

# Digital Discovery

Volume 5  
Number 1  
January 2026  
Pages 1-466

rsc.li/digitaldiscovery



ISSN 2635-098X



## Benchmarking self-driving labs

Cite this: *Digital Discovery*, 2026, 5, 14Adedire D. Adesiji,<sup>†a</sup> Jiashuo Wang,<sup>†b</sup> Cheng-Shu Kuo<sup>c</sup>  
and Keith A. Brown<sup>†abd</sup>

A key goal of modern materials science is accelerating the pace of materials discovery. Self-driving labs, or systems that select experiments using machine learning and then execute them using automation, are designed to fulfil this promise by performing experiments faster, more intelligently, more reliably, and with richer metadata than conventional means. This review summarizes progress in understanding the degree to which SDLs accelerate learning by quantifying how much they reduce the number of experiments required for a given goal. The review begins by summarizing the theory underlying two key metrics, namely acceleration factor AF and enhancement factor EF, which quantify how much faster and better an algorithm is relative to a reference strategy. Next, we provide a comprehensive review of the literature, which reveals a wide range of AFs with a median of 6, and that tends to increase with the dimensionality of the space, reflecting an interesting blessing of dimensionality. In contrast, reported EF values vary by over two orders of magnitude, although they consistently peak at 10–20 experiments per dimension. To understand these results, we perform a series of simulated Bayesian optimization campaigns that reveal how EF depends upon the statistical properties of the parameter space while AF depends on its complexity. Collectively, these results reinforce the motivation for using SDLs by revealing their value across a wide range of material parameter spaces and provide a common language for quantifying and understanding this acceleration.

Received 31st July 2025  
Accepted 3rd October 2025

DOI: 10.1039/d5dd00337g

rsc.li/digitaldiscovery

## 1. Introduction

The pace of research progress is in sharp focus due to pressing societal needs demanding the discovery of new materials.<sup>1</sup> The

field of autonomous experimentation (AE) is addressing this challenge by developing automated systems that increase the rate and reliability of experiments while also developing algorithms that select experiments to best achieve user-defined

<sup>a</sup>Department of Mechanical Engineering, Boston University, Boston, MA, USA.  
E-mail: brownka@bu.edu<sup>b</sup>Division of Materials Science & Engineering, Boston University, Boston, MA, USA<sup>c</sup>Department of Electrical and Computer Engineering, Boston University, Boston, MA, USA<sup>d</sup>Physics Department, Boston University, Boston, MA, USA

† These authors contributed equally.



From left to right:  
Cheng-Shu Kuo, Adedire D.  
Adesiji, Keith A. Brown, Jiashuo Wang

The KABlab is a research group at Boston University that sits between Mechanical Engineering, Materials Science and Engineering, and Physics. Adedire D. Adesiji is a fourth-year doctoral student in Mechanical Engineering. He is currently studying metal-organic framework (MOF)-polymer interactions for polymer membrane applications and building self-driving labs for polymer processing. Jiashuo Wang is a third-year doctoral student in Materials Science and Engineering. He is developing advanced algorithms to guide the operation of self-driving labs. Cheng-Shu Kuo is a senior undergraduate student in Electrical Engineering. He is interested in using automated systems for materials testing. Prof. Keith A. Brown leads the KABlab and studies methods to accelerate polymer discovery using self-driving labs.



goals.<sup>2-4</sup> The combination of these elements is termed a self-driving lab (SDL) (Fig. 1A), in which experiments are algorithmically selected and performed without human intervention.<sup>5</sup> Such systems have rapidly expanded from the first SDL for materials research less than a decade ago to now being common across materials, nanoscience, additive manufacturing, and chemistry.<sup>6-13</sup> The vanguard of this field has moved from demonstrations of these systems to using them for materials discoveries that have been forthcoming in areas such as lasing,<sup>14</sup> mechanics,<sup>15</sup> and battery materials.<sup>16</sup>

While SDLs are increasingly common, their value proposition has yet to be fully articulated, and different definitions and metrics have been proposed. Several of their virtues can be easily quantified and appreciated, such as how automation can allow additional experiments to be performed per unit time.<sup>17,18</sup>

A more subtle metric is how much they accelerate research, with reports ranging from  $2\times$  to  $1000\times$ .<sup>17</sup> One reason for this challenge is that quantifying the acceleration of research progress requires comparing the advanced strategy to some reference strategy, often necessitating additional experiments that do not directly contribute to the domain science being explored. Nevertheless, studies have established and explored different metrics that quantify the degree to which AE improves research outcomes. Two metrics that stand out are acceleration factor AF and enhancement factor EF that describe how much faster or better one process is relative to another (Fig. 1B).<sup>19,20</sup> These metrics are compatible with experimental campaigns as they do not require the parameter space to be fully explored or the optimum to be known. However, comparisons are not always possible because the values of these metrics are not always reported, they depend on the benchmark approach, and they depend sensitively on the details of the space being explored in a manner that has not been explored for materials.

In this paper, we review the existing experimental results that benchmark the acceleration inherent to SDLs and provide insight into how to interpret these metrics. We begin by defining EF and AF while providing the theoretical foundation for how these should behave in a typical active learning campaign. Next, we summarize the efforts in the community to provide experimental benchmarking. Finally, we perform basic simulations that provide context for interpreting EF in different parameter spaces. This review should help interpret acceleration values reported, provide guidance for the most impactful circumstances in which to apply active learning, and suggest future work in curating high-quality materials datasets for refining algorithms with direct application to materials science.

## 2. Theory

The canonical task for a materials or chemistry SDL is to run a campaign to optimize a measurable property  $y$  that depends on a set of parameters  $\vec{x}$ . Here,  $y$  can be a scalar or a vector with the latter being the purview of multi-objective optimization. Like the majority of benchmarking, we consider scalar objectives for simplicity and adopt the language of maximization, although the same logic applies to minimization tasks. The input parameter space has a finite dimensionality  $d$ , and the variables can represent compositions, processing conditions, other conditions of the experiment, or even latent variables found using unsupervised learning. With these definitions, the goal of the campaign is to identify the conditions  $\vec{x}^* = \arg \max(y(\vec{x}))$ . After experiment number  $n$  in the campaign, the progress towards this goal can be quantified by considering how close the current observed maximum  $y^*(n) = \max(y(\vec{x}_n))_n$  is to the true maximum  $y^* = \max(y(\vec{x}))$ . Interestingly, if the campaign proceeds by selecting experiments uniformly at random across  $\vec{x}$ , this average progress has a closed-form solution that depends upon the cumulative distribution function  $F_y(y)$ .<sup>19</sup> Specifically, the average performance after  $n$  experiments corresponds to the performance at which there is a 50% chance that no larger value has been observed, or

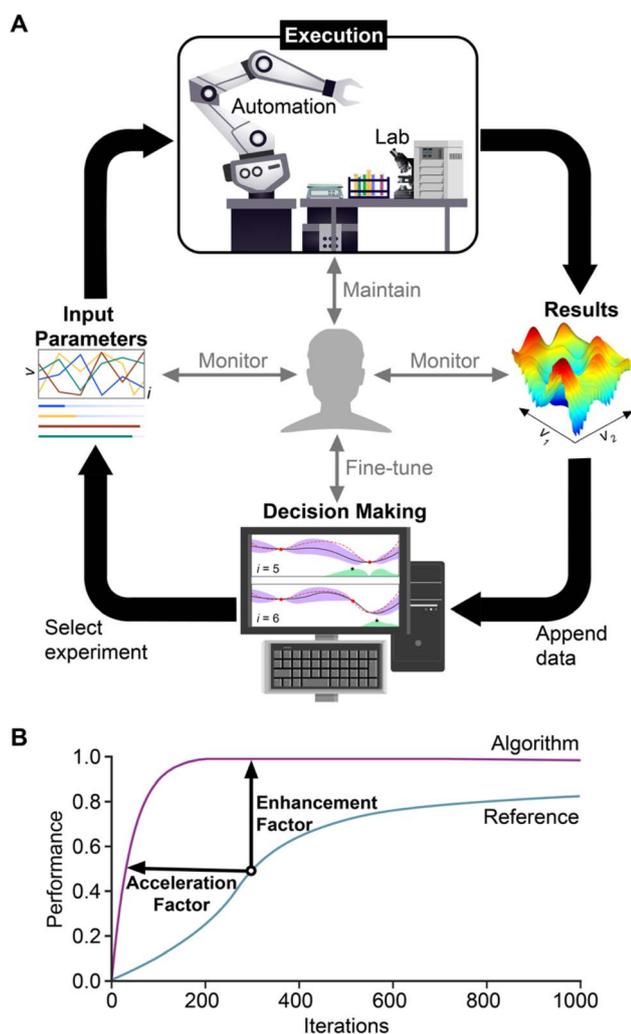


Fig. 1 (A) Schematic of the workflow of a self-driving lab (SDL). (B) Representative performance convergence plot, also known as a horse race plot, illustrating enhancement factor EF and acceleration factor AF. EF quantifies relative performance after a fixed number of experiments, while AF quantifies the reduction in the number of experiments required to reach a target performance. Both metrics are defined relative to a reference strategy, such as sampling the space uniformly at random.



$$\frac{1}{2} = F_y(y_{\text{rand}}^*(n))^n, \quad (1)$$

where  $y_{\text{rand}}^*(n)$  is the expected maximum performance from random sampling (*i.e.* we take an ensemble-median of the maximum of  $n$  samples drawn uniformly at random). At  $n = 1$ ,  $y_1 = \text{median}(y)$  and  $y_n$  approaches  $y^*(n)$  as  $n \rightarrow \infty$ . This simple analysis illustrates that the convergence of a simple decision-making policy depends intimately on the details of the parameter space.

While it is reasonable to derive closed-form solutions for expected convergence when the property space is known, for real materials systems,  $y$  is unknowable except through experiment. The nature of continuous variables and the presence of noise in measurements mean that ground truth will never be completely known. This makes it impossible to predict how fast convergence is expected or even when the process has fully converged. Thus, benchmarking learning using an SDL involves completing two campaigns, an active learning (AL) campaign designed to test the learning algorithm along with a reference campaign guided by a standard method. From a benchmarking perspective, the most relevant data available are the best performance observed in the first  $n$  experiments, defined as  $y_{\text{AL}}^*(n)$  for the AL campaign and  $y_{\text{ref}}^*(n)$  for the reference campaign. There are two main ways of comparing these sets of data.<sup>19,20</sup> The first metric is the acceleration factor (AF) that is defined as the ratio of  $n$  needed to achieve a given performance  $y_{\text{AF}}$ , namely,

$$\text{AF}(y_{\text{AF}}) = \frac{n_{\text{ref}}}{n_{\text{AL}}}, \quad (2)$$

where  $n_{\text{AL}}$  is the smallest  $n$  for which  $y_{\text{AL}}^*(n) \geq y_{\text{AF}}$  while  $n_{\text{ref}}$  satisfies the same condition for the reference campaign. Larger values of AF indicate a more efficient AL process. The second metric is the enhancement factor (EF) that is defined as the improvement in performance after a given number of experiments, namely

$$\text{EF}(n) = \frac{y_{\text{AL}}^*(n)}{y_{\text{ref}}^*(n)}. \quad (3)$$

EF presents an interesting limit when considering benchmarking using random sampling. Specifically, the very best outcome of an active learning campaign would be  $y^*$  while the worst median performance possible using random sampling would be  $\text{median}(y)$ , which is  $y_{\text{rand}}^*$  at  $n = 1$ . This leads us to define the contrast  $C$  of the property space as,

$$C = \frac{y^*}{\text{median}(y)}, \quad (4)$$

which defines the best possible EF that could be found when studying that property space. Between the two metrics EF and AF, EF is often more convenient to compute as it is defined *vs.*  $n$  and thus can be calculated at all points for reference and benchmark campaigns that have the same number of experiments.

The choice to define progress in terms of the maximum experimentally observed value rather than the maximum value predicted by a surrogate model deserves further discussion. Ultimately, these values will converge as a campaign progresses

due to most optimization algorithms naturally including exploitative steps. However, the surrogate model may differ greatly from experiment especially early in the campaign. Thus, in order to confidently assess progress using the surrogate model, one would have to perform an experiment using the parameters that correspond to the maximum value predicted by the surrogate model. This is readily accessible for analytical functions but would double the experimental budget for experimental campaigns. If such validations are desired, an algorithm can mix in purely exploitative steps, but these should count as experiments in the experimental budget.

### 3. Literature survey

As a goal of the SDL field is accelerating progress, much work has been dedicated to benchmarking the acceleration of these systems. To comprehensively consider the literature that benchmarks active learning, we began with a broad literature search (Fig. 2). We searched the Scopus database using the keywords “Bayesian optimization” combined with “benchmark.” As the field of optimization research extends far beyond its overlap with materials or SDLs, this search yielded considerable results with 4245 publications matching these keywords (Fig. 2A). The keyword “Bayesian optimization” was chosen due to its prevalent adoption for active learning in the field of material science while the term active learning is widely used for an unrelated method in education. Most studies outside materials science utilized analytical functions or look-up tables that are designed to be challenging to optimize and thus provide insight into comparisons between learning approaches. While this broad survey is useful for evaluating active learning strategies, our focus is to evaluate benchmarking using actual experimental materials datasets. To narrow down the search to those that involved benchmarking using self-driving labs, we conducted a search with the broad term “self-driving lab”, resulting in 111 studies. After examining each study, only 40% of these articles reported direct efforts to benchmark performance. These data are provided at <https://doi.org/10.5281/zenodo.17287854>.

While “Bayesian optimization” was used as a keyword in the initial Scopus search, this literature analysis includes all SDL papers that report benchmarking including those that do not use BO. That said, almost all SDL studies used BO. Reinforcement learning<sup>21</sup> and genetic algorithms<sup>22,23</sup> have also been used and their results are included in this analysis.

Having narrowed down the field to a targeted set of papers considering experimental materials data, we set out to more fully compare this subset of the literature. The reviewed literature spans a diverse range of material domains, including electrochemistry,<sup>19,24–28</sup> bulk materials discovery,<sup>23,24,29–35</sup> spectroscopy and imaging,<sup>24,36</sup> mechanics,<sup>37–41</sup> nanoparticle and quantum dot synthesis,<sup>21,42–46</sup> and solar cell or device optimization.<sup>22,24,32–34,47</sup> This diversity underscores the breadth of SDL applications and highlights the variety of experimental contexts in which AF and EF are reported. This breadth of methods makes establishing reproducible results very important. Indeed, reproducibility is a key challenge in SDLs and efforts promote reproducibility have employed computer



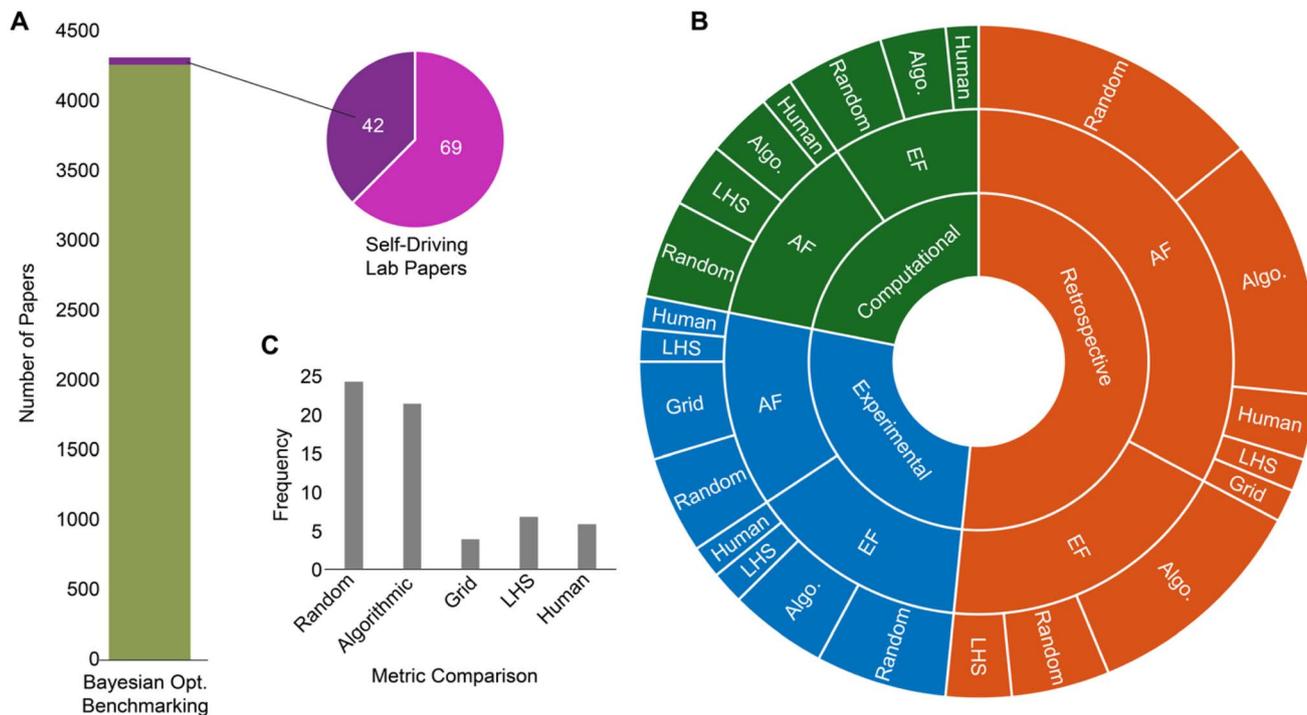


Fig. 2 Trends in SDL benchmarking studies: (A) Summary of the Bayesian optimization benchmarking studies. The pie chart details the studies that involve SDLs. (B) Sunburst diagram depicting benchmarking results from SDL studies. The inner ring depicts the benchmarking type (experimental, retrospective, and computational), the middle ring describes the reported metric, and the outer ring depicts the reference campaign (random sampling, Latin hypercube sampling - LHS, grid-based sampling, human-directed sampling, or algorithmic to reflect a different active learning process than Bayesian optimization). (C) Bar chart showing the number of SDL benchmarking studies that utilize each type of reference used for comparison.

science abstractions,<sup>48</sup> novel programming languages,<sup>49</sup> and knowledge graph-based approaches.<sup>50</sup> That said, the stochastic nature of active learning can lead to campaigns having different values of AF or EF even when performed in the same lab. Thus, while important, experimental benchmarking may not be the ideal method for evaluating reproducibility, at least on the basis of individual campaigns.

### 3.1 The source of the data

Benchmarking can be categorized by the source of the data, which falls into three categories.<sup>19–27,29–40,42–46,51–61</sup> Experimental benchmarking are studies that complete at least two independent campaigns of experiments comparing an AL strategy to a reference strategy using unique physical experiments. This is the most informative class of benchmarking as it captures both statistical and systematic sources of experimental variability. However, this may be impractical, as it requires additional experiments that can be resource-intensive or beyond the scope of a materials study. A more attainable category of benchmarking is retrospective, where tables of previously completed experiments are used as ground truth for simulated campaigns. This approach has the advantages of being faster and less resource-intensive while also featuring known optima. However, decision-making policies are forced to become discrete to align with the existing data, the parameter space is vastly constricted, and noise becomes embedded into the

system. Nevertheless, this approach is popular as a method to tune hyperparameters and compare algorithms. Computational benchmarking comprises running a campaign that queries an analytical function or computational model. This process can be fast, inexpensive, and the optima can be known for analytical functions. As such, these are extremely common in materials science and the broader optimization community for benchmarking AL algorithms. Here, we choose not to include benchmarking based on purely analytical functions and instead focus on studies that use data relevant to materials experiments, as these will provide the most direct articulation of the acceleration inherent to SDLs in materials research.

Retrospective analysis is the most common type of SDL benchmarking (Fig. 2B). For instance, Rohr *et al.* used a dataset of 2121 catalyst compositions collected using high-throughput experimentation spanning a six-dimensional electrocatalytic metal oxide space to benchmark various sequential learning models evaluated such as Gaussian process (GP), random forest (RF), and least-squares estimation (LE).<sup>19</sup> The analysis, which was conducted over 1000 learning cycles, revealed up to a 20-fold reduction in the number of experiments required to find top-performing oxygen evolution reaction catalysts, comparing GP to random sampling. The study also evaluated the effect of exploration-exploitation tuning and dataset type on model performance. Similarly, Liu *et al.* developed an SDL to optimize the open-air perovskite solar cell manufacturing process and



benchmarked its BO framework using a regression model trained on experimental data.<sup>34</sup> They ran 300 iteration steps comparing standard BO and BO with knowledge constraint against Latin hypercube sampling (LHS), factorial sampling with progressive grid subdivision (FS-PGS), and one-variable-at-a-time sampling (OVATS). The BO methods consistently outperformed the others, showing up to a 10-fold enhancement in power conversion efficiency relative to LHS and FS-PGS.

Experimental benchmarks, while less common, are the most representative of real-world variability and experimental constraints. For example, Liu *et al.* had a limited budget of less than 100 process conditions, which limited experimental benchmarking to only standard BO vs. LHS.<sup>34</sup> Within 85 process conditions, BO identified four times as many high-performing perovskite films as LHS. As a separate example, Wu *et al.* benchmarked the efficiency of a BO-guided gold tetrapod nanoparticle synthesis against random search over an experimental run of 30 iterations. The BO algorithm utilized in this work, Gryffin, uses a Bayesian neural network to construct a kernel regression surrogate model. The algorithm was benchmarked based on four hierarchical objectives related to the plasmonic response of the particles. While random sampling occasionally satisfied three of the objectives, it failed to meet the final objective within the experimental budget. One note about experimental benchmarking is that campaigns used for benchmarking often do not consider the experiments performed to establish the bounds of the parameter space or develop the SDL more generally. While this may be a substantial amount of work, it encourages researchers to use SDLs for prolonged campaigns to amortize this overhead.

Computational analyses, although sampled more selectively in this review due to our focus on benchmarking strategies that use experimental data, remain a valuable tool for comparing algorithmic strategies. Jiang *et al.* developed a chemical synthesis robot, AI-EDISON, for gold and silver nanoparticle synthesis with the goal of optimizing their optical properties.<sup>44</sup> As part of their workflow, they benchmarked AI-EDISON against random search in a simulated chemical space using PyDScat-GPU, a simulation tool based on discrete dipole approximation-based simulations. During a campaign with 200 steps, the algorithm outperformed random search by the 27th step, identifying samples from nine of ten spectral classes and completing all ten by the 78th step. In terms of mean fitness, which measures the similarity of a sample's spectrum to the target, AI-EDISON reached the performance achieved by 200 random steps in just 25 iterations guided by the algorithm. Annevelink *et al.* likewise developed a framework for electrochemical systems, AutoMAT, with input generation from atomic descriptors to continuum device simulations such as PyBaMM.<sup>25</sup> Compared to random search, AutoMAT found top-performing Li-metal electrolytes and nitrogen reduction reaction catalysts in 3 and 15 times fewer iterations respectively.

### 3.2 The nature of the reference campaign

A central consideration when benchmarking learning is deciding how to select experiments for the reference campaign.

We highlight the four most used reference methods. Random sampling involves choosing each experiment uniformly at random in the parameter space. Random sampling is simple to implement and will converge in a predictable manner, as described by eqn (1). Furthermore, the total number of experiments does not have to be chosen prior to the campaign, which facilitates analysis and data reuse. Grid-based sampling involves dividing the parameter space into uniformly spaced intervals. It is easy to implement and will provide a balanced view across parameter space, but at the cost of needing to specify the total number of experiments *a priori*. Latin hypercube sampling (LHS) combines the even distribution of grid sampling with the perturbations of random sampling to provide a balanced picture of parameter space while using any number of points. This is generally the preferred method for obtaining data when performing initial training campaigns. Like grid sampling, an LHS campaign cannot be stopped early without having a biased data distribution and relying on evenly distributed samples may over-sample flat regions while potentially missing areas with sharp transitions. Human-directed sampling is the non-SDL state of the art and provides a useful comparison when evaluating whether the algorithm is providing value. However, human-directed sampling is time-consuming and introduces variability and bias from individual decision-making. All four of these methods have been explored for benchmarking (Fig. 2C).

Across the reviewed SDL papers, which include 42 unique studies and 63 reported benchmarks, the most fundamental and widely adopted baseline is random sampling. MacLeod *et al.* evaluated their SDL, Ada, for multi-objective optimization of palladium film synthesis, balancing conductivity and annealing temperature.<sup>22</sup> In a simulated campaign using a model built from experimental data, Ada's q-expected hypervolume improvement (q-EHVI) strategy achieved twice the hypervolume of random sampling within 25 steps and reached a hypervolume achieved by 10 000 random samples in just 100 steps. Similarly, Bai *et al.* developed a platform to explore the copper antimony sulfide (Cu-Sb-S) compositional space for photo-electrocatalytic hydrogen evolution. In this experimental benchmarking study, the Bayesian optimizer revealed a Cu-Sb-S composition that exhibited 2.3 times greater catalytic activity than results from random sampling.

Many SDL studies compare performance between algorithms, which frequently includes variants of BO (*e.g.*, differing surrogate models, acquisition functions, or kernels),<sup>52</sup> as well as hybridized approaches involving evolutionary algorithms,<sup>22,23</sup> or reinforcement learning.<sup>21</sup> For instance, Ziomek *et al.* proposed a length scale balancing GP-UCB (LB-GP-UCB), a BO variant with an upper confidence bound (UCB) acquisition function that aggregates multiple GPs with different length scales to address the challenge of unknown kernel hyperparameters.<sup>41</sup> They retrospectively benchmarked the performance of LB-GP-UCB against adaptive GP-UCB (A-GP-UCB),<sup>62</sup> maximum likelihood estimation (MLE),<sup>63</sup> and Markov chain Monte Carlo (MCMC)<sup>64</sup> using the crossed barrel<sup>37</sup> and silver nanoparticle<sup>65</sup> datasets. For both datasets, LB-GP-UCB consistently found the



optimal solution with fewer experiments, specifically requiring 40% fewer trials than MLE and MCMC.

A relatively small number of studies reported performance relative to LHS and grid-based sampling. Gongora *et al.* developed the Bayesian experimental autonomous researcher (BEAR) to optimize the toughness of crossed barrel structures.<sup>37,60</sup> They benchmarked its performance against grid sampling, where the 4D design space was discretized into 600 points, each tested in triplicate. The BEAR running on a BO framework with an expected improvement (EI) acquisition function discovered higher-performing structures with 18 times fewer experiments. Also, Bateni *et al.* developed an SDL, Smart Dope, for space exploration and optimization of lead halide perovskite (LHP) quantum dots (QDs).<sup>42</sup> Using LHS, 150 initial experiments were conducted across the nine-dimensional space to generate training data for closed-loop optimization. Smart Dope, also running BO with an expected improvement acquisition function, achieved a photoluminescence quantum yield (PLQY) of 158% after just four closed-loop iterations, exceeding the 151% maximum obtained by LHS. This suggests that LHS and grid-based sampling's fixed intervals may over-represent flat regions while missing sharp transitions.

Human-directed sampling, where expert researchers select experimental conditions based on intuition and domain knowledge, also appears in the reviewed SDL literature, and it provides a useful comparison between SDLs and conventional experimentation. Nakayama *et al.* benchmarked BO against human-directed sampling using a one-dimensional model of synthesis temperature optimization.<sup>51</sup> Human experts required 13–14 trials to find the global optimum, while BO required only ten steps with the appropriate acquisition function and hyper-parameters. The search efficiency of BO demonstrated in this simple 1D case is expected to grow in higher-dimensional spaces where human intuition is more limited. Shields *et al.* benchmarked the performance of BO against 50 expert chemists using high-throughput experimental data covering a ten-dimensional parameter space for optimizing the yield of direct arylation of imidazoles.<sup>59</sup> To reduce bias, the performance was averaged across the 50 human participants and 50 runs of the Bayesian optimizer, each conducted over 100 steps. While humans achieved 15% higher yield in the first five experiments, by the 15th experiment, the average performance of the optimizer surpassed that of the humans. BO consistently achieved >99% yield within the experimental budget, and within the first 50 experiments, it discovered the global optimum that none of the experts found.

It should be noted that, as the field matures, future work may focus more on comparing advanced strategies to one another rather than comparing advanced algorithms to comparatively inefficient benchmarking approaches such as random sampling. While this is a valuable pursuit and highly relevant to accelerating materials discovery, it may make it challenging to compare the values reported by different studies. Fortunately, metrics such as AF and EF can be applied in a multiplicative fashion if compared at specific  $y$  or  $n$ , respectively. Thus, it may be possible to relate such advanced comparisons back to random sampling, which has the advantage of being

a deterministic function of the cumulative distribution function of a property space.

### 3.3 Meta analysis of reported benchmarking

To visualize the reported SDL benchmarking, we extracted AF from studies spanning a range of  $d$  (Fig. 3). Overall, the reported AF spanned a wide range, from 1.3 to 100, highlighting the variability in how effectively active learning accelerates research across different experimental domains. The median reported AF was 6. Interestingly, AF appeared to increase with increasing  $d$ , suggesting that the “curse of dimensionality” was managed more effectively by active learning than by random sampling. From a learning efficiency perspective, this suggests a “blessing of dimensionality” in which higher-dimensional spaces provide more incentive to use advanced learning algorithms. A summary of the AF values is provided in Table 1. To provide some notable examples, at the low end, an AF of 1.3 was observed in a 1D temperature-dependent synthesis optimization task, where the number of iterations required for BO to locate the global maximum was compared to that required by a human researcher.<sup>51</sup> At the high end, a multi-objective Bayesian optimization campaign for metallic thin-film synthesis in a 4D parameter space achieved an AF of 100 when benchmarked against random sampling.<sup>22</sup>

While AF is simple to report, it is subtle to interpret as it depends on the chosen performance threshold. Typically, this threshold corresponds either to a value defined by the researcher or the highest performance achieved during the campaign.<sup>22,32</sup> In contrast, EF is easy to calculate at each experiment, and it does not rely on a performance value, making it useful for tracking learning progress.

In order to visualize EF progression over the course of SDL campaigns, we extracted EF from reported performance trajectories (Fig. 4). We limited this analysis to studies that benchmarked against random sampling since this can serve as a common baseline. To enable comparison across studies with different  $d$ , we divided experiment number  $n$  by  $d$ . We focused specifically on experimental and retrospective benchmarking studies, as these are grounded in real experimental data. Examining the computed EF values, a consistent pattern emerges in which EF initially grows with  $n/d$ , reaches a peak,

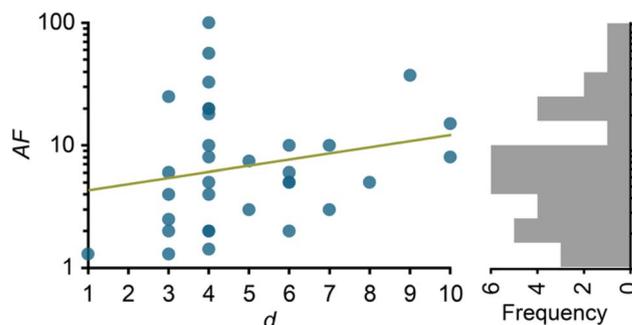


Fig. 3 Acceleration factor (AF) vs. input parameter space dimensionality  $d$  across benchmarking SDL studies, with corresponding AF frequency.



Table 1 Summary of reported AF from SDL benchmarking studies

Case	Source	AF	Type	Dimension	Comparison	Objective
1	Bateni <i>et al.</i> <sup>42</sup>	37.5	Experimental	9	GP-EI vs. LHS	Photoluminescence quantum yield
2	Cakan <i>et al.</i> <sup>32</sup>	2.5	Experimental	3	GP-EI vs. grid	Film photothermal stability
3	Fatehi <i>et al.</i> <sup>28</sup>	20	Experimental	4	GP-EI & GP-UCB vs. random search	Catalyst activity
4	Gongora <i>et al.</i> <sup>37</sup>	18	Experimental	4	GP-EI vs. grid (best grid performance as reference)	Structure toughness
5	Gongora <i>et al.</i> <sup>37</sup>	56.25	Experimental	4	GP-EI vs. grid (best BO performance within a time budget as reference)	Structure toughness
6	Gongora <i>et al.</i> <sup>38</sup>	10	Experimental	4	GP-EI (FEA informed) vs. GP-EI (uninformative prior)	Structure toughness
7	Wu <i>et al.</i> <sup>45</sup>	10	Experimental	7	Gryffin algorithm (BO based on kernel density estimation) vs. random search	Nanoparticle plasmonic response
8	Borg <i>et al.</i> <sup>29</sup>	2	Retrospective	3	RF-EI & RF-EV (expected value) vs. random search (identifying single target material)	Band gap of inorganics
9	Borg <i>et al.</i> <sup>29</sup>	4	Retrospective	3	RF-EI & RF-EV vs. random search (identifying five target materials)	Band gap of inorganics
10	Dave <i>et al.</i> <sup>26</sup>	1.3	Retrospective	3	Random search vs. human	Electrolyte ionic conductivity
11	Dave <i>et al.</i> <sup>26</sup>	6	Retrospective	3	GP-MLE vs. random search	Electrolyte ionic conductivity
12	Guay-Hottin <i>et al.</i> <sup>52</sup>	1.42	Retrospective	4	$\alpha$ - $\pi$ BO (GP-EI with dynamic hyperparameter tuning) vs. standard GP-EI	Structure toughness
13	Langner <i>et al.</i> <sup>33</sup>	33	Retrospective	4	Bayesian neural network (BNN) vs. grid	Film photostability
14	Liang <i>et al.</i> <sup>20</sup>	2	Retrospective	4	GP-ARD (automatic relevance detection)-LCB vs. random search	Structure toughness
15	Liang <i>et al.</i> <sup>20</sup>	8	Retrospective	4	RF-LCB (lower confidence bound) vs. random search	Structure toughness
16	Liang <i>et al.</i> <sup>20</sup>	4	Retrospective	4	GP-LCB (lower confidence bound) vs. random search	Structure toughness
17	Liu <i>et al.</i> <sup>34</sup>	61	Retrospective	6	Standard BO & knowledge-constrained BO vs. LHS	Film power conversion efficiency
18	Lookman <i>et al.</i> <sup>31</sup>	3	Retrospective	7	GP-EI vs. random search	Material electrostrain
19	Low <i>et al.</i> <sup>23</sup>	5	Retrospective	8	qNEHVI (q-noisy expected hypervolume improvement) vs. U-NSGA-III (unified non-dominated sorting genetic algorithm III)	Concrete slump & compressive strength
20	Low <i>et al.</i> <sup>23</sup>	20	Retrospective	4	qNEHVI vs. U-NSGA-III	Film conductivity & annealing temperature
21	MacLeod <i>et al.</i> <sup>22</sup>	100	Retrospective	4	qEHVI (q-expected hypervolume improvement) vs. random search	Film conductivity & annealing temperature
22	Rohr <i>et al.</i> <sup>19</sup>	10	Retrospective	6	RF-UCB & GP-UCB vs. random search	Catalyst activity
23	Rohr <i>et al.</i> <sup>19</sup>	5	Retrospective	6	LE (linear ensemble) vs. random search	Catalyst activity
24	Ros <i>et al.</i> <sup>35</sup>	5	Retrospective	6	GP-EI-Thompson sampling & vs. random search	Drug solubility
25	Thelen <i>et al.</i> <sup>27</sup>	5	Retrospective	4	GP-EI & GP-PI (probability of improvement) vs. random search	Battery cycle life
26	Thelen <i>et al.</i> <sup>27</sup>	2	Retrospective	4	GP-UCB vs. random search	Battery cycle life
27	Ament <i>et al.</i> <sup>24</sup>	25	Computational	3	GP-IGU (integrated gradient uncertainty) vs. random search	Phase boundary mapping
28	Annevelink <i>et al.</i> <sup>25</sup>	3	Computational	5	AutoMat-FUELS (forests with uncertainty estimates for learning sequentially) vs. random search	Catalyst activity
29	Annevelink <i>et al.</i> <sup>25</sup>	15	Computational	10	AutoMat-FUELS vs. random search	Battery cycle life
30	Jiang <i>et al.</i> <sup>44</sup>	7.41	Computational	5	Quality diversity (QD) algorithm vs. random search	Nanoparticle extinction spectra



Table 1 (Contd.)

Case	Source	AF	Type	Dimension	Comparison	Objective
31	Lei <i>et al.</i> <sup>60</sup>	8	Computational	10	BART (Bayesian additive regression trees) & BMARS (Bayesian multivariate adaptive regression splines) vs. standard BO	Crystal stacking fault energy
32	Lookman <i>et al.</i> <sup>31</sup>	2	Computational	6	GP-EI vs. RF + EI	LED quantum efficiency
33	Nakayama <i>et al.</i> <sup>51</sup>	1.3	Computational	1	GP-EI vs. human	Synthesis temperature

and then gradually declines. This indicates that the benefit from active learning is most important early in a campaign, where the algorithm can make rapid progress towards the chosen goal. At higher numbers of experiments, the diminishing marginal gains of active learning combined with the continual progress of random sampling mean that the benefit of active learning becomes less important. In other words, if enough of the parameter space will be sampled, the order in which it is sampled is not important. Interestingly, this peak in EF occurs at  $\sim 10$  to 20 experiments per dimension, which provides a useful reference point for the SDL community when planning campaigns. It is worth emphasizing that while EF measured for the SDLs did peak, it was not reported to be worse than random sampling at large numbers of experiments. In order to fully capture the acceleration inherent to SDLs, it would be useful to use multi-fidelity learning or early stopping criteria chosen using simulations.<sup>66,67</sup>

While the number of experiments at which EF peaked was relatively consistent, the peak value of EF varied substantially between studies. However, the difference in magnitude is largely due to two separate metrics both being described as EF. Studies shown as solid lines define EF based upon the enhancement in the property, as we define in Section 2. These studies all have magnitudes in range 1 to 2. The analysis in Section 2 reveals that the maximum attainable value for EF computed in this way is  $C$ , which depends on the property

space. For instance, Zhu *et al.*<sup>40</sup> using experimental design via Bayesian optimization package (EDBO)<sup>59</sup> and Li *et al.*<sup>39</sup> using graph-based Bayesian optimization with pseudo labeling (GBOPL), both benchmarked their algorithms on the crossed barrel dataset, to find modest maximum EF of 1.2 and 1.1, reflecting the narrower performance gap in this property space. This is similar to the EF of 1.2 observed in the experimental benchmarking study by Gongora *et al.*,<sup>37</sup> the source of the dataset. In contrast, studies shown as dashed lines define EF as the enhancement in the number of high-performing combinations of parameters that have been found. These studies have much larger magnitudes. For example, the largest EF observed in our analysis was 23, reported by Fatehi *et al.*,<sup>28</sup> who applied a Bayesian optimization framework with a UCB acquisition function to quantify the proportion of top-performing oxygen evolution reaction (OER) catalysts identified relative to random sampling, using the dataset by Rohr *et al.*<sup>19</sup> While these variations on EF ultimately quantify different things, choosing between them ultimately reflects the priority of the campaign.

## 4. Exploration of benchmarking metrics

While it is clear from the reported values of EF that this metric varies dramatically, it is not clear how this should be interpreted or whether this variation is due to differences in algorithms or the underlying parameter spaces. To explore this, we perform a series of simulated Bayesian optimization campaigns designed to illuminate how  $EF(n)$  depends on the underlying parameter space. In particular, we develop a simple two-dimensional parameter space that features a single Gaussian peak in the center of the space (Fig. 5A). The results of simulated BO campaigns in this space are reported as a horse race plot in which shaded regions depict the interquartile ranges from 100 independent campaigns (Fig. 5B). These are compared to campaigns based on sampling uniformly at random which center on the theoretical performance predicted by eqn (1). These campaigns were executed using the BoTorch package, and the code is shared at <https://doi.org/10.5281/zenodo.17287854>.

In a first round of simulations to explore the magnitude of  $\max(EF)$ , we performed optimization campaigns using five functions that differed only in their contrast  $C$  (Fig. 5A). As expected, all campaigns achieved a  $\max(EF)$  at similar  $n$  but exhibited very different magnitudes depending on the function

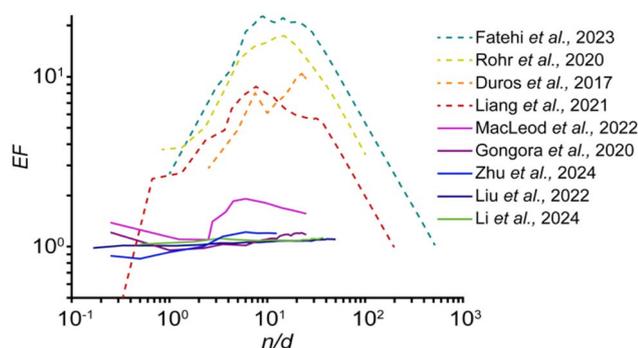


Fig. 4 EF vs. experiment number  $n$  normalized by input parameter space dimensionality  $d$ , extracted from performance-over-iteration data (relative to random sampling) in experimental and retrospective benchmarking SDL studies. Solid lines show EF based on measured property values and dashed lines show EF based on the number of high-performing candidates found.



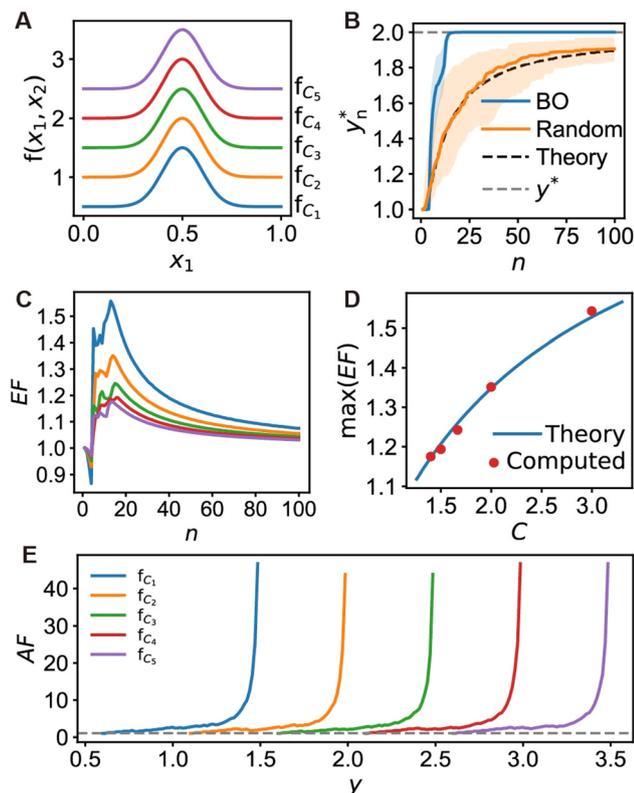


Fig. 5 Simulated Bayesian optimization (BO) campaigns to explore how the property space dictates convergence. (A) Five two-dimensional functions  $f$  under consideration that differ only in their contrast  $C = \max(f)/\text{median}(f)$ . While all are two-dimensional, they depend on  $x_1$  and  $x_2$  in the same way and  $x_2 = 0.5$  is shown. (B) Simulated horse race plot showing the convergence of BO and random sampling for function  $f_{C_2}$ . Theory corresponds to eqn (1). The shaded regions show interquartile ranges. (C) EF vs.  $n$  for the five functions shown in (A). (D)  $\max(\text{EF})$  relating BO and random sampling vs.  $C$ . Dashed line shows a fit to  $\max(\text{EF}) = (\alpha C + 1 - \alpha)/(\beta C + 1 - \beta)$ . (E) AF vs.  $y$  for the five functions shown in (A) showing that they stop at commensurate values. For all functions, AF is plotted until  $y_n^*$  is within 0.01 of  $y^*$  (i.e. surpassed the 99.94th percentile of the function).

(Fig. 5C). Indeed, the theoretical and computed  $\max(\text{EF})$  followed identical trends and monotonically increased with  $C$  (Fig. 5D). These points are fit to  $\max(\text{EF}) = (\alpha C + 1 - \alpha)/(\beta C + 1 - \beta)$ , which reflects the expected EF comparing two campaigns whose rate of convergence does not depend directly on  $C$ . This analysis confirms that while the complexity of the function dictates how many samples are needed to find the optima, its  $C$  bounds EF, partially explaining why the literature features such a wide range in reported  $\max(\text{EF})$ .

While EF clearly depends on many  $C$ , the progression of AF throughout a campaign along with its maximum does not (Fig. 5E). In particular, AF is found to monotonically increase throughout a campaign and reach a maximum when the learning algorithm has found  $y^*$ . Two facets of this trajectory make AF more suitable than EF as a metric for broad comparison. First, being equivariant with the output space is congruent with our expectations that shifting by a constant should not affect the quality of a learning algorithm. Second, being

monotonic means makes it easier to compare campaigns with a single value, namely  $\max(\text{AF})$ .

While the functions explored in Fig. 5 exhibited the same complexity, we sought to explore whether one can use simple statistics of a function to gain insight into how many experiments are needed to achieve optimum performance. In particular, we explore Lipschitz complexity  $L$ , which is defined as,<sup>68</sup>

$$L = \max|\nabla f|, \quad (5)$$

where  $|\nabla f|$  represents the magnitude of the gradient of the function  $f$  in which each independent variable has been normalized to fall between 0 and 1. We construct a family of functions with the same  $C$  but different  $L$  by changing the standard deviation  $\sigma_a$  and linear offset of a two-dimensional Gaussian (Fig. 6A). Unlike the case when only  $C$  is changed, each campaign requires different numbers of experiments to converge with sharper functions requiring more experiments (Fig. 6B). Interestingly, we find a linear relationship between  $L$  and  $n_{\text{AL}}^*$ , highlighting the challenge inherent to parameter spaces that appear to be needles in a haystack. Interestingly, the empirically observed best experiment number  $n_{\text{AL}}^*$  from the

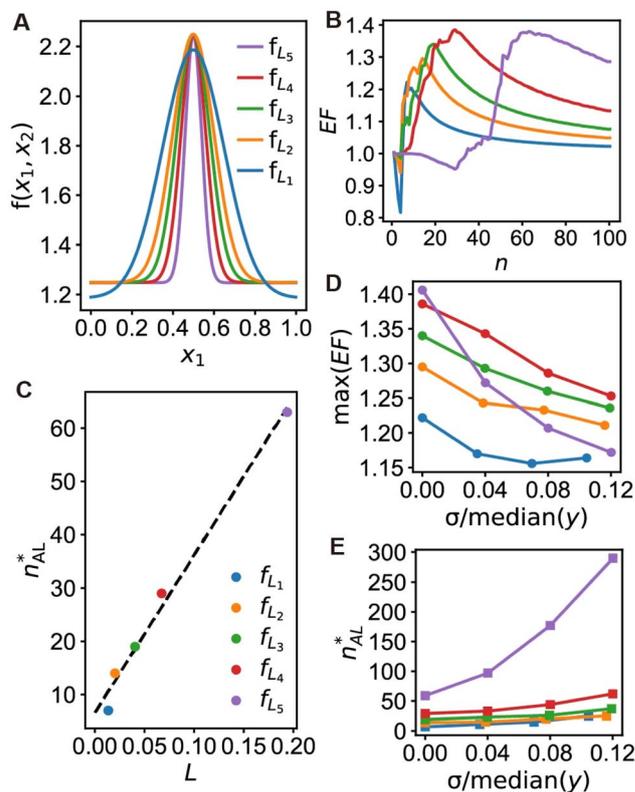


Fig. 6 Simulated BO campaigns to explore how property space complexity impacts learning. (A) Five two-dimensional parameter spaces  $f$  under consideration that differ only in their Lipschitz complexity  $L$ , as defined in eqn (5). While all are two-dimensional, they depend on  $x_1$  and  $x_2$  in the same manner and  $x_2 = 0.5$  is shown. (B) EF vs.  $n$  for the five functions shown in (A). (C) Optimum experiment number  $n_{\text{AL}}^*$  corresponding to  $\max(\text{EF})$  vs.  $L$ . The dashed line shows a linear fit. (D)  $\max(\text{EF})$  vs. noise standard deviation  $\sigma$  normalized by  $\text{median}(y)$ . (E)  $n_{\text{AL}}^*$  and vs.  $\sigma$  normalized by  $\text{median}(y)$ .



literature appears to be  $\sim 15/d$ , which amounts to 30 experiments in the present example. This suggests that the functions explored here share statistical features with the materials spaces previously studied. Importantly,  $\max(\text{EF})$  increases with  $L$ , highlighting that it is more impactful to use active learning in parameter spaces that are more difficult to learn.

The analytical spaces considered here are deterministic, while experimental parameter spaces will necessarily feature noise. In an effort to understand how the presence of noise will impact convergence, simulated BO campaigns were repeated for the functions shown in Fig. 6A with homoscedastic Gaussian noise with standard deviation  $\sigma$  added. Both  $\max(\text{EF})$  and  $n_{\text{AL}}^*$  had a smooth dependence on  $\sigma$  (Fig. 6D), with the most complex functions exhibiting drastic increases in  $n_{\text{AL}}^*$  (Fig. 6E). This result indicates that reducing noise becomes more important the more complex the parameter space. The observation that noise slows convergence is consistent with prior analytical work.<sup>69</sup> The range of noise was chosen to be analogous to the range of noise typically found in experimental systems. It is not expected that larger values of noise will change the trends observed, only that it would make the simulations take longer to converge.

While these heuristic simulations have focused on single objectives, many recent SDLs focus on multiple objectives simultaneously. That said, the most widely-used approach for multi-objective optimization is hypervolume optimization wherein the algorithm seeks to maximally improve the Pareto front balancing all objectives.<sup>22,23</sup> In many ways, once this type of problem has been transitioned into a scalar optimization (*i.e.* maximizing hypervolume), the same types of benchmarking could be done to compare performance of an active learning algorithm and a reference process. Simulations of such processes reveal similar non-monotonic behavior of EF and monotonically increasing AF,<sup>22</sup> suggesting that the principles studied here apply to multi-objective cases as well.

These heuristic simulations have provided context for how to interpret AF and EF values generated by SDL campaigns and guidance for how reparameterizing the input or output space may affect convergence. One reason why EF is an imperfect metric is that shifting the output space by a constant will change EF but not impact the actual learning rate. In contrast, applying a non-linear transform to the property space that reduces  $L$  is likely to accelerate convergence. Analogously, narrowing parameter space to focus on regions of interest will similarly reduce  $L$ , which provides a mechanism for understanding how approaches such as ZoMBI improve learning.<sup>70</sup> AF is likely a more useful metric for comparing algorithms, but it still depends on the length of campaigns and, being monotonic, it does not help experimentalists determine the point of diminishing returns for additional experiments.

## 5. Conclusions and future recommendations

Benchmarking SDLs is important because it provides part of the justification for developing and running these systems. As

a result, there have been significant efforts in the community to quantify performance. The two most reported metrics are the enhancement factor EF and the acceleration factor AF, which address the questions of how much better and how much faster, respectively. A systematic evaluation of the reported metrics reveals key insights:

(i) SDLs achieve top-performing results on average six times faster than random sampling, and this acceleration improves with the dimensionality of the parameter space.

(ii) The enhancement inherent to SDLs is reported to peak at 10–20 experiments per dimension of parameter space, with enhancement factors that vary tremendously depending on the space.

It is important to highlight that both of these outcomes depend intimately on the nature of the property spaces, but the fact that these all represent actual experimental materials datasets suggests that they are useful guidelines for the field. Further, simulated campaigns in analytical spaces reveal key features of how to interpret metrics, namely that EF can simply be related to the statistics of the parameter space such as its contrast, the complexity of the space determines the speed with which convergence can be expected, and that noise affects AF more than EF. Despite the simplicity of the heuristic simulations presented here, the fact that they required similar numbers of experiments to converge relative to what is seen in the SDL literature suggests that these functions share statistical features with studied materials systems. With the growing confidence and expertise present in the SDL field, researchers will undoubtedly explore much more complex spaces going forward. While the specific values in this study will hopefully be improved upon in the coming years as more advanced algorithms are employed, they nevertheless provide a valuable snapshot of the field and a useful tool to align progress. While there are many ways to parameterize a function that might be useful to contextualize benchmarking, we have focused on contrast as defined in eqn (4) and the Lipschitz constant. The former directly bounds EF and is a very straightforward property to compute while the latter is widely used in machine learning to evaluate models.<sup>68,71–74</sup> Other factors can play an important role in optimization such as multimodality (having multiple local optima) or anisotropy (having very different gradients in different directions). The presence of these and other factors emphasizes that the simulations shown here are heuristic and more in-depth study is needed. Addressing the materials challenges facing our society demands rapid progress and a thorough analysis of methods to accelerate this progress is necessary to move the field forward.

## Conflicts of interest

The authors declare no conflicts of interest.

## Data availability

The simulations used to create the plots in Fig. 5 and 6 and the data used to create the plots in Fig. 3 and 4 are available at: <https://doi.org/10.5281/zenodo.17287854>. These resources



include a Jupyter notebook that facilitates the exploration of more complicated parameter spaces.

## Acknowledgements

The authors thank Robert Brown and Peter Frazier for helpful conversations. The authors acknowledge the Hariri Institute at Boston University (2024-07-001), the National Science Foundation (DMR-2323728), and the Army Research Office (W911NF2420095) for supporting this research.

## References

- 1 J. J. de Pablo, N. E. Jackson, M. A. Webb, L.-Q. Chen, J. E. Moore, D. Morgan, R. Jacobs, T. Pollock, D. G. Schlom and E. S. Toberer, New frontiers for the materials genome initiative, *npj Comput. Mater.*, 2019, **5**(1), 1–23, DOI: [10.1038/s41524-019-0173-4](https://doi.org/10.1038/s41524-019-0173-4).
- 2 R. B. Canty, J. A. Bennett, K. A. Brown, T. Buonassisi, S. V. Kalinin, J. R. Kitchin, B. Maruyama, R. G. Moore, J. Schrier, M. Seifrid, S. Sun, T. Vegge and M. Abolhasani, Science acceleration and accessibility with self-driving labs, *Nat. Commun.*, 2025, **16**(1), 3856, DOI: [10.1038/s41467-025-59231-1](https://doi.org/10.1038/s41467-025-59231-1).
- 3 R. Vescovi, T. Ginsburg, K. Hippe, D. Ozgulbas, C. Stone, A. Stroka, R. Butler, B. Blaiszik, T. Brettin, K. Chard, M. Hereld, A. Ramanathan, R. Stevens, A. Vriza, J. Xu, Q. Zhang and I. Foster, Towards a modular architecture for science factories, *Digital Discovery*, 2023, **2**(6), 1980–1998, DOI: [10.1039/D3DD00142C](https://doi.org/10.1039/D3DD00142C).
- 4 E. Stach, B. DeCost, A. G. Kusne, J. Hattrick-Simpers, K. A. Brown, K. G. Reyes, J. Schrier, S. Billinge, T. Buonassisi, I. Foster, C. P. Gomes, J. M. Gregoire, A. Mehta, J. Montoya, E. Olivetti, C. Park, E. Rotenberg, S. K. Saikin, S. Smullin, V. Stanev and B. Maruyama, Autonomous experimentation systems for materials development: A community perspective, *Matter*, 2021, **4**(9), 2702–2726, DOI: [10.1016/j.matt.2021.06.036](https://doi.org/10.1016/j.matt.2021.06.036).
- 5 G. Tom, S. P. Schmid, S. G. Baird, Y. Cao, K. Darvish, H. Hao, S. Lo, S. Pablo-García, E. M. Rajaonson, M. Skreta, N. Yoshikawa, S. Corapi, G. D. Akkoc, F. Strieth-Kalthoff, M. Seifrid and A. Aspuru-Guzik, Self-driving laboratories for chemistry and materials science, *Chem. Rev.*, 2024, **124**(16), 9633–9732, DOI: [10.1021/acs.chemrev.4c00055](https://doi.org/10.1021/acs.chemrev.4c00055).
- 6 E. J. Kluender, J. L. Hedrick, K. A. Brown, R. Rao, B. Meckes, J. Du, L. Moreau, B. Maruyama and C. A. Mirkin, Catalyst discovery through megalibraries of nanomaterials, *Proc. Natl. Acad. Sci. U. S. A.*, 2019, **116**(1), 40–45, DOI: [10.1073/pnas.1815358116](https://doi.org/10.1073/pnas.1815358116).
- 7 J. Zhou, M. Luo, L. Chen, Q. Zhu, S. Jiang, F. Zhang, W. Shang and J. Jiang, A multi-robot–multi-task scheduling system for autonomous chemistry laboratories, *Digital Discovery*, 2025, **4**(3), 636–652, DOI: [10.1039/D4DD00031F](https://doi.org/10.1039/D4DD00031F).
- 8 M. Zaki, C. Prinz and B. Ruehle, A self-driving lab for nano- and advanced materials synthesis, *ACS Nano*, 2025, **19**(9), 9029–9041, DOI: [10.1021/acsnano.4c17504](https://doi.org/10.1021/acsnano.4c17504).
- 9 C. Wang, Y.-J. Kim, A. Vriza, R. Batra, A. Baskaran, N. Shan, N. Li, P. Darancet, L. Ward, Y. Liu, M. K. Y. Chan, S. K. R. S. Sankaranarayanan, H. C. Fry, C. S. Miller, H. Chan and J. Xu, Autonomous platform for solution processing of electronic polymers, *Nat. Commun.*, 2025, **16**(1), 1498, DOI: [10.1038/s41467-024-55655-3](https://doi.org/10.1038/s41467-024-55655-3).
- 10 T. Song, M. Luo, X. Zhang, L. Chen, Y. Huang, J. Cao, Q. Zhu, D. Liu, B. Zhang, G. Zou, G. Zhang, F. Zhang, W. Shang, Y. Fu, J. Jiang and Y. Luo, A multiagent-driven robotic AI chemist enabling autonomous chemical research on demand, *J. Am. Chem. Soc.*, 2025, **147**, 12534–12545, DOI: [10.1021/jacs.4c17738](https://doi.org/10.1021/jacs.4c17738).
- 11 A. Sanin, J. K. Flowers, T. H. Piotrowiak, F. Felsen, L. Merker, A. Ludwig, D. Bresser and H. S. Stein, Integrating automated electrochemistry and high-throughput characterization with machine learning to explore si–ge–sn thin-film lithium battery anodes, *Adv. Energy Mater.*, 2025, **15**(11), 2404961, DOI: [10.1002/aenm.202404961](https://doi.org/10.1002/aenm.202404961).
- 12 S. Putz, J. Döttling, T. Ballweg, A. Tschöpe, V. Biniyaminov and M. Franzreb, Self-driving lab for solid-phase extraction process optimization and application to nucleic acid purification, *Adv. Intell. Syst.*, 2025, **7**(1), 2400564, DOI: [10.1002/aisy.202400564](https://doi.org/10.1002/aisy.202400564).
- 13 K. Nishio, A. Aiba, K. Takihara, Y. Suzuki, R. Nakayama, S. Kobayashi, A. Abe, H. Baba, S. Katagiri, K. Omoto, K. Ito, R. Shimizu and T. Hitosugi, A digital laboratory with a modular measurement system and standardized data format, *Digital Discovery*, 2025, **4**, 1734–1742, DOI: [10.1039/D4DD00326H](https://doi.org/10.1039/D4DD00326H).
- 14 F. Strieth-Kalthoff, H. Hao, V. Rathore, J. Derasp, T. Gaudin, N. H. Angello, M. Seifrid, E. Trushina, M. Guy, J. Liu, X. Tang, M. Mamada, W. Wang, T. Tsagaantsooj, C. Lavigne, R. Pollice, T. C. Wu, K. Hotta, L. Bodo, S. Li, M. Haddadnia, A. Wołos, R. Roszak, C. T. Ser, C. Bozal-Ginesta, R. J. Hickman, J. Vestfrid, A. Aguilar-Granda, E. L. Klimareva, R. C. Sigerson, W. Hou, D. Gahler, S. Lach, A. Warzybok, O. Borodin, S. Rohrbach, B. Sanchez-Lengeling, C. Adachi, B. A. Grzybowski, L. Cronin, J. E. Hein, M. D. Burke and A. Aspuru-Guzik, Delocalized, asynchronous, closed-loop discovery of organic laser emitters, *Science*, 2024, **384**, 1–9, DOI: [10.1126/science.adk9227](https://doi.org/10.1126/science.adk9227).
- 15 K. L. Snapp, B. Verdier, A. E. Gongora, S. Silverman, A. D. Adesiji, E. F. Morgan, T. J. Lawton, E. Whiting and K. A. Brown, Superlative mechanical energy absorbing efficiency discovered through self-driving lab-human partnership, *Nat. Commun.*, 2024, **15**(1), 4290, DOI: [10.1038/s41467-024-48534-4](https://doi.org/10.1038/s41467-024-48534-4).
- 16 S. Matsuda, G. Lambard and K. Sodeyama, Data-driven automated robotic experiments accelerate discovery of multi-component electrolyte for rechargeable Li–O<sub>2</sub> batteries, *Cell Rep. Phys. Sci.*, 2022, **3**(4), 100832, DOI: [10.1016/j.xcrp.2022.100832](https://doi.org/10.1016/j.xcrp.2022.100832).
- 17 F. Delgado-Licona and M. Abolhasani, Research acceleration in self-driving labs: technological roadmap toward accelerated materials and molecular discovery, *Adv. Intell. Syst.*, 2023, **5**(4), 2200331, DOI: [10.1002/aisy.202200331](https://doi.org/10.1002/aisy.202200331).



- 18 L. Hung, J. A. Yager, D. Monteverde, D. Baiocchi, H.-K. Kwon, S. Sun and S. Suram, Autonomous laboratories for accelerated materials discovery: a community survey and practical insights, *Digital Discovery*, 2024, 3(7), 1273–1279, DOI: [10.1039/D4DD00059E](https://doi.org/10.1039/D4DD00059E).
- 19 B. Rohr, H. S. Stein, D. Guevarra, Y. Wang, J. A. Haber, M. Aykol, S. K. Suram and J. M. Gregoire, Benchmarking the acceleration of materials discovery by sequential learning, *Chem. Sci.*, 2020, 11(10), 2696–2706, DOI: [10.1039/c9sc05999g](https://doi.org/10.1039/c9sc05999g).
- 20 Q. H. Liang, A. E. Gongora, Z. K. Ren, A. Tiihonen, Z. Liu, S. J. Sun, J. R. Deneault, D. Bash, F. Mekki-Berrada, S. A. Khan, K. Hippalgaonkar, B. Maruyama, K. A. Brown, J. I. I. Fisher and T. Buonassisi, Benchmarking the performance of Bayesian optimization across multiple experimental materials science domains, *npj Comput. Mater.*, 2021, 7(1), 188, DOI: [10.1038/s41524-021-00656-9](https://doi.org/10.1038/s41524-021-00656-9).
- 21 A. A. Volk, R. W. Epps, D. T. Yonemoto, B. S. Masters, F. N. Castellano, K. G. Reyes and M. Abolhasani, AlphaFlow: autonomous discovery and optimization of multi-step chemistry using a self-driven fluidic lab guided by reinforcement learning, *Nat. Commun.*, 2023, 14(1), 1403, DOI: [10.1038/s41467-023-37139-y](https://doi.org/10.1038/s41467-023-37139-y).
- 22 B. P. MacLeod, F. G. L. Parlane, C. C. Rupnow, K. E. Dettelbach, M. S. Elliott, T. D. Morrissey, T. H. Haley, O. Proskurin, M. B. Rooney, N. Taherimakhosousi, D. J. Dvorak, H. N. Chiu, C. E. B. Waizenegger, K. Ocean, M. Mokhtari and C. P. Berlinguette, A self-driving laboratory advances the Pareto front for material properties, *Nat. Commun.*, 2022, 13(1), 995, DOI: [10.1038/s41467-022-28580-6](https://doi.org/10.1038/s41467-022-28580-6).
- 23 A. K. Y. Low, E. Vissol-Gaudin, Y. F. Lim and K. Hippalgaonkar, Mapping pareto fronts for efficient multi-objective materials discovery, *J. Mater. Inf.*, 2023, 3(2), 1–19, DOI: [10.20517/jmi.2023.02](https://doi.org/10.20517/jmi.2023.02).
- 24 S. Ament, M. Amsler, D. R. Sutherland, M. C. Chang, D. Guevarra, A. B. Connolly, J. M. Gregoire, M. O. Thompson, C. P. Gomes and R. B. van Dover, Autonomous materials synthesis via hierarchical active learning of nonequilibrium phase diagrams, *Sci. Adv.*, 2021, 7(51), 1–12, DOI: [10.1126/sciadv.abg4930](https://doi.org/10.1126/sciadv.abg4930).
- 25 E. Annevelink, R. Kurchin, E. Muckley, L. Kavalsky, V. I. Hegde, V. Sulzer, S. Zhu, J. K. Pu, D. Farina, M. Johnson, D. Gandhi, A. Dave, H. Y. Lin, A. Edelman, B. Ramsundar, J. Saal, C. Rackaukas, V. R. Shah, B. Meredig and V. Viswanathan, AutoMat: Automated materials discovery for electrochemical systems, *MRS Bull.*, 2022, 47(10), 1036–1044, DOI: [10.1557/s43577-022-00424-0](https://doi.org/10.1557/s43577-022-00424-0).
- 26 A. Dave, J. Mitchell, S. Burke, H. Y. Lin, J. Whitacre and V. Viswanathan, Autonomous optimization of non-aqueous Li-ion battery electrolytes via robotic experimentation and machine learning coupling, *Nat. Commun.*, 2022, 13(1), 5454, DOI: [10.1038/s41467-022-32938-1](https://doi.org/10.1038/s41467-022-32938-1).
- 27 A. Thelen, M. Zohair, J. Ramamurthy, A. Harkaway, W. M. Jiao, M. Ojha, M. Ul Ishtiaque, T. A. Kingston, C. L. Pint and C. Hu, Sequential Bayesian optimization for accelerating the design of sodium metal battery nucleation layers, *J. Power Sources*, 2023, 581, 1–14, DOI: [10.1016/j.jpowsour.2023.233508](https://doi.org/10.1016/j.jpowsour.2023.233508).
- 28 E. Fatehi, M. Thadani, G. Birsan and R. W. Black, *arXiv*, preprint, arXiv:2305.12541, 2023.
- 29 C. K. H. Borg, E. S. Muckley, C. Nyby, J. E. Saal, L. Ward, A. Mehta and B. Meredig, Quantifying the performance of machine learning models in materials discovery, *Digital Discovery*, 2023, 2(2), 327–338, DOI: [10.1039/d2dd00113f](https://doi.org/10.1039/d2dd00113f).
- 30 P. Honarmandi, V. Attari and R. Arroyave, Accelerated materials design using batch Bayesian optimization: A case study for solving the inverse problem from materials microstructure to process, *Comput. Mater. Sci.*, 2022, 210, 111417, DOI: [10.1016/j.commatsci.2022.111417](https://doi.org/10.1016/j.commatsci.2022.111417).
- 31 T. Lookman, P. V. Balachandran, D. Z. Xue and R. H. Yuan, Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design, *npj Comput. Mater.*, 2019, 5, 21, DOI: [10.1038/s41524-019-0153-8](https://doi.org/10.1038/s41524-019-0153-8).
- 32 D. N. Cakan, E. Oberholtz, K. Kaushal, S. P. Dunfield and D. P. Fenning, Bayesian optimization and prediction of the durability of triple-halide perovskite thin films under light and heat stressors, *Mater. Adv.*, 2025, 6(2), 598–606, DOI: [10.1039/d4ma00747f](https://doi.org/10.1039/d4ma00747f).
- 33 S. Langner, F. Häse, J. D. Perea, T. Stubhan, J. Hauch, L. M. Roch, T. Heumueller, A. Aspuru-Guzik and C. J. Brabec, Beyond ternary opv: high-throughput experimentation and self-driving laboratories optimize multicomponent systems, *Adv. Mater.*, 2020, 32(14), 1907801, DOI: [10.1002/adma.201907801](https://doi.org/10.1002/adma.201907801).
- 34 Z. Liu, N. Rolston, A. C. Flick, T. W. Colburn, Z. K. Ren, R. H. Dauskardt and T. Buonassisi, Machine learning with knowledge constraints for process optimization of open-air perovskite solar cell manufacturing, *Joule*, 2022, 6(4), 834–849, DOI: [10.1016/j.joule.2022.03.003](https://doi.org/10.1016/j.joule.2022.03.003).
- 35 H. Ros, M. Cook and D. Shorthouse, Efficient discovery of new medicine formulations using a semi-self-driven robotic formulator, *Digital Discovery*, 2025, 4(8), 2263–2272, DOI: [10.1039/D5DD00171D](https://doi.org/10.1039/D5DD00171D).
- 36 R. K. Vasudevan, K. P. Kelley, J. Hinkle, H. Funakubo, S. Jesse, S. V. Kalinin and M. Ziatdinov, Autonomous experiments in scanning probe microscopy and spectroscopy: choosing where to explore polarization dynamics in ferroelectrics, *ACS Nano*, 2021, 15(7), 11253–11262, DOI: [10.1021/acsnano.0c10239](https://doi.org/10.1021/acsnano.0c10239).
- 37 A. E. Gongora, B. W. Xu, W. Perry, C. Okoye, P. Riley, K. G. Reyes, E. F. Morgan and K. A. Brown, A Bayesian experimental autonomous researcher for mechanical design, *Sci. Adv.*, 2020, 6(15), 1–6, DOI: [10.1126/sciadv.aaz1708](https://doi.org/10.1126/sciadv.aaz1708).
- 38 A. E. Gongora, K. L. Snapp, E. Whiting, P. Riley, K. G. Reyes, E. F. Morgan and K. A. Brown, Using simulation to accelerate autonomous experimentation: A case study using mechanics, *iScience*, 2021, 24(4), 1–10, DOI: [10.1016/j.isci.2021.102262](https://doi.org/10.1016/j.isci.2021.102262).
- 39 G. Y. Li and X. N. Jin, Mechanical design parameter optimization through graph-based bayesian optimization



- and pseudo labeling, *IEEE Int Con Auto Sc*, 2024, 2955–2960, DOI: [10.1109/Case59546.2024.10711759](https://doi.org/10.1109/Case59546.2024.10711759).
- 40 M. J. Zhu, A. Mroz, L. F. Gui, K. E. Jelfs, A. Bemporad, E. A. D. Chanona and Y. S. Lee, Discrete and mixed-variable experimental design with surrogate-based approach, *Digital Discovery*, 2024, 3(12), 2589–2606, DOI: [10.1039/d4dd000113c](https://doi.org/10.1039/d4dd000113c).
- 41 J. Ziomek, M. Adachi and M. A. Osborne, *arXiv*, preprint, arXiv:2410.10384, 2024.
- 42 F. Bateni, S. Sadeghi, N. Orouji, J. A. Bennett, V. S. Punati, C. Stark, J. Y. Wang, M. C. Rosko, O. Chen, F. N. Castellano, K. G. Reyes and M. Abolhasani, Smart dope: a self-driving fluidic lab for accelerated development of doped perovskite quantum dots, *Adv. Energy Mater.*, 2024, 14(1), 2302303, DOI: [10.1002/aeam.202470001](https://doi.org/10.1002/aeam.202470001).
- 43 R. W. Epps, M. S. Bowen, A. A. Volk, K. Abdel-Latif, S. Y. Han, K. G. Reyes, A. Amassian and M. Abolhasani, Artificial chemist: an autonomous quantum dot synthesis bot, *Adv. Mater.*, 2020, 32(30), 2001626, DOI: [10.1002/adma.202001626](https://doi.org/10.1002/adma.202001626).
- 44 Y. B. Jiang, D. Salley, A. Sharma, G. Keenan, M. Mullin and L. Cronin, An artificial intelligence enabled chemical synthesis robot for exploration and optimization of nanomaterials, *Sci. Adv.*, 2022, 8(40), 1–11, DOI: [10.1126/sciadv.abo2626](https://doi.org/10.1126/sciadv.abo2626).
- 45 T. Y. Wu, S. Kheiri, R. J. Hickman, H. C. Tao, T. C. Wu, Z. B. Yang, X. Ge, W. Zhang, M. Abolhasani, K. Liu, A. Aspuru-Guzik and E. Kumacheva, Self-driving lab for the photochemical synthesis of plasmonic nanoparticles with targeted structural and optical properties, *Nat. Commun.*, 2025, 16(1), 1473, DOI: [10.1038/s41467-025-56788-9](https://doi.org/10.1038/s41467-025-56788-9).
- 46 S. Sadeghi, F. Bateni, T. Kim, D. Y. Son, J. A. Bennett, N. Orouji, V. S. Punati, C. Stark, T. D. Cerra, R. Awad, F. Delgado-Licona, J. E. Xu, N. Mukhin, H. Dickerson, K. G. Reyes and M. Abolhasani, Autonomous nanomanufacturing of lead-free metal halide perovskite nanocrystals using a self-driving fluidic lab, *Nanoscale*, 2024, 16(2), 580–591, DOI: [10.1039/d3nr05034c](https://doi.org/10.1039/d3nr05034c).
- 47 C. Tamura, H. Job, H. Chang, W. Wang, Y. Liang and S. Sun, *ChemRxiv*, 2025, preprint, DOI: [10.26434/chemrxiv-2025-11bzs](https://doi.org/10.26434/chemrxiv-2025-11bzs).
- 48 R. B. Canty and M. Abolhasani, Reproducibility in automated chemistry laboratories using computer science abstractions, *Nat. Synth.*, 2024, 3(11), 1327–1339, DOI: [10.1038/s44160-024-00649-8](https://doi.org/10.1038/s44160-024-00649-8).
- 49 R. Rauschen, M. Guy, J. E. Hein and L. Cronin, Universal chemical programming language for robotic synthesis repeatability, *Nat. Synth.*, 2024, 3(4), 488–496, DOI: [10.1038/s44160-023-00473-6](https://doi.org/10.1038/s44160-023-00473-6).
- 50 J. Bai, S. Mosbach, C. Taylor, D. Karan, K. Lee, S. Rihm, J. Akroyd, A. Lapkin and M. Kraft, A dynamic knowledge graph approach to distributed self-driving laboratories, *Nat. Commun.*, 2024, 15(1), 462, DOI: [10.1038/s41467-023-44599-9](https://doi.org/10.1038/s41467-023-44599-9).
- 51 R. Nakayama, R. Shimizu, T. Haga, T. Kimura, Y. Ando, S. Kobayashi, N. Yasuo, M. Sekijima and T. Hitosugi, Tuning of Bayesian optimization for materials synthesis: simulation of the one-dimensional case, *Sci. Technol. Adv. Mater.:Methods*, 2022, 2(1), 119–128, DOI: [10.1080/27660400.2022.2066489](https://doi.org/10.1080/27660400.2022.2066489).
- 52 R. Guay-Hottin, L. Kardassevitch, H. Pham, G. Lajoie and M. Bonizzato, Robust prior-biased acquisition function for human-in-the-loop Bayesian optimization, *Knowl.-Based Syst.*, 2025, 311(1), 113039, DOI: [10.1016/j.knosys.2025.113039](https://doi.org/10.1016/j.knosys.2025.113039).
- 53 V. Duros, J. Grizou, W. M. Xuan, Z. Hosni, D. L. Long, H. N. Miras and L. Cronin, Human versus robots in the discovery and crystallization of gigantic polyoxometalates, *Angew. Chem., Int. Ed.*, 2017, 56(36), 10815–10820, DOI: [10.1002/anie.201705721](https://doi.org/10.1002/anie.201705721).
- 54 J. Grizou, L. J. Points, A. Sharma and L. Cronin, A curious formulation robot enables the discovery of a novel protocell behavior, *Sci. Adv.*, 2020, 6(5), 1–10, DOI: [10.1126/sciadv.aay4237](https://doi.org/10.1126/sciadv.aay4237).
- 55 Y. Bai, Z. H. J. Khoo, R. Made, H. Q. Xie, C. Y. J. Lim, A. D. Handoko, V. Chellappan, J. J. Cheng, F. X. Wei, Y. F. Lim and K. Hippalgaonkar, Closed-loop multi-objective optimization for Cu-Sb-S photo-electrocatalytic materials' discovery, *Adv. Mater.*, 2024, 36(2), 2304269, DOI: [10.1002/adma.202304269](https://doi.org/10.1002/adma.202304269).
- 56 L. Kavalsky, V. I. Hegde, E. Muckley, M. S. Johnson, B. Meredig and V. Viswanathan, By how much can closed-loop frameworks accelerate computational materials discovery?, *Digital Discovery*, 2023, 2(4), 1112–1125, DOI: [10.1039/d2dd000133k](https://doi.org/10.1039/d2dd000133k).
- 57 Q. Liang, A. E. Gongora, Z. Ren, A. Tiihonen, Z. Liu, S. Sun, J. R. Deneault, D. Bash, F. Mekki-Berrada, S. A. Khan, K. Hippalgaonkar, B. Maruyama, K. A. Brown, J. Fisher Iii and T. Buonassisi, Benchmarking the performance of Bayesian optimization across multiple experimental materials science domains, *npj Comput. Mater.*, 2021, 7(1), 188, DOI: [10.1038/s41524-021-00656-9](https://doi.org/10.1038/s41524-021-00656-9).
- 58 F. Conrad, M. Mälzer, M. Schwarzenberger, H. Wiemer and S. Ihlenfeldt, Benchmarking AutoML for regression tasks on small tabular data in materials design, *Sci. Rep.*, 2022, 12(1), 19350, DOI: [10.1038/s41598-022-23327-1](https://doi.org/10.1038/s41598-022-23327-1).
- 59 B. J. Shields, J. Stevens, J. Li, M. Parasram, F. Damani, J. I. M. Alvarado, J. M. Janey, R. P. Adams and A. G. Doyle, Bayesian reaction optimization as a tool for chemical synthesis, *Nature*, 2021, 590(7844), 89–96, DOI: [10.1038/s41586-021-03213-y](https://doi.org/10.1038/s41586-021-03213-y).
- 60 B. W. Lei, T. Q. Kirk, A. Bhattacharya, D. Pati, X. N. Qian, R. Arroyave and B. K. Mallick, Bayesian optimization with adaptive surrogate models for automated experimental design, *npj Comput. Mater.*, 2021, 7(1), 194, DOI: [10.1038/s41524-021-00662-x](https://doi.org/10.1038/s41524-021-00662-x).
- 61 D. A. Cohn, Z. Ghahramani and M. I. Jordan, Active learning with statistical models, *Artif. Intell. Res.*, 1996, 4, 129–145, DOI: [10.1613/jair.295](https://doi.org/10.1613/jair.295).
- 62 F. Berkenkamp, A. P. Schoellig and A. Krause, No-regret Bayesian optimization with unknown hyperparameters, *J. Mach. Learn. Res.*, 2019, 20(50), 1–24, <https://www.jmlr.org/papers/volume20/18-213/18-213.pdf>.



- 63 D. C. Liu and J. Nocedal, On the limited memory BFGS method for large scale optimization, *Math. Program*, 1989, **45**(1), 503–528, DOI: [10.1007/BF01589116](https://doi.org/10.1007/BF01589116).
- 64 M. D. Hoffman and A. Gelman, The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo, *J. Mach. Learn. Res.*, 2014, **15**(1), 1593–1623, <https://www.jmlr.org/papers/volume15/hoffman14a/hoffman14a.pdf>.
- 65 F. Mekki-Berrada, Z. Ren, T. Huang, W. K. Wong, F. Zheng, J. Xie, I. P. S. Tian, S. Jayavelu, Z. Mahfoud and D. Bash, Two-step machine learning enables optimized nanoparticle synthesis, *npj Comput. Mater.*, 2021, **7**, 55, DOI: [10.1038/s41524-021-00520-w](https://doi.org/10.1038/s41524-021-00520-w).
- 66 T. Savage, N. Basha, J. McDonough, J. Krassowski, O. Matar and E. A. del Rio Chanona, Machine learning-assisted discovery of flow reactor designs, *Nat. Chem. Eng.*, 2024, **1**(8), 522–531, DOI: [10.1038/s44286-024-00099-1](https://doi.org/10.1038/s44286-024-00099-1).
- 67 V. Sabanza-Gil, R. Barbano, D. Pacheco Gutiérrez, J. S. Luterbacher, J. M. Hernández-Lobato, P. Schwaller and L. Roch, Best practices for multi-fidelity Bayesian optimization in materials and molecular research, *Nat. Comput. Sci.*, 2025, **5**, 572–581, DOI: [10.1038/s43588-025-00822-9](https://doi.org/10.1038/s43588-025-00822-9).
- 68 G. R. Wood and B. Zhang, Estimation of the Lipschitz constant of a function, *J. Global Optim.*, 1996, **8**(1), 91–103, DOI: [10.1007/BF00229304](https://doi.org/10.1007/BF00229304).
- 69 A. A. Volk and M. Abolhasani, Performance metrics to unleash the power of self-driving labs in chemistry and materials science, *Nat. Commun.*, 2024, **15**(1), 1378, DOI: [10.1038/s41467-024-45569-5](https://doi.org/10.1038/s41467-024-45569-5).
- 70 A. E. Siemenn, Z. Ren, Q. Li and T. Buonassisi, Fast Bayesian optimization of Needle-in-a-Haystack problems using zooming memory-based initialization (ZoMBI), *npj Comput. Mater.*, 2023, **9**(1), 79, DOI: [10.1038/s41524-023-01048-x](https://doi.org/10.1038/s41524-023-01048-x).
- 71 A. Virmaux and K. Scaman, Lipschitz regularity of deep neural networks: analysis and efficient estimation, *Adv Neural Inf. Process Syst.*, 2018, **31**, 1–10.
- 72 O. L. Mangasarian and T.-H. Shiau, Lipschitz continuity of solutions of linear inequalities, programs and complementarity problems, *SIAM J. Control Optim.*, 1987, **25**(3), 583–595, DOI: [10.1137/0325033](https://doi.org/10.1137/0325033).
- 73 A. Ambroladze, E. Parrado-Hernández and J. Shawe-Taylor, Complexity of pattern classes and the Lipschitz property, *Theor. Comput. Sci.*, 2007, **382**(3), 232–246, DOI: [10.1016/j.tcs.2007.03.047](https://doi.org/10.1016/j.tcs.2007.03.047).
- 74 T. Boult and K. Sikorski, Complexity of computing topological degree of Lipschitz functions in n dimensions, *J. Complex*, 1986, **2**(1), 44–59, DOI: [10.1016/0885-064X\(86\)90022-1](https://doi.org/10.1016/0885-064X(86)90022-1).

