ORIGINAL RESEARCH **Open access**

Benchmarking Without Illusion: A Conceptual Critique of Performance Comparisons in Materials AI

Javier Ruiz[1*], Maria Gonzalez[1], Lucia Torres[2]

Abstract

In the rapidly evolving field of applied artificial intelligence (AI) for materials science, benchmarking serves as a cornerstone for evaluating model performance and guiding research trajectories. However, this paper advances a conceptual critique that unveils the inherent illusions embedded within conventional performance comparisons, which often obscure the nuanced realities of materials discovery and prediction. By synthesizing recent literature, we highlight how benchmarking practices can perpetuate misconceptions about model efficacy, generalizability, and alignment with real-world materials challenges. The critique centers on the interaction dynamics among data representations, evaluation metrics, and contextual factors, revealing feedback structures that amplify epistemic distortions. We propose a novel conceptual framework that reinterprets benchmarking as a multi-layered system of steering logics, in which trade-offs among precision, robustness, and interpretability shape the interpretive landscape of AI-driven insights into materials. This framework emphasizes systems-level insights into how illusory superiority emerges from mismatched expectations and overlooked interdependencies. Through analytical implications, we explore how recalibrating these dynamics could foster more transparent and ethically grounded performance assessments. Ultimately, the paper advocates for an integrative approach that prioritizes conceptual interpretations over superficial metrics, offering epistemic reasoning to navigate the complexities of materials AI without succumbing to benchmarking illusions. This conceptual reevaluation has the potential to refine the field's theoretical underpinnings, promoting advancements that are both innovative and reliable.

Keywords Materials AI, Interaction dynamics, Epistemic reasoning, Conceptual frameworks, Benchmarking illusions, Performance critique

*Correspondence:
Javier Ruiz
javier.ruiz@outlook.com
[1] Department of Materials Informatics and Data Modeling, Faculty of Engineering, University of Granada, Granada, Spain
[2] Department of AI Materials Simulation, Faculty of Engineering, University of Seville, Seville, Spain

Introduction

The integration of artificial intelligence into materials science has transformed the landscape of discovery, design, and optimization processes [1, 2]. Over the past decade, AI techniques have been increasingly applied to predict material properties, simulate behaviors, and accelerate innovation across energy storage, catalysis, and advanced composites. This surge is driven by the promise of overcoming traditional limitations in experimental throughput and computational expense, enabling researchers to explore vast chemical spaces with unprecedented efficiency. Yet, as the field matures, the mechanisms for assessing AI model performance—predominantly through benchmarking—have come under scrutiny for their potential to introduce subtle distortions in how progress is perceived and pursued [3-5].

Benchmarking, in essence, involves standardized comparisons of AI models against curated datasets using predefined metrics, aiming to provide objective measures of success. In materials AI, benchmarks often focus on tasks such as property prediction, structure generation, or inverse design, drawing on repositories that encompass diverse material classes. Recent developments have led to the establishment of comprehensive benchmark suites that aggregate multiple tasks, facilitating broader evaluations. However, these practices are not without conceptual challenges. The illusion of unequivocal progress arises when benchmarks fail to capture the intricate interplay between model assumptions and the heterogeneous nature of materials data, leading to interpretations that overemphasize quantitative gains while sidelining qualitative nuances [6, 7].

One key concern is the epistemic foundation of benchmarking. Materials science data is inherently noisy, sparse, and context-dependent, reflecting experimental variabilities, measurement uncertainties, and domain-specific constraints. When AI models are benchmarked, the selection of datasets can inadvertently privilege certain representations, creating an illusion of robustness that does not translate across different material systems. For instance, benchmarks dominated by crystalline structures may undervalue models suited for amorphous or defective materials, distorting the perceived utility of approaches. This dynamic underscores a feedback loop in which dataset curation influences model development, which in turn reinforces the benchmark's authority, potentially entrenching biases in the field's trajectory [8, 9].

Furthermore, performance metrics themselves contribute to these illusions. Common metrics such as mean absolute error or accuracy scores provide simplified views of model capability, but they often mask trade-offs in predictive fidelity across varying scales or conditions. In materials AI, where predictions inform high-stakes applications like sustainable energy solutions, such simplifications can lead to misguided priorities. The conceptual interpretation here reveals a steering logic: metrics guide research toward optimizable outcomes, yet they may diverge from the holistic goals of materials innovation, fostering an environment where short-term gains eclipse long-term interpretive depth [10, 11].

The literature illustrates these issues through various lenses. Studies have examined how benchmarking frameworks evolve, highlighting the need for more inclusive evaluations that account for uncertainty and transferability. Yet, a gap persists in conceptually dissecting the illusions that arise from these comparisons. This paper addresses this void by offering a purely theoretical critique, focusing on analytical implications rather than empirical validation. By interrogating systems-level insights into benchmarking, we aim to uncover how interactions among components—data, models, metrics, and community norms—generate epistemic distortions [12, 13].

At the heart of this critique is the recognition that benchmarking is not merely an evaluative tool but a constitutive element of the knowledge production process in materials AI. It shapes what is deemed valuable, influences funding allocations, and defines success narratives. However, when illusions permeate these processes, the field risks pursuing phantom advancements that dissolve under closer conceptual scrutiny. For example, the rapid adoption of graph neural networks in materials prediction has been extensively benchmarked, yet critiques suggest that performance gains may stem from data artifacts rather than intrinsic model strengths, illustrating a trade-off between sophistication and transparency [14, 15].

This introduction sets the stage for a deeper exploration. We first synthesize the theoretical background, drawing on recent scholarship to map the evolution of the literature. Subheadings will delineate key themes, from the foundations of benchmarking to emergent critiques. Subsequently, we propose a conceptual framework that integrates these elements, providing interpretive tools to navigate illusions. Through this lens, the paper contributes to a more reflexive understanding of performance comparisons, encouraging the materials AI community to embrace epistemic humility in its pursuits. The framework's emphasis on feedback structures and steering logics offers a pathway to reinterpret benchmarks not as definitive arbiters but as dialogic instruments that evolve with the field's conceptual maturity [16, 17]. To make the critique analytically tractable, the principal forms of benchmarking illusion, together with their generative mechanisms and practical manifestations, are synthesized in **Table 1**.

Table 1. Benchmarking “illusions” in materials AI: sources, mechanisms, and observable manifestations.

Illusion category (what appears true)	Primary driver (where it originates)	Mechanism (how the illusion forms)	What it looks like in papers/benchmarks	Systems-level consequence	What y framev revea
Illusory superiority reminds us that "Model A is better"	Dataset composition + metric narrowness	Performance uplift reflects distributional convenience rather than capability	Small MAE gains interpreted as broad progress; architecture claims exceed evidence	Convergent research toward benchmark-friendly solutions	Superio condition represent and m alignm
Illusory generalizability "works across materials"	Homogeneous or over-curated benchmarks	Narrow coverage masks regime shifts (defects, amorphous phases, extremes)	Cross-domain claims built on in-domain splits; weak stress tests	Epistemic gaps in underrepresented material classes	Generali is constr not infe witho contextua
Illusory robustness "stable under noise/variance"	Preprocessing conventions + train/test leakage risks	Sanitized pipelines suppress real experimental variability	Apparent stability disappears with small perturbations or alternate splits	Misleading readiness for deployment	Robust must assesse perturba aware, c sensit protoc
Illusory objectivity "benchmark is neutral"	Benchmark governance + community norms	Benchmark design choices become invisible defaults	Benchmark becomes "standard"; alternative tasks marginalized	Norm-setting power that defines "progress"	Benchn operat constitu episte infrastru
Illusory completeness "metric captures performance"	Overreliance on error metrics	Single-axis scoring collapses multi-objective material requirements	Error-only ranking substitutes for relevance, safety, and interpretability	Overspecification for easy-to-optimize targets	Evaluatio be multi- and pur align
Illusory applicability "ready for real-world use"	Lack of contextual embeddings (synthesis/processing/operation)	Missing real-world constraints inflate utility	Strong results on idealized data; failures in operational regimes	Misallocation of trust and resources	Applica requires conte modelin uncerta visibi

Illusory progress, "the field is advancing"	Feedback loops between benchmarks and model tuning	Benchmark chasing yields local improvements without conceptual breadth	Rapid leaderboard turnover; brittle rank changes	Innovation stagnation masked as momentum	Progress be evalua breac transfer epistemic —not r
Illusory equity "state-of-the-art is accessible"	Resource asymmetries (compute + private data)	Dominance by well-resourced groups becomes performance "truth"	Benchmark leadership correlates with infrastructure	Inequitable agenda-setting	Equity benchma variab governa must ac for struc imbala

In summary, the illusions in benchmarking stem from overlooked interdependencies that distort the interpretive landscape. By foregrounding these dynamics, this manuscript invites a reevaluation of how performance is conceptualized, paving the way for more robust theoretical foundations in applied AI for materials science. The ensuing sections will elaborate on these ideas, building toward a cohesive critique that enriches the discourse without resorting to empirical assertions [18].

Theoretical Background and Literature Synthesis

Foundations of benchmarking in materials AI

The advent of AI in materials science has necessitated robust mechanisms for performance assessment, with benchmarking emerging as a pivotal practice. Since 2020, the field has witnessed the development of standardized datasets and evaluation protocols designed to facilitate fair comparisons among diverse AI architectures. These foundations rest on the premise that quantifiable metrics can distill complex model behaviors into comparable indices, enabling researchers to identify promising directions. Literature from this period underscores the role of benchmarks in democratizing access to materials data, allowing for reproducible evaluations across institutions [19, 20].

Central to these foundations is the curation of datasets that represent the breadth of material phenomena. Efforts have focused on compiling high-quality repositories encompassing crystal structures, electronic properties, and thermodynamic stabilities, often leveraging high-throughput computations. The conceptual interpretation of these datasets reveals them as interpretive constructs rather than neutral artifacts, in which choices about inclusion criteria shape the benchmarking narrative. Interaction dynamics between data sources and AI models highlight how such foundations can inadvertently favor certain material classes, creating systems-level biases that propagate through performance comparisons [1, 21].

Moreover, selecting evaluation metrics is a core component of benchmarking. Metrics such as root-mean-square error and area under the curve are used to gauge predictive accuracy, but their application in materials contexts introduces trade-offs. For instance, metrics optimized for bulk properties may inadequately capture surface or interface behaviors, leading to epistemic mismatches. Recent syntheses emphasize the need for metric diversity to reflect the multifaceted nature of materials challenges. Yet, the literature reveals a persistent reliance on conventional measures that foster illusions of comprehensiveness [5, 22].

Evolution of performance comparisons

Performance comparisons in materials AI have evolved significantly, transitioning from ad hoc evaluations to structured benchmarking initiatives. This evolution is marked by the introduction of multi-task benchmarks that test models across

varied prediction scenarios, aiming to assess versatility. Scholarly works document this shift, noting how comparisons have become integral to validating novel algorithms, such as those based on transformers or equivariant networks [23, 24].

The conceptual critique in this evolution points to feedback structures in which comparative results influence subsequent model refinements. As models are tuned to excel on specific benchmarks, the field converges toward optimized but potentially narrow solutions. Systems-level insights suggest that this dynamic can create an illusion of collective progress, masking underlying limitations in generalizability. Literature syntheses highlight cases where performance rankings fluctuate with minor dataset perturbations, underscoring the fragility of comparisons [7, 25].

Ethical reasoning enters the discourse here, as the evolution of comparisons raises questions about equity in resource allocation. Institutions with access to superior computational infrastructure may dominate benchmarks, distorting the interpretive view of what constitutes state-of-the-art. This trade-off between accessibility and rigor invites a reevaluation of how comparisons steer the field's priorities, emphasizing the need for inclusive frameworks that mitigate such disparities [11, 26].

Foundations of benchmarking in materials AI

The increasing integration of artificial intelligence into materials science has intensified the need for systematic mechanisms of performance assessment, positioning benchmarking as a central epistemic practice. Since 2020, the field has seen the consolidation of standardized datasets and evaluation protocols to enable fair, reproducible comparisons across heterogeneous AI architectures. These foundations rest on the assumption that complex model behaviors can be rendered comparable through quantifiable metrics, thereby supporting claims of progress and guiding methodological choice. Contemporary literature emphasizes benchmarking as a means of democratizing access to materials data and evaluation pipelines, facilitating cross-institutional reproducibility and methodological transparency [19, 20].

At the core of these foundations lies dataset curation. Substantial efforts have been devoted to assembling high-quality repositories that capture diverse materials phenomena, including crystal structures, electronic properties, and thermodynamic stabilities, often derived from high-throughput computational workflows. Conceptually, however, such datasets function less as neutral repositories than as interpretive constructs: decisions regarding inclusion criteria, preprocessing, and representational scope actively shape the benchmarking landscape. Interaction dynamics between curated data and learning architectures reveal how these foundations can systematically privilege particular material classes or property regimes, embedding biases that propagate through downstream performance comparisons and influence perceived model superiority [1, 21].

Equally foundational is the selection of evaluation metrics. Common measures such as root-mean-square error or area under the curve are routinely used to quantify predictive accuracy, yet their application in materials contexts introduces nontrivial trade-offs. Metrics optimized for bulk properties, for example, may inadequately capture surface phenomena, interfacial effects, or rare-event behaviors, producing epistemic mismatches between evaluation criteria and scientific objectives. While recent syntheses advocate greater metric diversity to reflect the multidimensional nature of materials challenges, the literature continues to rely heavily on conventional measures, fostering an illusion of evaluative completeness that masks unresolved conceptual limitations [5, 22].

Evolution of performance comparisons

Performance comparisons in materials AI have evolved from largely ad hoc evaluations toward increasingly structured benchmarking initiatives. This transition is marked by the emergence of multi-task and multi-property benchmarks designed to assess model versatility across heterogeneous prediction scenarios. Scholarly accounts document how comparative performance has become integral to validating emerging architectures, including transformer-based models and symmetry-aware or equivariant networks [23, 24].

A conceptual critique of this evolution highlights feedback structures linking benchmark outcomes to subsequent model development. As architectures are progressively optimized to excel on established benchmarks, the field tends to converge toward narrowly tuned solutions. Systems-level analyses suggest that this dynamic can generate an illusion of cumulative progress, wherein incremental metric improvements obscure persistent limitations in robustness and generalizability. Evidence from comparative studies shows that performance rankings can shift substantially under minor dataset perturbations, underscoring the fragility of many benchmarking claims [7, 25].

Ethical considerations further complicate this evolution. Benchmark leadership increasingly correlates with access to large-scale computational resources, enabling well-resourced institutions to dominate comparative narratives. This asymmetry distorts collective perceptions of state-of-the-art performance and raises concerns regarding equity, inclusivity, and the allocation of scientific attention. The resulting trade-off between rigor and accessibility motivates calls for benchmarking frameworks that explicitly address structural disparities rather than implicitly reinforcing them [11, 26].

Critiques of benchmarking practices

Recent literature has increasingly challenged the uncritical authority attributed to benchmarking in materials AI. Since 2020, scholars have identified systematic biases in datasets, particularly the overrepresentation of well-studied or computationally convenient materials, which skew performance assessments and narrow the apparent scope of model competence. From this perspective, benchmarks function as epistemic instruments that, while operationally useful, can perpetuate illusions of objectivity when their underlying assumptions remain unexamined. Reliance on simulated or synthetic data further exacerbates this issue by introducing potential disconnects from experimental realities, leading to inflated expectations of real-world applicability [13, 27].

A parallel line of critique addresses metric misalignment. When evaluation measures fail to align with downstream scientific or engineering objectives, benchmarking can reinforce models that optimize for tractable yet scientifically shallow targets. Analytical accounts reveal feedback loops in which easily measurable attributes are privileged over holistic relevance, systematically marginalizing complex phenomena such as metastability, defect dynamics, or extreme-condition behavior. Systems-level reviews indicate that such practices may inadvertently suppress innovation in precisely those regimes where materials AI could offer the greatest epistemic value [15, 28].

Beyond technical considerations, the social dimensions of benchmarking have attracted growing scrutiny. Community norms surrounding leaderboard performance and comparative visibility exert pressure on researchers to prioritize benchmark dominance, sometimes at the expense of conceptual rigor or interpretability. These dynamics raise ethical concerns about knowledge production and dissemination, prompting integrative syntheses that call for steering logics balancing transparency, pluralism, and performance claims [17, 29].

Synthesis of conceptual gaps

Across foundational developments, evolutionary trajectories, and critical analyses, several conceptual gaps emerge. While benchmarking infrastructures provides a scaffold for comparison, they also embed implicit trade-offs and interpretive shortcuts that sustain epistemic illusions. The interaction dynamics among datasets, metrics, and models form a tightly coupled system in which biases and limitations are rarely isolated but often obscured. Systems-level perspectives suggest that addressing these gaps requires moving beyond isolated evaluations toward relational interpretations that explicitly acknowledge interdependencies and feedback effects [3, 9].

This synthesis points to the need to reframe benchmarking as a dialogic and reflexive process rather than a terminal validation mechanism. By integrating analytical and ethical perspectives, the literature converges on the need for conceptual frameworks that expose and interrogate the illusions embedded in comparative practices. The following section builds on this synthesis to articulate such a framework, positioning benchmarking as an evolving epistemic practice rather than a definitive arbiter of progress [20, 22].

Proposed conceptual framework

The proposed conceptual framework reimagines benchmarking in materials AI as a dynamic ecosystem characterized by interlocking components and emergent behaviors. At its core, the framework interprets performance comparisons through the lens of interaction dynamics, in which data curation, model architectures, metric selections, and contextual embeddings engage in continuous feedback loops. These dynamics generate steering logics that guide research trajectories, often amplifying illusions when misalignments occur. For instance, data curation acts as an entry point, influencing how models internalize material representations, while metrics provide evaluative feedback that reinforces or challenges these internalizations [4, 8].

Systems-level insights reveal the framework as a multi-layered construct: the foundational layer encompasses data and model interplays, the intermediary layer involves metric-driven interpretations, and the overarching layer incorporates epistemic and ethical considerations. Trade-offs manifest across layers, such as between computational efficiency and representational fidelity, shaping the overall interpretive landscape. This integrative approach avoids reductive views, instead emphasizing how illusions emerge from overlooked synergies, like the co-evolution of benchmarks and community expectations [12, 18].

Analytical implications extend to how the framework can recalibrate perceptions of progress. By foregrounding feedback structures, it highlights scenarios in which apparent performance advantages dissolve under contextual shifts, promoting a more nuanced understanding of model contributions. Ethical reasoning within this framework advocates transparency in these dynamics, urging the field to confront epistemic distortions arising from unexamined dependencies [23, 27]. The feedback-oriented benchmarking framework structure is shown in **Figure 1**.

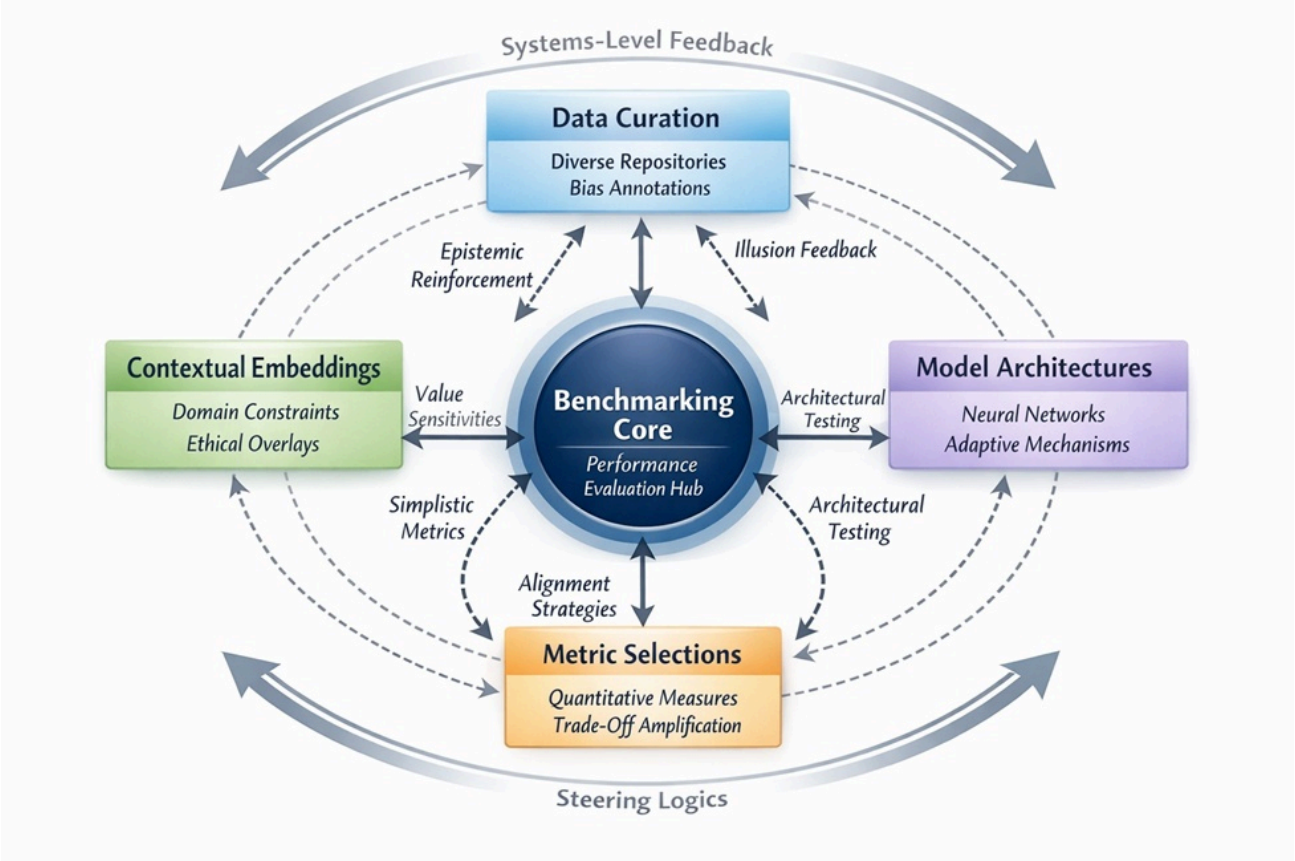

Figure 1. Cyclic benchmarking framework: Interconnected core with data, models, metrics, and contextual nodes.

This framework's value lies in its ability to offer conceptual interpretations that transcend traditional comparisons, fostering an environment in which AI advances in materials are evaluated holistically. By elucidating these elements, it provides tools for navigating illusions without prescriptive mandates [25, 29].

Building on **Figure 1**, **Table 2** maps the framework's layers to their interaction dynamics, dominant trade-offs, and practical levers for reducing benchmarking illusions.

Table 2. "Benchmarking as an ecosystem": framework layers, interaction dynamics, trade-offs, and mitigation levers

Framework layer	Core elements (what belongs here)	Key interaction dynamics	Dominant trade-offs	Typical failure mode (illusion generated)	"Without illusion" lever (conceptual mitigation)
Foundational layer: Data–Model coupling	Dataset scope, representational choices, splits, preprocessing, and architecture inductive biases	Data distributions select which inductive biases appear strong	Coverage vs cleanliness; realism vs standardization	Illusory generalizability/robustness from narrow, sanitized distributions	Stress tests across regimes; perturbation-aware splits; coverage reporting as epistemic metadata
Intermediary layer: Metric-driven interpretation	Error metrics, ranking rules, aggregation strategies, multi-task scoring	Metrics steer optimization and determine what "counts" as success	Accuracy vs interpretability; parsimony vs expressivity; single vs multi-objective scoring	Illusory completeness/superiority from single-axis ranking	Metric portfolios aligned to task purpose; uncertainty + calibration visibility; decision-relevant evaluation
Context layer: Deployment realism	Synthesis constraints, defects, process history, operating conditions, data provenance	Context shifts reconfigure what performance means	Controlled evaluation vs operational validity	Illusory applicability by ignoring real-world variability	Context embeddings; scenario-based evaluation; explicit domain shift characterization
Social layer: Community norms and incentives	Leaderboards, publication incentives, resource inequality, benchmark governance	Incentives create benchmark-chasing feedback loops	Speed vs rigor; spectacle vs transparency; openness vs competitive advantage	Illusory progress/equity via dominance and convergence	Governance norms (reporting standards, reproducibility); inclusivity criteria; compute-normalized comparisons
Overarching layer: Epistemic +	Claims discipline, uncertainty ethics, interpretive	How narratives about "SOTA"	Innovation vs reliability; openness vs	Illusory objectivity when benchmarks become unquestioned arbiters	Reflexive benchmarking: treat as

ethical steering	humility, accountability	shape funding, adoption, and risk	misuse; rigor vs accessibility		provisional; mandate limitations; align claims with epistemic warrant

Analytical implications

The proposed conceptual framework yields several analytical implications for interpreting performance comparisons in materials AI, particularly when examined through the lens of interaction dynamics and embedded trade-offs. By treating benchmarking as an ecosystem of interdependent components rather than a neutral evaluative instrument, the framework clarifies how data curation actively shapes perceptions of model robustness. When benchmark datasets disproportionately emphasize specific material attributes—such as electronic bandgaps in semiconductors—comparative outcomes may inflate apparent model superiority within those domains while obscuring deficiencies in others, including mechanical performance under extreme conditions. This imbalance generates a steering logic that channels research attention toward data-rich regimes, thereby reinforcing epistemic gaps in underrepresented material classes and phenomena [1, 3].

The framework's emphasis on feedback structures further exposes how metric selection can systematically distort interpretive conclusions. Evaluation schemes that prioritize predictive accuracy alone often marginalize dimensions such as interpretability, causal coherence, and physical plausibility, leading to trade-offs that favor opaque, highly optimized models over those that offer mechanistic insight. At a systems level, this suggests the need for recalibrated evaluative strategies that jointly consider multiple metrics. In the context of materials AI, such recalibration may involve complementing error-based measures with indicators of uncertainty propagation or robustness, thereby aligning benchmarking practices more closely with the probabilistic and heterogeneous nature of materials phenomena [6, 10].

Epistemic reasoning within the framework reveals an additional layer of implication rooted in community norms. Benchmark rankings increasingly function as symbolic markers of progress, creating reinforcing feedback loops in which models that perform well on established benchmarks attract disproportionate resources, attention, and validation. Over time, this dynamic can entrench dominant architectures while marginalizing alternative approaches that may offer complementary strengths but lack benchmark visibility. Conceptually, this underscores the importance of interpretive flexibility, framing benchmarks as provisional and context-dependent rather than definitive arbiters of merit. Such a stance introduces ethical considerations into performance narratives, particularly regarding whose contributions are amplified and whose are rendered peripheral [12, 15].

Interaction dynamics between model architectures and real-world contextual factors further suggest that benchmarking-related illusions may impede innovation in complex or hybrid materials systems. When evaluations fail to account for practical variabilities—such as synthesis imperfections, processing-induced heterogeneity, or environmental fluctuations—comparative analyses tend to overestimate model applicability. This misalignment is especially consequential in domains such as battery materials and photocatalysis, where operational conditions deviate substantially from the idealized conditions used in training data. Systems-level insights from the framework indicate that incorporating richer contextual embeddings into benchmarking protocols could improve the fidelity of performance comparisons, fostering interpretations that more effectively bridge theoretical prediction and practical utility [18, 22].

Trade-offs related to scalability constitute another critical analytical implication. As materials AI models increase in architectural complexity and computational demand, benchmarking practices must adapt accordingly to avoid introducing new distortions, including overfitting to benchmark-specific artifacts or privileging resource-intensive approaches by default. From an analytical standpoint, this highlights the need for dynamic evaluation regimes that evolve alongside technological advances, ensuring that performance comparisons remain meaningful rather than ossified representations of past capabilities [25, 29].

Collectively, these analytical implications argue for a shift away from narrow, metric-centric evaluations toward holistic assessment practices that integrate technical performance with epistemic robustness and ethical awareness. Interpreting benchmarking through this multifaceted lens enables researchers to more effectively identify and mitigate evaluative illusions, supporting the development of a materials AI ecosystem that values conceptual depth, contextual validity, and long-term scientific utility alongside computational efficiency.

Results and Discussion

The critique advanced in this manuscript invites a broader discussion on the role of benchmarking in shaping the trajectory of materials AI. At its core, the conceptual framework highlights how illusions in performance comparisons stem from unexamined interdependencies, prompting a reevaluation of how the field constructs knowledge. Interaction dynamics, for instance, reveal that benchmarking is not a neutral process but one embedded in social and institutional contexts, where choices in dataset and metric design reflect underlying priorities. This discussion extends to ethical reasoning, as illusions can lead to inequitable resource distribution, favoring well-resourced groups and potentially stifling diverse innovations [4, 11].

Systems-level insights further enrich the discourse by illustrating feedback structures that perpetuate certain paradigms. In materials science, where AI is applied to critical challenges like sustainable materials development, such structures risk aligning research with benchmark-optimized outcomes rather than real-world needs. Conceptual interpretations suggest that incorporating adaptive mechanisms, such as periodic benchmark revisions, could counteract this and foster a more responsive ecosystem. Trade-offs between generalizability and specificity also warrant discussion; while specialized benchmarks drive targeted advancements, they may fragment the field, complicating cross-domain applications [7, 13].

Moreover, the framework's emphasis on epistemic distortions opens avenues for interdisciplinary dialogue. Drawing parallels with other AI domains, such as computational biology, underscores shared challenges in performance evaluation, suggesting that lessons from materials AI could inform broader AI governance. Ethically, this implies a responsibility to transparently communicate benchmarking limitations, ensuring that stakeholders, from policymakers to industry practitioners, base decisions on robust interpretations rather than illusory metrics [17, 20].

In discussing these elements, the manuscript contributes to a reflexive turn in materials AI, where critiquing benchmarking practices becomes integral to progress. By prioritizing integrative approaches over competitive rankings, the field can cultivate a culture that values conceptual depth, ultimately enhancing the reliability and impact of AI-driven materials insights [23, 26].

Conclusion

This conceptual critique underscores the illusions inherent in benchmarking practices in materials AI and advocates a framework that interprets performance comparisons through interaction dynamics, feedback structures, and epistemic reasoning. By revealing trade-offs and offering systems-level insights, the paper encourages a shift toward more transparent, integrative evaluations, fostering advancements that align with the complex realities of materials science. Ultimately, embracing this perspective can steer the field away from superficial illusions, toward a more ethically grounded and conceptually rich future.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 15 Mar 2025 Revised: 10 Apr 2025 Accepted: 16 May 2025
Published online: 18 July 2025

Rights and permissions

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Choudhary K, Wines D, Li K, Garrity KF, Gupta V, Romero AH, et al. JARVIS-leaderboard: a large scale benchmark of materials design methods. arXiv preprint arXiv:2306.11688. 2024.

Merchant A, Batzner S, Schoenholz SS, Aykol M, Cheon G, Cubuk ED. Scaling deep learning for materials discovery. Nature. 2023;624(7990):80-5.
https://doi.org/10.1038/s41586-023-06735-9.

Dunn A, Dagdelen J, Boulais O, Lunha A, Poeppelmeier KR, Zunger A, et al. Benchmarking materials property prediction methods: the matbench test set and automatminer reference algorithm. npj Comput Mater. 2020;6(1):138.
https://doi.org/10.1038/s41524-020-00406-3.

Jain A. Machine learning in materials research: developments over the last decade and challenges for the future. Curr Opin Solid State Mater Sci. 2024;28(2):101142.
https://doi.org/10.1016/j.cossms.2024.101142.

Zheng YB, et al. Advancing materials science through next-generation machine learning. Curr Opin Solid State Mater Sci. 2024;28(2):101143.

Yu GL, et al. Machine learning models in phononic metamaterials. Curr Opin Solid State Mater Sci. 2024;28(2):101144.
https://doi.org/10.1016/j.cossms.2024.101144.

Peng P, Fang Z. Pushing the limits of multifunctional metasurface by deep learning. Curr Opin Solid State Mater Sci. 2024;28(2):101145.
https://doi.org/10.1016/j.cossms.2024.101145.

Rho J, et al. Mapping information and light: trends of ai-enabled metaphotonics. Curr Opin Solid State Mater Sci. 2024;28(2):101146.
https://doi.org/10.1016/j.cossms.2024.101146.

Fung V, Ganesh P, Sumpter BG. Benchmarking graph neural networks for materials chemistry. npj Comput Mater. 2021;7:84.
https://doi.org/10.1038/s41524-021-00552-0.

Choudhary K, DeCost B. Evolution of artificial intelligence for application in contemporary materials science. npj Comput Mater. 2022;8:73.
https://doi.org/10.1038/s41524-022-00734-6.

Mazurek J, et al. Ai benchmarking for science: efforts from the MLCommons science working group. In: International conference on supercomputing. Springer; 2023:1-12.
https://doi.org/10.1007/978-3-031-32041-5_1.

Abolhasani M, Brown KA. Role of ai in experimental materials science. MRS Bull. 2023;48(2):134-41.

Maqsood H. The future of material scientists in an age of artificial intelligence. Adv Sci. 2024;11(16):2401401.
https://doi.org/10.1002/advs.202401401.

Hügle M, Omoumi P, van Laar JM, Boedecker J, Hügle T. Applied machine learning and artificial intelligence in rheumatology. Rheumatol Adv Pract. 2020;4(1):rkaa005.

Qin C, Chen X, Wang C, Wu P, Chen X, Cheng Y, et al. SciHorizon: benchmarking ai-for-science readiness from scientific data to large language models. arXiv preprint arXiv:2503.13503. 2024.

Merchant A, et al. Artificial intelligence driving materials discovery? perspective on the article: scaling deep learning for materials discovery. Chem Mater. 2024;36(3):1021-3.
https://doi.org/10.1021/acs.chemmater.4c00643.

Poluektova VA, Poluektov MA. Artificial intelligence in materials science and modern concrete technologies: analysis of possibilities and prospects. Inorg Mater Appl Res. 2024;15(5):1187-98.

Hu B, et al. LLM4Mat-bench: benchmarking large language models for materials property prediction. Mach Learn Sci Technol. 2023;4(4):045027.
https://doi.org/10.1088/2632-2153/add3bb.

Choudhary K, Wines D, Li K, Garrity KF, Gupta V, Romero AH, et al. Artificial intelligence in materials science and engineering: current landscape, key challenges, and future trajectories. J Mater Process Technol. 2024;326:118370.
https://doi.org/10.1016/j.jmatprotec.2024.118370.

Goga AS. Integrating artificial intelligence in nanomaterials science: pathways to revolutionary materials discovery and design. Ethics and risks. In: International conference on innovative research; 2024.

Reeves-McLaren N, Christensen SM. Data integrity in materials science in the era of ai: balancing accelerated discovery with responsible science and innovation. J Mater Chem A. 2026;14(1):276-83.

Maqsood H, et al. Using artificial intelligence to accelerate materials development. MRS Bull. 2023;48:456-62.
https://doi.org/10.1557/s43577-023-00532-7.

Chávez-Angel E, et al. Application of artificial intelligence in the materials science, with a special focus on fuel cells and electrolyzers. Next Energy. 2024;4:100090.
https://doi.org/10.1016/j.nxener.2024.100090.

Jain A, et al. Artificial intelligence-powered materials science. Nano-Micro Lett. 2024;16:1634. https://doi.org/10.1007/s40820-024-01634-8.

Wassel AR, El-Sawy ER, El-Mahalawy AM. Unveiling of novel synthesized coumarin derivative for efficient self-driven hybrid organic/inorganic photodetector applications. Mater Today Sustain. 2024;26:100737.

Fung V, et al. Materials researchers put machine-learning performance to the test. Chem Eng News. 2021;99(13).

Thiyagalingam J, von Laszewski G, Yin J, Emani M, Papay J, Barrett G, et al. Ai benchmarking for science: efforts from the MLCommons science working group. In: Anzt H, Bienz A, Luszczek P, Baboulin M, eds. High performance computing. ISC high performance 2022 international workshops. Springer; 2022:1-12. https://doi.org/10.1007/978-3-031-23220-6_4.

Baranwal BS, Jones LO, Maind A, Balande UT, Khandare DG, Goyal HZ, et al. The AI revolution in chemistry: shaping the future of materials and biomedical sciences. ChemRxiv. 2025 Aug 25. https://doi.org/10.26434/chemrxiv-2025-8lrf0.

Choudhary K, et al. Current state and benchmarking of generative artificial intelligence for additive manufacturing. In: ASME 2024 international design engineering technical conferences and computers and information in engineering conference; 2024.