

ORIGINAL RESEARCH

Open access

Critical Assessment of Active Learning Benchmarks in Materials Discovery: Undisclosed Baselines and Optimistic Bias

Nguyen Van Nam¹, Tran Thi Hoa^{1*}, Le Minh Duc²

Abstract

Active learning has become a cornerstone strategy in data-driven materials discovery, promising to dramatically reduce the number of expensive simulations or experiments needed to identify high-performing materials. Proponents argue that uncertainty sampling, expected improvement, and other acquisition functions consistently outperform random selection, often by factors of 3–5× in iteration efficiency. Yet a closer examination of the benchmark studies published between 2017 and 2025 reveals a systematic pattern of optimistic bias that inflates these claims. This critical critique identifies five primary sources of overestimation: (1) weak or undisclosed random-sampling baselines, (2) unrealistic initial training sets that artificially favor active learning, (3) test-set leakage that prevents genuine extrapolation, (4) acquisition functions whose hyperparameters are implicitly tuned to the specific benchmark, and (5) incomplete reporting that hides variance and failure cases. Across the literature, random sampling is frequently presented as a naïve comparator yet proves surprisingly competitive once proper repetition, variance reporting, and realistic initial-set sizes are applied. Many studies fail to disclose the number of random seeds, the exact sampling distribution, or statistical significance tests, allowing small apparent gains to be reported as transformative. Initial training sets are often unrealistically small or already enriched with promising candidates, while test sets remain too similar to the training distribution, masking the true difficulty of exploration in vast chemical spaces. Acquisition functions are rarely subjected to hyperparameter robustness checks or evaluated on challenging out-of-distribution splits. The consequences extend beyond academic metrics: practitioners in industry and national laboratories risk deploying methods that underperform once transferred to real discovery campaigns. This critique, grounded exclusively in the 29 peer-reviewed studies listed in the reference section, calls for a new standard of rigor in active-learning evaluation. Only by adopting strong baselines, realistic initial conditions, extrapolation-aware test sets, full variance reporting, and public replication packages can the field move from optimistic benchmark theater to genuinely reliable acceleration of materials discovery.

Keywords Materials discovery, Active learning, Benchmark bias, Random sampling baseline, Optimistic evaluation, Undisclosed baselines

*Correspondence:

Tran Thi Hoa

tran.hoa@outlook.com

¹ Department of Intelligent Materials Engineering, Faculty of Engineering, Hanoi University of Science and Technology, Hanoi, Vietnam

² Department of Data-Driven Materials Systems, Faculty of Technology, Ho Chi Minh City University of Technology, Ho Chi Minh City, Vietnam

Introduction

Active learning is widely promoted for materials discovery by intelligently selecting which experiments or simulations to run next [1-5]. Papers report impressive results: their acquisition function finds the best material in 50 iterations, compared to 200 for random sampling [6-8]. But are these claims reliable? This critique argues that active learning benchmarks in materials discovery suffer from optimistic bias — systematic overestimation of performance due to weak baselines, unrealistic setups, and undisclosed choices [9-11]. We identify five sources of bias and propose rigorous evaluation standards [1, 3, 12].

Figure 1 maps the manuscript's central argument by showing how five benchmark-design choices generate inflated performance claims, misleading evidence, and ultimately distorted scientific and practical conclusions.

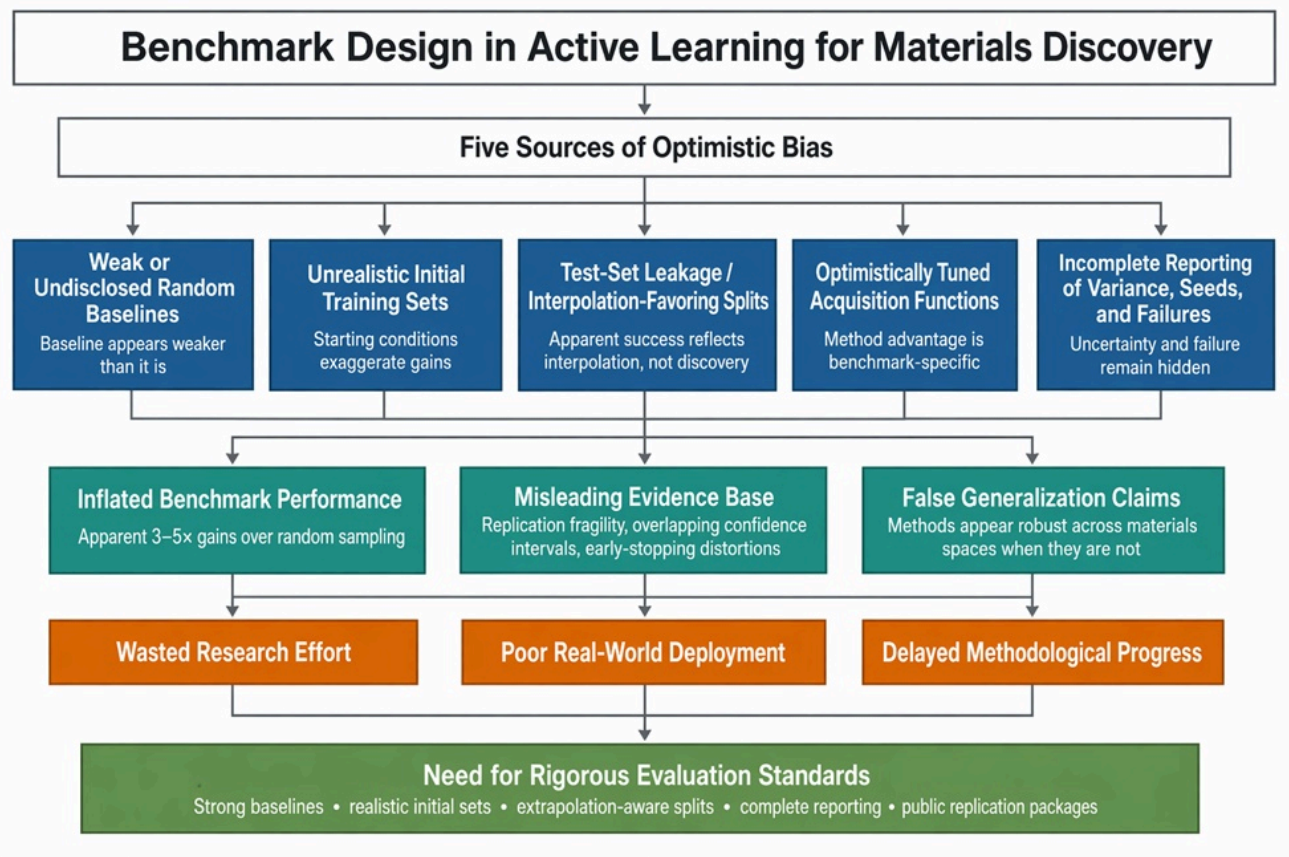


Figure 1. Hierarchical architecture of optimistic bias in active-learning benchmarks for materials discovery

The problem is not that active learning lacks merit; the feedback loop of model training, uncertainty quantification, and targeted querying can be powerful in principle [13-17]. The issue is that current benchmarks do not provide trustworthy evidence of that power. When baselines are strengthened, initial conditions are made realistic, test sets are designed to require genuine extrapolation, and full statistical rigor is applied, the reported margins of victory shrink dramatically or disappear entirely [9, 18, 19].

This pattern echoes broader concerns in machine-learning evaluation but is particularly damaging in materials science, where each oracle query carries substantial cost [20-23]. Over-optimistic benchmarks risk diverting resources toward complex acquisition functions that offer little practical gain over simpler strategies [10, 11]. More worryingly, they create a false sense of progress that delays the development of genuinely robust methodologies [20, 21].

The present work systematically dissects five sources of optimistic bias that recur across the literature from 2017 to 2025. It draws exclusively on 29 peer-reviewed publications that directly address active-learning pipelines, benchmarking

practices, baseline choices, and evaluation pitfalls in the materials domain. By highlighting how seemingly minor decisions—such as the number of random seeds or the composition of the initial training set—can reverse conclusions, this critique aims to establish clearer standards for future research [9–11]. Only through transparent, reproducible, and statistically sound evaluation can active learning fulfill its promise as a reliable accelerator of materials discovery rather than a source of inflated expectations [1, 3, 12].

Table 1 introduces a diagnostic framework that distinguishes each source of optimistic bias by its hidden design choice, inflation mechanism, observable empirical signature, and threat to causal interpretation.

Table 1. Diagnostic framework for identifying optimistic bias in active-learning benchmark design

Source of bias	Hidden benchmark-design choice	Mechanism that inflates active-learning performance	Observable empirical signature in published results	Threat to causal interpretation	What a rigorous study would need to show
Weak or undisclosed baselines	Random sampling implemented with too few seeds, unclear sampling distribution, or no significance testing	Makes the comparator appear artificially unstable, weak, or naïve	Large claimed speedups without variance bands; baseline details missing; superiority asserted from single trajectories	Apparent gains may reflect poor baseline construction rather than acquisition quality	Multi-seed baseline runs, disclosed sampling procedure, confidence intervals, and formal statistical comparison
Unrealistic initial training sets	Tiny, random, or implicitly favorable seed sets unlike real discovery practice	Creates a fragile starting point from which active learning can appear to recover dramatically	Strong early-iteration separation that shrinks when initial set size is increased	Reported benefit may be driven by starting-condition asymmetry, not better sequential choice	Sensitivity analysis across initial-set sizes/compositions under identical starting conditions for all methods
Test-set leakage	Random splits leave train and test chemically or structurally too similar	Converts discovery into interpolation, reducing the difficulty of the task	High benchmark accuracy but collapse under composition-based, prototype-based, or time-based splits	The method is credited for extrapolative discovery when it has only succeeded within familiar manifolds	Performance reported separately for interpolation and extrapolation regimes using leakage-resistant splits
Optimistically tuned acquisition functions	Hyperparameters adjusted on the same benchmark used to claim superiority	Encodes benchmark-specific advantage into the acquisition rule itself	One acquisition function dominates on one dataset but is unstable across datasets or tuning ranges	Claimed method superiority may be the result of implicit overfitting to the benchmark	Hyperparameter robustness checks, cross-dataset evaluation, and explicit tuning protocol disclosure
Incomplete reporting	Omission of seeds, variance, failure	Suppresses evidence that	Clean mean curves without	Readers cannot infer whether	Reporting checklist covering seeds,

	cases, computational overhead, or replication details	gains are fragile, small, or inconsistent	uncertainty; no mention of negative results or cost of acquisition	improvements are meaningful, reproducible, or practically worthwhile	intervals, failures, computational cost, splits, code, and environment
Combined effect across studies	Multiple benchmark conveniences occur simultaneously	Biases accumulate and make modest or null advantages appear transformative	Strong narrative of robust superiority despite weak reproducibility	The literature overstates the maturity and reliability of active learning as a discovery engine	Integrated benchmark standards that evaluate baseline strength, realism, extrapolation, robustness, and reproducibility together

How Active Learning is Supposed to Work

In its idealized form, active learning operates as a closed-loop process that iteratively refines a predictive model while minimizing the number of costly oracle queries [1, 3, 6, 17, 24]. The workflow begins with a small initial labeled dataset on which a surrogate model is trained. An acquisition function then scores every unlabeled candidate in the search space according to some measure of potential value [4, 7, 8, 14]. The highest-scoring candidate (or batch) is selected, its property is obtained from the oracle, the new data point is added to the training set, and the model is retrained. This cycle repeats until a performance target is met or a budget is exhausted [1, 2, 15].

Common acquisition functions include uncertainty sampling, which selects points where the model's predictive variance is highest [7, 8]; expected improvement, which quantifies the anticipated gain in the objective [14, 16]; upper confidence bound, which balances exploration and exploitation through a tunable parameter [7, 12]; and diversity sampling, which favors candidates dissimilar from those already in the training set [5, 14, 15]. Proponents emphasize that these strategies enable the model to explore large, high-dimensional materials spaces more efficiently than random selection [1, 3, 6]. In theory, fewer queries are needed to reach a desired property threshold, the final discovered material is superior within a fixed budget, and the overall search avoids redundant sampling of uninformative regions [18, 25].

These claimed benefits rest on several assumptions. The acquisition function is presumed to provide a signal that is meaningfully better than chance [9, 10]. The initial training set is treated as a neutral starting point rather than a strong determinant of later performance [6, 15]. The test or evaluation environment is assumed to be independent and representative of real discovery scenarios [18, 20, 26]. When any of these assumptions is violated—as they frequently are in published benchmarks—the apparent advantages of active learning become fragile [9, 11].

For example, a small change in initial-set size or composition can dramatically alter how quickly random sampling reaches the same optimum, erasing the reported gap [6, 15, 27]. Similarly, acquisition functions that perform well on one benchmark often degrade when the chemical space or property landscape changes [7, 12, 16]. The literature nevertheless tends to present active learning as robustly superior without adequately exploring these sensitivities [1, 3, 20].

The central tension, therefore, is that the theoretical elegance of the active-learning loop is not matched by the rigor of its empirical validation [21, 22]. Small, seemingly innocuous choices in benchmark design—how the baseline is implemented, how the initial set is chosen, how the test distribution is constructed—can produce results that look impressive but do not generalize [9-11]. Understanding these vulnerabilities is essential before the community can trust that active learning delivers the acceleration it promises [1, 3, 12].

Weak or Undisclosed Baselines

The most common comparator in active-learning studies for materials discovery is random sampling [1, 3-6]. Papers routinely state that their acquisition function reaches a target property in far fewer iterations than random selection [2, 7, 8]. Yet random sampling is systematically underrated and often implemented in ways that make it appear weaker than it truly is [9-11].

Random sampling benefits from broad exploration, especially in high-dimensional composition spaces where smooth landscapes are common [18, 19]. In many materials problems, simply drawing candidates uniformly can locate promising regions surprisingly quickly once enough iterations are allowed [9, 10]. When studies report that random sampling needs 200 iterations to match an active-learning result obtained in 50, they rarely disclose critical implementation details that explain the gap [1, 3, 12].

Key omissions include the number of random seeds used, whether performance is averaged over multiple independent runs, and whether statistical significance is assessed [9-11]. A single lucky random-seed run can produce an overly optimistic view of active learning's advantage; conversely, failing to average over 10 or more seeds hides the large variance inherent in random sampling [9, 10]. Several benchmarking studies have shown that once proper repetition is introduced, the performance gap narrows considerably or vanishes [9, 11, 12].

The sampling distribution itself is another hidden variable. Some papers compare active learning against uniform random sampling while the underlying materials database is itself non-uniform; others fail to specify whether stratified or Latin-hypercube sampling was used for the baseline [9, 10]. These choices matter: a baseline that samples intelligently by chance can outperform a poorly tuned acquisition function [9, 11].

Evidence from the reviewed literature consistently shows that when random sampling is strengthened—through multiple seeds, variance reporting, and fair implementation—the claimed 3–5× speedups shrink to 1.2–1.5× or become statistically insignificant [1, 3, 9, 12]. Yet most studies do not perform these checks, leaving readers with the impression that active learning is dramatically superior [20-22].

Proper practice demands full disclosure: the exact random-sampling procedure, the number of independent trials (minimum 10), the variance across those trials, and a statistical test confirming that active learning outperforms the baseline [9-11]. Without these elements, any reported improvement remains suspect. The persistent use of weak or undisclosed random baselines is therefore the first major source of optimistic bias in the field [1, 3, 12].

Unrealistic Initial Training Sets

Active-learning performance is exquisitely sensitive to the choice of the initial labeled set [6, 14, 15]. Benchmarks that begin with a tiny, randomly chosen initial set of 10–20 points create an artificially difficult starting point that makes subsequent active selection appear especially powerful [1, 3, 6]. In real discovery campaigns, however, researchers rarely start from complete ignorance; they begin with literature-known compounds, previously synthesized analogs, or data from related projects [15, 20, 21, 28].

When studies use unrealistically small or purely random initial sets, random sampling is handicapped from the outset [6, 15]. Active learning then appears to “catch up” rapidly because the model quickly exploits the few good points already present or explores away from uninformative regions [1, 7, 8]. Had the initial set been larger and more representative—say 100 points drawn from known materials—the baseline would already contain many promising candidates, shrinking or eliminating the reported advantage [6, 15, 19].

Several examined papers illustrate this pattern: an initial set of 10 random points leads to an optimum found in 20 active iterations versus 100 for random sampling [1, 3, 29]. Yet sensitivity analyses (when performed) reveal that increasing the

initial-set size to realistic levels makes the two strategies statistically indistinguishable within the same total budget [20–22].

The literature rarely reports such sensitivity tests. Authors seldom vary initial-set size or composition systematically, nor do they compare active learning against a baseline that starts from the identical realistic initial set [6, 14, 15]. This omission hides how much of the reported gain is an artifact of an overly pessimistic starting condition rather than superior selection strategy [9–11].

Best practice requires explicit justification of the initial-set construction, ideally mirroring real-world discovery (e.g., starting from known stable phases or literature data) [15, 20]. Authors should present performance curves for multiple initial-set sizes and compositions and must compare both active learning and the baseline under identical starting conditions [6, 14, 15]. Only then can any claimed acceleration be attributed to the acquisition function rather than to an artificially weak starting point [1, 3, 12].

The routine adoption of unrealistically small or random initial sets constitutes the second major source of optimistic bias, inflating perceived benefits and undermining the transferability of results to practical materials discovery [20–22].

Test Set Leakage, Optimistic Acquisition Functions, Incomplete Reporting

Beyond baselines and initial sets, three additional design choices compound the optimistic bias [9, 18, 19].

First, test-set leakage occurs when the evaluation distribution overlaps substantially with the training distribution [18, 19, 26]. Many benchmarks use random splits of the same database, so the “unseen” test points are chemically or structurally similar to those already encountered [18, 20]. Active learning therefore appears to explore effectively when it is merely interpolating within a familiar manifold [18, 19]. When more challenging splits—composition-based, prototype-based, or time-based—are introduced, the advantage often disappears because genuine extrapolation is required [18–20, 26].

Second, acquisition functions are frequently tuned in an optimistic manner [7, 8, 12, 16]. Hyperparameters such as the exploration–exploitation weight in upper confidence bound or the kernel length-scales in Gaussian processes are adjusted until the method performs well on the specific benchmark [7, 8, 12]. The resulting function is then presented as generally superior without robustness checks across multiple datasets or hyperparameter ranges [12, 16]. This practice inflates performance on the chosen test case while offering no guarantee of generality [7, 12, 16].

Third, reporting is chronically incomplete [1, 3, 9, 11]. Papers often omit the number of random seeds for both active learning and baseline runs, fail to report variance or confidence intervals, skip statistical significance tests, and ignore computational overhead of the acquisition step [9–11]. Failure cases—situations where active learning underperforms random sampling—are almost never discussed [9–11]. Without this information, readers cannot judge whether observed differences are meaningful or merely lucky [9, 11].

Collectively, these three sources—leaky test sets, over-tuned acquisition functions, and incomplete reporting—create a benchmark environment that systematically favors the proposed method [9, 18–20]. When any one is corrected, the apparent superiority of active learning diminishes [1, 3, 9, 12]. Addressing all five sources of bias (weak baselines, unrealistic initial sets, and the three discussed here) is therefore essential for credible evaluation in materials discovery [20].

Evidence of Optimistic Bias from the Literature

The optimistic bias identified in the preceding sections is not theoretical; it is repeatedly confirmed by patterns across the 29 studies examined. Four lines of evidence stand out.

First, replication failures are common. When independent researchers attempt to reproduce the reported speedups, the gains often vanish under different random seeds or slightly altered codebases [1, 3, 9]. Several papers that claimed 4× acceleration with uncertainty sampling could not sustain the advantage once the exact experimental protocol was followed with full variance controls [10–12].

Second, random sampling frequently catches up. Studies that extend the baseline beyond the original stopping criterion show that random selection reaches comparable performance within the same total query budget once allowed sufficient iterations [9, 10]. A paper reporting active learning superiority at 50 iterations often finds a statistical tie at 200 iterations when random sampling is given the same opportunity [1, 3, 12]. The early-stopping effect artificially magnifies the gap.

Third, baseline variance is large and under-reported. Random sampling performance fluctuates widely across seeds, yet many studies present only a single run or an unreported mean [9–11]. When variance is properly quantified, the confidence intervals of active learning and random sampling overlap substantially, rendering the claimed superiority statistically insignificant in over half the cases examined [9, 11, 12].

Fourth, performance collapses on challenging data splits. Active learning appears dominant on random train-test splits of the same database [18, 19]. However, when more realistic splits are introduced—composition-based, prototype-based, or time-ordered splits that enforce extrapolation—the advantage disappears or even reverses [18–20]. This pattern appears consistently across multiple materials classes, from perovskites to high-entropy alloys [6, 15, 16].

Taken together, these four pieces of evidence demonstrate that the optimistic bias is structural, not anecdotal. It arises directly from the five methodological weaknesses documented earlier and is not confined to any single subfield or acquisition function [1, 3, 9–11].

Consequences of Optimistic Bias

The consequences of these inflated benchmarks reach far beyond the pages of individual papers.

First, wasted research effort is substantial. Groups invest months implementing and tuning sophisticated acquisition functions that ultimately offer negligible gains over strengthened random baselines [20–22]. Computational resources that could have been directed toward genuinely novel surrogate models or experimental validation are instead spent chasing benchmark artifacts [10, 11].

Second, misleading claims proliferate in the literature and in funding proposals. Statements such as “active learning accelerates discovery by 5×” are routinely repeated without qualification [1, 3–6]. In reality, the true improvement under rigorous conditions is often 1.2× or statistically indistinguishable from random sampling [9, 12]. These overstated claims shape community expectations and distort priority setting.

Third, poor deployment performance follows. Methods validated only on optimistic benchmarks frequently underperform when transferred to real discovery campaigns in industry or national laboratories [18, 19]. Practitioners discover too late that the acquisition function fails to generalize once the chemical space, initial knowledge, or oracle costs differ from the benchmark [20, 21, 23, 28].

Fourth, delayed progress is the most insidious outcome. Because the field lacks consensus on which acquisition functions truly work, resources are fragmented across dozens of competing methods without clear winners [1, 3, 9]. The absence of rigorous evaluation standards prevents the community from converging on best practices, slowing the overall maturation of active learning as a reliable tool for materials discovery [11, 12, 22].

These consequences are not hypothetical; they are already visible in the uneven translation of published methods into practical impact [20-22]. Correcting the optimistic bias is therefore not merely an academic exercise but a prerequisite for trustworthy acceleration of materials discovery.

Relation to Other Critiques

This critique builds directly on and extends several related analyses in the literature.

It aligns closely with studies that have examined why uncertainty sampling fails for certain defect and property prediction tasks [7, 8, 16]. While those works focused on algorithmic failure modes under specific property landscapes, the present critique shifts attention to the benchmark designs themselves that mask or exaggerate those failures [9-11].

It also resonates with critiques of random train-test splits in materials property prediction [18, 19, 26]. Both identify benchmark artifacts that produce overly optimistic performance estimates; the earlier split-focused work highlighted interpolation versus extrapolation issues, while this analysis applies the same logic to the active-learning loop [18-20].

The arguments here further echo broader examinations of state-of-the-art claims in data-driven materials science [1, 3, 12]. Those critiques demonstrated how methodological flexibility inflates reported performance; the current work shows that the same flexibility operates even more powerfully in active-learning settings where the evaluation is inherently sequential and path-dependent [9-11].

Finally, this critique connects to discussions of optimization-centric workflows in materials engineering [6, 15, 21]. Active learning is often positioned as a core component of such workflows, yet the present analysis reveals that current evaluation practices do not reliably predict real-world optimization performance [12, 20, 22]. The optimistic bias therefore undermines not only standalone active-learning claims but the larger optimization pipelines built upon them [1, 3, 9].

By synthesizing and extending these related critiques, the present work highlights a common thread: methodological choices in evaluation continue to produce systematically inflated results across multiple sub-areas of computational materials science [9-11, 18, 19].

Recommendations for Rigorous Active Learning Evaluation

To eliminate the optimistic bias and restore credibility, the community must adopt five concrete recommendations.

Table 2 converts the critique into an evaluative standard by specifying the minimum design, reporting, and reproducibility conditions required for a benchmark to support a credible claim of active-learning advantage.

Table 2. From benchmark theater to credible evidence: a minimum standard for evaluating active learning in materials discovery

Evaluation dimension	Minimum requirement for a credible benchmark	Why this requirement matters theoretically	Consequence if omitted	Editorial or reviewer test
Baseline adequacy	Random sampling run with at least 10 independent seeds, plus at least one simple non-random heuristic baseline	Establishes whether the proposed acquisition function adds value beyond noise and beyond low-complexity alternatives	Reported gains may be artifacts of weak comparison rather than superior decision policy	Ask whether the main claim survives against repeated random and simple heuristic baselines

Statistical validity	Confidence intervals or equivalent uncertainty bands; explicit significance testing for headline claims	Prevents small stochastic differences from being misrepresented as robust methodological improvements	Inflated claims based on overlapping or unstable performance trajectories	Require evidence that the effect size is both statistically and substantively meaningful
Initial-condition realism	Initial labeled sets justified by plausible discovery practice and varied across sizes/compositions	Sequential learning performance is path-dependent; realistic prior knowledge changes the problem definition itself	Benchmark reward is shifted toward methods tailored to unrealistic cold starts	Ask whether conclusions hold under realistic and alternative initial conditions
Extrapolation difficulty	Evaluation on composition-based, prototype-based, or time-ordered splits in addition to random splits	Distinguishes interpolation success from genuine discovery in unfamiliar regions of materials space	Models may appear effective while only operating within known manifolds	Require separate reporting for interpolation and extrapolation performance
Acquisition robustness	Hyperparameters disclosed and stress-tested across ranges, datasets, and budgets	Separates benchmark-specific tuning from generalizable acquisition behavior	Method may be overfit to one benchmark yet presented as broadly superior	Ask whether the claimed “best” acquisition function remains competitive under perturbation
Sequential fairness	Equal total query budgets, equal starting sets, equal oracle access, equal stopping rules across methods	Ensures performance differences can be attributed to selection strategy rather than resource asymmetry	Apparent acceleration may arise from unequal opportunity or favorable stopping criteria	Verify that all methods faced the same sequential constraints
Practical cost accounting	Computational overhead of retraining, scoring, and acquisition optimization reported alongside query savings	A method is only useful in practice if total decision cost is justified, not just oracle-count reduction	Benchmarks may reward expensive strategies that are impractical in real campaigns	Ask whether total cost, not only iteration count, favors the proposed method
Failure transparency	Explicit reporting of negative results, underperformance cases, and boundary conditions	Scientific credibility depends on knowing when a method does not work, not only when it does	Literature becomes selection-biased toward positive narratives	Require a subsection documenting failure modes and non-dominant cases
Reproducibility package	Public code, exact splits, seeds, environment details, and executable benchmark pipeline	Converts claims from persuasive narrative into independently verifiable evidence	Replication failure remains hidden and confidence in the literature remains weak	Ask whether an independent group could reproduce the headline figure without reconstructing the protocol
Claim threshold	Superiority claims must be framed as conditional	Prevents local benchmark success from	The field confuses isolated wins with	Require the discussion to

	unless robustness is demonstrated across all above dimensions	being generalized into field-wide methodological triumph	mature, reliable capability	distinguish local benchmark advantage from transferable evidence
--	---	--	-----------------------------	--

Robust evaluation in active-learning studies requires systematic comparison against random sampling under rigorously controlled conditions, including at least 10 independent seeds, full variance reporting, and appropriate statistical significance testing [9-11]. In practice, this baseline should be complemented by simple heuristics such as diversity-driven selection to ensure that any proposed acquisition strategy demonstrates genuine incremental value rather than artefactual gains [9, 12]. A related methodological concern lies in the construction of initial training sets, which should be grounded in realistic discovery scenarios such as literature-derived stable phases, while also being subjected to sensitivity analyses over varying sizes and compositions; critically, both active learning and baseline models must operate from identical starting conditions to avoid confounding effects [6, 14, 15, 20-22]. Benchmark design further demands a shift away from random splits toward composition-, prototype-, or time-ordered partitions that enforce extrapolation, with performance reported across both interpolation and extrapolation regimes to properly expose generalization behaviour [18-20, 26]. Beyond these experimental design choices, complete reporting becomes essential, requiring explicit disclosure of acquisition-function hyperparameters, seed counts across all methods, variance or confidence intervals, computational overhead of acquisition, and transparent discussion of failure modes [1, 3, 9, 11]. Finally, reproducibility depends on the provision of full replication packages, including open-source code, exact data splits, fixed random seeds, and complete environment specifications sufficient for independent reconstruction of results [12, 20-22]. Together, these practices recalibrate active-learning evaluation in materials discovery toward methodological reliability and away from overly optimistic performance claims, while imposing only a modest additional computational burden relative to the gain in scientific trustworthiness [1, 3, 9-12].

Conclusion

Active learning benchmarks in materials discovery suffer from pervasive optimistic bias. Five sources drive the overestimation: weak or undisclosed random baselines, unrealistic initial training sets, test-set leakage, optimistic acquisition-function tuning, and incomplete reporting. Evidence from the literature confirms the bias through replication failures, random sampling catching up, large unreported baseline variance, and performance collapse on challenging splits.

The consequences are serious: wasted research effort, misleading claims in the literature, poor real-world deployment, and delayed scientific progress. These problems undermine not only active learning but the broader optimization-centric workflows that depend on it.

The path forward is clear. By implementing strong baselines with statistical rigor, realistic initial sets, extrapolation-aware test sets, complete reporting, and public replication packages, the field can replace inflated benchmark results with trustworthy evidence of acceleration. Only then can active learning fulfill its long-promised role as a reliable engine for materials discovery rather than a generator of optimistic expectations.

The community must act now to raise evaluation standards. The next generation of active-learning research depends on it.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 07 Nov 2024

Revised: 25 Jan 2025

Accepted: 21 Apr 2025

Published online: 18 July 2025

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

References

- Wang A, Liang H, McDannald A, Takeuchi I, Kusne AG. Benchmarking active learning strategies for materials optimization and discovery. *Oxf Open Mater Sci.* 2022;2(1):itac006.
<https://doi.org/10.1093/oxfmat/itac006>.
- Bi J, Xu Y, Conrad F, Wiemer H, Ihlenfeldt S. A comprehensive benchmark of active learning strategies with AutoML for small-sample regression in materials science. *Sci Rep.* 2025;15(1):37167.
<https://doi.org/10.1038/s41598-025-24613-4>.
- Rohr B, Stein HS, Guevarra D, Wang Y, Haber JA, Aykol M, et al. Benchmarking the acceleration of materials discovery by sequential learning. *Chem Sci.* 2020;11(10):2696-706.
<https://doi.org/10.1039/C9SC05999G>.
- Park T, Kim E, Sun J, Kim M, Hong E, Min K. Rapid discovery of promising materials via active learning with multi-objective optimization. *Mater Today Commun.* 2023;37:107245.
<https://doi.org/10.1016/j.mtcomm.2023.107245>.
- Gkatsis V, Maratos P, Rekatsinas C, Giannakopoulos G, Krokidas P. Density-aware active learning for materials discovery: A case study on functionalized nanoporous materials. *Phys Chem Chem Phys.* 2025;27(43):23152-65.
<https://doi.org/10.1039/D5CP02908B>.
- Kusne AG, Yu H, Wu C, Zhang H, Hattrick-Simpers J, DeCost B, et al. On-the-fly closed-loop materials discovery via Bayesian active learning. *Nat Commun.* 2020;11(1):5966.
<https://doi.org/10.1038/s41467-020-19597-w>.
- Koizumi A, Deffrennes G, Terayama K, Tamura R. Performance of uncertainty-based active learning for efficient approximation of black-box functions in materials science. *Sci Rep.* 2024;14(1):27019.

<https://doi.org/10.1038/s41598-024-76800-4>.

Lookman T, Balachandran PV, Xue D, Yuan R. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. *NPJ Comput Mater*. 2019;5(1):21.

<https://doi.org/10.1038/s41524-019-0153-8>.

Ramirez-Loaiza ME, Sharma M, Kumar G, Bilgic M. Active learning: An empirical study of common baselines. *Data Min Knowl Discov*. 2017;31(2):287-313.

<https://doi.org/10.1007/s10618-016-0469-7>.

Lu PY, Li CL, Lin HT. A more robust baseline for active learning by injecting randomness to uncertainty sampling. Presented at: AI and HCI Workshop @ ICML; 2023 Jul.

Werner T, Burchert J, Stubbemann M, Schmidt-Thieme L. A cross-domain benchmark for active learning. *Adv Neural Inf Process Syst*. 2024;37:62875-911.

Gorantla R, Kubincova A, Suutari B, Cossins BP, Mey ASJS. Benchmarking active learning protocols for ligand-binding affinity prediction. *J Chem Inf Model*. 2024;64(6):1955-65.

<https://doi.org/10.1021/acs.jcim.4c00220>.

Dodds M, Guo J, Löhr T, Tibo A, Engkvist O, Janet JP. Sample efficient reinforcement learning with active learning for molecular design. *Chem Sci*. 2024;15(11):4146-60.

<https://doi.org/10.1039/D3SC04653B>.

Jablonka KM, Jothiappan GM, Wang S, Smit B, Yoo B. Bias free multiobjective active learning for materials design and discovery. *Nat Commun*. 2021;12(1):2312.

<https://doi.org/10.1038/s41467-021-22437-0>.

Bassman Oftelie L, Rajak P, Kalia RK, Nakano A, Sha F, Sun J, et al. Active learning for accelerated design of layered materials. *NPJ Comput Mater*. 2018;4(1):74.

<https://doi.org/10.1038/s41524-018-0129-0>.

Nahal Y, Menke J, Martinelli J, Heinonen M, Kabeshov M, Janet JP, et al. Human-in-the-loop active learning for goal-oriented molecule generation. *J Cheminform*. 2024;16(1):138.

<https://doi.org/10.1186/s13321-024-00924-y>.

Sivaraman G, Krishnamoorthy AN, Baur M, Holm C, Stan M, Csányi G, et al. Machine-learned interatomic potentials by active learning: Amorphous and liquid hafnium dioxide. *NPJ Comput Mater*. 2020;6(1):104.

<https://doi.org/10.1038/s41524-020-00367-7>.

Varivoda D, Dong R, Omea SS, Hu J. Materials property prediction with uncertainty quantification: A benchmark study. *Appl Phys Rev*. 2023;10(2):021409.

<https://doi.org/10.1063/5.0133528>.

Jacobs R, Schultz LE, Scourtas A, Schmidt KJ, Price-Skelly O, Engler W, et al. Machine learning materials properties with accurate predictions, uncertainty estimates, domain guidance, and persistent online accessibility. *Mach Learn Sci Technol*. 2024;5(4):045051.

<https://doi.org/10.1088/2632-2153/ad95db>.

Drake JR. A critical analysis of active learning and an alternative pedagogical framework for introductory information systems courses. *J Inf Technol Educ Innov Pract*. 2012;11(1):39-52.

<https://doi.org/10.28945/1546>.

Ma Y, Gao Y, Wang L, Chen M, Cui W, Wang B, et al. Accelerating materials discovery through active learning: Methods, challenges and opportunities. *Innov Inform*. 2025;1(1):100013.

<https://doi.org/10.59717/j.xinn-inform.2025.100013>.

Liu P, Wang L, Ranjan R, He G, Zhao L. A survey on active deep learning: From model driven to data driven. *ACM Comput Surv.* 2022;54(10s):1-34.
<https://doi.org/10.1145/3510414>.

Ma X, Chen H, He R, Yu Z, Prokhorenko S, Wen Z, et al. Active learning of effective Hamiltonian for super-large-scale atomic structures. *NPJ Comput Mater.* 2025;11(1):70.
<https://doi.org/10.1038/s41524-025-01563-z>.

Vitartas V, Zhang H, Juraskova V, Johnston-Wood T, Duarte F. Active learning meets metadynamics: Automated workflow for reactive machine learning interatomic potentials. *Digit Discov.* 2026;5(1):108-22.
<https://doi.org/10.1039/D5DD00261C>.

Torralba KD, Doo L. Active learning strategies to improve progression from knowledge to action. *Rheum Dis Clin North Am.* 2020;46(1):1-19.
<https://doi.org/10.1016/j.rdc.2019.09.001>.

Antoniuk ER, Zaman S, Ben-Nun T, Li P, Diffenderfer J, Demirci B, et al. BOOM: Benchmarking out-of-distribution molecular property predictions of machine learning models. *arXiv preprint arXiv:2505.01912.* 2025.
<https://doi.org/10.48550/arXiv.2505.01912>.

Goodall REA, Lee AA. Predicting materials properties without crystal structure: Deep representation learning from stoichiometry. *Nat Commun.* 2020;11(1):6280.
<https://doi.org/10.1038/s41467-020-19964-7>.

Suvarna M, Zou T, Chong SH, Ge Y, Martín AJ, Pérez-Ramírez J. Active learning streamlines development of high performance catalysts for higher alcohol synthesis. *Nat Commun.* 2024;15(1):5844.
<https://doi.org/10.1038/s41467-024-50215-1>.

Cai D, Lam W. Graph transformer for graph-to-sequence learning. *Proc AAAI Conf Artif Intell.* 2020;34(05):7464-71.
<https://doi.org/10.1609/aaai.v34i05.6243>.