Digital Discovery

Accepted Manuscript

This article can be cited before page numbers have been issued, to do this please use: A. S. Nair and L. Foppa, *Digital Discovery*, 2026, DOI: 10.1039/D6DD00081A.

ROYAL SOCIETY OF CHEMISTRY

rsc.li/digitaldiscovery

View Article Online
DOI: 10.1039/D6DD00081A

A Critical Examination of Active Learning Workflows in Materials Science

Akhil S. Nair[1, 2, *] and Lucas Foppa[1, 3]

[1] *The NOMAD Laboratory at the Fritz Haber Institute of the Max Planck Society, Faradayweg 4-6, 14195 Berlin, Germany*
[2] *Institut für Chemie und Biochemie, Freie Universität Berlin, Arnimallee 22, 14195, Berlin, Germany*
[3] *Molecular Simulations from First Principles e.V., Akazienstr. 3A, 10823 Berlin, Germany*

Active learning (AL) is an increasingly important approach for data-efficient machine learning (ML) in materials science. It is widely used, from building training datasets to guiding autonomous materials discovery platforms. However, the performance of AL workflows depends on a number of often implicit design choices that are rarely examined systematically. Here, we critically analyze commonly used AL strategies in materials science, highlighting overlooked assumptions, hidden biases, and methodological limitations across different applications. Based on this, we provide practical guidelines to enhance the efficiency and reliability of AL workflows for materials science applications.

I. INTRODUCTION

The application of machine learning (ML) in materials science and engineering has expanded rapidly, enabling progress across a wide range of tasks, including high-throughput screening, accelerated materials property prediction and autonomous experimentation [1–4]. Central to many of these efforts is the challenge of learning complex mappings between structure, composition, processing conditions, and materials properties, which are typically non-linear and only partially understood. By learning directly from data rather than relying on explicit physical models, ML methods provide a flexible framework for capturing such complex relationships and have demonstrated transformative potential in materials science [5, 6]. However, their success is fundamentally constrained by the availability, accuracy and quality of the training data. Despite advances in high-throughput experimentation and simulation frameworks [7, 8], generating consistent, high-fidelity materials data remains time- and resource-intensive. As a result, many materials science applications operate in regimes better characterized as data-scarce rather than data-rich [9]. Moreover, materials datasets are often not "statistically representative", i.e., they tend to be shaped by prior domain knowledge, biased curation strategies, or feasibility constraints, and frequently fail to adequately represent the broader space of potentially interesting materials. As a result, ML models trained on such data are prone to unreliable performance when applied to unexplored or underrepresented regions [10, 11].

Active machine learning (hereafter referred to as active learning, AL) has emerged as a promising framework for addressing these challenges in ML-driven materials science. As formalized by Cohn et al. [12], "AL studies the closed-loop phenomenon of a learner selecting actions or making queries that influence what data are added to its training set". In AL, the training dataset is iteratively updated during the learning process, as new data points are selected and labeled at each step. This is in contrast with static approaches which execute a predetermined design without incorporating feedback from incoming data or real-time experimental observations. By adaptively expanding the dataset, AL prioritizes the acquisition of the most relevant data points [13]. While often conceptualized as a data-efficient labelling strategy, the applications of AL in materials science are broader. To illustrate the diversity of its use cases, we highlight two representative settings that are discussed in detail in this perspective: (i) efficient data acquisition for training ML models, and (ii) optimization-driven workflows for materials discovery. For data acquisition tasks, the goal of AL is to achieve broad and representative coverage of the relevant regions of the design space, enabling the ML models trained on the acquired data to generalize reliably across diverse conditions [14, 15]. In this context, the design space refers to all possible sets of inputs that define materials, for instance, all possible atomic arrangements and compositions. In autonomous platforms focused on materials discovery, the goal of AL is to sample specific, high-value regions of the design space. Here, AL strategies aim to balance exploration of previously unobserved regions with exploitation of ML model predictions and uncertainties to efficiently guide the search toward optimal candidates [16–18]. The success of AL in such cases is defined by a specific discovery objective, such as identifying materials with target properties below a given threshold value with a minimal number of evaluations [19, 20].

While AL workflows are now routinely applied in materials science [17, 21, 22], their design choices vary widely, even for closely related tasks. This diversity partly reflects the aforementioned breadth of materials science applications and is therefore neither surprising nor inherently problematic. However, it also introduces substantial inconsistency in how AL workflows are set up, executed, and assessed, making it difficult to compare performance or draw general conclusions across studies. For example, for the data acquisition task, an AL workflow may focus to prioritize unfamiliar data points (e.g. novel compositions or structures) [23], reduce bi-

* akhil.sugathan.nair@fu-berlin.de

View Article Online
DOI: 10.1039/D6DD00081A 2

ases in the initial dataset [24], improve predictive performance of ML models [15], or decrease model uncertainty [25]. While these objectives are often interrelated, practitioners typically evaluate the effectiveness of AL using only a subset of metrics, most commonly improvements in model accuracy, without systematically assessing whether coverage of critical regions of the design space has also improved. Similarly, in materials discovery applications, the choice of surrogate models [10], acquisition functions (vida infra) [26], and uncertainty quantification methods [27] can strongly influence outcomes, yet these choices are often made heuristically and evaluated using application-specific criteria that hinder cross-study comparison. Moreover, a fundamental question remains insufficiently addressed: "to what extent does AL provide benefits beyond those achievable through simpler data selection strategies based on human intuition, prior domain knowledge, or hand-crafted rules?." In other words, it is often unclear whether the observed performance gains stem from a principled, algorithmic AL framework or could instead be achieved by informed, human-guided selection of data points without an explicit AL loop. These issues highlight the need for modular AL workflows that enable consistent evaluation and comparison across applications, while still allowing flexibility to accommodate domain-specific objectives. Without such structure, AL workflows risk being inefficient, difficult to interpret, or misdirected, potentially leading to unnecessary computational or experimental cost and convergence toward suboptimal solutions.

In this perspective, we critically examine AL workflows commonly employed in materials science, focusing on two key applications of data acquisition and optimization-driven materials discovery. We begin with a brief overview of the AL methodology and analyze the strengths and limitations of tools and techniques used at different stages of AL workflows. We do not aim to provide a comprehensive review here and instead refer readers to existing review articles for the broader context [21, 28]. Our focus is to raise awareness of practitioner-driven biases in the design choices that can impact the performance of AL workflows. Building on this analysis, we propose guidelines to support the rigorous design, assessment, and interpretation of AL workflows in materials science.

II. ACTIVE LEARNING METHODOLOGY

An AL algorithm involves performing data acquisition iteratively and adaptively, with the goal of selecting the most important data points for labelling under a limited evaluation budget. In materials science, the "labels" are typically obtained from an oracle such as a high-fidelity simulation, a physical experiment, or expert annotation (Fig. 1a), all of which are costly in terms of time or resources. An AL workflow therefore seeks to maximize learning efficiency by prioritizing which data points to evaluate next, rather than relying on sampling performed randomly [29]. It is important to distinguish AL from the classical framework of statistical design of experiments (DoE), which is widely used in engineering and industrial applications [30]. Traditional DoE methods typically rely on predefined, space-filling or statistically optimal designs (e.g., Latin hypercube sampling [31]) that aim to efficiently explore the design space. In contrast, AL is inherently sequential and adaptive, selecting new samples based on information gained from previously acquired data. Because AL constructs a sequence of targeted queries, the framework is often referred to in the literature as "query learning" [32, 33] or "sequential learning" [34, 35]. In addition, different AL settings can be also distinguished depending on how unlabeled data are accessed or generated. In "query synthesis", new candidate inputs are generated during the learning process rather than selected from a predefined dataset [36], whereas "stream-based AL" assumes that data arrive sequentially and must either be selected for labeling or discarded [37]. However, in this perspective, we will use the term "active learning" and stick to discussing pool-based AL [38] where it is assumed that a relatively larger pool of unlabelled data is available for labelling, because of its simplicity and its widely adopted usage in the field. Note that while the design space represents the total theoretical bounds of exploration and can be continuous, vast regions of it may not correspond to physically realizable or stable materials. A pool-based AL framework circumvents the need to map such abstract coordinates back to unique, valid materials, a task that is often non-trivial.

Algorithm 1 outlines the core logic of pool-based AL, which can be tailored to accommodate specific objectives. In a supervised learning setting, let $\mathcal{D}_{\text{init}} \subset \mathbb{X} \times \mathbb{Y}$ denote the initially available labelled dataset (hereafter referred to as "seed data"), and let $\mathcal{U} \subset \mathbb{X}$ represent the pool of unlabeled data (hereafter referred to as "pool data"). ere, $\mathbb{X}$ represents the design space and $\mathbb{Y}$ the target space of quantities of interest (e.g., band gaps, formation energies). A surrogate ML model M is trained to approximate an unknown target function $f : \mathbb{X} \to \mathbb{Y}$, mapping $x \in \mathbb{X}$ to $y \in \mathbb{Y}$. Crucially, the surrogate model should provide not only predictions but also reliable estimates of uncertainty, which quantify the model's confidence in its predictions at a given point. Common choices for surrogate models in AL include Gaussian processes (GP) [39], random forests (RF) [40], and neural networks [41], each offering different trade-offs in terms of predictive accuracy, training cost, and interpretability (see Table II). At each iteration, a sampling strategy Q evaluates the pool data and selects a batch of k candidates $\mathcal{X}_{\text{batch}} \subseteq \mathcal{U}$. This selection is typically guided by the model's predictions and/or uncertainty estimates, with the goal of identifying regions of the design space where additional labels are expected to be most relevant. Note that Q can be deterministic , yielding a fixed set of top-ranked candidates [20] or stochastic, where candidates are sampled

View Article Online
DOI: 10.1039/D6DD00081A 3

based on a probability distribution [14]. The selected candidates are then evaluated then evaluated by an oracle $\mathcal{O}$ to obtain corresponding labels, and the labeled dataset is updated accordingly. The selected candidates are removed from the unlabeled pool, and the model is retrained on the updated dataset. This process is repeated until a stopping criterion is satisfied, such as a performance threshold or exhaustion of a query budget B.

Algorithm 1 Pool-Based Active Learning

Require: Initial labeled dataset $\mathcal{D}_{\text{init}} \subset \mathbb{X} \times \mathbb{Y}$, Unlabeled pool $\mathcal{U} \subset \mathbb{X}$, Oracle $\mathcal{O} : \mathbb{X} \to \mathbb{Y}$, Surrogate model M, Batch size k, Sampling strategy $Q : (M, \mathcal{U}, k) \to \mathcal{X} \subseteq \mathcal{U}$, Total budget B, Stopping criterion $S : (M, B)$

Train M on $\mathcal{D}_{\text{init}}$
while $S =$ **False do**
  Select batch: $\mathcal{X}_{\text{batch}} \leftarrow Q(M, \mathcal{U}, k)$
  for all $x \in \mathcal{X}_{\text{batch}}$ **do**
    Obtain label $y \leftarrow \mathcal{O}(x)$
    Update dataset: $\mathcal{D} \leftarrow \mathcal{D}_{\text{init}} \cup \{(x, y)\}$
  end for
  Update unlabeled pool: $\mathcal{U} \leftarrow \mathcal{U} \setminus \mathcal{X}_{\text{batch}}$
  Retrain M on $\mathcal{D}$
end while
return $(\mathcal{D}, M)$

III. ACTIVE LEARNING APPLICATIONS AND CHALLENGES

In this section, we highlight the key challenges in employing AL workflows for materials science problems through two representative application domains: (i) efficient data acquisition for training ML models (ii) optimization-driven materials discovery. It is to be noted that here, the term "materials discovery" refers broadly to identifying materials with desirable properties, whether among previously synthesized, computationally generated, or yet-to-be-synthesized candidates. In this sense, discovery entails jointly establishing material identity and properties within a vast, largely unexplored materials space. For both i) and ii), we discuss, based on existing literature, how AL can be useful and identify critical challenges that remain unaddressed or are frequently overlooked. We provide practical guidelines and recommendations for addressing these challenges for future AL-driven materials research in the Outlook section.

A. Efficient Data Acquisition for Training ML Models

Redundancy problem in materials data : Materials datasets are often curated based on prior knowledge (e.g. well-known materials with desirable properties) or ease of access (e.g. from existing data repositories). These practices can introduce substantial redundancy and lead to biased coverage of the broader materials space. Recent work by Li et al. [42] demonstrated that widely used computational materials databases, including the Materials Project (MP) [43] and OQMD [44], contain significant data redundancy. Using a data-pruning strategy, they showed that removing a large fraction of these redundant entries neither degraded in-distribution model performance nor improved out-of-distribution (OOD) generalization for ML models trained on the full datasets (Fig. 2a). While such redundancy is perhaps unsurprising given the high-throughput nature of computational database generation, biased sampling can persist even when subsets of these databases are selectively curated. For example, in many materials discovery studies [6, 45], candidate materials are preferentially chosen near the convex hull of a given compositional phase diagram, which can distort an ML model's representation of the underlying stability landscape. Indeed, Bartel et al. [46] showed that ML models can accurately predict formation energies yet still perform poorly in classifying stable versus unstable materials, particularly in sparsely sampled compositional spaces such as underrepresented quaternary systems in MP. In such scenarios, AL offers a promising alternative by adaptively focusing on data in the underexplored regions of the materials space. By doing so, AL can enable the construction of smaller, more relevant datasets that mitigate redundancy while maintaining or even improving predictive performance.

Inadequate sampling strategy : One of the key components of an AL workflow is the *sampling strategy*, which is used to indicate which datapoints to be selected from an unlabelled pool. Many of the AL workflows adapted in materials science [15, 47] use *informativeness*, i.e., ability of a sample to improve the model performance, as the sole sampling criterion. However, an effective AL sampling strategy is inherently multi-faceted and cannot be fully captured by informativeness alone (Fig. 1b). A key complementary criterion is *representativeness*, which assesses whether a queried sample reflects the structure of the unlabeled data distribution [48]. This is particularly important to prevent sampling extreme outliers (e.g. highly distorted structures) that are not statistically representing the remaining, relevant design space of interest. While informativeness and representativeness are conceptually distinct, they are often complementary when AL performance is evaluated over the full search space. However, in the presence of a significant distribution mismatch between the labelled seed data and the unlabeled pool, the two criteria may diverge. In such cases, representativeness favors sampling from high-density regions of the pool distribution, whereas informativeness prioritizes regions that are underrepresented in the seed data, leading to potentially differing sampling. Various methods have been proposed by the ML community to include the representativeness factor [49, 50], or jointly optimize informativeness and representativeness [51, 52],

View Article Online
DOI: 10.1039/D6DD00081A 4

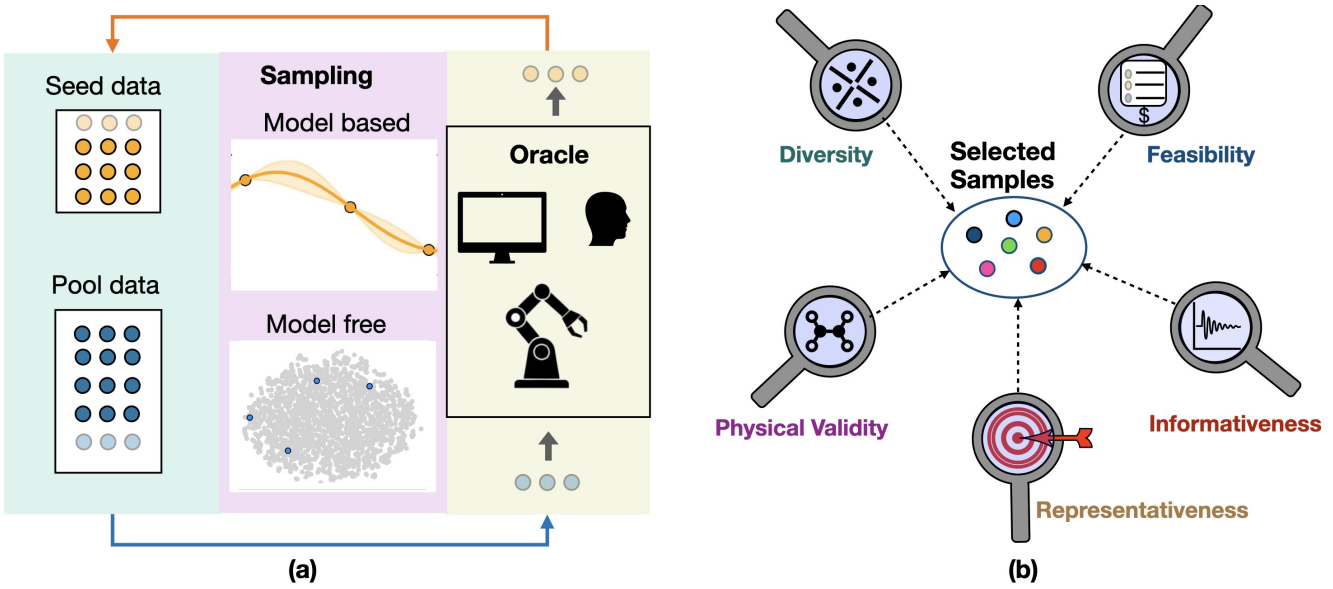
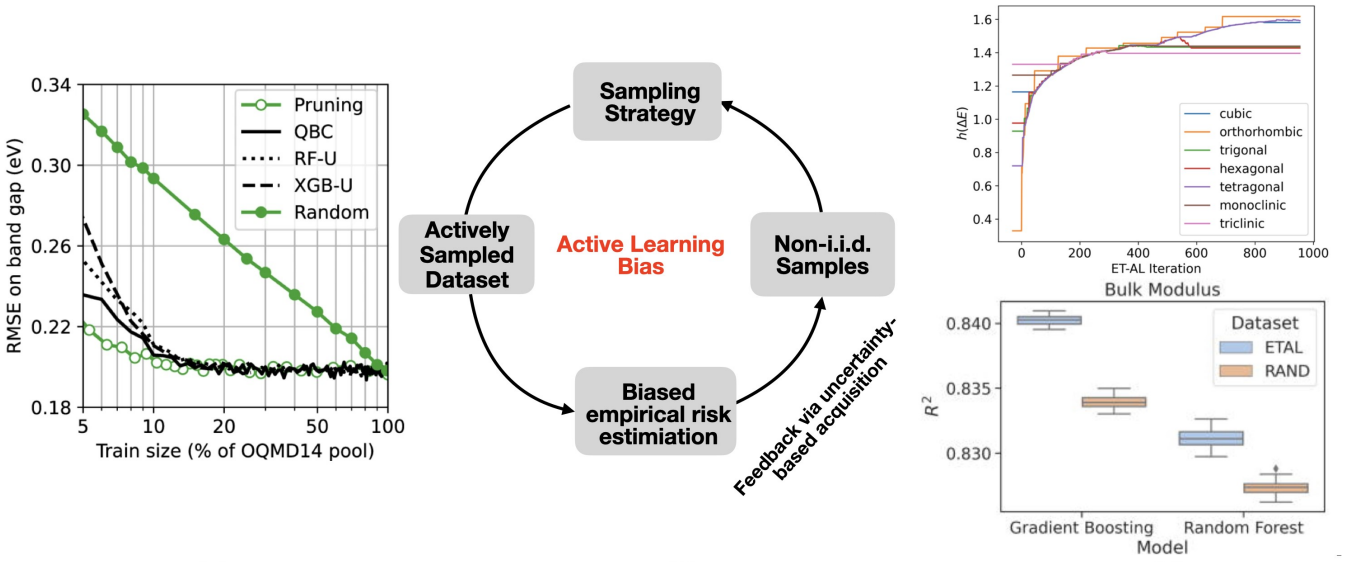


FIG. 1: Schematic representation of (a) an AL workflow where both model-based and model-free AL strategies can be employed to acquire samples from pool data and update the seed data by interacting with an oracle, (b) various factors that need to be considered while designing the sampling strategy for AL.

yet most AL workflows in materials science remain focused on informativeness due to the ease of monitoring via proxies such as model test errors. Beyond these standard criteria, additional materials-science-specific factors need to be considered for sampling. These include: (i) *diversity*, which ensures that selected samples are sufficiently distinct relative to both the seed data and between themselves to avoid redundancy [53, 54] (ii) *physical validity*, ensuring that samples are chemically and physically meaningful, for example by excluding crystal structures under extreme conditions and (iii) *feasibility*, which accounts for practical constraints such as computational cost, favoring candidates that provide maximal information without incurring excessive expense (e.g. extremely large system sizes for simulations). In addition, for AL campaigns for which a reasonably large seed data are already available, pre-existing biases can persist, such as overrepresentation of certain chemistries, phases, or structural motifs. If these biases are not identified and addressed at the outset, AL may inadvertently reinforce them, leading to a sampling that further skews the dataset. Failure to consider these criteria while designing the sampling strategy can therefore severely limit the reliability of AL in materials science.

The ill-addressed active learning bias : While AL can mitigate the redundancy problem, it can paradoxically introduce a new form of bias, termed as active learning bias (ALB) [55, 56]. This arises because during AL, samples are no longer drawn independently and identically distributed (i.i.d.), a fundamental assumption underlying ML model training. As a result, actively curated training sets may deviate substantially from the application-relevant data distribution in the design space of interest (Fig. 2b). This deviation has important implications for model training and evaluation. Standard empirical risk minimization, which optimizes model parameters by minimizing the average loss over the training data, implicitly assumes that the training set is representative of the target distribution. When this assumption is violated, as is often the case in AL, model performance measured on finite or randomly curated test sets may reflect optimization with respect to a biased objective rather than genuine generalization across the materials domain. In materials science applications, this can manifest as models that perform well on actively sampled configurations while failing to generalize to unexplored chemistries or structures. Statistical corrections have been proposed to mitigate ALB, such as reweighting the training loss by the inverse probability of sample acquisition [57], thereby partially restoring consistency with the underlying data distribution. However, such approaches have not yet been systematically explored in materials science. Moreover, reweighting introduces additional subtleties: in overparameterized models, including deep neural networks, ALB can sometimes act as an implicit form of regularization, reducing overfitting and even improving apparent generalization [58]. While this effect may be beneficial in practice, it complicates the interpretation of AL performance, as improvements may stem from sampling-induced regularization rather than principled coverage of the design space. Because the magnitude and impact of ALB depend strongly on the mismatch between the seed dataset, the unlabeled pool, and the target application domain, careful attention to data distributions is essential. Incorporating explicit analysis of distributional coverage using tools such as dimen-

View Article Online
DOI: 10.1039/D6DD00081A 5



(top panel) and improved performance of ML models trained on such actively learned dataset compared to randomly sampled dataset (bottom panel). Adapted with permission from Ref. [24], Copyright 2023 American Institute of Physics.

FIG. 2: (a) Active learning reduces redundancy in materials datasets: Performance of XGBoost (XGB) and Random Forests (RF) models on band gap prediction trained on datasets obtained by uncertainty guided active learning, pruning, and random sampling from the OQMD14 dataset. Comparable accuracy is achieved with only 10% of the data, highlighting substantial redundancy in the dataset, Adapted with permission from Ref. [42], Copyright 2023 Springer (b) schematic representation of active learning bias induced by sampling of data points do not following i.i.d assumption, (c) information-entropy guided active learning (ETAL) minimizing the large structure-stability bias by improving the coverage of less symmetric crystal systems in the JARVIS dataset

sionality reduction [59] or density estimation [60, 61] can therefore provide critical context for evaluating AL outcomes and for designing more robust, application-aware AL workflows [62, 63]. Without such considerations, AL strategies risk reinforcing hidden biases, limiting transferability, and overstating the effectiveness during the workflow deployment.

Model-based vs. Model-free active learning : While most AL workflows employed in materials science are built around surrogate ML models and consequently face the challenges outlined above, strategies which do not involve training a surrogate model could be adapted as an alternative. These approaches are typically formulated under unsupervised settings, enabling sampling to target specific objectives without relying on model predictions or uncertainty estimates, in contrast to model-based approaches that operate in a supervised setting. In this work, we refer to such approaches as "model-free". A key advantage of model-free strategies is their conceptual simplicity and flexibility. They are particularly useful for scenarios involving a very small amount of seed data, where surrogate models have limited predictive accuracy, and their uncertainty estimates may be unreliable. Zhang et al. [24] employed an information-entropy-based AL workflow to mitigate structure–stability bias in computational crystal databases, where low-symmetry structures are often underrepresented. By prioritizing structurally informative samples, measured using information entropy, their approach improved the coverage in crystallographic space and yielded ML models with superior predictive performance compared to random sampling (Fig. 2c). Similarly, Schwalbe-Koda et al. [64] demonstrated that atomistic information entropy, computed directly from local atomic descriptors, can serve as a model-free proxy for uncertainty of machine learning interatomic potentials (MLIPs), guiding molecular dynamics (MD) simulations. Beyond informativeness, model-free AL can explicitly enhance representativeness and diversity, two criteria that are often weakly controlled in model-based AL workflows. Density-based strategies, such as clustering or kernel density estimation, promote sampling from statistically significant regions of the design space, thereby preserving global coverage and facilitating the representative sampling [68, 72]. On the other hand, similarity-based model-free strategies emphasize improving diversity and minimizing redundancy. These methods are often implemented using some distance-based metrics, defined over feature, descriptor, latent, or embedding spaces [42, 73, 74] and select samples that are maximally distinct from one another and from the existing seed data, promoting diversity. It has to be noted that the distinction between model-based and model-free approaches becomes less clear when distances are computed in latent spaces derived from trained models, as these representa-

View Article Online
DOI: 10.1039/D6DD00081A 6

Sampling Criterion	Methods	Central Idea	Application in Materials Science
Informativeness	Entropy-guided	Select samples that maximize information entropy computed from structural or descriptor distributions, enabling bias reduction and improved coverage without model uncertainty estimates.	Curation of bias-minimized crystal structure datasets [24]; Model-free uncertainty for MLIP-driven MD [64, 65]
Representativeness	Clustering	Partition the unlabeled pool into clusters and select representative samples (e.g. cluster centroids) to preserve global coverage of the distribution and avoid oversampling statistical outliers.	Discovery of perovskite oxides for oxygen evolution catalysis [66]; Training data generation for MLIPs [67]
	Density estimation	Prioritize samples located in high-density regions of the unlabeled data distribution to ensure representativeness.	Functionalized nanoporous materials (MOFs/COFs) property prediction [68]; Assessing out-of-distribution performance of ML models [60]
Diversity	Distance-based	Select samples that maximize the minimum distance to the seed data in feature, output, latent, or an embedding space to avoid redundancy and maximize diversity.	Discovery of high-entropy oxides for H_2 production [69]; Surface structure exploration for catalysis [70]
	Physical metric-based	Select batches of samples that are mutually dissimilar while also distinct from the existing labeled data based on a physical or chemically motivated metric	Developing accurate property-prediction models for structure-property mapping of microstructures [71]

TABLE I: Examples of model-free active learning strategies and representative applications in materials science.

tions implicitly depend on the model, even if the sampling criterion itself does not directly rely on the model. Alternatively, diversity can also be enforced using physically or chemically motivated similarity measures, for example, based on composition, local coordination environments, structural topology, bonding motifs, or symmetry classes [71, 74]. It has been shown that when labeled data are scarce, similarity-based model-free methods can outperform model-based AL due to their robustness against less accurate surrogate models [75]. However, these approaches are not without limitations. High-dimensional vector spaces used to represent materials data may suffer from the *curse of dimensionality*, and outcomes of distance-based sampling are sensitive to the choice of representations (e.g. elemental-property-based features or SOAP descriptors [76]), similarity metrics (e.g. Euclidean or Mahalanobis distances), and analytical choices such as centroid-based versus nearest-neighbor selection [77]. Additionally, since such approaches do not leverage predictive models, they may lack adaptivity to variations in the underlying response surface, potentially leading to inefficient sampling in regions where the target property varies non-uniformly across the design space [78]. Recent benchmarking by Bi et al. [79] further suggests that, on average, model-free strategies underperform model-based AL when evaluated across diverse materials datasets, as they lack explicit mechanisms to capture the relationship between samples and target properties. Despite these limitations, model-free AL remains practically valuable; it avoids the computational overhead of training and retraining surrogate models (e.g. ensembles of ML models) and remains effective when initial datasets are small or highly biased. These observations highlight an unresolved dilemma in AL design: whether to prioritize model-based or model-free strategies, and motivate hybrid approaches that combine the robustness of model-free sampling with the task-awareness of model-based methods [80, 81]. Representative model-free AL strategies and their applications in materials science are summarized in Table I.

View Article Online
DOI: 10.1039/D6DD00081A 7

Model Type	Data Efficiency	UQ	Interpretability	Cost	Application in Materials Science
Gaussian Processes	High	Principled (exact Bayesian posterior)	Limited in high-dimension	High ($\mathcal{O}(N^3)$)	Phase-change memory material for photonic switching devices [78] (exp); layered materials with suitable electronic properties [17] (comp)
Random Forests	Moderate	Heuristic (ensemble-based)	Moderate with feature importance	Low	Biochar syntehsis for CO_2 capture [82] (exp); screening of inorganic materials [83] (comp)
Gradient Boosting Methods	Moderate	Heuristic (ensemble-based)	Moderate with feature importance	Moderate	High-entropy oxides for H_2 production [69] (exp); power factor prediction of thermoelectrics [47] (comp)
Bayesian Neural Networks	Low	Heuristic (approximate posterior)	Low	High	Optimal parameters for chemical reactions [84] (exp); van der Waals heterostructures with suitable bandgaps [85] (comp)
Support Vector Regression	Moderate	Limited	Limited	Moderate	Shape memory alloys with low thermal hysteresis [20] (exp); Piezoelectric materials screening [27] (comp)
Deep ensembles	Low	Heuristic (inter-model predictive variance)	Low	High	Crystal structure prediction [86] (comp); MLIP assisted material simulations [87] (comp)
Symbolic regression	High	Limited	High (analytical equations)	High	Screening of acid-stable oxides for electrocatalysis [19] (comp)

TABLE II: Common surrogate models and their properties relevant to Bayesian Optimization based active learning applications. The *exp* and *comp* in parenthesis of literature references indicates whether the works do involve experiments or purely computational simulations, respectively.

B. Optimization-driven Materials Discovery

Interplay between surrogate models and sampling strategy : The dominant AL practices in both computational and experimental settings for materials discovery are black-box optimization (BBO), with Bayesian Optimization (BO) [88] being the most widely adopted approach [89]. BO's ability to navigate complex search spaces under data-scarce conditions has led to numerous success stories in materials discovery [90–92]. Notable examples include the identification of Pb-free $BaTiO_3$-based piezoelectrics with enhanced electrostrictive strain [93], NiTi-based shape memory alloys with low thermal hysteresis [20], and efficient high-entropy alloy catalysts [92]. The key components of a BO-driven AL (BO-AL) include: (i) a surrogate model to approximate the expensive objective function mapping the materials property to a set of given input parameters (ii) an acquisition function (analogous to sampling strategy in non-BO AL) to guide sample selection, and (iii) an oracle for labelling new data points. While BO traditionally relies on probabilistic surrogate models such as GP for their principled uncertainty estimates, non-Bayesian models have also been adopted in materials applications [94, 95]. This is often motivated by practical considerations, including the poor scalability of GP in high-dimensional spaces and the superior extrapolation performance observed with certain non-Bayesian models on specific datasets [96]. Nevertheless, the criteria used to select a suitable surrogate model for BO-AL in materials science are not thoroughly discussed. Since no surrogate model demonstrates universally optimal performance across all problem settings, surrogate model selection is inherently task-dependent, with different models exhibiting varying performance across applications. For instance, Lim et al. [94] demonstrated that GP with carefully selected kernels outperformed alternative models, including RF, on experimental materials datasets. In contrast, Liang et al. [95] found that RF-based BO outperformed GP-BO using standard isotropic kernels and performed comparably to GP with anisotropic kernels. Table II and III summarize some of the characteristics of various surrogate models and acquisition strategies used in BO-AL for materials science applications, respectively.

Importantly, the performance of BO-AL is often not governed just by the surrogate model or acquisition function in isolation, but by their combined behaviour. A highly accurate surrogate model may still perform poorly if paired with a suboptimal sampling strategy. For example, expected improvement (EI), a popular acquisition function used in BO-AL for material property optimization, depends on the current best observation, which might be an unreliable benchmark if the seed data is strongly biased, misleading the optimization trajectory. Therefore, evaluating surrogate model–acquisition function combinations through after-the-fact (AFT) AL trials (where pool data is already labelled but excluded from seed data) can help identify optimal configurations [20], though these analyses may not always generalize to all pool datasets due to distribution shifts. As illustrated by Boley et al. [10] (Fig. 3a), in the AFT-AL experiment for discovering perovskites with high bulk modulus, RF with both pure exploitation (XT) and EI acquisition functions perform similarly to that of GP with EI, but only as good as random sampling. However, in real AL runs (using DFT calculations as ground truth), GP with EI clearly outperforms the others, highlighting that surrogate model–acquisition function selection

View Article Online
DOI: 10.1039/D6DD00081A 8

Acquisition Strategy	Exploration-Exploitation Balance	UQ Dependency	Application in Materials Science
Random Sampling (RS)	Exploration only (uninformed)	None	Baseline for benchmarking AL workflows in optimization-driven materials discovery [34, 95]
Pure Exploitation (XT)	Exploitation only	Low (Mean prediction)	Identifying perovskites with high bulk modulus [10]; benchmarked against EI in shape memory alloy design [20]
Probability of Improvement (PI)	Balanced; tunable via ξ	High	Identifying materials with low lattice thermal conductivity [97]; discovery of materials with the high melting points [98]
Expected Improvement (EI)	Balanced; tunable via ξ	High	Discovery of stable materials [99]; accelerating synthesis of superconducting materials [100]
Upper Confidence Bound (UCB)	Exploration-leaning; tunable via β	High	Mg alloy design [101]; parametrization of DFT for accurate bandstructure prediction [102]
Probability of Feasibility (POF)	Exploitation (constraint-driven)	High	Computational discovery of acid-stable oxides [19]
Thompson Sampling (TS)	Exploration-leaning (stochastic)	High (posterior sampling)	Atomic structure determination [103]; crystal structure prediction [104]
Knowledge Gradient (KG)	Balanced	High	Optimizing processing conditions of hybrid organic–inorganic perovskites [105]
Expected Hypervolume Improvement (EHVI / qNEHVI)	Balanced (Pareto front-focused)	High	Nanoparticle synthesis [106]; additive manufacturing [107]

TABLE III: Common acquisition strategies and their properties relevant to Bayesian Optimization-based active learning applications in materials science.

based solely on initial data may be misleading due to distributional shifts and the underrepresentation of high-performing materials. Although dynamic switching between acquisition functions has been proposed [108], it remains underexplored in BO-AL workflows applied for materials discovery. It is to be noted that such a sensitivity to sampling strategy is a general challenge in AL, even beyond BO frameworks. In particular, the choice of representation plays a critical role in shaping the trajectory and performance of AL [109, 110]. We have recently shown that efficient feature selection on-the-fly at each AL iteration can significantly enhance the performance of BO-AL campaigns compared to high-dimensional representations, which are not updated during AL iterations [111]. Also, different representations induce distinct geometries of the design space, which further determine how similarity and diversity are quantified. This becomes especially important when the sampling strategy relies on geometric notions (e.g., distance-based approaches), as the representation directly influences sampling behaviour and the overall efficiency of AL. For instance, a materials representation based on global, composition-only descriptors (e.g., Magpie [112]) is fundamentally distinct from one based on local atomic environments (e.g., SOAP). A distance-based sampling strategy navigating a compositionally defined space will naturally prioritize exploring broad chemical families and diverse elemental combinations. This approach is ideal for AL workflows targeting exceptional materials when the underlying property landscape is predominantly composition-driven, such as screening vast compositional spaces for novel high-entropy alloys. Conversely, the same strategy operating within a local structural descriptor space will prioritize exploring local atomic environments which could be more suitable for AL focused on diversifying local environments when training MLIPs, where capturing subtle structural variations is vital for accurate force and energy predictions. Hence, the choice of representation must therefore be strictly aligned with both the specific objective of the AL campaign and the underlying physics of the target property landscape.

For many applications, BO-AL is preferred to be applied in batched or parallel settings where a number of samples (instead of one as in standard BO) are selected, which could leverage the availability of oracles with parallel execution capabilities (e.g. high performance computing facilities, high-throughput synthesis platforms) [22, 84]. However, standard acqusition functions are limited by their nonadditivity and do not account for information overlap between simultaneously queried samples. To address this limitation, specialized batch acquisition functions have been developed that explicitly account for correlations between candidate points. Notable examples include multipoint Expected Improvement (qEI) [114], which generalizes EI to jointly evaluate a batch of points and the recently proposed multipoint Probability of Optimality (qPO) [115], which maximizes the joint likelihood that the true optimum is contained within the selected batch. For large batches common in materials screening (10-100 samples), hybrid methods combining uncertainty-based acquisition with explicit diversity mechanisms such as determinantal point processes [116] or clustering-based selection have been shown to improve coverage of the design space and reduce redundancy. It is important to note that while batching is desirable from a resource management perspective, it

View Article Online
DOI: 10.1039/D6DD00081A 9

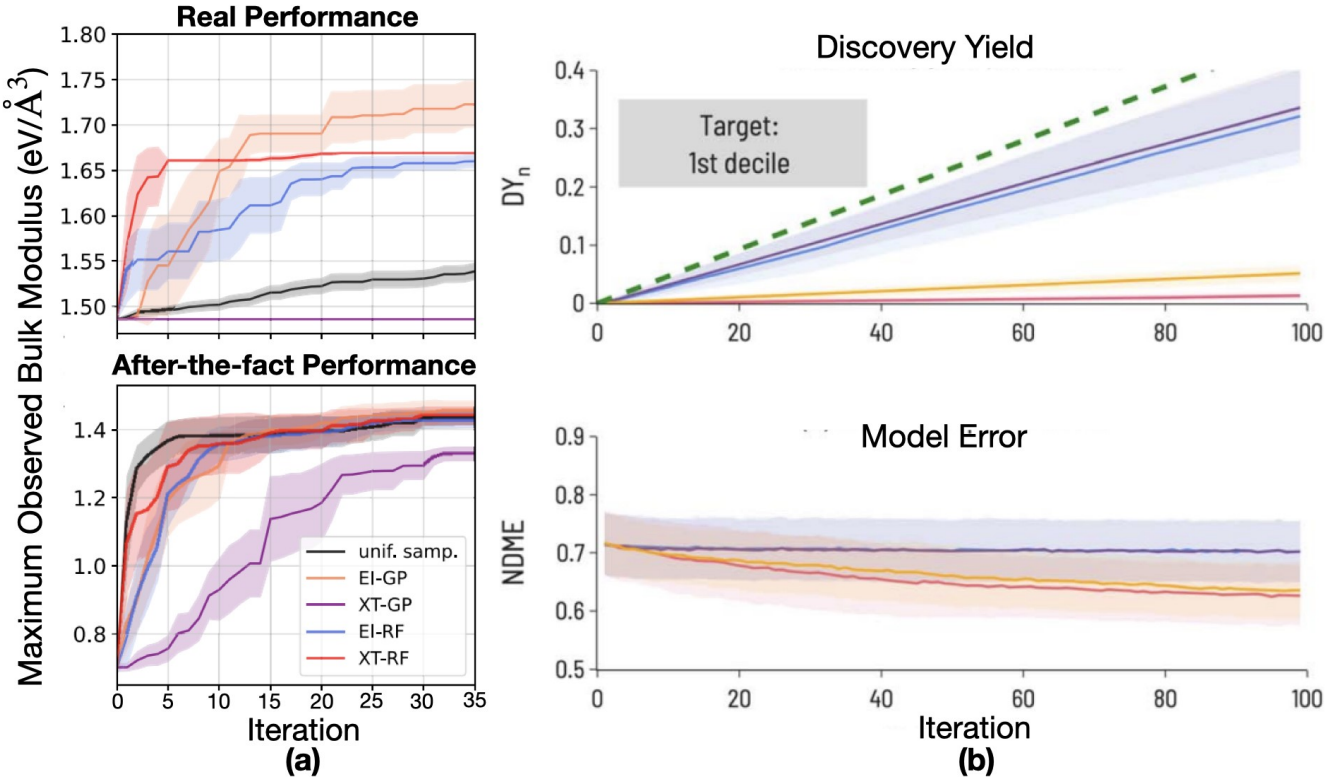


FIG. 3: (a) Impact of surrogate model and acquisition function choices on active learning for identifying perovskites with high bulk modulus: Real AL runs (top panel) show varied efficiencies across GP and RF models with exploitation (XT), expected improvement (EI), and uniform (random) sampling acquisition functions, compared to retrospective seed-only baselines (bottom panel). Adapted with permission from Ref. [10], Copyright 2024 Institute of Physics. (b) Lack of correlation between model performance and materials discovery in Bayesian optimization based AL for identifying high-bandgap materials. Colored curves denote acquisition functions (blue: expected value, violet: EI, red: maximum uncertainty, yellow: random sampling). Discovery yield (DY) (see Table IV for definition) improves, but non-dimensional model error (NDME = RMSE/standard deviation of holdout set) does not consistently decrease. Adapted with permission from Ref. [113], Copyright 2023 Royal Society of Chemistry

may degrade sampling efficiency if diversity and information gain are not explicitly accounted for, as acquisition functions can otherwise fail to correctly rank candidates in terms of marginal utility. This challenge is particularly pronounced in materials science applications, where multiscale structure–property relationships and heterogeneous synthesis or characterization pathways can further amplify redundancy and bias in batch selection. Nevertheless, platforms such as Pheonix [84], ChemOS [117], Olympous [118] etc. have made success in enabling the application of BO-AL with robotic experimentation or lab automation systems to enable parallel acquisition and execution, useful for chemistry and materials science. In addition to batch selection, incorporating evaluation cost into the acquisition strategy is critical for real-world applications, as the cost of evaluating candidates can vary by orders of magnitude. This can be addressed either through cost-aware acquisition functions [119], such as Expected Improvement per Unit Cost (EIpu) [120] which prioritize candidates offering high utility relative to their evaluation cost, or through multi-fidelity approaches [121], where each candidate can be evaluated at multiple levels of fidelity with different cost–accuracy trade-offs. Such approaches enable more efficient utilization of limited resources.

Unreliable uncertainty quantification : In AL strategies with an exploration component such as those used in BO, uncertainty quantification (UQ) plays a pivotal role. This is based on the notion that the surrogate models are most error-prone in regions where their predictions are least confident, and hence, acquiring data from such regions is often the most informative. Reliable UQ helps in identifying these regions by assigning a confidence interval to predictions, thus guiding the sampling strategy toward high-impact data points [122]. However, two core challenges hinder this process: (i) most ML models lack inherent UQ capabilities (see Table II), and (ii) many UQ techniques produce unreliable estimates which are often underconfident or overconfident relative to actual prediction errors [123, 124]. Unreliable UQ can misguide AL by undervaluing informative samples, potentially leading to premature convergence or excessive exploitation of suboptimal regions. For instance, in BO-AL with EI as the acquisition function, overconfident uncertainty estimates shrink the exploration term, causing the algorithm to overlook uncertain but informative re-

View Article Online
DOI: 10.1039/D6DD00081A

gions in favor of already well-explored ones.

The reliability of UQ in ML remains an open research challenge in materials science, with relatively few studies that critically benchmark UQ methods [127, 128]. Existing efforts typically evaluate the quality of uncertainty estimates based on hold-out test sets using metrics based on the uncertainty distribution [129], correlation with prediction errors [127], or computational cost [128]. These studies consistently show that no single UQ method outperforms across all scenarios, with strong dependencies on surrogate model and dataset characteristics. Nevertheless, calibration approaches such as scaling uncertainty estimates with respect to residuals can enhance their reliability [129, 130]. Openly available tools like uncertainty-toolbox [131] and UQLab [132] facilitate such calibration processes. However, the effectiveness of these calibration methods within AL workflows remains largely unexplored, with existing studies still in their early stages [133]. Since such approaches require labeled data, they are constrained to the currently known samples and hence cannot guarantee improved uncertainty estimates on the unlabeled pool data where accurate UQ is most critical. This highlights the need for more standardized, AL-focused metrics to assess and compare UQ methods for materials property prediction and optimization tasks.

Since one of the reasons for unreliable UQ is that the surrogate models are primarily trained to minimize prediction error without providing default uncertinay estimates, a range of alternative, model-agnostic UQ strategies have been proposed to address this challenge. These rely on heuristic measures of uncertainty, such as distance in feature [134] or latent spaces [73] and are not specifically dependent on predictive model architecture. It has been demonstrated that latent space distance provides more reliable uncertainty estimates (by better capturing of residuals) for both artificial [73] and graph [135] neural networks, indicating that such model-agnostic methods could be beneficial to AL. Similarly, integrating Domain of Applicability (DoA) analysis [59, 60] with UQ in AL workflows allows for the identification and expansion of the model's reliable prediction regime, enabling more effective sample selection near or beyond the current DoA to improve model robustness [136]. However, these methods are more mature in cheminformatics, and lack standardized implementation in materials science, due to the absence of universally accepted DoA frameworks [60, 137]. Also, a fair comparison between these model-agnostic vs. model-based UQ methods for AL performance for materials discovery has not been done to the best of our knowledge, which is also necessary for improving the reliability of AL practices in materials science.

Lack of standardized performance evaluation metrics : A rigorous and accurate evaluation of AL performance is essential for assessing its efficiency and determining when to terminate the AL loop. This is especially critical in closed-loop, automated experimentation platforms with limited human intervention, where failure to recognize diminishing returns can result in substantial resource consumption. Various performance metrics have been proposed to quantify AL efficiency in optimization-driven workflows for materials discovery. However, no single metric universally captures all aspects of performance, and even metrics which are expected to be complementary could show diverging trends for performance estimation throughout AL iterations. The most popular metrics used in assessing the efficacy of AL are based on comparison with random sampling [94, 125], which involves sampling data points from the pool without the use of a surrogate model-derived insight and hence also can be referred to as "passive learning". Rohr et al. [34] proposed quantitative metrics such as the acceleration and enhancement factors (see Table IV), measuring the fraction of promising materials found and the iterations required to find them relative to random sampling, which have been used to benchmark surrogate models and acquisition strategies for AFT materials discovery campaigns [95, 138]. However, using random sampling as a baseline may not be considered a universal performance metric in materials science, where experimental choices are typically guided by prior knowledge rather than random selection. Moreover, AL can underperform random sampling if the surrogate model's inductive bias, i.e., the set of foundational assumptions about the underlying data, does not accurately reflect the true property landscape, thereby misguiding the search process [139]. In such cases, acquisition strategies like uncertainty sampling can become 'trapped' in intrinsically noisy or physically irrelevant regions of the design space that offer little contribution to the model's generalizability. This was recently found to be the case for training MLIPs [140], where random sampling achieved superior predictive accuracy because the AL oversampled high-energy and distorted configurations. This introduced systematic energy offsets and compromised the model's performance within the physically relevant regions of interest.

Another common evaluation criterion in optimization-driven AL workflows is tracking model performance over AL iterations. However, Borg et al. [113] observed that improvements in model accuracy do not necessarily correlate with better discovery rates of high-performing materials. As shown in Fig. 3b, AFT BO-AL campaigns targeting high band gap materials achieved high discovery rates, even without noticeable improvements in surrogate model accuracy. This suggests that the popular notion of "model is getting better" may not qualify as an accurate evaluation criteria for AL for materials discovery. This arises because global accuracy metrics (e.g., model error measured on hold out test sets) primarily reflect performance in the bulk of the data distribution, while successful discovery depends on the model's ability to prioritize candidates in sparsely populated, high-performing regions. Similarly, Koizumi et al.[14] demonstrated that the AL performance can vary substantially with changes in the material property, descriptor dimension and size of seed data even when using a fixed choice

View Article Online
DOI: 10.1039/D6DD00081A

Metric	Definition	Limitation
Discovery yield [34], [113]	$\frac{\text{No. of Materials with FOM at } i^{\text{AL}}}{\text{Total No. of Materials with FOM in Pool}}$	Requires apriori knowledge of interesting materials in the pool
Acceleration factor [95],[113]	$\frac{\text{No. of } i^{\text{AL}} \text{ for FOM}}{\text{No. of } i^{\text{RS}} \text{ for FOM}}$	Not (always) a fair baseline
Enhancement factor [95],[113]	$\frac{\text{FOM(AL)}}{\text{FOM(RS)}}$	Not (always) a fair baseline
Decision efficiency [34]	$2 \cdot \frac{\#\{\text{samples with } \mathrm{i}_{\text{FOM}} \leq \mathrm{i}_{\text{FOM}}^{(\text{selected})}\}}{N} - 1$	Dependence on pool quality
Model accuracy [42], [125]	Error of surrogate model estimated on a (holdout) test set	Prone to overestimation due to active learning bias
Model uncertainty [19], [126]	Uncertainty of surrogate model	Prone to misleading due to unreliable uncertainties

TABLE IV: Evaluation metrics used for assessing AL performance for optimization-driven materials discovery. i^{AL} is the active leaning iteration(s) and i^{RS} is the random sampling iterations. The figure of merit (FOM) is a general quantitative metric measuring the AL efficiency and could indicate the desired material property value, the number of promising materials in a target range, etc. The previous studies employed the metrics for performance evaluation of AL are indicated.

for surrogate model and sampling strategy. Kim et al. [141] emphasized that beyond the model and seed data, the quality of the candidate pool, particularly the fraction of promising materials, critically affects AL performance evaluation. They pointed out that the efficiency of optimization-driven AL workflows depend on how likely candidates in the pool are to outperform those in the initial seed data and proposed metrics such as *predicted fraction of improved candidates* to quantify the pool quality. It is important to note that these insights are derived from post hoc studies with fully labeled datasets and may not directly translate to real-time AL workflows involving partial and potentially noisy data acquisition. Nevertheless, the use of diverse datasets spanning different properties, surrogate models, and acquisition functions, along with repeated trials to gather statistics suggests that these trends may hold more broadly, including for AL workflows involving real-time experiments and simulations.

In BO-AL frameworks focused on multi-fidelity or multi-objective optimization, performance metrics must be adapted beyond those used in single-fidelity and single-objective tasks. In the multi-fidelity setting, data acquisition comes with differing costs and fidelities, making it essential to quantify not just predictive accuracy, but also cost-efficiency. Acquisition functions that explicitly incorporate cost information across fidelity levels such as multi-fidelity Expected Improvement, multi-fidelity Maximum Entropy Search [142], and Targeted Variance Reduction [143], have been proposed for this purpose. Despite their promise, the performance of multi-fidelity AL is highly sensitive to the cost ratio and informativeness across the different fideltities. Jacobs et al. [144] demonstrated that low-fidelity data improves optimization only when it is significantly cheaper (e.g., $\leq$5% of high-fidelity cost) and partially available upfront. This underscores the need for performance metrics that jointly account for data cost and the degree of information transfer between fidelity levels. Steps in this direction include the recently proposed discount metric, which measures how much cheaper a multi-fidelity optimization campaign is compared to a single-fidelity one to achieve a comparable quality of solution [121]. In multi-objective AL, the aim is to identify candidate materials that perform well across multiple properties. Most studies have centered on Pareto optimization, where the goal is to approximate the Pareto front of non-dominated solutions [145, 146], using acquisition strategies such as Expected Hypervolume Improvement (EHVI) [147, 148], Pareto AL [149], and scalarization-based approaches such as ParEGO [146]. As the number of objectives grows, however, computing the hypervolume improvement becomes increasingly expensive and coverage of the Pareto front becomes sparse, limiting the efficiency of these approaches. Furthermore, materials properties are often unevenly distributed, and may vary substantially in measurement cost, which are not accounted for in standard multi-objective formulations. These challenges collectively highlight the pressing need for task-aware and cost-sensitive AL performance metrics that go beyond standard error, accuracy, or hypervolume, and that reflect the practical realities of materials design.

IV. OUTLOOK

AL has shown substantial promise for advancing data-driven materials science, yet its broader impact remains

View Article Online
DOI: 10.1039/D6DD00081A

limited by unresolved challenges in the design, execution, and evaluation of AL workflows. While this perspective has highlighted several aspects in this direction, we conclude by summarizing these into a set of core challenges and outlining future directions that we believe are most critical for advancing the field.

- **Towards systematic selection of design choices :** AL workflows involve many interacting design choices including the surrogate model, acquisition function, sampling strategy, UQ method, and stopping criterion, yet these components are rarely tuned systematically for a specific materials task. In analogy to ML model selection, these workflow components should be treated as *hyperparameters* that require careful optimization before deploying AL in resource-intensive discovery campaigns. An important next step is the development of large-scale benchmarking studies that systematically compare combinations of these choices across diverse materials datasets and target properties. Such benchmarks could establish practical guidelines and serve as reference points for future AL applications in related materials design problems. In parallel, open and modular software frameworks that support automated optimization of these hyperparameters would reduce reliance on ad hoc decisions and lower the barrier to adoption for practitioners.
- **Improved integration of domain knowledge and practical constraints :** Generic AL frameworks often overlook the domain-specific constraints inherent to materials science, such as physical validity of candidate structures, synthesis feasibility and cost. This limitation is particularly pronounced in complex, non-uniform design spaces, where the underlying property landscape may vary significantly across regions, and uninformed sampling can fail to concentrate effort where it is most needed. When such considerations are neglected, AL may waste valuable evaluation budget on impractical or infeasible candidates, reducing the efficiency gains. Future AL workflows should incorporate domain knowledge and practical constraints, for example, through physics-informed surrogate models or sampling strategies that directly encode feasibility constraints [78, 150]. Additionally, physically motivated representations enable AL to be grounded on the relevant structural and physicochemical characteristics of the materials. Such integration not only enhances data efficiency but also improves the reliability and interpretability of AL outcomes, particularly in scenarios with limited data.
- **Leveraging emerging AI paradigms :** Several recent advances in AI offer promising pathways to address persistent limitations of current AL workflows. The unreliability of surrogate models in early AL stages could be alleviated by incorporating pretrained or foundational models trained on broad materials or even synthetic datasets, which provide more reliable uncertainty estimates under low-data regimes [151]. However, care must be taken as these models can introduce systematic inductive biases; if the pretraining distribution is poorly aligned with the target domain, the AL process may become trapped in over-represented regions of the design space. For high-dimensional and complex design spaces where conventional surrogate-based AL struggles, reformulating the sampling problem as sequential decision-making through reinforcement learning offers a route to more adaptive exploration policies [152]. In parallel, recent advances in generative models [153, 154] open opportunities to go beyond pool-based AL, where a predefined set of unlabeled candidates is required, which may not be feasible for many materials discovery applications. In this setting, AL can be combined with generative models to sample new candidates by exploring an open design space, which are then decoded into material compositions and structures using generative models such as variational autoencoders [155]. However, it is important to ensure that such generated candidates are physically plausible and chemically meaningful by enforcing strict physicochemical constraints. Realizing the full potential of these approaches also requires a stronger integration between the materials science and ML communities, as many advances relevant to AL, such as advanced UQ methods [156] developed by the later remain underexplored for AL for materials science applications.

DATA AVAILABILITY

No data is newly generated during this work.

AUTHOR CONTRIBUTIONS

A.S.N proposed the idea, conducted the literature study and prepared the first draft of the paper. All the authors discussed and commented on the manuscript.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

ACKNOWLEDGEMENT

A.S.N thanks Matthias Scheffler for helpful discussions, and Beate Paulus and the German Research Foun-

View Article Online
DOI: 10.1039/D6DD00081A

dation (DFG) for support through the Walter-Benjamin Fellowship Program (project No. 540316537).

REFERENCES

.

[1] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh, Machine learning for molecular and materials science, Nature **559**, 547 (2018).
[2] R. Vasudevan, G. Pilania, and P. V. Balachandran, Machine learning for materials design and discovery, Journal of Applied Physics **129** (2021).
[3] S. Axelrod, D. Schwalbe-Koda, S. Mohapatra, J. Damewood, K. P. Greenman, and R. Gómez-Bombarelli, Learning matter: Materials design with machine learning and atomistic simulations, Accounts of Materials Research **3**, 343 (2022).
[4] B. G. Sumpter, R. K. Vasudevan, T. Potok, and S. V. Kalinin, A bridge for accelerating materials by design, NPJ Computational Materials **1**, 1 (2015).
[5] E. O. Pyzer-Knapp, J. W. Pitera, P. W. Staar, S. Takeda, T. Laino, D. P. Sanders, J. Sexton, J. R. Smith, and A. Curioni, Accelerating materials discovery using artificial intelligence, high performance computing and robotics, npj Computational Materials **8**, 84 (2022).
[6] A. Merchant, S. Batzner, S. S. Schoenholz, M. Aykol, G. Cheon, and E. D. Cubuk, Scaling deep learning for materials discovery, Nature **624**, 80 (2023).
[7] W. F. Maier, K. Stoewe, and S. Sieg, Combinatorial and high-throughput materials science, Angewandte chemie international edition **46**, 6016 (2007).
[8] S. Curtarolo, G. L. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, and O. Levy, The high-throughput highway to computational materials design, Nature materials **12**, 191 (2013).
[9] P. Xu, X. Ji, M. Li, and W. Lu, Small data machine learning in materials science, npj Computational Materials **9**, 42 (2023).
[10] S. Bauer, P. Benner, T. Bereau, V. Blum, M. Boley, C. Carbogno, R. Catlow, G. Dehm, S. Eibl, R. Ernstorfer, *et al.*, Roadmap on data-centric materials science, Modelling and Simulation in Materials Science and Engineering (2024).
[11] P. Karande, B. Gallagher, and T. Y.-J. Han, A strategic approach to machine learning for material science: how to tackle real-world challenges and avoid pitfalls, Chemistry of Materials **34**, 7650 (2022).
[12] D. A. Cohn, Z. Ghahramani, and M. I. Jordan, Active learning with statistical models, Journal of artificial intelligence research **4**, 129 (1996).
[13] D. Cohn, L. Atlas, and R. Ladner, Improving generalization with active learning, Machine learning **15**, 201 (1994).
[14] A. Koizumi, G. Deffrennes, K. Terayama, and R. Tamura, Performance of uncertainty-based active learning for efficient approximation of black-box functions in materials science, Scientific Reports **14**, 27019 (2024).
[15] A. Jose, E. Devijver, N. Jakse, and R. Poloni, Informative training data for efficient property prediction in metal–organic frameworks by active learning, Journal of the American Chemical Society **146**, 6134 (2024).
[16] H. Jang, W. Lee, H.-J. Kim, S. Cha, H. Shin, W. B. Lee, M.-W. Oh, Y. S. Jung, and Y. Kim, Active learning-guided accelerated discovery of ultra-efficient high-entropy thermoelectrics, Advanced Materials **38**, e15054 (2026).
[17] L. Bassman Oftelie, P. Rajak, R. K. Kalia, A. Nakano, F. Sha, J. Sun, D. J. Singh, M. Aykol, P. Huck, K. Persson, *et al.*, Active learning for accelerated design of layered materials, npj Computational Materials **4**, 74 (2018).
[18] K. Tran and Z. W. Ulissi, Active learning across intermetallics to guide discovery of electrocatalysts for co2 reduction and h2 evolution, Nature Catalysis **1**, 696 (2018).
[19] A. S. Nair, L. Foppa, and M. Scheffler, Materials-discovery workflow guided by symbolic regression for identifying acid-stable oxides for electrocatalysis, npj Computational Materials **11**, 1 (2025).
[20] D. Xue, P. V. Balachandran, J. Hogden, J. Theiler, D. Xue, and T. Lookman, Accelerated search for materials with targeted properties by adaptive design, Nature communications **7**, 1 (2016).
[21] T. Lookman, P. V. Balachandran, D. Xue, and R. Yuan, Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design, npj Computational Materials **5**, 21 (2019).
[22] O. Ozbayram, D. Olsen, M. Annamaraju, A. E. Robertson, A. Venkatraman, S. R. Kalidindi, M. Zhou, and L. Graham-Brady, Batch active learning for microstructure–property relations in energetic materials, Mechanics of Materials **205**, 105308 (2025).
[23] M. Hu, Q. Tan, R. Knibbe, M. Xu, G. Liang, J. Zhou, J. Xu, B. Jiang, X. Li, M. Ramajayam, *et al.*, Designing unique and high-performance al alloys via machine learning: Mitigating data bias through active learning, Computational Materials Science **244**, 113204 (2024).
[24] H. Zhang, W. W. Chen, J. M. Rondinelli, and W. Chen, Et-al: entropy-targeted active learning for bias mitigation in materials data, Applied Physics Reviews **10** (2023).
[25] K. Kang, T. A. Purcell, C. Carbogno, and M. Scheffler, Accelerating the training and improving the reliability of machine-learned interatomic potentials for strongly anharmonic materials through active learning, arXiv preprint arXiv:2409.11808 (2024).
[26] A. Wang, H. Liang, A. McDannald, I. Takeuchi, and A. G. Kusne, Benchmarking active learning strategies

View Article Online
DOI: 10.1039/D6DD00081A

for materials optimization and discovery, Oxford Open Materials Science **2**, itac006 (2022).
[27] Y. Tian, R. Yuan, D. Xue, Y. Zhou, X. Ding, J. Sun, and T. Lookman, Role of uncertainty estimation in accelerating materials development via active learning, Journal of Applied Physics **128** (2020).
[28] M. Kulichenko, B. Nebgen, N. Lubbers, J. S. Smith, K. Barros, A. E. Allen, A. Habib, E. Shinkle, N. Fedik, Y. W. Li, *et al.*, Data generation for machine learning interatomic potentials and beyond, Chemical Reviews **124**, 13681 (2024).
[29] K. Konyushkova, R. Sznitman, and P. Fua, Learning active learning from data, Advances in neural information processing systems **30** (2017).
[30] T. Lookman, P. V. Balachandran, D. Xue, J. Hogden, and J. Theiler, Statistical inference and adaptive design for materials discovery, Current Opinion in Solid State and Materials Science **21**, 121 (2017).
[31] M. Stein, Large sample properties of simulations using latin hypercube sampling, Technometrics **29**, 143 (1987).
[32] C. Campbell, N. Cristianini, A. Smola, *et al.*, Query learning with large margin classifiers, in *ICML*, Vol. 20 (2000) p. 0.
[33] J.-N. Hwang, J. J. Choi, S. Oh, R. Marks, *et al.*, Query-based learning applied to partially trained multilayer perceptrons, IEEE Transactions on Neural Networks **2**, 131 (1991).
[34] 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, Chemical science **11**, 2696 (2020).
[35] H. Khosravi, T. Olajire, A. S. Raihan, and I. Ahmed, A data driven sequential learning framework to accelerate and optimize multi-objective manufacturing decisions, Journal of Intelligent Manufacturing , 1 (2024).
[36] D. Angluin, Queries and concept learning, Machine learning **2**, 319 (1988).
[37] L. Atlas, D. Cohn, and R. Ladner, Training connectionist networks with queries and selective sampling, Advances in neural information processing systems **2** (1989).
[38] D. D. Lewis, A sequential algorithm for training text classifiers: Corrigendum and additional data, in *Acm Sigir Forum*, Vol. 29 (ACM New York, NY, USA, 1995) pp. 13–19.
[39] C. Williams and C. Rasmussen, Gaussian processes for regression, Advances in neural information processing systems **8** (1995).
[40] L. Breiman, Random forests, Machine learning **45**, 5 (2001).
[41] Y. LeCun, Y. Bengio, and G. Hinton, Deep learning, nature **521**, 436 (2015).
[42] K. Li, D. Persaud, K. Choudhary, B. DeCost, M. Greenwood, and J. Hattrick-Simpers, Exploiting redundancy in large materials datasets for efficient machine learning with less data, Nature Communications **14**, 7283 (2023).
[43] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, *et al.*, Commentary: The materials project: A materials genome approach to accelerating materials innovation, APL materials **1** (2013).
[44] S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl, and C. Wolverton, The open quantum materials database (oqmd): assessing the accuracy of dft formation energies, npj Computational Materials **1**, 1 (2015).
[45] P. Lyngby and K. S. Thygesen, Data-driven discovery of 2d materials by deep generative models, npj Computational Materials **8**, 232 (2022).
[46] C. J. Bartel, A. Trewartha, Q. Wang, A. Dunn, A. Jain, and G. Ceder, A critical examination of compound stability predictions from machine-learned formation energies, npj computational materials **6**, 97 (2020).
[47] Y. Sheng, Y. Wu, J. Yang, W. Lu, P. Villars, and W. Zhang, Active learning for the power factor prediction in diamond-like thermoelectric materials, npj Computational Materials **6**, 171 (2020).
[48] B. Settles, Active learning literature survey, arXiv preprint arXiv:1010.1010 (2009).
[49] Z. Wang and J. Ye, Querying discriminative and representative samples for batch mode active learning, ACM Transactions on Knowledge Discovery from Data (TKDD) **9**, 1 (2015).
[50] X. Li, D. Kuang, and C. X. Ling, Active learning for hierarchical text classification, in *Pacific-Asia conference on knowledge discovery and data mining* (Springer, 2012) pp. 14–25.
[51] B. Du, Z. Wang, L. Zhang, L. Zhang, W. Liu, J. Shen, and D. Tao, Exploring representativeness and informativeness for active learning, IEEE transactions on cybernetics **47**, 14 (2015).
[52] S.-J. Huang, R. Jin, and Z.-H. Zhou, Active learning by querying informative and representative examples, Advances in neural information processing systems **23** (2010).
[53] V. Zaverkin, D. Holzmüller, I. Steinwart, and J. Kästner, Exploring chemical and conformational spaces by batch mode deep active learning, Digital Discovery **1**, 605 (2022).
[54] K. Brinker, Incorporating diversity in active learning with support vector machines, in *Proceedings of the 20th international conference on machine learning (ICML-03)* (2003) pp. 59–66.
[55] D. J. MacKay, Information-based objective functions for active data selection, Neural computation **4**, 590 (1992).
[56] S. Dasgupta, Two faces of active learning, Theoretical computer science **412**, 1767 (2011).
[57] S. Farquhar, Y. Gal, and T. Rainforth, On statistical bias in active learning: How and when to fix it, arXiv preprint arXiv:2101.11665 (2021).
[58] C. Murray, J. U. Allingham, J. Antorán, and J. M. Hernández-Lobato, Addressing bias in active learning with depth uncertainty networks... or not, in *I (Still) Can't Believe It's Not Better! Workshop at NeurIPS 2021* (PMLR, 2022) pp. 59–63.
[59] J. Hu, D. Liu, N. Fu, and R. Dong, Realistic material property prediction using domain adaptation based machine learning, Digital Discovery **3**, 300 (2024).
[60] L. E. Schultz, Y. Wang, R. Jacobs, and D. Morgan, A general approach for determining applicability domain of machine learning models, npj Computational Materials **11**, 95 (2025).
[61] C. Zeni, A. Anelli, A. Glielmo, and K. Rossi, Exploring the robust extrapolation of high-dimensional machine learning potentials, Physical Review B **105**, 165141

View Article Online
DOI: 10.1039/D6DD00081A

(2022).
[62] X. Yang, Y. Liu, C. Mi, and X. Wang, Active learning kriging model combining with kernel-density-estimation-based importance sampling method for the estimation of low failure probability, Journal of Mechanical Design **140**, 051402 (2018).
[63] S. Xiong, J. Azimi, and X. Z. Fern, Active learning of constraints for semi-supervised clustering, IEEE Transactions on Knowledge and Data Engineering **26**, 43 (2013).
[64] D. Schwalbe-Koda, S. Hamel, B. Sadigh, F. Zhou, and V. Lordi, Model-free estimation of completeness, uncertainties, and outliers in atomistic machine learning using information theory, Nature Communications **16**, 4014 (2025).
[65] A. PA Subramanyam and D. Perez, Information-entropy-driven generation of material-agnostic datasets for machine-learning interatomic potentials, npj Computational Materials **11**, 218 (2025).
[66] J. Moon, W. Beker, M. Siek, J. Kim, H. S. Lee, T. Hyeon, and B. A. Grzybowski, Active learning guides discovery of a champion four-metal perovskite oxide for oxygen evolution electrocatalysis, Nature Materials **23**, 108 (2024).
[67] J. Qi, T. W. Ko, B. C. Wood, T. A. Pham, and S. P. Ong, Robust training of machine learning interatomic potentials with dimensionality reduction and stratified sampling, npj Computational Materials **10**, 43 (2024).
[68] V. Gkatsis, P. Maratos, C. Rekatsinas, G. Giannakopoulos, and P. Krokidas, Density-aware active learning for materials discovery: A case study on functionalized nanoporous materials, Physical Chemistry Chemical Physics (2025).
[69] S. Nie, Y. Xiang, L. Wu, G. Lin, Q. Liu, S. Chu, and X. Wang, Active learning guided discovery of high entropy oxides featuring high h2-production, Journal of the American Chemical Society **146**, 29325 (2024).
[70] H. Jung, L. Sauerland, S. Stocker, K. Reuter, and J. T. Margraf, Machine-learning driven global optimization of surface adsorbate geometries, npj Computational Materials **9**, 114 (2023).
[71] H. Liu, B. Yucel, B. Ganapathysubramanian, S. R. Kalidindi, D. Wheeler, and O. Wodo, Active learning for regression of structure–property mapping: the importance of sampling and representation, Digital Discovery **3**, 1997 (2024).
[72] Y. Kim and B. Shin, In defense of core-set: A density-aware core-set selection for active learning, in *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining* (2022) pp. 804–812.
[73] J. P. Janet, C. Duan, T. Yang, A. Nandy, and H. J. Kulik, A quantitative uncertainty metric controls error in neural network-driven chemical discovery, Chemical science **10**, 7913 (2019).
[74] Q. Li, N. Fu, S. S. Omee, and J. Hu, Md-hit: Machine learning for material property prediction with dataset redundancy control, npj Computational Materials **10**, 245 (2024).
[75] X. Zhan, H. Liu, Q. Li, and A. B. Chan, A comparative survey: Benchmarking for pool-based active learning., in *IJCAI* (2021) pp. 4679–4686.
[76] A. P. Bartók, R. Kondor, and G. Csányi, On representing chemical environments, Physical Review B—Condensed Matter and Materials Physics **87**, 184115 (2013).
[77] S. S. Omee, N. Fu, R. Dong, M. Hu, and J. Hu, Structure-based out-of-distribution (ood) materials property prediction: a benchmark study, npj Computational Materials **10**, 144 (2024).
[78] A. G. Kusne, H. Yu, C. Wu, H. Zhang, J. Hattrick-Simpers, B. DeCost, S. Sarker, C. Oses, C. Toher, S. Curtarolo, *et al.*, On-the-fly closed-loop materials discovery via bayesian active learning, Nature communications **11**, 5966 (2020).
[79] J. Bi, Y. Xu, F. Conrad, H. Wiemer, and S. Ihlenfeldt, A comprehensive benchmark of active learning strategies with automl for small-sample regression in materials science, Scientific Reports **15**, 37167 (2025).
[80] A. Jose, J. P. A. de Mendonça, E. Devijver, N. Jakse, V. Monbet, and R. Poloni, Regression tree-based active learning, Data Mining and Knowledge Discovery **38**, 420 (2024).
[81] S. Kee, E. Del Castillo, and G. Runger, Query-by-committee improvement with diversity and density in batch active learning, Information Sciences **454**, 401 (2018).
[82] X. Yuan, M. Suvarna, J. Y. Lim, J. Pérez-Ramírez, X. Wang, and Y. S. Ok, Active learning-based guided synthesis of engineered biochar for co2 capture, Environmental Science & Technology **58**, 6628 (2024).
[83] K. Min and E. Cho, Accelerated discovery of novel inorganic materials with desired properties using active learning, The Journal of Physical Chemistry C **124**, 14759 (2020).
[84] F. Hase, L. M. Roch, C. Kreisbeck, and A. Aspuru-Guzik, Phoenics: a bayesian optimizer for chemistry, ACS central science **4**, 1134 (2018).
[85] M. Fronzi, O. Isayev, D. A. Winkler, J. G. Shapter, A. V. Ellis, P. C. Sherrell, N. A. Shepelin, A. Corletto, and M. J. Ford, Active learning in bayesian neural networks for bandgap predictions of novel van der waals heterostructures, Advanced Intelligent Systems **3**, 2100080 (2021).
[86] S. S. Hessmann, K. T. Schütt, N. W. Gebauer, M. Gastegger, T. Oguchi, and T. Yamashita, Accelerating crystal structure search through active learning with neural networks for rapid relaxations, npj Computational Materials **11**, 44 (2025).
[87] L. Zhang, D.-Y. Lin, H. Wang, R. Car, and W. E, Active learning of uniformly accurate interatomic potentials for materials simulation, Physical Review Materials **3**, 023804 (2019).
[88] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. De Freitas, Taking the human out of the loop: A review of bayesian optimization, Proceedings of the IEEE **104**, 148 (2015).
[89] Y. Wu, A. Walsh, and A. M. Ganose, Race to the bottom: Bayesian optimisation for chemical problems, Digital Discovery **3**, 1086 (2024).
[90] A. Deshwal, C. M. Simon, and J. R. Doppa, Bayesian optimization of nanoporous materials, Molecular Systems Design & Engineering **6**, 1066 (2021).
[91] 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 specification, Computational Materials Science **210**, 111417 (2022).

View Article Online
DOI: 10.1039/D6DD00081A

[92] J. K. Pedersen, C. M. Clausen, O. A. Krysiak, B. Xiao, T. A. Batchelor, T. Löffler, V. A. Mints, L. Banko, M. Arenz, A. Savan, *et al.*, Bayesian optimization of high-entropy alloy compositions for electrocatalytic oxygen reduction, Angewandte Chemie **133**, 24346 (2021).

[93] R. Yuan, Z. Liu, P. V. Balachandran, D. Xue, Y. Zhou, X. Ding, J. Sun, D. Xue, and T. Lookman, Accelerated discovery of large electrostrains in batio3-based piezoelectrics using active learning, Advanced materials **30**, 1702884 (2018).

[94] Y.-F. Lim, C. K. Ng, U. Vaitesswar, and K. Hippalgaonkar, Extrapolative bayesian optimization with gaussian process and neural network ensemble surrogate models, Advanced Intelligent Systems **3**, 2100101 (2021).

[95] Q. Liang, A. E. Gongora, Z. Ren, A. Tiihonen, Z. Liu, S. Sun, J. R. Deneault, D. Bash, F. Mekki-Berrada, S. A. Khan, *et al.*, Benchmarking the performance of bayesian optimization across multiple experimental materials science domains, npj Computational Materials **7**, 188 (2021).

[96] R. Moriconi, M. P. Deisenroth, and K. Sesh Kumar, High-dimensional bayesian optimization using low-dimensional feature spaces, Machine Learning **109**, 1925 (2020).

[97] A. Seko, A. Togo, H. Hayashi, K. Tsuda, L. Chaput, and I. Tanaka, Prediction of low-thermal-conductivity compounds with first-principles anharmonic lattice-dynamics calculations and bayesian optimization, Phys. Rev. Lett **115**, 205901 (2015).

[98] A. Seko, T. Maekawa, K. Tsuda, and I. Tanaka, Machine learning with systematic density-functional theory calculations: Application to melting temperatures of single-and binary-component solids, Physical Review B **89**, 054303 (2014).

[99] Y. Zuo, M. Qin, C. Chen, W. Ye, X. Li, J. Luo, and S. P. Ong, Accelerating materials discovery with bayesian optimization and graph deep learning, Materials Today **51**, 126 (2021).

[100] A. Ishii, S. Kikuchi, A. Yamanaka, and A. Yamamoto, Application of bayesian optimization to the synthesis process of bafe2 (as, p) 2 polycrystalline bulk superconducting materials, Journal of Alloys and Compounds **966**, 171613 (2023).

[101] M. Ghorbani, M. Boley, P. Nakashima, and N. Birbilis, An active machine learning approach for optimal design of magnesium alloys using bayesian optimisation, Scientific Reports **14**, 8299 (2024).

[102] M. Yu, S. Yang, C. Wu, and N. Marom, Machine learning the hubbard u parameter in dft+ u using bayesian optimization, npj computational materials **6**, 180 (2020).

[103] T. Ueno, T. D. Rhone, Z. Hou, T. Mizoguchi, and K. Tsuda, Combo: An efficient bayesian optimization library for materials science, Materials discovery **4**, 18 (2016).

[104] T. Yamashita, N. Sato, H. Kino, T. Miyake, K. Tsuda, and T. Oguchi, Crystal structure prediction accelerated by bayesian optimization, Physical Review Materials **2**, 013803 (2018).

[105] H. C. Herbol, M. Poloczek, and P. Clancy, Cost-effective materials discovery: Bayesian optimization across multiple information sources, Materials Horizons **7**, 2113 (2020).

[106] S. R. Chitturi, A. Ramdas, Y. Wu, B. Rohr, S. Ermon, J. Dionne, F. H. d. Jornada, M. Dunne, C. Tassone, W. Neiswanger, *et al.*, Targeted materials discovery using bayesian algorithm execution, npj Computational Materials **10**, 156 (2024).

[107] J. I. Myung, J. R. Deneault, J. Chang, I. Kang, B. Maruyama, and M. A. Pitt, Multi-objective bayesian optimization: a case study in material extrusion, Digital Discovery **4**, 464 (2025).

[108] C. Benjamins, E. Raponi, A. Jankovic, K. van der Blom, M. L. Santoni, M. Lindauer, and C. Doerr, Pi is back! switching acquisition functions in bayesian optimization, arXiv preprint arXiv:2211.01455 (2022).

[109] S. B. Torrisi, M. Z. Bazant, A. E. Cohen, M. G. Cho, J. S. Hummelshøj, L. Hung, G. Kamat, A. Khajeh, A. Kolluru, X. Lei, *et al.*, Materials cartography: A forward-looking perspective on materials representation and devising better maps, APL Machine Learning **1** (2023).

[110] M. Tenorio, M. H. Rahman, A. Mannodi-Kanakkithodi, and J. Chapman, Out-of-distribution machine learning for materials discovery: Challenges and opportunities, Chemical Physics Reviews **7** (2026).

[111] A. S. Nair, L. Foppa, and M. Scheffler, Interpretable bayesian optimization for catalyst discovery, Faraday Discussions (2026).

[112] L. Ward, A. Agrawal, A. Choudhary, and C. Wolverton, A general-purpose machine learning framework for predicting properties of inorganic materials, npj Computational Materials **2**, 16028 (2016).

[113] C. K. 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 **2**, 327 (2023).

[114] Y. Tenne and C.-K. Goh, *Computational intelligence in expensive optimization problems*, Vol. 2 (Springer Science & Business Media, 2010).

[115] J. Fromer, R. Wang, M. Manjrekar, A. Tripp, J. M. Hernández-Lobato, and C. W. Coley, Batched bayesian optimization by maximizing the probability of including the optimum, Journal of Chemical Information and Modeling **65**, 4808 (2025).

[116] T. Kathuria, A. Deshpande, and P. Kohli, Batched gaussian process bandit optimization via determinantal point processes, Advances in neural information processing systems **29** (2016).

[117] L. M. Roch, F. Häse, C. Kreisbeck, T. Tamayo-Mendoza, L. P. Yunker, J. E. Hein, and A. Aspuru-Guzik, Chemos: orchestrating autonomous experimentation, Science Robotics **3**, eaat5559 (2018).

[118] F. Häse, M. Aldeghi, R. J. Hickman, L. M. Roch, M. Christensen, E. Liles, J. E. Hein, and A. Aspuru-Guzik, Olympus: a benchmarking framework for noisy optimization and experiment planning, Machine Learning: Science and Technology **2**, 035021 (2021).

[119] E. H. Lee, V. Perrone, C. Archambeau, and M. Seeger, Cost-aware bayesian optimization, arXiv preprint arXiv:2003.10870 (2020).

[120] J. Snoek, H. Larochelle, and R. P. Adams, Practical bayesian optimization of machine learning algorithms, Advances in neural information processing systems **25** (2012).

[121] V. Sabanza-Gil, R. Barbano, D. Pacheco Gutiérrez, J. S. Luterbacher, J. M. Hernández-Lobato, P. Schwaller, and

View Article Online
DOI: 10.1039/D6DD00081A

L. Roch, Best practices for multi-fidelity bayesian optimization in materials and molecular research, Nature Computational Science **5**, 572 (2025).

[122] F. Grasselli, S. Chong, V. Kapil, S. Bonfanti, and K. Rossi, Uncertainty in the era of machine learning for atomistic modeling, Digital Discovery **4**, 2654 (2025).

[123] P. Pernot, Calibration in machine learning uncertainty quantification: beyond consistency to target adaptivity, APL Machine Learning **1** (2023).

[124] Y. Hwang, W. Jo, J. Hong, and Y. Choi, Overcoming overconfidence for active learning, IEEE Access (2024).

[125] 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 **2**, 1112 (2023).

[126] L. Kavalsky, V. I. Hegde, B. Meredig, and V. Viswanathan, A multiobjective closed-loop approach towards autonomous discovery of electrocatalysts for nitrogen reduction, Digital discovery **3**, 999 (2024).

[127] D. Varivoda, R. Dong, S. S. Omee, and J. Hu, Materials property prediction with uncertainty quantification: A benchmark study, Applied Physics Reviews **10** (2023).

[128] K. Tran, W. Neiswanger, J. Yoon, Q. Zhang, E. Xing, and Z. W. Ulissi, Methods for comparing uncertainty quantifications for material property predictions, Machine Learning: Science and Technology **1**, 025006 (2020).

[129] C. J. Gruich, V. Madhavan, Y. Wang, and B. R. Goldsmith, Clarifying trust of materials property predictions using neural networks with distribution-specific uncertainty quantification, Machine Learning: Science and Technology **4**, 025019 (2023).

[130] G. Palmer, S. Du, A. Politowicz, J. P. Emory, X. Yang, A. Gautam, G. Gupta, Z. Li, R. Jacobs, and D. Morgan, Calibration after bootstrap for accurate uncertainty quantification in regression models, npj Computational Materials **8**, 115 (2022).

[131] Y. Chung, I. Char, H. Guo, J. Schneider, and W. Neiswanger, Uncertainty toolbox: an open-source library for assessing, visualizing, and improving uncertainty quantification, arXiv preprint arXiv:2109.10254 (2021).

[132] C. Lataniotis, S. Marelli, and B. Sudret, Uqlab 2.0 and uqcloud: open-source vs. cloud-based uncertainty quantification, in *SIAM Conference on Uncertainty Quantification (SIAM UQ 2022)* (ETH Zurich, Institute of Structural Engineering, 2022).

[133] A. Thomas-Mitchell, G. Hawe, and P. L. Popelier, Calibration of uncertainty in the active learning of machine learning force fields, Machine Learning: Science and Technology **4**, 045034 (2023).

[134] V. Korolev, I. Nevolin, and P. Protsenko, A universal similarity based approach for predictive uncertainty quantification in materials science, Scientific Reports **12**, 14931 (2022).

[135] J. Musielewicz, J. Lan, M. Uyttendaele, and J. R. Kitchin, Improved uncertainty estimation of graph neural network potentials using engineered latent space distances, The Journal of Physical Chemistry C **128**, 20799 (2024).

[136] S. Zhong, D. R. Lambeth, T. K. Igou, and Y. Chen, Enlarging applicability domain of quantitative structure–activity relationship models through uncertainty-based active learning, ACS ES&T Engineering **2**, 1211 (2022).

[137] C. Sutton, M. Boley, L. M. Ghiringhelli, M. Rupp, J. Vreeken, and M. Scheffler, Identifying domains of applicability of machine learning models for materials science, Nature communications **11**, 4428 (2020).

[138] A. Palizhati, S. B. Torrisi, M. Aykol, S. K. Suram, J. S. Hummelshøj, and J. H. Montoya, Agents for sequential learning using multiple-fidelity data, Scientific reports **12**, 4694 (2022).

[139] J. N. Fuhg, A. Fau, and U. Nackenhorst, State-of-the-art and comparative review of adaptive sampling methods for kriging, Archives of Computational Methods in Engineering **28**, 2689 (2021).

[140] N. Stolte, J. Daru, H. Forbert, D. Marx, and J. Behler, Random sampling versus active learning algorithms for machine learning potentials of quantum liquid water, Journal of Chemical Theory and Computation **21**, 886 (2025).

[141] Y. Kim, E. Kim, E. Antono, B. Meredig, and J. Ling, Machine-learned metrics for predicting the likelihood of success in materials discovery, npj Computational Materials **6**, 131 (2020).

[142] S. Takeno, H. Fukuoka, Y. Tsukada, T. Koyama, M. Shiga, I. Takeuchi, and M. Karasuyama, Multi-fidelity bayesian optimization with max-value entropy search and its parallelization, in *International Conference on Machine Learning* (PMLR, 2020) pp. 9334–9345.

[143] C. Fare, P. Fenner, M. Benatan, A. Varsi, and E. O. Pyzer-Knapp, A multi-fidelity machine learning approach to high throughput materials screening, npj Computational Materials **8**, 257 (2022).

[144] R. Jacobs, P. E. Goins, and D. Morgan, Role of multi-fidelity data in sequential active learning materials discovery campaigns: case study of electronic bandgap, Machine Learning: Science and Technology **4**, 045060 (2023).

[145] V. Trinquet, M. L. Evans, C. J. Hargreaves, P.-P. De Breuck, and G.-M. Rignanese, Optical materials discovery and design with federated databases and machine learning, Faraday Discussions **256**, 459 (2025).

[146] P. Xu, Y. Ma, W. Lu, M. Li, W. Zhao, and Z. Dai, Multi-objective optimization in machine learning assisted materials design and discovery, Journal of Materials Informatics **5**, N (2025).

[147] K. Park, C. Song, J. Park, and S. Ryu, Multi-objective bayesian optimization for the design of nacre-inspired composites: optimizing and understanding biomimetics through ai, Materials Horizons **10**, 4329 (2023).

[148] B. P. MacLeod, F. G. Parlane, C. C. Rupnow, K. E. Dettelbach, M. S. Elliott, T. D. Morrissey, T. H. Haley, O. Proskurin, M. B. Rooney, N. Taherimakhsousi, *et al.*, A self-driving laboratory advances the pareto front for material properties, Nature communications **13**, 995 (2022).

[149] K. M. Jablonka, G. M. Jothiappan, S. Wang, B. Smit, and B. Yoo, Bias free multiobjective active learning for materials design and discovery, Nature communications **12**, 2312 (2021).

[150] H. A. Doan, G. Agarwal, H. Qian, M. J. Counihan, J. Rodríguez-López, J. S. Moore, and R. S. Assary, Quantum chemistry-informed active learning to accelerate the design and discovery of sustainable energy storage materials, Chemistry of Materials **32**, 6338 (2020).

View Article Online
DOI: 10.1039/D6DD00081A

[151] J. Hu, R. Dong, Y. Feng, M. Hu, and J. Hu, Foundation-model surrogates enable data-efficient active learning for materials discovery, arXiv preprint arXiv:2603.12567 (2026).
[152] Y. Xian, X. Ding, X. Jiang, Y. Zhou, J. Sun, D. Xue, and T. Lookman, Unlocking the black box beyond bayesian global optimization for materials design using reinforcement learning, npj Computational Materials **11**, 1 (2025).
[153] H. Metni, L. Ruple, L. N. Walters, L. Torresi, J. Teufel, H. Schopmans, J. Östreicher, Y. Zhang, M. Neubert, Y. Koide, *et al.*, Generative models for crystalline materials, Advanced Materials , e23620 (2026).
[154] H. Park, Z. Li, and A. Walsh, Has generative artificial intelligence solved inverse materials design?, Matter **7**, 2355 (2024).
[155] R. Xin, E. M. Siriwardane, Y. Song, Y. Zhao, S.-Y. Louis, A. Nasiri, and J. Hu, Active-learning-based generative design for the discovery of wide-band-gap materials, The Journal of Physical Chemistry C **125**, 16118 (2021).
[156] S. Lahlou, M. Jain, H. Nekoei, V. I. Butoi, P. Bertin, J. Rector-Brooks, M. Korablyov, and Y. Bengio, Deup: Direct epistemic uncertainty prediction, arXiv preprint arXiv:2102.08501 (2021).

View Article Online
DOI: 10.1039/D6DD00081A

Data Availability Statement

No data is newly generated during this work.