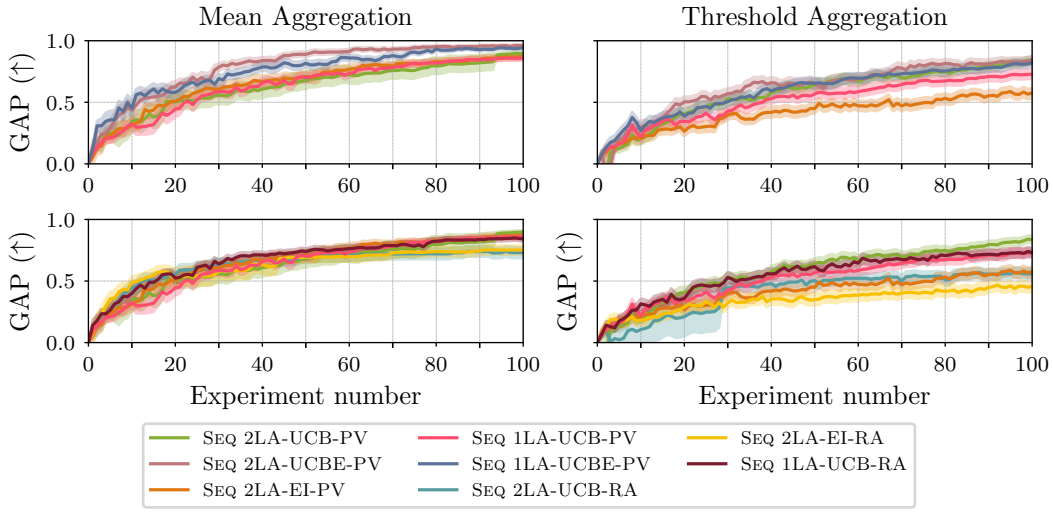


Figure 5: Optimization trajectories using sequential acquisition strategies. The top row shows the variation of α_x, while the bottom row shows the variation of α_w). Trajectories are averaged over four augmented benchmark problems.

evaluated. For this purpose, we compare multiple acquisition functions α_x for selecting $\mathbf{x}_{\text{next}} \in \mathcal{X}$, as formalized in Appendix A.4 and Section 3.2. Overall, the empirical results (Figure 5, top half) indicate largely similar optimization behavior for the different α_x. However, it can be observed that a higher degree of exploration has a positive effect on optimization performance, e.g., when comparing the baseline method SEQ 1LA-UCB-PV (α_x: UCB with $\beta = 0.5$) with the more exploratory SEQ 1LA-UCBE-PV (α_x: UCB with $\beta = 5.0$). While systematic investigations into the generalizability of this finding are ongoing, we hypothesize that it can be attributed to the partial monitoring scenario, which leads to larger predictive uncertainties, and therefore less efficient exploitation. Surprisingly, the use of two-step-lookahead acquisition functions for α_x, which should conceptually be well-suited for the partial monitoring scenario (Section 2.2.2), did not lead to significant improvements compared to their one-step-lookahead counterparts (e.g., comparing SEQ 1LA-UCB-PV with SEQ 2LA-UCB-PV and SEQ 2LA-EI-PV). Yet, the trend that more exploratory α_x lead to improved optimization behavior can also be observed for two-step-lookahead acquisition functions. However, we find that, especially for the threshold aggregation function (Figure 5), Expected Improvement (EI) shows significantly decreased optimization performance, which may be attributed to the partial monitoring scenario, and the resulting uncertainty in estimating the current optimum.

Similarly, we observe only a small influence of the choice of α_w (Figure 5, bottom half). In particular, an uncertainty-driven acquisition of α_w, as used by Angello et al. (2022), shows only slightly

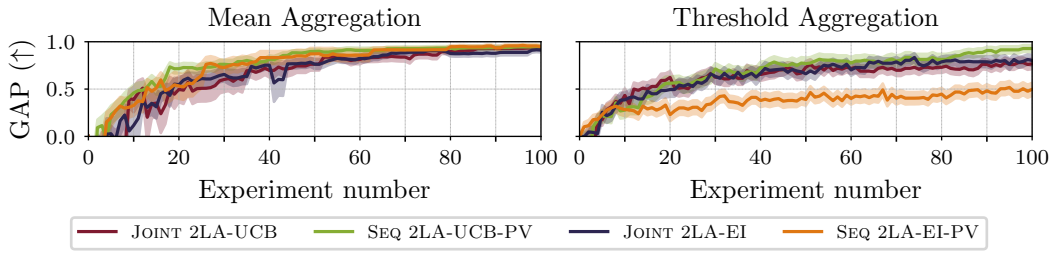


Figure 6: Optimization trajectories using sequential and joint two-step lookahead acquisition strategies. Note that, owing to computational cost constraints, the trajectories are only averaged over the N,S-Acetal formation and Deoxyfluorination reaction augmented benchmark problems.

improved optimization performance over a fully random acquisition of $\mathbf{w}_{\text{next}}$ (compare SEQ 1LA-UCB-PV and SEQ 1LA-UCB-RA). Notably, the difference becomes more pronounced for two-step lookahead acquisition policies (SEQ 2LA-UCB-PV and SEQ 2LA-UCB-RA). These findings indicate that, in the partial monitoring scenario, predictive uncertainties are not used effectively in *myopic* decision making, but their accurate propagation can improve *hyperopic* decisions. However, in the case of sequentially acquiring $\mathbf{x}_{\text{next}}$ and $\mathbf{w}_{\text{next}}$, this ability to effectively harness uncertainties for α_w does not lead to empirical performance improvements over the one-step lookahead acquisition policies. This could be attributed to the decoupling of $\mathcal{X}$ and $\mathcal{W}_{\text{train}}$.

Therefore, we evenutally benchmark acquisitions strategies that recommend $\mathbf{x}_{\text{next}}$ and $\mathbf{w}_{\text{next}}$ through a joint optimization over $\mathcal{X} \times \mathcal{W}_{\text{train}}$, as originally proposed by Toscano-Palmerin & Frazier (2018) in the context of BO with expensive integrands. Figure 6 shows a comparison of different joint acquisition strategies to the sequential strategy discussed above. Empirically, we find that jointly optimizing for $\mathbf{x}_{\text{next}}$ and $\mathbf{w}_{\text{next}}$ does not lead to improved optimization performance, both when using EI and UCB as the acquisition function. However, we find that, in the case of joint acquisition, the discrepancies between EI and UCB that are observed in the sequential case, are no longer present, showcasing that the algorithm proposed by Toscano-Palmerin & Frazier (2018) can be applied in this context. However, given the increased computational cost of joint optimization, our empirical findings suggest that the algorithmically simpler sequential acquisition strategy with one-step lookahead acquisition functions is well-suited for generality-oriented optimization for chemical reactions, and performs on par with more advanced algorithmic approaches.

5 CONCLUSION

In this work, we extend global optimization frameworks to the identification of general and transferable optima, exemplified by the real-world problem of chemical reaction condition optimization. Systematic analysis of common reaction optimization benchmarks supports the hypothesis that optimization over multiple related tasks can yield more general optima, particularly in scenarios with a low the density of high-outcome experiments across the search space. We provide augmented versions of these benchmarks to reflect these real-life considerations. For BO aimed at identifying general optima, we find that a simple and cost-effective strategy —– sequentially optimizing one-step-lookahead acquisition functions over $\mathcal{X}$ and $\mathcal{W}$ — is well-suited, and performs on par with more complex policies involving two-step lookahead acquisition. Our analyses indicate that the choice of explorative acquisition function for sampling $\mathcal{X}$ is the most influential factor in achieving successful generality-oriented optimization, likely due to the partial optimization nature of the problem. While our findings mark an important step towards applying generality-oriented optimization in chemical laboratories, they also highlight the continued need for benchmark problems that accurately reflect real-world scenarios (Liang et al., 2021). We believe that such benchmarks, along with systematic evaluations of chemical reaction representations, are essential for a principled usage of generality-oriented optimization. Building on our results, we anticipate that generality-oriented optimization will see increasing adoption in chemistry and beyond, contributing to the development of more robust, applicable and sustainable reactions.

REFERENCES

Anonymized Repository - Anonymous GitHub. URL https://anonymous.4open.science/r/general-bayesopt.

Rajeev Agrawal. Sample mean based index policies by O(log n) regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27(4):1054–1078, December 1995. ISSN 0001-8678, 1475-6064. doi: 10.2307/1427934. URL https://www.cambridge.org/core/journals/advances-in-applied-probability/article/sample-mean-based-index-policies-by-olog-n-regret-for-the-multiarmed-bandit-proble F79B49DC58E1070F6DFBE6F5D6DFD6FE.

Derek T. Ahneman, Jesús G. Estrada, Shishi Lin, Spencer D. Dreher, and Abigail G. Doyle. Predicting reaction performance in C–N cross-coupling using machine learning. *Science*, 360(6385): 186–190, April 2018. doi: 10.1126/science.aar5169. URL https://www.science.org/doi/10.1126/science.aar5169. Publisher: American Association for the Advancement of Science.

Nicholas H. Angello, Vandana Rathore, Wiktor Beker, Agnieszka Wołos, Edward R. Jira, Rafał Roszak, Tony C. Wu, Charles M. Schroeder, Alán Aspuru-Guzik, Bartosz A. Grzybowski, and Martin D. Burke. Closed-loop optimization of general reaction conditions for heteroaryl Suzuki-Miyaura coupling. *Science*, 378(6618):399–405, October 2022. doi: 10.1126/science.adc8743. URL https://www.science.org/doi/10.1126/science.adc8743. Publisher: American Association for the Advancement of Science.

Maximilian Balandat, Brian Karrer, Daniel Jiang, Samuel Daulton, Ben Letham, Andrew G Wilson, and Eytan Bakshy. BoTorch: A Framework for Efficient Monte-Carlo Bayesian Optimization. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21524–21538. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/f5b1b89d98b7286673128a5fb112cb9a-Abstract.html.

Wiktor Beker, Rafał Roszak, Agnieszka Wołos, Nicholas H. Angello, Vandana Rathore, Martin D. Burke, and Bartosz A. Grzybowski. Machine Learning May Sometimes Simply Capture Literature Popularity Trends: A Case Study of Heterocyclic Suzuki–Miyaura Coupling. *Journal of the American Chemical Society*, 144(11):4819–4827, March 2022. ISSN 0002-7863. doi: 10.1021/jacs.1c12005. URL https://doi.org/10.1021/jacs.1c12005. Publisher: American Chemical Society.

Isaiah O. Betinol, Junshan Lai, Saumya Thakur, and Jolene P. Reid. A Data-Driven Workflow for Assigning and Predicting Generality in Asymmetric Catalysis. *Journal of the American Chemical Society*, 145(23):12870–12883, June 2023. ISSN 0002-7863. doi: 10.1021/jacs.3c03989. URL https://doi.org/10.1021/jacs.3c03989. Publisher: American Chemical Society.

Ilija Bogunovic, Jonathan Scarlett, Stefanie Jegelka, and Volkan Cevher. Adversarially Robust Optimization with Gaussian Processes. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://papers.nips.cc/paper_files/paper/2018/hash/60243f9b1ac2dba11ff8131c8f4431e0-Abstract.html.

Alexander Buitrago Santanilla, Erik L. Regalado, Tony Pereira, Michael Shevlin, Kevin Bateman, Louis-Charles Campeau, Jonathan Schneeweis, Simon Berritt, Zhi-Cai Shi, Philippe Nantermet, Yong Liu, Roy Helmy, Christopher J. Welch, Petr Vachal, Ian W. Davies, Tim Cernak, and Spencer D. Dreher. Nanomole-scale high-throughput chemistry for the synthesis of complex molecules. *Science*, 347(6217):49–53, January 2015. doi: 10.1126/science.1259203. URL https://www.science.org/doi/full/10.1126/science.1259203. Publisher: American Association for the Advancement of Science.

Adam D. Clayton, Jamie A. Manson, Connor J. Taylor, Thomas W. Chamberlain, Brian A. Taylor, Graeme Clemens, and Richard A. Bourne. Algorithms for the self-optimisation of chemical reactions. *Reaction Chemistry & Engineering*, 4(9):1545–1554, August 2019. ISSN 2058-9883. doi: 10.1039/C9RE00209J. URL https://pubs.rsc.org/en/content/articlelanding/2019/re/c9re00209j. Publisher: The Royal Society of Chemistry.

Erik Daxberger, Anastasia Makarova, Matteo Turchetta, and Andreas Krause. Mixed-Variable Bayesian Optimization. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pp. 2633–2639, July 2020. doi: 10.24963/ijcai.2020/365. URL http://arxiv.org/abs/1907.01329. arXiv:1907.01329 [cs, stat].

Peter Frazier, Warren Powell, and Savas Dayanik. The Knowledge-Gradient Policy for Correlated Normal Beliefs. *INFORMS Journal on Computing*, 21(4):599–613, November 2009. ISSN 1091-9856, 1526-5528. doi: 10.1287/ijoc.1080.0314. URL https://pubsonline.informs.org/doi/10.1287/ijoc.1080.0314.

Peter I. Frazier, Warren B. Powell, and Savas Dayanik. A knowledge-gradient policy for sequential information collection. *SIAM Journal on Control and Optimization*, 47(5):2410–2439, 2008. ISSN 0363-0129. doi: 10.1137/070693424. URL https://collaborate.princeton.edu/en/publications/a-knowledge-gradient-policy-for-sequential-information-collection. Publisher: Society for Industrial and Applied Mathematics Publications.

Simone Gallarati, Puck van Gerwen, Ruben Laplaza, Lucien Brey, Alexander Makaveev, and Clemence Corminboeuf. A genetic optimization strategy with generality in asymmetric organocatalysis as a primary target. *Chemical Science*, 15(10):3640–3660, 2024. doi: 10.1039/D3SC06208B. URL https://pubs.rsc.org/en/content/articlelanding/2024/sc/d3sc06208b. Publisher: Royal Society of Chemistry.

Jacob Gardner, Geoff Pleiss, Kilian Q Weinberger, David Bindel, and Andrew G Wilson. GPyTorch: Blackbox Matrix-Matrix Gaussian Process Inference with GPU Acceleration. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://papers.nips.cc/paper_files/paper/2018/hash/27e8e17134dd7083b050476733207ea1-Abstract.html.

Roman Garnett. *Bayesian Optimization*. Cambridge University Press, 2023.

Tobias Gensch, Gabriel dos Passos Gomes, Pascal Friederich, Ellyn Peters, Théophile Gaudin, Robert Pollice, Kjell Jorner, AkshatKumar Nigam, Michael Lindner-D'Addario, Matthew S. Sigman, and Alán Aspuru-Guzik. A Comprehensive Discovery Platform for Organophosphorus Ligands for Catalysis. *Journal of the American Chemical Society*, 144(3):1205–1217, January 2022a. ISSN 0002-7863. doi: 10.1021/jacs.1c09718. URL https://doi.org/10.1021/jacs.1c09718. Publisher: American Chemical Society.

Tobias Gensch, Sleight R. Smith, Thomas J. Colacot, Yam N. Timsina, Guolin Xu, Ben W. Glasspoole, and Matthew S. Sigman. Design and Application of a Screening Set for Monophosphine Ligands in Cross-Coupling. *ACS Catalysis*, 12(13):7773–7780, July 2022b. doi: 10.1021/acscatal.2c01970. URL https://doi.org/10.1021/acscatal.2c01970. Publisher: American Chemical Society.

Ryan-Rhys Griffiths, Leo Klarner, Henry B. Moss, Aditya Ravuri, Sang Truong, Samuel Stanton, Gary Tom, Bojana Rankovic, Yuanqi Du, Arian Jamasb, Aryan Deshwal, Julius Schwartz, Austin Tripp, Gregory Kell, Simon Frieder, Anthony Bourached, Alex Chan, Jacob Moss, Chengzhi Guo, Johannes Durholt, Saudamini Chaurasia, Felix Strieth-Kalthoff, Alpha A. Lee, Bingqing Cheng, Alán Aspuru-Guzik, Philippe Schwaller, and Jian Tang. GAUCHE: A Library for Gaussian Processes in Chemistry, February 2023. URL http://arxiv.org/abs/2212.04450. arXiv:2212.04450 [cond-mat, physics:physics].

Jeff Guo, Bojana Ranković, and Philippe Schwaller. Bayesian Optimization for Chemical Reactions. *CHIMIA*, 77(1/2):31–38, February 2023. ISSN 2673-2424. doi: 10.2533/chimia.2023.31. URL https://www.chimia.ch/chimia/article/view/2023_31. Number: 1/2.

S. S. Gupta and K. J. Miescke. Bayesian look ahead one stage sampling allocations for selecting the largest normal mean. *Statistical Papers*, 35(1):169–177, December 1994. ISSN 1613-9798. doi: 10.1007/BF02926410. URL https://doi.org/10.1007/BF02926410.

Andreas T. Güntner, Sebastian Abegg, Karsten Königstein, Philipp A. Gerber, Arno Schmidt-Trucksäss, and Sotiris E. Pratsinis. Breath Sensors for Health Monitoring. *ACS Sensors*, 4(2):

268–280, February 2019. doi: 10.1021/acssensors.8b00937. URL https://doi.org/10.1021/acssensors.8b00937. Publisher: American Chemical Society.

Jeremy J. Henle, Andrew F. Zahrt, Brennan T. Rose, William T. Darrow, Yang Wang, and Scott E. Denmark. Development of a Computer-Guided Workflow for Catalyst Optimization. Descriptor Validation, Subset Selection, and Training Set Analysis. *Journal of the American Chemical Society*, 142(26):11578–11592, July 2020. ISSN 0002-7863. doi: 10.1021/jacs.0c04715. URL https://doi.org/10.1021/jacs.0c04715. Publisher: American Chemical Society.

José Miguel Hernández-Lobato, James Requeima, Edward O. Pyzer-Knapp, and Alán Aspuru-Guzik. Parallel and Distributed Thompson Sampling for Large-scale Accelerated Exploration of Chemical Space. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1470–1479. PMLR, July 2017. URL https://proceedings.mlr.press/v70/hernandez-lobato17a.html. ISSN: 2640-3498.

Hisham Husain, Vu Nguyen, and Anton van den Hengel. Distributionally Robust Bayesian Optimization with φ-divergences. *Advances in Neural Information Processing Systems*, 36:20133–20145, December 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/3feb8ed3c33c3310b45f80be7dfef707-Abstract-Conference.html.

Florian Häse, Matteo Aldeghi, Riley J. Hickman, Loïc M. Roch, Melodie Christensen, Elena Liles, Jason E. Hein, and Alán Aspuru-Guzik. Olympus: a benchmarking framework for noisy optimization and experiment planning. *Machine Learning: Science and Technology*, 2 (3):035021, July 2021. ISSN 2632-2153. doi: 10.1088/2632-2153/abedc8. URL https://dx.doi.org/10.1088/2632-2153/abedc8. Publisher: IOP Publishing.

Shali Jiang, Henry Chai, Javier Gonzalez, and Roman Garnett. BINOCULARS for efficient, nonmyopic sequential experimental design. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 4794–4803. PMLR, November 2020. URL https://proceedings.mlr.press/v119/jiang20b.html. ISSN: 2640-3498.

Donald R. Jones, Matthias Schonlau, and William J. Welch. Efficient Global Optimization of Expensive Black-Box Functions. *Journal of Global Optimization*, 13(4):455–492, December 1998. ISSN 1573-2916. doi: 10.1023/A:1008306431147. URL https://doi.org/10.1023/A:1008306431147.

Leslie Pack Kaelbling. Associative Reinforcement Learning: A Generate and Test Algorithm. *Machine Learning*, 15(3):299–319, June 1994a. ISSN 1573-0565. doi: 10.1023/A:1022642026684. URL https://doi.org/10.1023/A:1022642026684.

Leslie Pack Kaelbling. Associative Reinforcement Learning: Functions in k-DNF. *Machine Learning*, 15(3):279–298, June 1994b. ISSN 1573-0565. doi: 10.1023/A:1022689909846. URL https://doi.org/10.1023/A:1022689909846.

Johannes Kirschner, Ilija Bogunovic, Stefanie Jegelka, and Andreas Krause. Distributionally Robust Bayesian Optimization. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pp. 2174–2184. PMLR, June 2020a. URL https://proceedings.mlr.press/v108/kirschner20a.html. ISSN: 2640-3498.

Johannes Kirschner, Tor Lattimore, and Andreas Krause. Information Directed Sampling for Linear Partial Monitoring, February 2020b. URL http://arxiv.org/abs/2002.11182. arXiv:2002.11182 [cs, stat].

Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, and Vincent Fortuin. Promises and Pitfalls of the Linearized Laplace in Bayesian Optimization, July 2023. URL http://arxiv.org/abs/2304.08309. arXiv:2304.08309 [cs, stat].

Agustinus Kristiadi, Felix Strieth-Kalthoff, Marta Skreta, Pascal Poupart, Alán Aspuru-Guzik, and Geoff Pleiss. A Sober Look at LLMs for Material Discovery: Are They Actually Good for Bayesian Optimization Over Molecules?, May 2024. URL http://arxiv.org/abs/2402.05015. arXiv:2402.05015 [cs].

Gregory Landrum. RDKit: Open-source cheminformatics, 2023. URL http://www.rdkit.org.

Tor Lattimore. Minimax Regret for Partial Monitoring: Infinite Outcomes and Rustichini's Regret, February 2022. URL http://arxiv.org/abs/2202.10997. arXiv:2202.10997 [cs, math].

Tor Lattimore and Andras Gyorgy. Mirror Descent and the Information Ratio. In *Proceedings of Thirty Fourth Conference on Learning Theory*, pp. 2965–2992. PMLR, July 2021. URL https://proceedings.mlr.press/v134/lattimore21b.html. ISSN: 2640-3498.

Tor Lattimore and Csaba Szepesvari. An Information-Theoretic Approach to Minimax Regret in Partial Monitoring, May 2019. URL http://arxiv.org/abs/1902.00470. arXiv:1902.00470 [cs, math, stat].

Tor Lattimore and Csaba Szepesvári. Cleaning up the neighborhood: A full classification for adversarial partial monitoring. In *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, pp. 529–556. PMLR, March 2019. URL https://proceedings.mlr.press/v98/lattimore19a.html. ISSN: 2640-3498.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 1 edition, July 2020. ISBN 978-1-108-57140-1 978-1-108-48682-8. doi: 10.1017/9781108571401. URL https://www.cambridge.org/core/product/identifier/9781108571401/type/book.

Yucen Lily Li, Tim G. J. Rudner, and Andrew Gordon Wilson. A Study of Bayesian Neural Network Surrogates for Bayesian Optimization, May 2024. URL http://arxiv.org/abs/2305.20028. arXiv:2305.20028 [cs, stat].

Qiaohao Liang, Aldair E. Gongora, Zekun Ren, Armi Tiihonen, Zhe Liu, Shijing Sun, James R. Deneault, Daniil Bash, Flore Mekki-Berrada, Saif A. Khan, Kedar Hippalgaonkar, Benji Maruyama, Keith A. Brown, John Fisher III, and Tonio Buonassisi. Benchmarking the performance of Bayesian optimization across multiple experimental materials science domains. *npj Computational Materials*, 7(1):1–10, November 2021. ISSN 2057-3960. doi: 10.1038/s41524-021-00656-9. URL https://www.nature.com/articles/s41524-021-00656-9. Publisher: Nature Publishing Group.

H. L. Morgan. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *Journal of Chemical Documentation*, 5(2):107–113, May 1965. ISSN 0021-9576. doi: 10.1021/c160017a018. URL https://doi.org/10.1021/c160017a018. Publisher: American Chemical Society.

J. Močkus. On Bayesian Methods for Seeking the Extremum. In G. I. Marchuk (ed.), *Optimization Techniques IFIP Technical Conference: Novosibirsk, July 1–7, 1974*, pp. 400–404. Springer, Berlin, Heidelberg, 1975. ISBN 978-3-662-38527-2. doi: 10.1007/978-3-662-38527-2_55. URL https://doi.org/10.1007/978-3-662-38527-2_55.

Thanh Nguyen, Sunil Gupta, Huong Ha, Santu Rana, and Svetha Venkatesh. Distributionally Robust Bayesian Quadrature Optimization. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pp. 1921–1931. PMLR, June 2020. URL https://proceedings.mlr.press/v108/nguyen20a.html. ISSN: 2640-3498.

Matthew K. Nielsen, Derek T. Ahneman, Orestes Riera, and Abigail G. Doyle. Deoxyfluorination with Sulfonyl Fluorides: Navigating Reaction Space with Machine Learning. *Journal of the American Chemical Society*, 140(15):5004–5008, April 2018. ISSN 0002-7863. doi: 10.1021/jacs.8b01523. URL https://doi.org/10.1021/jacs.8b01523. Publisher: American Chemical Society.

AkshatKumar Nigam, Robert Pollice, Mario Krenn, Gabriel dos Passos Gomes, and Alán Aspuru-Guzik. Beyond generative models: superfast traversal, optimization, novelty, exploration and discovery (STONED) algorithm for molecules using SELFIES. *Chemical Science*, 12(20):7079–7090, May 2021. ISSN 2041-6539. doi: 10.1039/D1SC00231G. URL https://pubs.rsc.org/en/content/articlelanding/2021/sc/d1sc00231g. Publisher: The Royal Society of Chemistry.

Cesar N. Prieto Kullmer, Jacob A. Kautzky, Shane W. Krska, Timothy Nowak, Spencer D. Dreher, and David W. C. MacMillan. Accelerating reaction generality and mechanistic insight through additive mapping. *Science*, 376(6592):532–539, April 2022. doi: 10.1126/science.abn1885. URL https://www.science.org/doi/full/10.1126/science.abn1885. Publisher: American Association for the Advancement of Science.

Siriphan Promjun and Nopadon Sahachaisaeree. Factors Determining Athletic Footwear Design: A Case of Product Appearance and Functionality. *Procedia - Social and Behavioral Sciences*, 36: 520–528, January 2012. ISSN 1877-0428. doi: 10.1016/j.sbspro.2012.03.057. URL https://www.sciencedirect.com/science/article/pii/S1877042812005241.

Debanjan Rana, Philipp M. Pflüger, Niklas P. Hölter, Guangying Tan, and Frank Glorius. Standardizing Substrate Selection: A Strategy toward Unbiased Evaluation of Reaction Generality. *ACS Central Science*, 10(4):899–906, April 2024. ISSN 2374-7943. doi: 10.1021/acscentsci.3c01638. URL https://doi.org/10.1021/acscentsci.3c01638. Publisher: American Chemical Society.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

Jonas Rein, Soren D. Rozema, Olivia C. Langner, Samson B. Zacate, Melissa A. Hardy, Juno C. Siu, Brandon Q. Mercado, Matthew S. Sigman, Scott J. Miller, and Song Lin. Generality-oriented optimization of enantioselective aminoxyl radical catalysis. *Science*, 380(6646):706–712, May 2023. doi: 10.1126/science.adf6177. URL https://www.science.org/doi/10.1126/science.adf6177. Publisher: American Association for the Advancement of Science.

Aldo Rustichini. Minimizing Regret: The General Case. *Games and Economic Behavior*, 29(1): 224–243, October 1999. ISSN 0899-8256. doi: 10.1006/game.1998.0690. URL https://www.sciencedirect.com/science/article/pii/S089982569890690X.

Frederik Sandfort, Felix Strieth-Kalthoff, Marius Kühnemund, Christian Beecks, and Frank Glorius. A Structure-Based Platform for Predicting Chemical Reactivity. *Chem*, 6(6):1379–1390, June 2020. ISSN 2451-9294. doi: 10.1016/j.chempr.2020.02.017. URL https://www.sciencedirect.com/science/article/pii/S2451929420300851.

Stefan P. Schmid, Leon Schlosser, Frank Glorius, and Kjell Jorner. Catalysing (organo-)catalysis: Trends in the application of machine learning to enantioselective organocatalysis. *Beilstein Journal of Organic Chemistry*, 20(1):2280–2304, September 2024. ISSN 1860-5397. doi: 10.3762/bjoc.20.196. URL https://www.beilstein-journals.org/bjoc/articles/20/196. Publisher: Beilstein-Institut.

Tobias Schnitzer, Martin Schnurr, Andrew F. Zahrt, Nader Sakhaee, Scott E. Denmark, and Helma Wennemers. Machine Learning to Develop Peptide CatalystsSuccesses, Limitations, and Opportunities. *ACS Central Science*, 10(2):367–373, February 2024. ISSN 2374-7943. doi: 10.1021/acscentsci.3c01284. URL https://doi.org/10.1021/acscentsci.3c01284. Publisher: American Chemical Society.

Benjamin J. Shields, Jason Stevens, Jun Li, Marvin Parasram, Farhan Damani, Jesus I. Martinez Alvarado, Jacob M. Janey, Ryan P. Adams, and Abigail G. Doyle. Bayesian reaction optimization as a tool for chemical synthesis. *Nature*, 590(7844):89–96, February 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03213-y. URL https://www.nature.com/articles/s41586-021-03213-y. Publisher: Nature Publishing Group.

Jason M. Stevens, Jun Li, Eric M. Simmons, Steven R. Wisniewski, Stacey DiSomma, Kenneth J. Fraunhoffer, Peng Geng, Bo Hao, and Erika W. Jackson. Advancing Base Metal Catalysis through Data Science: Insight and Predictive Models for Ni-Catalyzed Borylation through Supervised Machine Learning. *Organometallics*, 41(14):1847–1864, July 2022. ISSN 0276-7333. doi: 10.1021/acs.organomet.2c00089. URL https://doi.org/10.1021/acs.organomet.2c00089. Publisher: American Chemical Society.

Felix Strieth-Kalthoff, Frederik Sandfort, Marius Kühnemund, Felix R. Schäfer, Herbert Kuchen, and Frank Glorius. Machine Learning for Chemical Reactivity: The Importance of Failed Experiments. *Angewandte Chemie International Edition*, 61(29): e202204647, 2022. ISSN 1521-3773. doi: 10.1002/anie.202204647. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.202204647. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.202204647.

Kevin Swersky, Jasper Snoek, and Ryan P Adams. Multi-Task Bayesian Optimization. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper/2013/hash/f33ba15effa5c10e873bf3842afb46a6-Abstract.html.

William R. Thompson. On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika*, 25(3/4):285–294, 1933. ISSN 0006-3444. doi: 10.2307/2332286. URL https://www.jstor.org/stable/2332286. Publisher: [Oxford University Press, Biometrika Trust].

Gary Tom, Stefan P. Schmid, Sterling G. Baird, Yang Cao, Kourosh Darvish, Han Hao, Stanley Lo, Sergio Pablo-García, Ella M. Rajaonson, Marta Skreta, Naruki Yoshikawa, Samantha Corapi, Gun Deniz Akkoc, Felix Strieth-Kalthoff, Martin Seifrid, and Alán Aspuru-Guzik. Self-Driving Laboratories for Chemistry and Materials Science. *Chemical Reviews*, 124(16):9633–9732, August 2024. ISSN 0009-2665. doi: 10.1021/acs.chemrev.4c00055. URL https://doi.org/10.1021/acs.chemrev.4c00055. Publisher: American Chemical Society.

Saul Toscano-Palmerin and Peter I. Frazier. Bayesian Optimization with Expensive Integrands, March 2018. URL http://arxiv.org/abs/1803.08661. arXiv:1803.08661.

Corin C. Wagen, Spencer E. McMinn, Eugene E. Kwan, and Eric N. Jacobsen. Screening for generality in asymmetric catalysis. *Nature*, 610(7933):680–686, October 2022. ISSN 1476-4687. doi: 10.1038/s41586-022-05263-2. URL https://www.nature.com/articles/s41586-022-05263-2. Publisher: Nature Publishing Group.

Jason Y. Wang, Jason M. Stevens, Stavros K. Kariofillis, Mai-Jan Tom, Dung L. Golden, Jun Li, Jose E. Tabora, Marvin Parasram, Benjamin J. Shields, David N. Primer, Bo Hao, David Del Valle, Stacey DiSomma, Ariel Furman, G. Greg Zipp, Sergey Melnikov, James Paulson, and Abigail G. Doyle. Identifying general reaction conditions by bandit optimization. *Nature*, 626(8001):1025–1033, February 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07021-y. URL https://www.nature.com/articles/s41586-024-07021-y. Publisher: Nature Publishing Group.

Brian Williams, Thomas Santner, and William Notz. Sequential Design of Computer Experiments to Minimize Integrated Response Functions. *Statistica Sinica*, 10:1133–1152, October 2000.

Jing Xie, Peter I. Frazier, Sethuraman Sankaran, Alison Marsden, and Saleh Elmohamed. Optimization of computationally expensive simulations with Gaussian processes and parameter uncertainty: Application to cardiovascular surgery. In *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 406–413, October 2012. doi: 10.1109/Allerton.2012.6483247. URL https://ieeexplore.ieee.org/abstract/document/6483247.

Andrew F. Zahrt, Jeremy J. Henle, Brennan T. Rose, Yang Wang, William T. Darrow, and Scott E. Denmark. Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning. *Science*, 363(6424):eaau5631, January 2019. doi: 10.1126/science.aau5631. URL https://www.science.org/doi/10.1126/science.aau5631. Publisher: American Association for the Advancement of Science.

A Appendix

A.1 Bayesian Optimization for Generality

A.1.1 Aggregation Functions

The aggregation function is a user-defined property that determines how the "set optimum" is calculated across objective functions. Through the choice of the set optimum, prior knowledge and preferences about the specific optimization problem at hand can be included. In this work, the following aggregation functions are evaluated:

Mean Aggregation

$$\phi(f(\mathbf{x};\mathbf{w}),\mathcal{W}) = \frac{1}{|\mathcal{W}|}\sum_{\mathbf{w}\in\mathcal{W}} f(\mathbf{x};\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{x};\mathbf{w}_i) \tag{4}$$

Threshold Aggregation

$$\phi\big(f(\mathbf{x};\mathbf{w}),\mathcal{W}\big) = \sum_{\mathbf{w}\in\mathcal{W}} \sigma\big(f(\mathbf{x};\mathbf{w}) - f_{\text{thr}}\big) = \sum_{i=1}^{n} \sigma\big(f(\mathbf{x};\mathbf{w}_i) - f_{\text{thr}}\big) \tag{5}$$

Conceivably, other aggregation functions also have practical use-cases, for example:

Mean Squared Error (MSE) Aggregation

$$\phi\big(f(\mathbf{x};\mathbf{w}),\mathcal{W}\big) = -\frac{1}{|\mathcal{W}|}\sum_{\mathbf{w}\in\mathcal{W}} \big(f_{\text{opt}}(\mathbf{x};\mathbf{w}) - f(\mathbf{x};\mathbf{w})\big)^2 = -\frac{1}{n}\sum_{i=1}^{n} \big(f_{\text{opt},i} - f(\mathbf{x};\mathbf{w}_i)\big)^2 \tag{6}$$

Minimum Aggregation

$$\phi\big(f(\mathbf{x};\mathbf{w}),\mathcal{W}\big) = \min_{\mathbf{w}_i\in\mathcal{W}} f(\mathbf{x};\mathbf{w}_i) \tag{7}$$

The above definitions assume that all $f(\mathbf{x};\mathbf{w}_i)$ have the same range, and that the optimization problem is formulated as maximization problem.

A.1.2 Acquisition Functions and the Sample Average Approximation

For the evaluation of posterior distributions, and the calculation of acquisition function values, we use the sample-average approximation, as introduced by Balandat et al. (2020). From a posterior distribution at time point k, $p\big(g_k(\mathbf{x})\big)$, M posterior samples $\zeta_m(\mathbf{x}) \sim p\big(g_k(\mathbf{x})\big)$ are drawn. These posterior samples can be used to estimate the posterior distribution, and to calculate acquisition function values as expectation values $\mathbb{E}_M$ over all M samples.

Herein, we use the following common acquisition functions:

- Upper Confidence Bound: $\text{UCB}(\mathbf{x}) = \mathbb{E}_M\big(\zeta_m(\mathbf{x})\big) + \beta \cdot \mathbb{E}_M\big(\zeta_m(\mathbf{x}) - \mathbb{E}_M(\zeta_m(\mathbf{x}))\big)$.
- Expected Improvement: $\text{EI}(\mathbf{x}) = \mathbb{E}_M\big(\zeta_m(\mathbf{x}) - f^*\big)$, where f^* is the best value observed so far.
- Posterior Variance: $\text{PV}(\mathbf{x}) = \mathbb{E}_M\big(\zeta_m(\mathbf{x}) - \mathbb{E}_M(\zeta_m(\mathbf{x}))\big)$.
- Random Selection, where the acquisition function value is a random number.

Moreover, we evaluate the optimization performance using a primitive implementation of two-step lookahead acquisition functions α^* (see Algorithm 4). The acquisition function value of α^* at a location $\mathbf{x_0}$ is estimated as follows: For each of the M posterior samples $\zeta_m(\mathbf{x_0}) \sim p\big(g_k(\mathbf{x_0})\big)$, a fantasy posterior distribution $p'\big(\phi(g_{k+1}(\mathbf{x_0}))\big)$ is generated by conditioning the posterior on the

new observation $(\mathbf{x_0}, \zeta_M(\mathbf{x_0}))$ and aggregation. From this fantasy posterior distribution, the values of the inner acquisition function α_m can be computed and optimized over $\mathbf{x} \in \mathcal{X}$. The final value of the two-step lookahead acquisition function is returned as $\alpha^*(\mathbf{x}_0) = \frac{1}{M} \sum_{m=1}^{M} \alpha_m$.

Algorithm 4 Two-step lookahead acquisition function using the sample average approximation.

Input:
input space $\mathcal{X}$
location $\mathbf{x_0}$ at which to evaluate the two-step lookahead acquisition function
aggregation function $\phi\big(f(\mathbf{x}; \mathbf{w}), \mathcal{W}\big)$
posterior distribution $p\big(g_k(\mathbf{x}) \mid \mathcal{D}\big)$
one-step lookahead acquisition function $\alpha(\mathbf{x})$

1: draw M posterior samples $\zeta_m(\mathbf{x_0}) \sim p\big(g_k(\mathbf{x_0})\big)$
2: empty set of fantasy acquisition function values $\mathcal{A} = \{\}$
3: **for** $m = 1, \ldots, M$ **do**
4: compute fantasy posterior $p'(\mathbf{x}) = p\Big(\phi\big(g_{k+1}(\mathbf{x}) \mid (\mathcal{D} \cup (\mathbf{x_0}, \zeta_m(\mathbf{x_0})))\big)\Big)$
5: optimize one-step-lookahead acquisition function $\alpha_m = \max_{\mathbf{x} \in \mathcal{X}} \alpha(p'(\mathbf{x}))$
6: update $\mathcal{A} = \mathcal{A} \cup \{\alpha_m\}$
7: **end for**
8: **return** $\alpha^*(\mathbf{x}_0) = \frac{1}{M} \sum_{m=1}^{M} \alpha_m$

A.1.3 Benchmarked Optimization Strategies for Selecting $\mathbf{x}_{\text{next}}$ and $\mathbf{w}_{\text{next}}$

Herein, we outline the use of the benchmarked optimization strategies for generality-oriented optimization. The discussed optimization strategies describe different variations of how to pick the next experiments $\mathbf{x}_{\text{next}}$ and $\mathbf{w}_{\text{next}}$.

Following the SAA (Balandat et al., 2020) outlined above, we estimate the predictive posterior distribution $p\big(\phi(\mathbf{x}) \mid \mathcal{D}\big)$ as follows: For each $\mathbf{w}_i \in \mathcal{W}$, M (typically $M = 512$ for one-step lookahead strategies and $M = 3$ for two-step lookahead strategies to reduce computational costs) samples $\zeta_{im}(\mathbf{x}) \sim p\big(g_k(\mathbf{x}, \mathbf{w}_i)\big)$ are drawn from the posterior distribution of the surrogate model. Aggregating over all $\mathbf{w}_i$ yields M samples $\zeta_m(\mathbf{x}) \sim p\big(\phi(\mathbf{x}) \mid \mathcal{D}\big)$ from the posterior distribution over $\phi(\mathbf{x})$, which can be used for calculating the acquisition function values using the sample-based acquisition function logic, as described in Appendix A.1.2. With this, we implement and benchmark the acquisition policies in Table 2.

The sequential acquisition is described in Algorithm 2 and refers to a strategy in which $\mathbf{x}_{\text{next}}$ and $\mathbf{w}_{\text{next}}$ are selected sequentially. In the first step, $\mathbf{x}_{\text{next}}$ is selected by optimizing an $\mathbf{x}$-specific acquisition function α_x over $\mathbf{x} \in \mathcal{X}$. With the selected $\mathbf{x}_{\text{next}}$ in hand, $\mathbf{w}_{\text{next}}$ is then selected by optimizing an independent, $\mathbf{w}$-specific acquisition function over $\mathbf{w} \in \mathcal{W}$. With $\alpha_x = \text{PI}$ (Probability of Improvement) and $\alpha_w = \text{PV}$, this would correspond to the strategy described in (Angello et al., 2022). In contrast, the joint acquisition, as outlined in Algorithm 3, refers to a strategy in which $\mathbf{x}_{\text{next}}$ and $\mathbf{w}_{\text{next}}$ are selected jointly through optimization of a two-step lookahead acquisition function (see Algorithm 4 and Appendix A.1.2).

A.2 Benchmark Problem Details

A.2.1 Original Benchmark Problems

Four chemical reaction benchmarks have been considered in this work: Reactant conversion optimization for Pd-catalyzed C–heteroatom couplings (Buitrago Santanilla et al., 2015), enantioselectivity optimization for a N,S-Acetal formation (Zahrt et al., 2019), yield optimization for a borylation reaction (Stevens et al., 2022; Wang et al., 2024) and yield optimization for deoxyfluorination reaction (Nielsen et al., 2018; Wang et al., 2024). Since it has been well-demonstrated that these problems can be effectively modeled by regression approaches (Zahrt et al., 2019; Ahneman et al., 2018; Sandfort et al., 2020), we trained a random forest regressor on each dataset, which was used as

Table 2: Nomenclature and description of the all benchmarked acquisition strategies and acquisition functions, as discussed in the main text and the Appendix. Each experiment is named according to the acquisition strategy used, followed by specifications of the used acquisition functions α_x and α_w or α for sequential and joint acquisitions, respectively. As an example, a sequential two-step lookahead acquisition strategy with an Upper Confidence Bound as α_x and Posterior Variance as α_w, is referred to as SEQ 2LA-UCB-PV.

Acquisition Strategy	Acquisition Function
SEQ 1LA: Sequential acquisition of $\mathbf{x}_{\text{next}}$ and $\mathbf{w}_{\text{next}}$, each using a one-step lookahead acquisition function. The final $\hat{\mathbf{x}}$ is selected greedily.	UCB: Upper confidence bound ($\beta = 0.5$).
SEQ 2LA: Sequential acquisition of $\mathbf{x}_{\text{next}}$ and $\mathbf{w}_{\text{next}}$, each using a two-step lookahead acquisition function. The final $\hat{\mathbf{x}}$ is selected greedily.	UCBE: Upper confidence bound ($\beta = 5$).
JOINT 2LA: Joint acquisition of $\mathbf{x}_{\text{next}}$ and $\mathbf{w}_{\text{next}}$ using a two-step lookahead acquisition function. The final $\hat{\mathbf{x}}$ is selected greedily.	EI: Expected Improvement.
BANDIT: Multi-armed bandit algorithm as implemented by Wang et al. (2024).	PV: Posterior Variance.
RANDOM: Random selection of the final $\hat{\mathbf{x}}$.	RA: Random acquisition.
	SINGLE: Selection of the same substrate ($\mathbf{w}$) for every iteration.
	COMPLETE: Selection of every substrate (i.e. every $\mathbf{w} \in \mathcal{W}$) for a selected $\mathbf{x}_{\text{next}}$.

the ground truth for all benchmark experiments (Häse et al., 2021). In the following, the benchmark problems are described briefly.

Pd-catalyzed carbon-heteroatom coupling

The Pd-catalyzed carbon-heteroatom coupling benchmark is concerned with the reaction of different nucleophiles with 3-bromopyridine (Figure 7). In total, 16 different nucleophiles were tested in a nanoscale high-throughput experimentation platform. As reaction conditions, bases (six different bases) and catalysts (16 different catalysts) were varied. In total, the benchmark consists of 1536 different experiments, for which the conversion is reported.

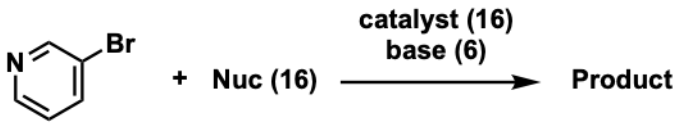


Figure 7: Reaction diagram of the Pd-catalyzed carbon-heteroatom coupling, where 3-bromopyridine reacts with a nucleophile. Reaction conditions include a catalyst and a base. The numbers indicate the amount of different species in the benchmark.

The average conversion is 2.05%, whereas the maximum conversion is 39.81% (Figure 8). The average of the average conversion of each condition is 2.05%, while the maximum of the average conversion of the conditions is 7.60% (Figure 8). The catalyst-base combination with the highest average conversion is shown in Figure 8.

With respect to the threshold aggregation function, the chosen threshold was 7.50%. The average number of substrates with a conversion above this threshold are 1.615, while the maximum number of substrates is 7 (Figure 9). The catalyst-base combination with the highest number of substrates with a conversion above the threshold is the same as shown in Figure 8.

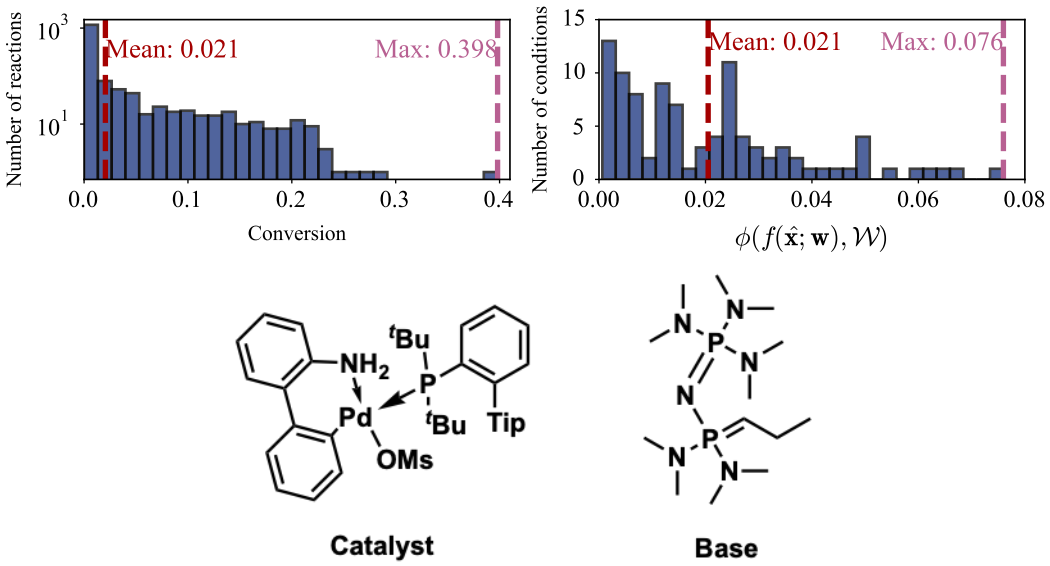


Figure 8: Top left: Distribution of the conversion for the Pd-catalyzed carbon-heteroatom coupling in the original benchmark. Top right: Distribution of the average conversion for each catalyst-base combination for the Pd-catalyzed carbon-heteroatom coupling in the original benchmark. Bottom: Catalyst-base combination with the highest average conversion in the original benchmark. Tip = 2,4,6-triisopropylphenyl.

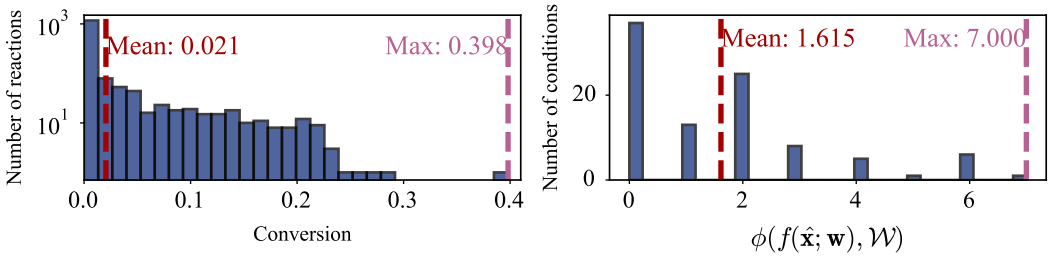


Figure 9: Left: Distribution of the conversion for the Pd-catalyzed carbon-heteroatom coupling in the original benchmark. Right: Distribution of the number of substrates with a conversion above the specified threshold for each catalyst-base combination for the Pd-catalyzed carbon-heteroatom coupling in the original benchmark.

N,S-Acetal formation

The N,S-Acetal formation benchmark is concerned with the nucleophilic addition of different thiols to imines, catalyzed by chiral phosphoric acids (CPAs) (see Figure 10). In total, five different imines and five different thiols were tested in manual experiments. As reaction conditions, 43 different CPA catalysts were considered. In total, the benchmark consists of 1075 different experiments, for which $\Delta\Delta G^{\ddagger}$, as a measure of the enantioselectivity, is reported.

The average $\Delta\Delta G^{\ddagger}$ is 0.988 kcal/mol, whereas the maximum $\Delta\Delta G^{\ddagger}$ is 3.135 kcal/mol (see Figure 11). The average of the average $\Delta\Delta G^{\ddagger}$ for each condition is 0.988 kcal/mol, while the maximum of the average $\Delta\Delta G^{\ddagger}$ for all conditions is 2.395 kcal/mol (see Figure 11). The catalyst with the highest average $\Delta\Delta G^{\ddagger}$ is shown in Figure 11.

With respect to the threshold aggregation function, the chosen threshold was 2.0 kcal/mol. The average number of substrates with $\Delta\Delta G^{\ddagger}$ above this threshold are 1.907, while the maximum number of substrates is 17 (Figure 12). The catalyst with the highest number of substrates with $\Delta\Delta G^{\ddagger}$ above the threshold is the same as shown in Figure 11.

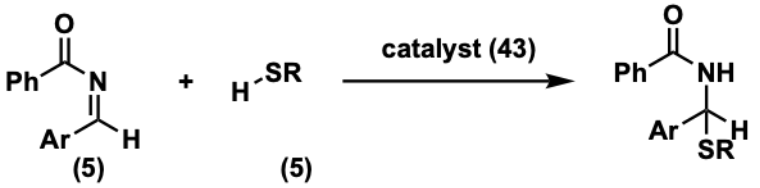
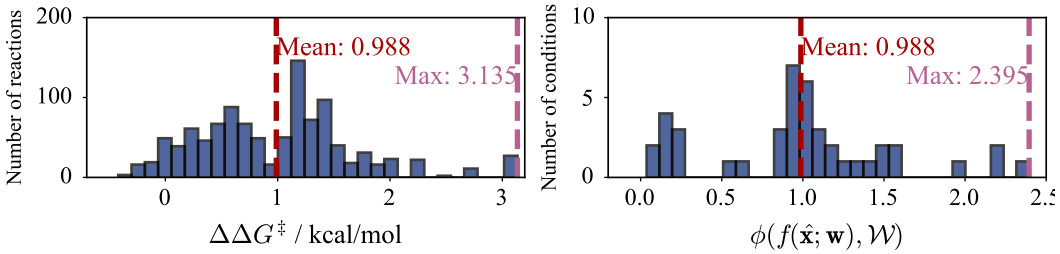


Figure 10: Reaction diagram of the N,S-Acetal formation, where an imine reacts with a thiol. Reaction conditions include a catalyst. The numbers indicate the amount of different species in the benchmark.



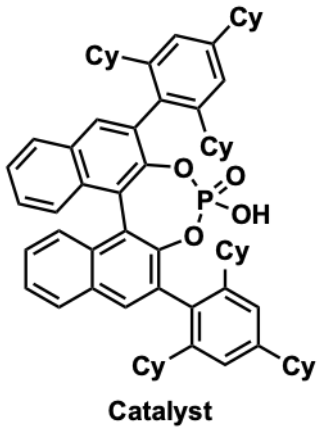


Figure 11: Top left: Distribution of $\Delta\Delta G^{\ddagger}$ for the N,S-Acetal formation in the original benchmark. Top right: Distribution of the average $\Delta\Delta G^{\ddagger}$ for each catalyst for the N,S-Acetal formation in the original benchmark. Bottom: Catalyst with the highest average $\Delta\Delta G^{\ddagger}$ in the original benchmark. Cy = Cyclohexyl

Borylation reaction

The borylation reaction benchmark is concerned with the Ni-catalyzed borylation of different aryl electrophiles (aryl chlorides, aryl bromides, and aryl sulfamates) (Figure 13). In total, 33 different aryl electrophiles were tested. As reaction conditions, ligands (23 different ligands), and solvents (2 different solvents) were varied. In total, the benchmark consists of 1518 different experiments, for which the yield is reported.

The average yield is 45.5%, whereas the maximum yield is 100.0% (Figure 14). The average of the average yield of each condition is 45.5%, while the maximum of the average yield of the conditions is 65.4% (Figure 14). The ligand-solvent combination with the highest average yield is shown in Figure 14.

With respect to the threshold aggregation function, the chosen threshold was 90%. The average number of substrates with a yield above this threshold are 1.457, while the maximum number of substrates is 5 (Figure 15). The ligand-solvent combination with the highest number of substrates with a yield above the threshold is the same as shown in Figure 14. However, the shown ligand-solvent combination is only one of four combinations.

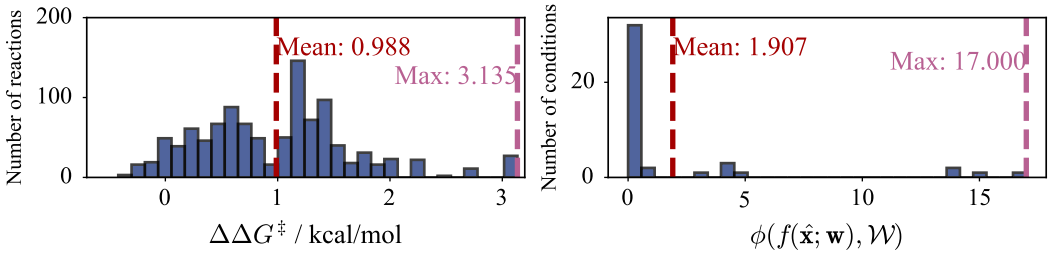


Figure 12: Left: Distribution of $\Delta\Delta G^{\ddagger}$ for the N,S-Acetal formation in the original benchmark. Right: Distribution number of substrates with a $\Delta\Delta G^{\ddagger}$ above the specified threshold for each catalyst for the N,S-Acetal formation in the original benchmark.

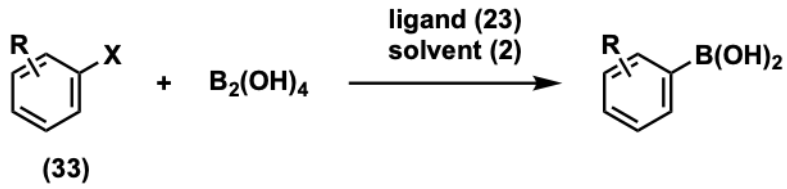


Figure 13: Reaction diagram of the borylation reaction, where different aryl electrophiles are borylated. Reaction conditions include a ligand, and a solvent. The numbers indicate the amount of different species in the benchmark.

Deoxyfluorination reaction

The deoxyfluorination reaction benchmark is concerned with the transformation of different alcohols into the corresponding fluorides (Figure 16). In total, 37 different alcohols were tested. As reaction conditions, sulfonyl fluorides (fluoride sources, five different fluorides) and bases (four different bases) were varied. In total, the benchmark consists of 740 different experiments, for which the yield is reported.

The average yield is 40.4%, whereas the maximum yield is 100.6% (Figure 17). The yield larger than 100% is contained in the originally published dataset. The average of the average yield of each condition is 40.4%, while the maximum of the average yield of the conditions is 57.2% (Figure 17). The fluoride-base combination with the highest average yield is shown in Figure 17.

With respect to the threshold aggregation function, the chosen threshold was 90%. The average number of substrates with a yield above this threshold are 1.400, while the maximum number of substrates is 5 (Figure 18). The fluoride-base combination with the highest number of substrates with a yield above the threshold is shown in Figure 18.

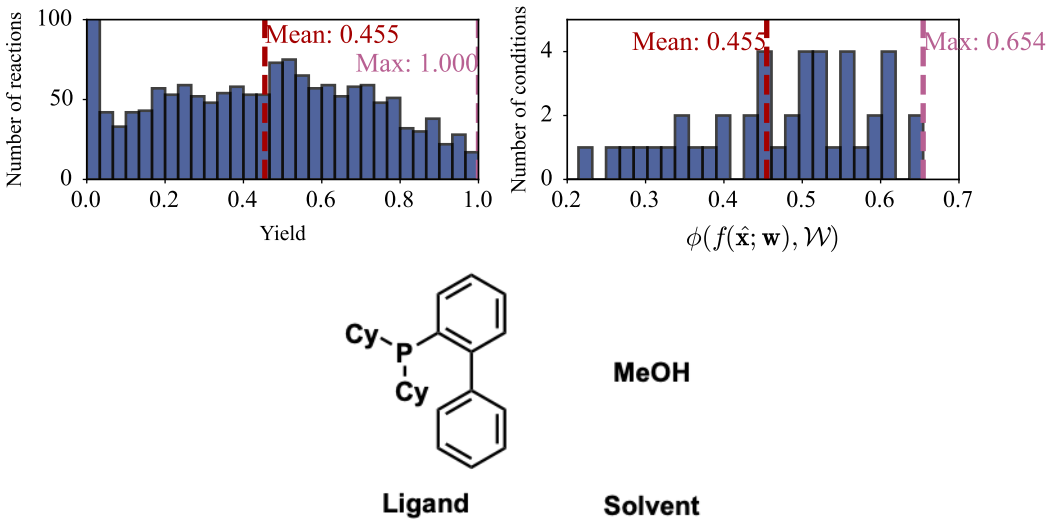


Figure 14: Top left: Distribution of the yield for the borylation reaction in the original benchmark. Top right: Distribution of the average yield for each ligand-solvent combination for the borylation reaction in the original benchmark. Bottom: Ligand-solvent combination with the highest average yield in the original benchmark. Cy = Cyclohexyl.

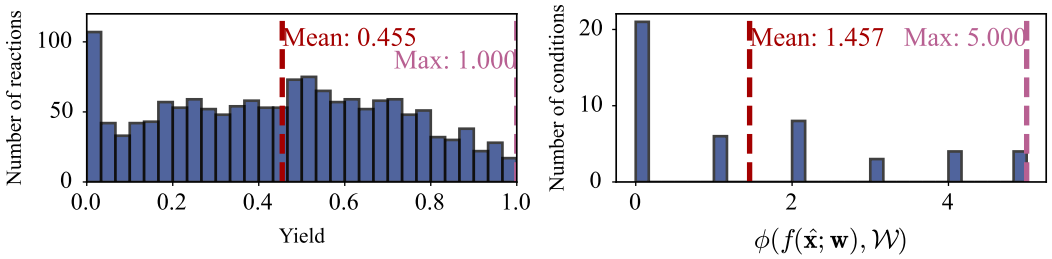


Figure 15: Left: Distribution of the yield for the borylation reaction in the original benchmark. Right: Distribution of the number of substrates with a yield above the specified threshold for each ligand-solvent combination for the borylation reaction in the original benchmark.

A.2.2 AUGMENTATION

Since the described benchmarks consist of a high number of high-outcome experiments (the respective search spaces were rationally designed by expert chemists), we augment them with more negative examples to make them more relevant to real-world optimization campaigns. New substrates are generated by mutating the originally reported substrates via the STONED algorithm (Nigam et al., 2021). In a first filtering step, new substrates were removed if they had a Tanimoto similarity to the original substrate smaller than 0.75 (0.6 for the borylation reaction to obtain a reasonable number of additinal substrates) or if they did not possess the functional groups required for the reaction. To ensure that the benchmark is augmented with negative examples, random forests are fitted to the original benchmarks (see above). The mean absolute errors (MAEs), root mean square errors (RMSEs) and r^2 score (r^2), Spearman's rank correlation coefficient (Spearman's ρ) of the random forest regressors fitted to and evaluated on the original benchmarks are shown in Table 3. In addition, to evaluate the predictive utility of the random forest regressors, we perform 5-fold cross validation on the original benchmark. The MAE, RMSE, r^2 and Spearman's ρ of the 5-fold cross validation are reported in Table 4. Even though the predictive performance on the CV does not achieve a high Spearman's rank coefficient, the comparably low MAEs and RMSEs, as well as high r^2 values suggest that they are a reasonable oracle. Newly generated substrates were incorporated if the average reaction outcome over all reported reaction conditions is below a defined threshold. The

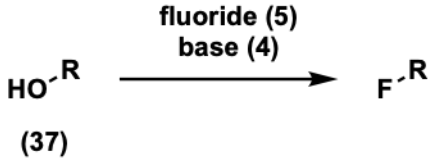


Figure 16: Reaction diagram of the deoxyfluorination reaction, where an alcohol is transformed to the corresponding fluoride. Reaction conditions include a fluoride source, and a base. The numbers indicate the amount of different species in the benchmark.

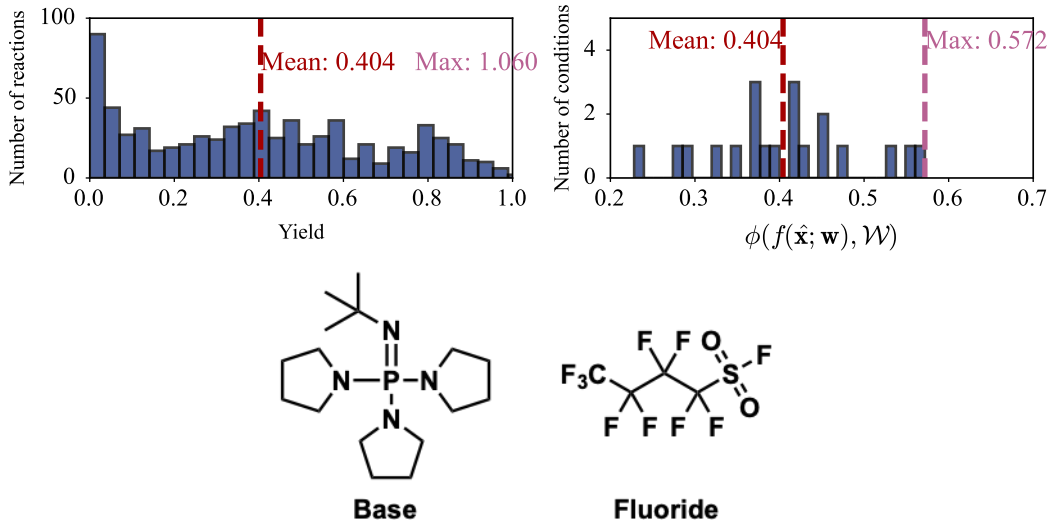


Figure 17: Top left: Distribution of the yield for the deoxyfluorination reaction in the original benchmark. Top right: Distribution of the average yield for each fluoride-base combination for the deoxyfluorination reaction in the original benchmark. Bottom: Fluoride-base combination with the highest average yield in the original benchmark.

chosen thresholds are 1.0% for the Pd-catalyzed carbon-heteroatom coupling, 0.7 kcal/mol for the N,S-Acetal formation, 12% for the borylation reaction, and 5% for the deoxyfluorination reaction. If a substrate passed these filters, the reactions with all different reported conditions were added, with reaction outcomes being taken from as predicted from the random forest emulator.

Table 3: MAE, RMSE, r^2, and Spearman's ρ of random forest regressors fitted to and evaluated on the original benchmark problems.

Benchmark problem	MAE	RMSE	r^2	Spearman's ρ
Pd-catalyzed coupling	3.16×10^{-3}	8.75×10^{-3}	0.966	0.898
N,S-Acetal formation	4.95×10^{-2} kcal/mol	7.39×10^{-2} kcal/mol	0.989	0.994
Borylation reaction	3.62×10^{-2}	4.92×10^{-2}	0.966	0.987
Deoxyfluorination	2.13×10^{-2}	3.38×10^{-2}	0.986	0.993

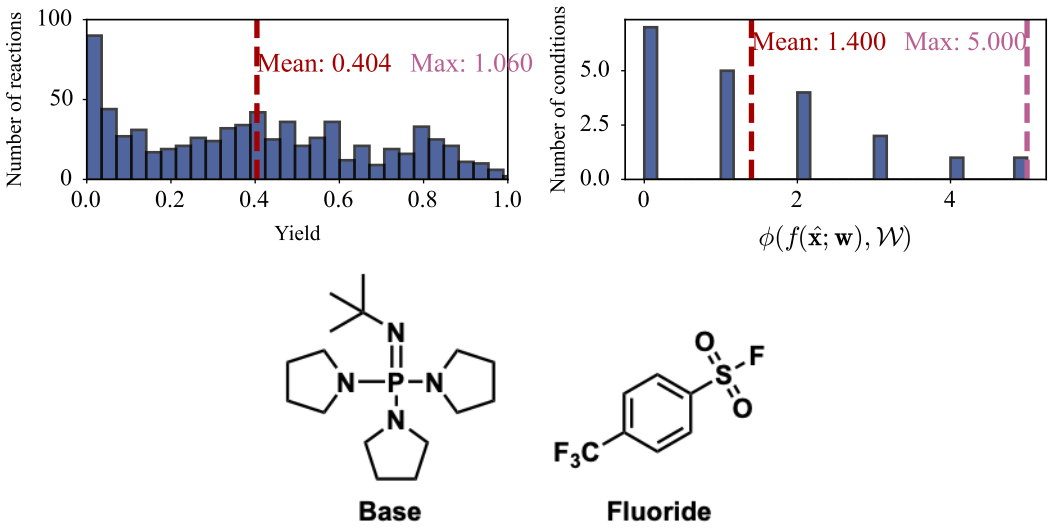


Figure 18: Top left: Distribution of the yield for the deoxyfluorination reaction in the original benchmark. Top right: Distribution of the number of substrates with a yield above the specified threshold for each fluoride-base combination for the deoxyfluorination reaction in the original benchmark. Bottom: Fluoride-base combination with the highest number of substrate with a yield above the threshold in the original benchmark.

Table 4: MAE, RMSE, r^2, and Spearman's rank correlation coefficient with their standard errors of random forest regressors in a 5-fold cross validation on the original benchmark problems.

Benchmark problem	MAE	RMSE	r^2	Spearman's ρ
Pd-catalyzed coupling	$(9.3 \pm 0.7) \times 10^{-3}$	$(2.44 \pm 0.18) \times 10^{-2}$	0.73 ± 0.03	0.429 ± 0.007
N,S-Acetal formation	$(1.43 \pm 0.07) \times 10^{-1}$ kcal/mol	$(2.11 \pm 0.10) \times 10^{-1}$ kcal/mol	0.908 ± 0.010	0.474 ± 0.007
Borylation reaction	$(1.04 \pm 0.03) \times 10^{-1}$	$(1.39 \pm 0.04) \times 10^{-1}$	0.729 ± 0.013	0.425 ± 0.009
Deoxyfluorination	$(5.96 \pm 0.14) \times 10^{-2}$	$(8.42 \pm 0.15) \times 10^{-2}$	0.913 ± 0.004	0.478 ± 0.003

A.2.3 Augmented Benchmark Problems

Pd-catalyzed carbon-heteroatom coupling

Augmentation increases the number of different nucleophiles from 16 to 31 (see Figure 19). Combined with the 96 reported reaction condition combinations, the augmented dataset consists of 2976 reactions, for which the conversion is reported.

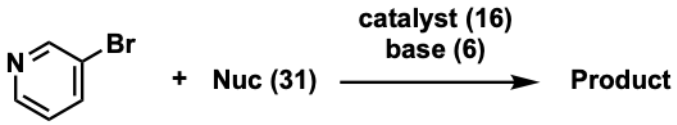


Figure 19: Reaction diagram of the Pd-catalyzed carbon-heteroatom coupling, where 3-bromopyridine reacts with a nucleophile. Reaction conditions include a catalyst and a base. The numbers indicate the amount of different species in the augmented benchmark.

Augmentation decreased the average conversion from 2.05% to 1.34%, whereas the maximum conversion remained the same at 39.81% (see Figure 20). The average of the average conversion of each

condition is decreased from 2.05% to 1.34%, and the maximum of the average conversion of each condition is also decreased from 7.60% to 6.00% (see Figure 20). The catalyst-base combination with the highest average conversion is unaffected by the augmentation and shown in Figure 20.

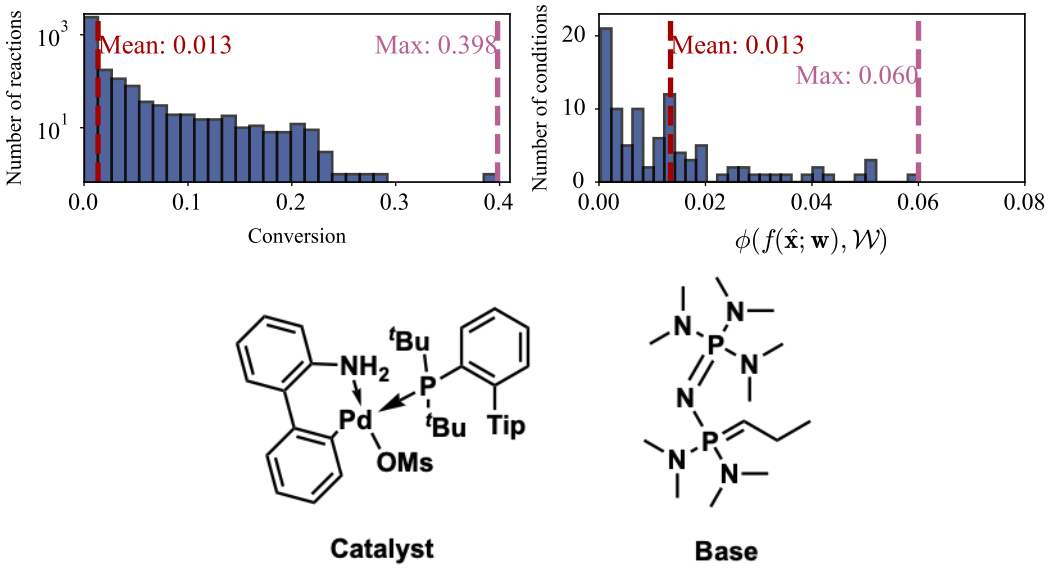


Figure 20: Top left: Distribution of the conversion for the Pd-catalyzed carbon-heteroatom coupling in the augmented benchmark. Top right: Distribution of the average conversion for each catalyst-base combination for the Pd-catalyzed carbon-heteroatom coupling in the augmented benchmark. Bottom: Catalyst-base combination with the highest average conversion in the augmented benchmark. Tip = 2,4,6-triisopropylphenyl.

With respect to the threshold aggregation function, the chosen threshold was 7.50%. The average number of substrates with a conversion above this threshold are 1.646, while the maximum number of substrates is 8 (Figure 21). The catalyst-base combination with the highest number of substrates with a conversion above the threshold is the same as shown in Figure 20.

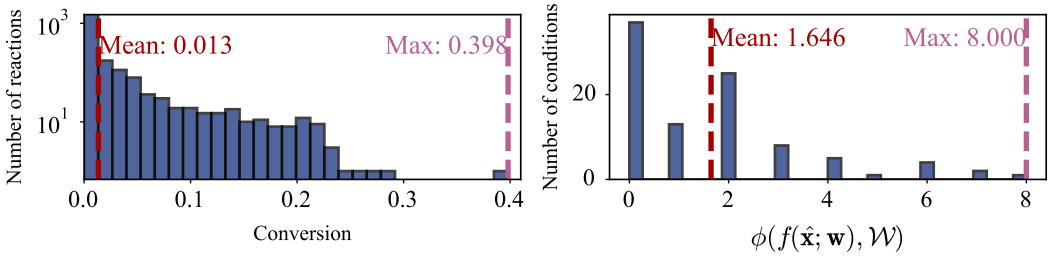


Figure 21: Left: Distribution of the conversion for the Pd-catalyzed carbon-heteroatom coupling in the augmented benchmark. Right: Distribution of the number of substrates with a conversion above the specified threshold for each catalyst-base combination for the Pd-catalyzed carbon-heteroatom coupling in the augmented benchmark.

N,S-Acetal formation

Augmentation increases the number of thiols from five to 13, while the number of imines remained constant at five (see Figure 22). Combined with the 43 reported reaction conditions, the augmented benchmark consists of 2795 reactions, for which $\Delta\Delta G^{\ddagger}$ is reported.

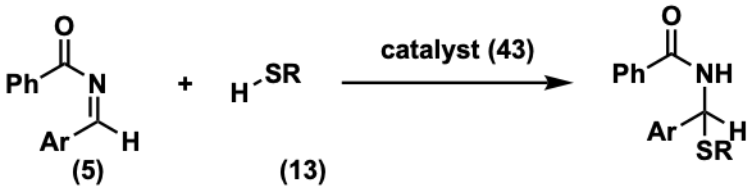


Figure 22: Reaction diagram of the N,S-Acetal formation, where an imine reacts with a thiol. Reaction conditions include a catalyst. The numbers indicate the amount of different species in the augmented benchmark.

Augmentation decreased the average $\Delta\Delta G^{\ddagger}$ from 0.988 kcal/mol to 0.757 kcal/mol, whereas the maximum $\Delta\Delta G^{\ddagger}$ was slightly decreased from 3.135 kcal/mol to 3.114 kcal/mol (see Figure 23). This decrease is due to the fact that the augmented benchmark only contains values are taken as predicted by the random forest emulator (to investigate optimization performance, the random forest emulator is taken for both the original and augmented benchmarks). Through augmentation, the average of the average $\Delta\Delta G^{\ddagger}$ of each condition decreased from 0.988 kcal/mol to 0.757 kcal/mol, while the maximum of the average $\Delta\Delta G^{\ddagger}$ of all conditions decreased as well from 2.395 kcal/mol to 1.969 kcal/mol (see Figure 23). The catalyst with the highest average $\Delta\Delta G^{\ddagger}$ is unaffected by the augmentation and shown in Figure 23.

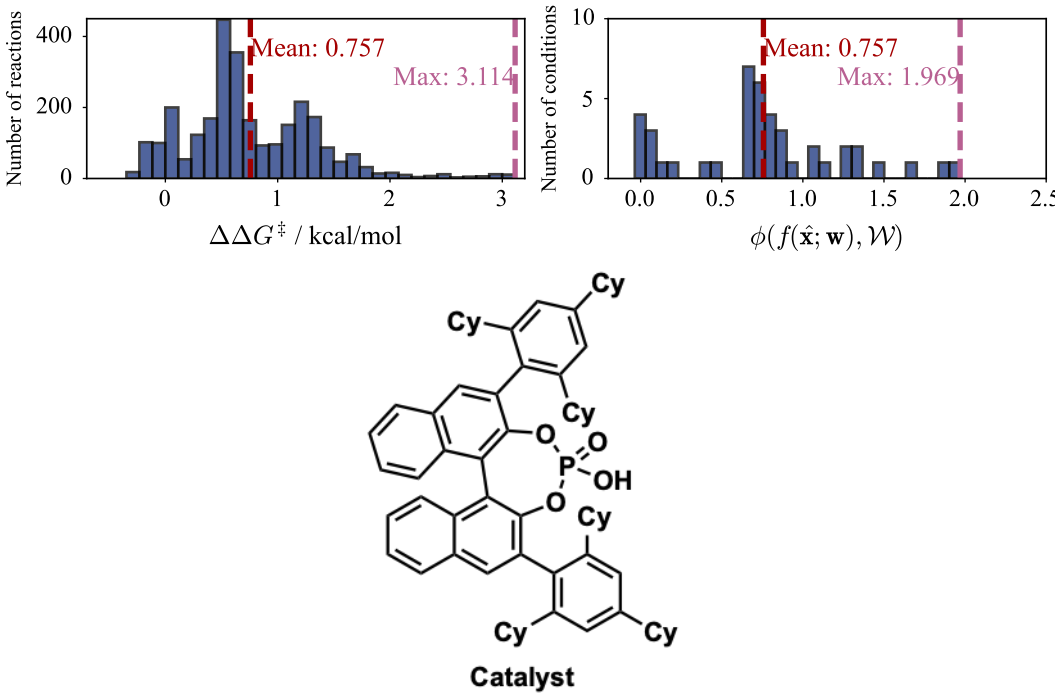


Figure 23: Top left: Distribution of $\Delta\Delta G^{\ddagger}$ for the N,S-Acetal formation in the augmented benchmark. Top right: Distribution of the average $\Delta\Delta G^{\ddagger}$ for each catalyst for the N,S-Acetal formation in the augmented benchmark. Bottom: Catalyst with the highest average $\Delta\Delta G^{\ddagger}$ in the augmented benchmark. Cy = Cyclohexyl.

With respect to the threshold aggregation function, the chosen threshold was 2.0 kcal/mol. The average number of substrates with $\Delta\Delta G^{\ddagger}$ above this threshold are 1.814, while the maximum number of

substrates is 16 (Figure 24). The catalyst with the highest number of substrates with $\Delta\Delta G^{\ddagger}$ above the threshold is the same as shown in Figure 23.

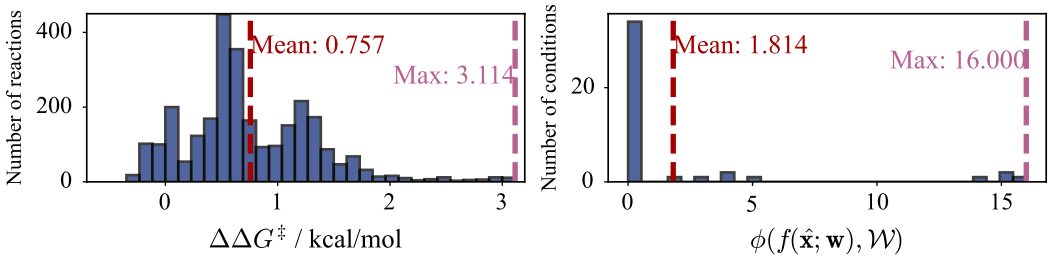


Figure 24: Left: Distribution of $\Delta\Delta G^{\ddagger}$ for the N,S-Acetal formation in the augmented benchmark. Right: Distribution number of substrates with a $\Delta\Delta G^{\ddagger}$ above the specified threshold for each catalyst for the N,S-Acetal formation in the augmented benchmark.

Borylation reaction

Augmentation increases the number of different aryl electrophiles from 33 to 75 (see Figure 25). Combined with the 46 reported reaction condition combinations, the augmented dataset consists of 3450 reactions, for which the yield is reported.

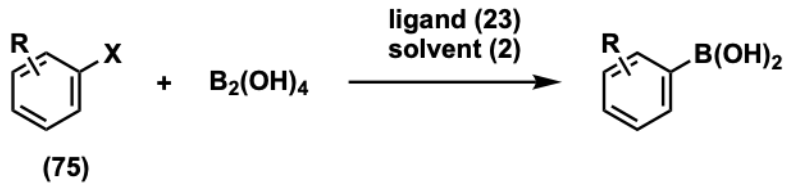


Figure 25: Reaction diagram of the borylation reaction, where an aryl electrophile is borylated via a nickel catalyst. Reaction conditions include a ligand, and a solvent. The numbers indicate the amount of different species in the augmented benchmark.

Augmentation decreased the average yield from 45.5% to 26.2%, whereas the maximum yield remained the same at 100.0% (see Figure 26). The average of the average yield of each condition is decreased from 45.5% to 26.2%, and the maximum of the average yield of each condition is also decreased from 65.4% to 38.4% (see Figure 26). The ligand-solvent combination with the highest average yield is unaffected by dataset and augmentation and shown in Figure 26.

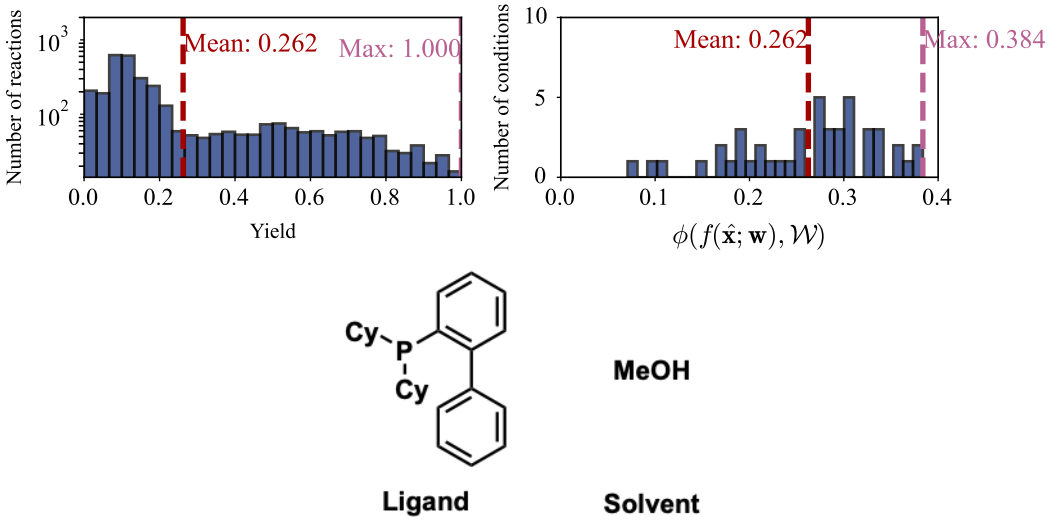


Figure 26: Top left: Distribution of the yield for the borylation reaction in the augmented benchmark. Top right: Distribution of the average yield for each ligand-solvent combination for the borylation reaction in the augmented benchmark. Bottom: Ligand-solvent combination with the highest average yield in the augmented benchmark. Cy = Cyclohexyl.

With respect to the threshold aggregation function, the chosen threshold was 90%. The average number of substrates with a yield above this threshold are 1.457, while the maximum number of substrates is 5 (Figure 27). Several ligand-solvent combinations provide the highest number of substrates with a yield above the threshold, one of them is shown in Figure 26. The ligand-solvent combinations are unaffected by the augmentation.

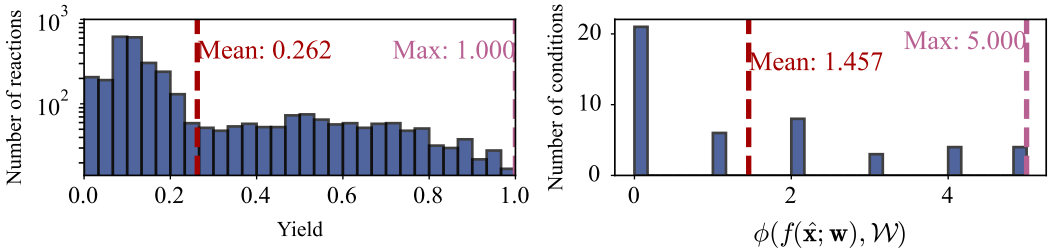


Figure 27: Left: Distribution of the yield for the borylation reaction in the augmented benchmark. Right: Distribution of the number of substrates with a yield above the specified threshold for each ligand-solvent combination for the borylation reaction in the augmented benchmark.

Deoxyfluorination reaction

Augmentation increases the number of different alcohols from 37 to 54 (see Figure 28). Combined with the 20 reported reaction condition combinations, the augmented dataset consists of 1080 reactions, for which the yield is reported.

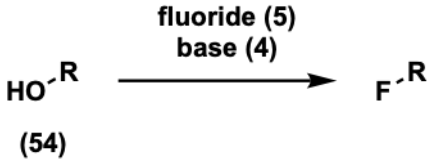


Figure 28: Reaction diagram of the deoxyfluorination reaction, where an alcohol is converted to the corresponding fluoride. Reaction conditions include a fluoride source and a base. The numbers indicate the amount of different species in the augmented benchmark.

Augmentation decreased the average yield from 40.4% to 28.9%, whereas the maximum yield remained the same at 100.6% (see Figure 29). The yield larger than 100% is contained in the originally published dataset. The average of the average yield of each condition is decreased from 40.4% to 28.9%, and the maximum of the average yield of each condition is also decreased from 57.2% to 43.8% (see Figure 29). The fluoride-base combination with the highest average yield is unaffected by augmentation and shown in Figure 29.

With respect to the threshold aggregation function, the chosen threshold was 90%. The average number of substrates with a yield above this threshold are 1.400, while the maximum number of substrates is 5 (Figure 30). The fluoride-base combination with the highest number of substrates with a yield above the threshold is also unaffected by augmentation and shown in Figure 30.

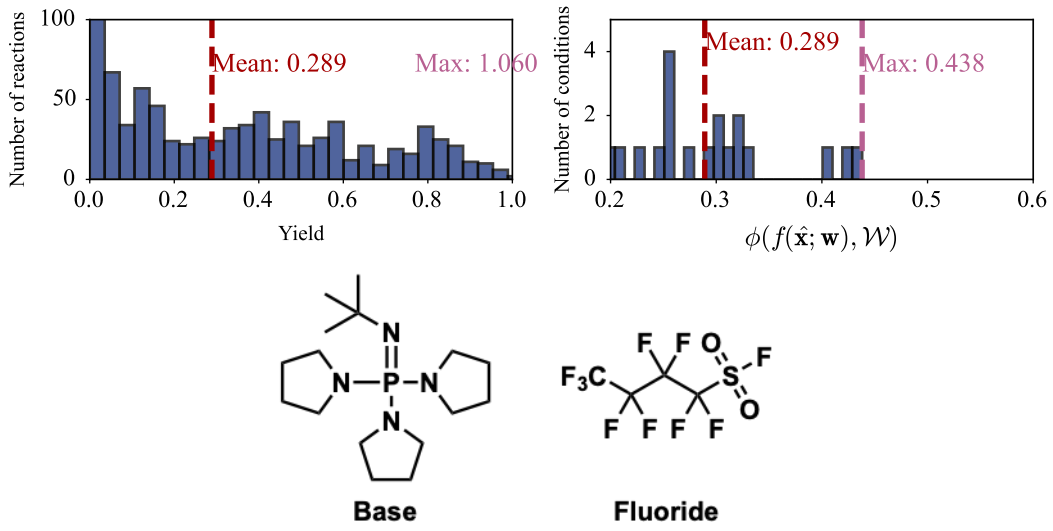


Figure 29: Top left: Distribution of the yield for the deoxyfluorination reaction in the augmented benchmark. Top right: Distribution of the average yield for each fluoride-base combination for the deoxyfluorination reaction in the augmented benchmark. Bottom: Fluoride-base combination with the highest average yield in the augmented benchmark.

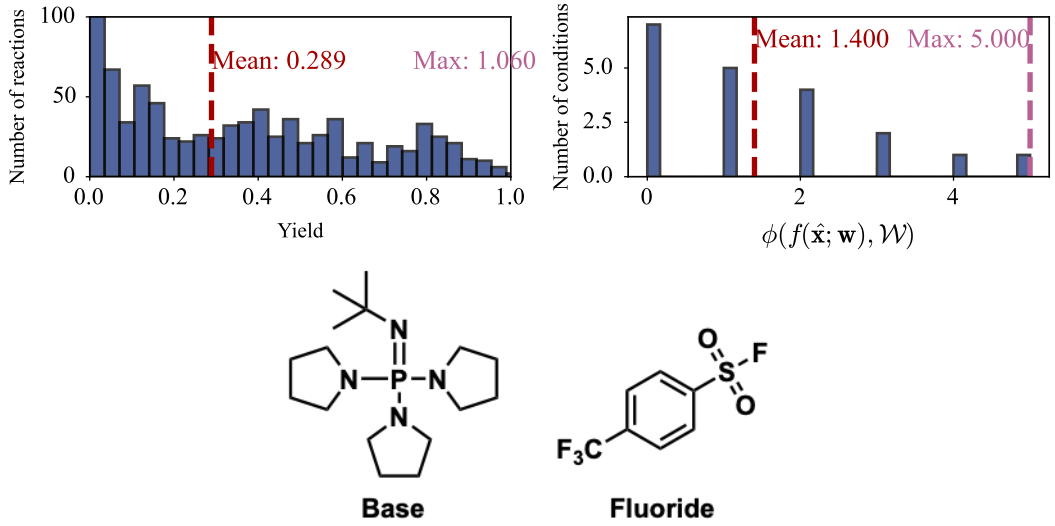


Figure 30: Top left: Distribution of the yield for the deoxyfluorination reaction in the augmented benchmark. Top right: Distribution of the number of substrates with a yield above the specified threshold for each fluoride-base combination for the deoxyfluorination reaction in the augmented benchmark. Bottom: Fluoride-base combination with the highest number of substrate with a yield above the threshold in the augmented benchmark.

A.3 Grid Search for Analyzing Benchmark Problems

To analyse the utility of considering multiple substrates in an optimization campaign, we performed exhaustive grid search on the described benchmark problems. For each problem, the substrates were split into an initial train and test set among the substrates. In total, thirty different train/test splits were performed. The obtained train set was further subsampled into smaller training sets with varying sizes to investigate the influence on the number of substrates. Sampling among the substrates in the train set was performed either through random sampling, farthest point sampling or "Average Sampling", where the required number of substrates was chosen as the substrates with

the highest average Tanimoto similarity to all other train substrates. For each subsampled training set, the most general conditions were identified via exhaustive grid search. The general reaction outcome, as specified by the aggregation function, is evaluated for these conditions on the held-out test set. Further, this general reaction outcome was scaled from 0 to 1 to give a dataset independent generality score, where 0 is the worst possible general reaction outcome for the given test set and 1 is the best possible general reaction outcome for the test set. Hence, this score should be maximized. For the different benchmark problems, we report this generality score, where we also compare the behaviour of the original and augmented problems. Below, the results of the described data analysis are shown for the benchmark problems not shown in the main text.

A.4 Details on BO for Generality Benchmarking

To identify whether BO for generality, as described above, can efficiently identify the general optima, we conducted several benchmarking runs on the described benchmark problems. On each problem, we perform benchmarking for multiple optimization strategies, as listed in Table 2.

In each optimization campaign, we used a single-task GP regressor, as implemented in *GPyTorch* (Gardner et al., 2018), with a TanimotoKernel as implemented in *Gauche* (Griffiths et al., 2023). Molecules were represented using Morgan Fingerprints (Morgan, 1965) with 1024 bits and a radius of 2. Fingerprints were generated using RDKit (Landrum, 2023). It is noteable that, while such a representation was chosen due to its suitability for broad chemical spaces, more specific representations such as descriptors might be able to improve the optimization performance.

The acquisition policies were benchmarked on all benchmark problems with differently sampled substrates for each optimization run. For each benchmark, we selected the train set randomly, consisting of twelve nucleophiles in the Pd-catalyzed carbon-heteroatom coupling benchmark, three imines and three thiols in the N,S-Acetal formation benchmark, twentyfive alcohols in the Deoxyfluorination reaction, and twenty aryl halides in the Borylation reaction. Thirty independent optimization campaigns were performed for each. The generality of the proposed general conditions at each step during the optimization is shown.

A.5 Details on Bandit Algorithm Benchmarking

The benchmarking of BANDIT (Wang et al., 2024) was performed across the benchmark problems using their proposed UCB1TUNED algorithm with differently sampled substrates for the optimization. For each benchmark, we selected the train set randomly, consisting of twelve nucleophiles in the Pd-catalyzed carbon-heteroatom coupling benchmark, three imines and three thiols in the N,S-Acetal formation benchmark, twentyfive alcohols in the Deoxyfluorination reaction, and twenty aryl halides in the Borylation reaction. Thirty independent optimization campaigns were performed for each. To ensure fair comparison, the ground truth was set to be the proxy function calculated for each dataset. To select the optimum $\mathbf{x}$ value at each step k, we relied on the authors definition of the best arm as the most sampled arm at step k.

A.6 Additional Results and Discussion

A.6.1 Additional Results on the Dataset Analysis for Utility of Generality-oriented Optimization

In addition to analysing the utility of generality-oriented optimization for ϕ as the mean aggregation, which is shown in Figure 3, we also perform a similar analysis for ϕ as the threshold aggregation, where the chosen thresholds are as described in Appendix A.2. The results of this analysis are shown in Figure 31. Similar to the case where ϕ is the mean aggregation, we observe that in the majority of benchmark problems, more general reaction conditions are obtained by considering multiple substrates. The only exemption to this observation is the Deoxyfluorination reaction benchmark, a benchmark with a particularly low number of conditions with a high threshold aggregation value (see Figure 30). In addition, we also observe a highly similar behaviour of the original and augmented benchmarks, which is due to the addition of low-performing reactions in the augmentation, which only slightly influences the results of the threshold (i.e. number of high-performing reactions).

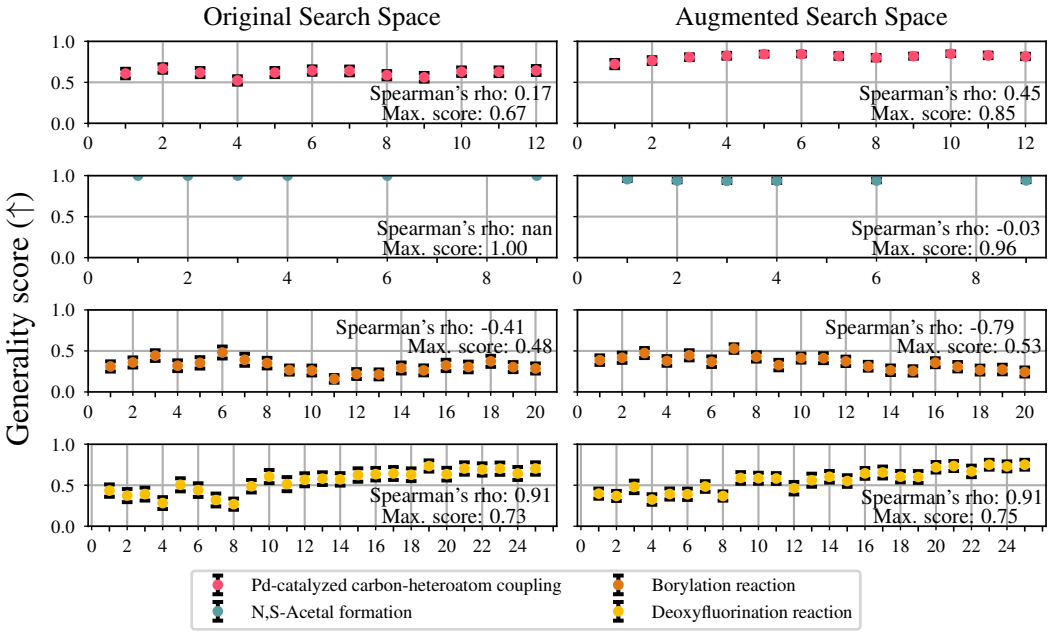


Figure 31: Normalized test-set generality score as determined by exhaustive grid search for the four benchmarks on the original (left) and augmented (right) problems for the threshold aggregation. Average and standard error are taken from thirty different train/test substrates splits.

Furthermore, we studied how different sampling techniques among the train set substrates influence the obtained generality scores. As sampling techniques, we used random sampling, farthest point sampling and "average sampling", as outlined in Appendix A.3. For ϕ as the mean aggregation, the results for the four different benchmarks are shown in Figure 32, Figure 33, Figure 34, and Figure 35. For ϕ as the threshold aggregation, the results for the two different benchmarks are shown in Figure 36, Figure 37, Figure 38, and Figure 39. Throughout the different benchmarks and aggregation functions, we observe that the generality score obtained through using the sampled train substrates are highly similar and no method clearly outperforms the others. It is particularly notable that farthest point sampling did not outperform other sampling techniques, as this strategy is commonly used to select chemicals to broadly cover chemical space (Henle et al., 2020; Gensch et al., 2022a;b; Schnitzer et al., 2024). We hypothesize that this method insensitivity is due to the low number of substrates chosen for the train set, which was chosen to still reflect realistic experimental cases.

A.6.2 Additional Results on the Benchmarking on the Augmented Benchmarks

In addition to the experiments shown in the main text, we benchmarked the sequential one-step and two-step lookahead functions where either a single substrate is selected or in the complete monitoring case. For both the one-step and two-step lookahead acquisition strategies we observe a significant loss in optimization efficiency for generality-oriented optimization, when only a single substrate is considered (see Figure 40). This is expected, as the constant observation of only one substrate does provide limited information into how different substrates might react, which is unsuitable for generality-oriented optimization. Similarly, the results shown in Figure 41 clearly demonstrate that a complete monitoring scenario is not optimally efficient for generality-oriented optimization. We hypothesize that this is because the $\mathcal{X}$ can be more efficiently explored, as not every substrate has to be tested for a specific set of reaction conditions. This underlines the utility of improved and efficient decision-making algorithms in complex optimization scenarios.

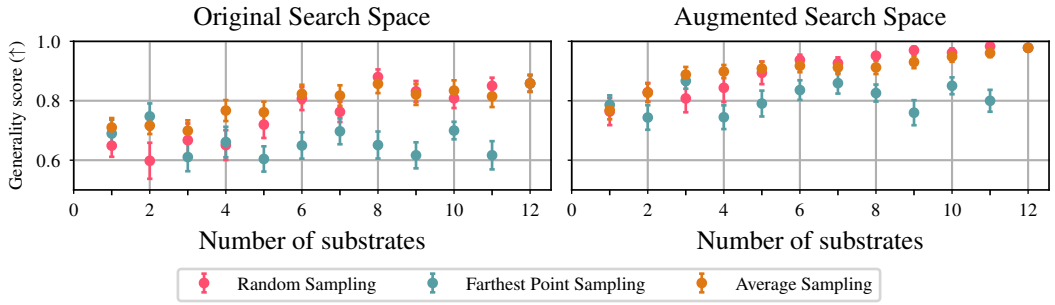


Figure 32: Generality score as determined by exhaustive grid search for the Pd-catalyzed carbon-heteroatom coupling benchmark on the original (left) and augmented (right) problems for the mean aggregation as ϕ. Average and standard error are taken from thirty different train/test substrates splits.

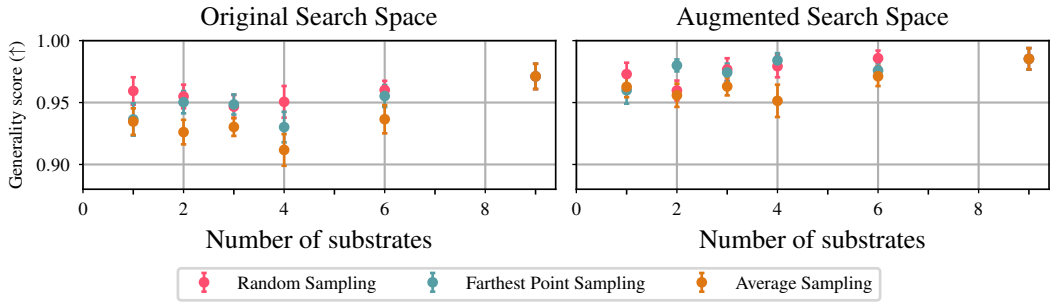


Figure 33: Generality score as determined by exhaustive grid search for the N,S-Acetal formation benchmark on the original (left) and augmented (right) problems for the mean aggregation as ϕ. Average and standard error are taken from thirty different train/test substrates splits.

A.6.3 Additional Results on the Benchmarking on the Original Benchmarks

In addition to the results described above, we also benchmark the strategies described in Table 2 on the original benchmarks. In general, we observe highly similar results compared to the augmented benchmarks that have already been discussed. This emphasizes that, while augmentation of established benchmarks remains necessary to reflect real-world conditions, the conclusions on algorithmic performances remain largely unaffected from the biases within the benchmarks. A high robustness in optimization performance on benchmark distribution further increases the utility of generality-oriented optimization in the laboratory.

Specifically, we find that, similar to the augmented benchmarks, the SEQ 1LA-UCB-PV strategy shows a significantly better optimization performance than other algorithms published in the chemical domain (see Figure 42). Comparing multiple one-step and two-step lookahead acquisition strategies, with varying α_x again emphasizes that both strategies perform similarly and that an explorative acquisition of $\mathbf{x}_{\text{next}}$ is crucial for successful generality-oriented optimization (see Figure 43). Confirming results from the augmented benchmarks, we also observe that a variation in α_w does not affect the optimization performance of the one-step lookahead acquisition strategy, while a random acquisition of $\mathbf{w}_{\text{next}}$ leads to less efficient optimizations for two-step lookahead strategies (see Figure 43). In addition, we also confirm the surprising empirical observation that a joint acquisition of $\mathbf{x}_{\text{next}}$ and $\mathbf{w}_{\text{next}}$ does not yield to a significantly improved optimization performance compared to a sequential acquisition (see Figure 44).

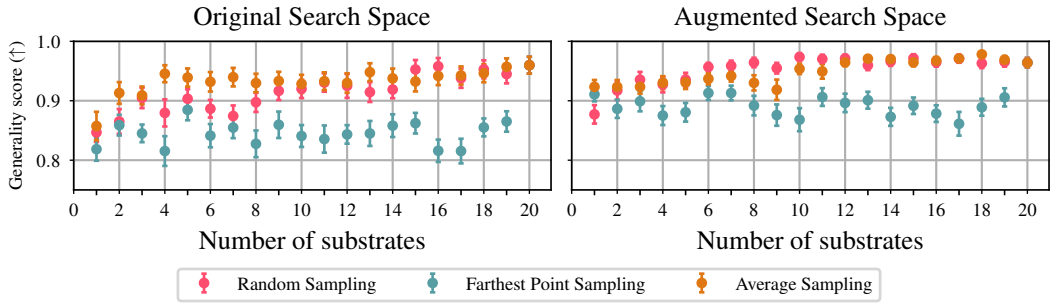


Figure 34: Generality score as determined by exhaustive grid search for the Borylation reaction benchmark on the original (left) and augmented (right) problems for the mean aggregation as ϕ. Average and standard error are taken from thirty different train/test substrates splits.

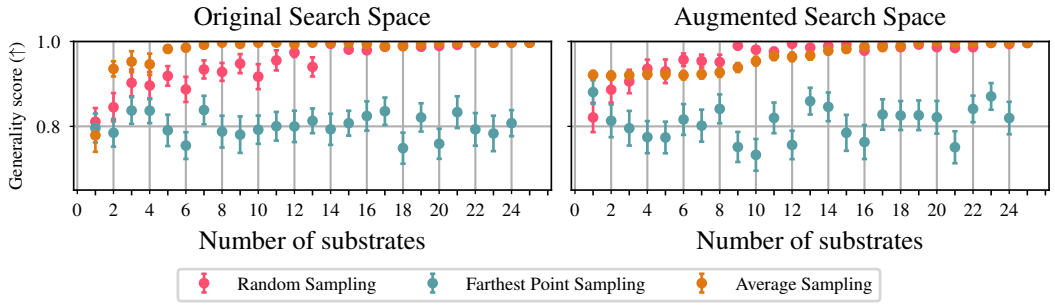


Figure 35: Generality score as determined by exhaustive grid search for the Deoxyfluorination reaction benchmark on the original (left) and augmented (right) problems for the mean aggregation as ϕ. Average and standard error are taken from thirty different train/test substrates splits.

Lastly, we also demonstrate that a generality-oriented optimization with a single substrate and in the complete monitoring case leads to suboptimal optimization performance, as shown in Figure 45 and Figure 46, respectively.

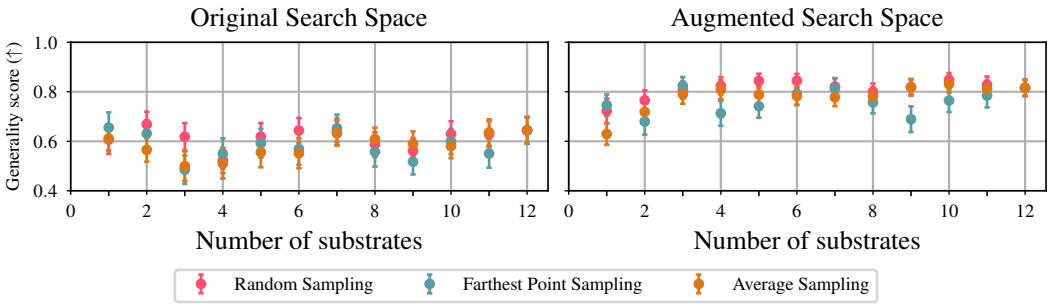


Figure 36: Generality score as determined by exhaustive grid search for the Pd-catalyzed carbon-heteroatom coupling benchmark on the original (left) and augmented (right) problems for the threshold aggregation as ϕ. Average and standard error are taken from thirty different train/test substrates splits.

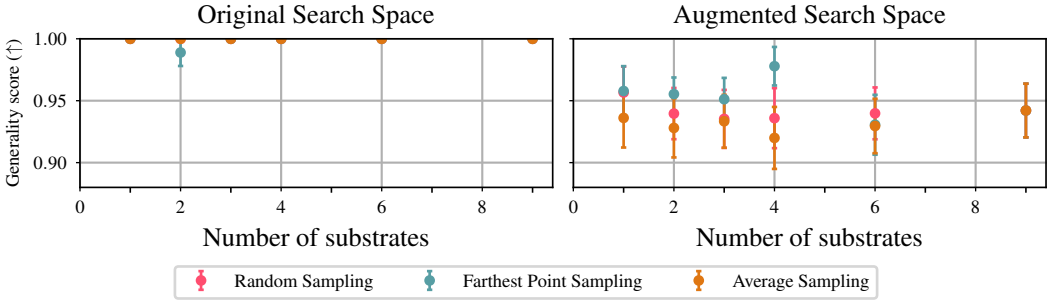


Figure 37: Generality score as determined by exhaustive grid search for the N,S-Acetal formation benchmark on the original (left) and augmented (right) problems for the threshold aggregation as ϕ. Average and standard error are taken from thirty different train/test substrates splits.

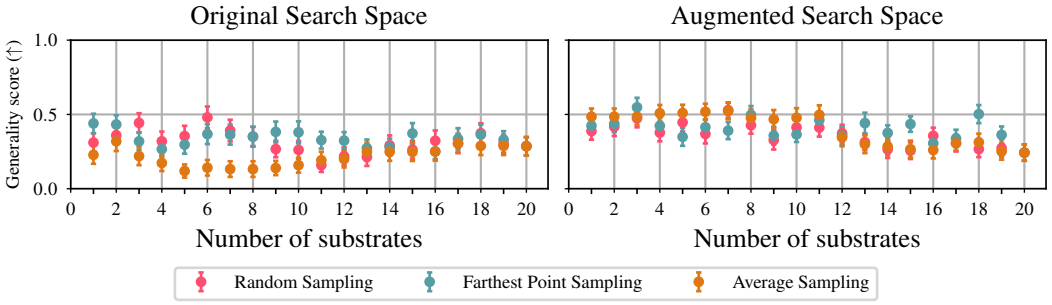


Figure 38: Generality score as determined by exhaustive grid search for the Borylation reaction benchmark on the original (left) and augmented (right) problems for the threshold aggregation as ϕ. Average and standard error are taken from thirty different train/test substrates splits.

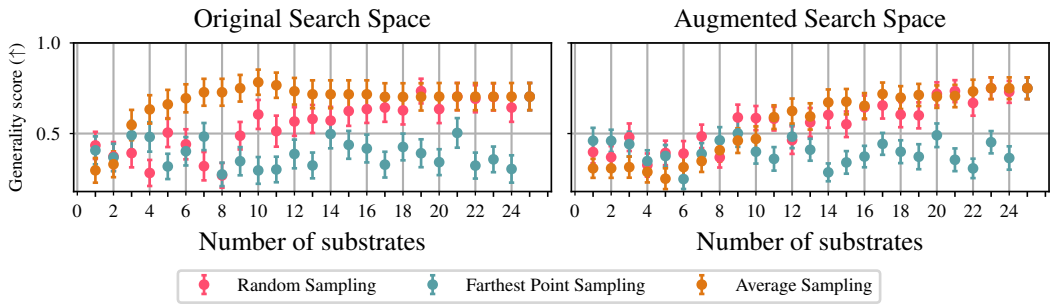


Figure 39: Generality score as determined by exhaustive grid search for the Deoxyfluorination reaction benchmark on the original (left) and augmented (right) problems for the threshold aggregation as ϕ. Average and standard error are taken from thirty different train/test substrates splits.

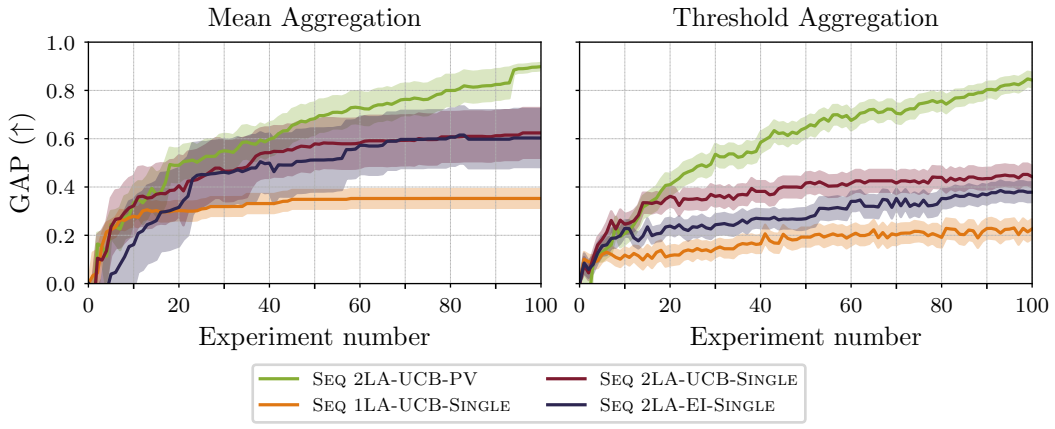


Figure 40: Optimization trajectories of different algorithms used for generality-oriented optimization considering multiple or a single substrate. The trajectories are averaged over all augmented benchmark problems with the mean (left) and threshold (right) aggregations. Optimization algorithms are described in Table 1.

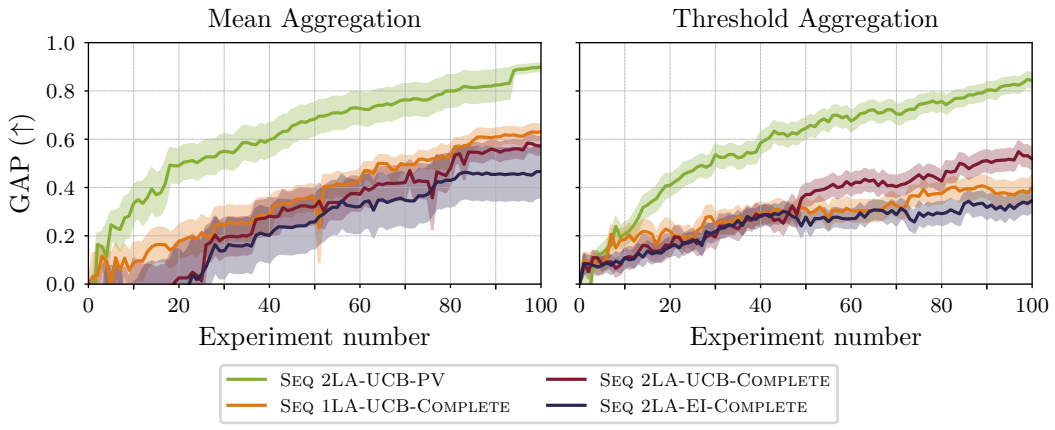


Figure 41: Optimization trajectories of different algorithms used for generality-oriented optimization considering the partial or complete monitoring case, respectively. The trajectories are averaged over all augmented benchmark problems with the mean (left) and threshold (right) aggregations. Optimization algorithms are described in Table 1.

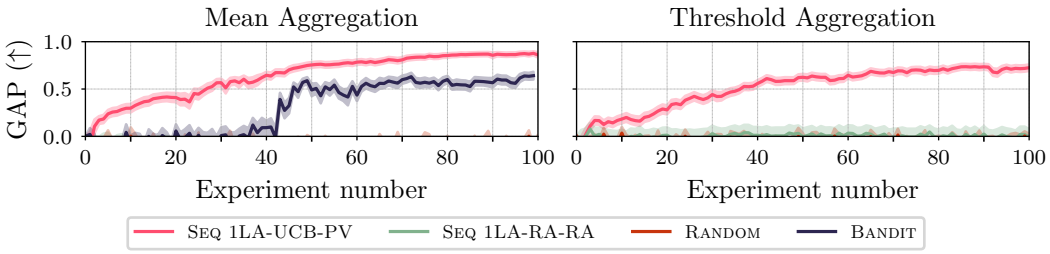


Figure 42: Optimization trajectories of different algorithms used for generality-oriented optimization in the chemical domain. The trajectories are averaged over all original benchmark problems with the mean (left) and threshold (right) aggregations. Optimization algorithms are described in Table 1.

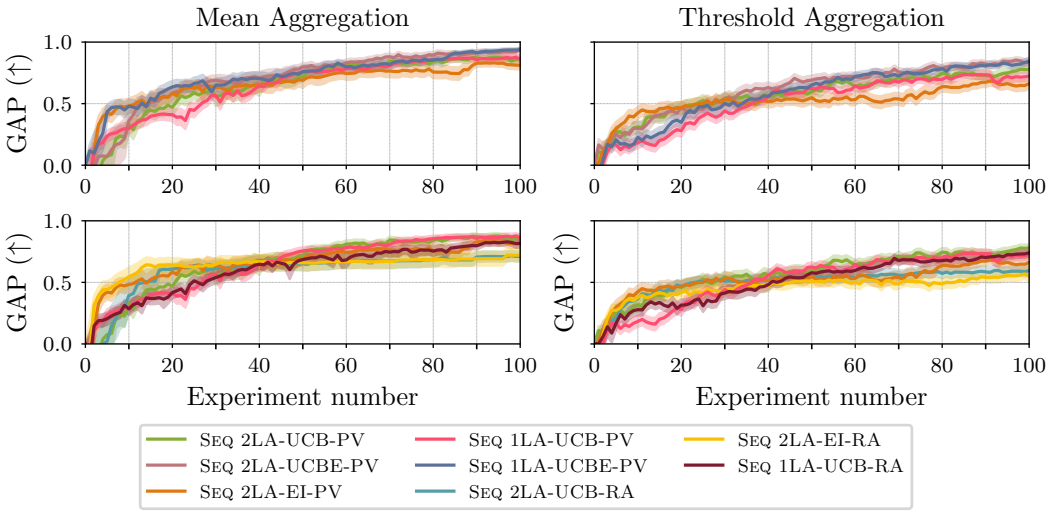


Figure 43: Optimization trajectories of different sequential acquisition strategies for generality-oriented optimization. The top row shows the influence of variation of the acquisition strategy of $\mathbf{x}_{\text{next}}$ (i.e., variation of α_x), while the bottom row shows the influence of variation of the acquisition strategy of $\mathbf{w}_{\text{next}}$ (i.e., variation of α_w). The trajectories are averaged over all original benchmark problems with the mean (left) and threshold (right) aggregations. Optimization algorithms are described in Table 1.

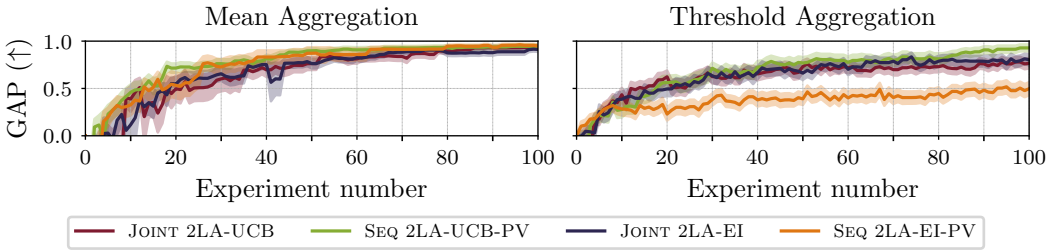


Figure 44: Optimization trajectories of sequential and joint two-step lookahead acquisition strategies for generality-oriented optimization. The trajectories are averaged over the N,S-Acetal formation and Deoxyfluorination reaction original benchmark problems with the mean (left) and threshold (right) aggregations. Optimization algorithms are described in Table 1.

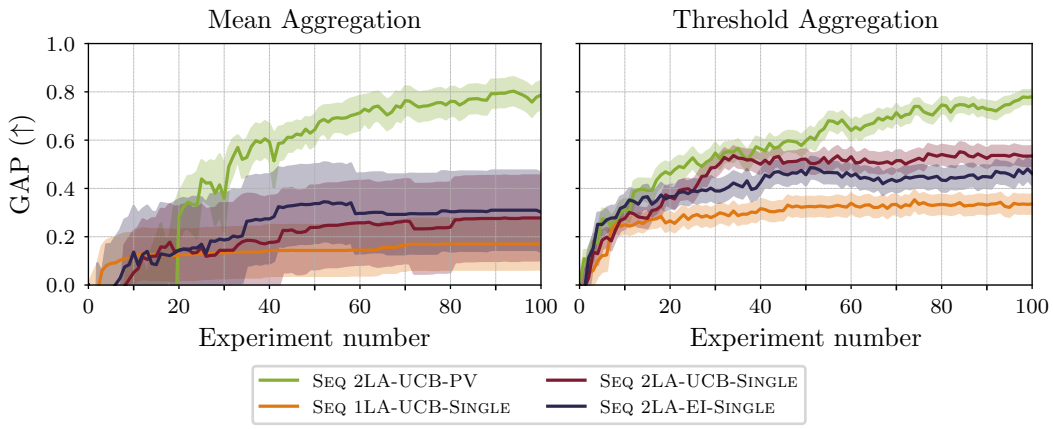


Figure 45: Optimization trajectories of different algorithms used for generality-oriented optimization considering multiple or a single substrate. The trajectories are averaged over all original benchmark problems with the mean (left) and threshold (right) aggregations. Optimization algorithms are described in Table 1.

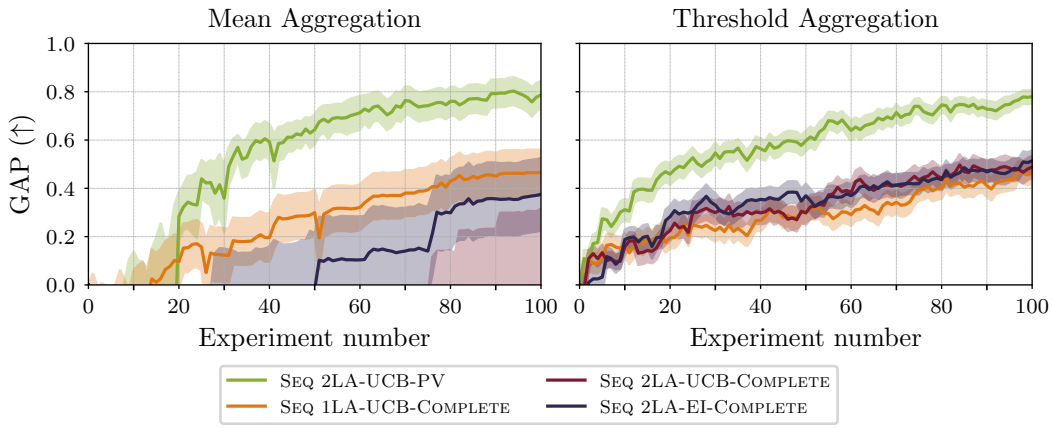


Figure 46: Optimization trajectories of different algorithms used for generality-oriented optimization considering the partial or complete monitoring case, respectively. The trajectories are averaged over all original benchmark problems with the mean (left) and threshold (right) aggregations. Optimization algorithms are described in Table 1.