REVIEW Open access

Benchmarking Practices in Materials Artificial Intelligence — What Is Measured and What Is Missed

Fatima Zahra Amrani[1*], Youssef Benali[1], Samira El-Haddad[2]

Abstract

Materials artificial intelligence (MAI) has revolutionized the discovery, design, and optimization of new materials by leveraging machine learning algorithms to analyze complex datasets and predict properties with high accuracy. However, the rapid proliferation of MAI tools has raised critical questions about benchmarking practices, which are essential for evaluating model performance, ensuring reproducibility, and addressing ethical concerns. This narrative review examines current benchmarking frameworks in MAI, highlighting what is effectively measured—such as predictive accuracy and computational efficiency—and what is often overlooked —such as data bias, interpretability, fairness, and ethical implications. Drawing on recent advances in frameworks such as JARVIS-Leaderboard and Matbench, the review discusses challenges in data quality, reproducibility, and the integration of explainable AI (XAI) methods. It also explores active learning strategies for optimizing materials discovery under limited data conditions and proposes directions for more inclusive and transparent benchmarking. By synthesizing insights from diverse studies, this review aims to guide future MAI research toward robust, equitable, and ethically sound practices that accelerate innovation while mitigating risks.

Keywords Interpretability, Data bias, Materials artificial intelligence, Benchmarking frameworks, Reproducibility, Ethical AI

*Correspondence:
Fatima Zahra Amrani
fatima.amrani@gmail.com
[1] Department of Computational Materials Science, Faculty of Engineering, Mohammed V University, Rabat, Morocco
[2] Department of Artificial Intelligence in Engineering Systems, Faculty of Engineering, University of Fez, Fez, Morocco

Introduction

The integration of artificial intelligence (AI) into materials science—commonly referred to as materials artificial intelligence (MAI) or materials informatics—has fundamentally reshaped how materials are discovered, designed, and optimized. By synergistically combining computational modeling, machine learning (ML) algorithms, and increasingly large and diverse materials databases, MAI enables rapid prediction of material properties, acceleration of synthesis pathways, and identification of previously unexplored compounds with tailored functionalities [1, 2]. This data-driven paradigm represents a departure from traditional trial-and-error approaches, which are often time-consuming, costly, and ill-suited for navigating the vast chemical and structural spaces inherent to modern materials research.

The growing adoption of MAI is motivated in large part by pressing global challenges, including the development of sustainable energy technologies, next-generation electronic and photonic devices, and advanced biomedical materials [3, 4]. In these domains, the ability to efficiently screen millions of candidate materials and rapidly converge on optimal solutions is critical. AI-driven frameworks have already demonstrated remarkable success, such as high-throughput screening of stable perovskite compositions for photovoltaic applications and the discovery of high-performance alloys

with enhanced strength–ductility trade-offs, achieving reductions in experimental cost and time by orders of magnitude [5, 6]. These successes underscore the transformative potential of MAI as a core methodology in materials science.

Despite this progress, significant concerns remain regarding the reliability, reproducibility, and comparability of MAI models, largely stemming from inconsistencies in benchmarking practices [7]. Benchmarking plays a central role in evaluating model performance by providing standardized datasets, evaluation protocols, and performance metrics that allow meaningful comparison across methods [8]. However, current benchmarking efforts are often narrowly focused on aggregate quantitative metrics—such as mean absolute error (MAE), root mean squared error (RMSE), or area under the receiver operating characteristic curve (AUC)—while neglecting critical qualitative dimensions. These include dataset bias, uncertainty quantification, model interpretability, reproducibility of results, and broader ethical considerations related to fairness and transparency [9, 10]. As a consequence, benchmarks may incentivize incremental improvements in numerical scores without adequately reflecting real-world applicability, potentially leading to overoptimistic claims, systematic biases in materials predictions, and barriers to cumulative scientific progress [11, 12].

In this context, the objectives of this review are threefold. First, we provide a comprehensive overview of existing benchmarking frameworks in MAI, with particular emphasis on widely adopted platforms such as Matbench and the JARVIS-Leaderboard. Second, we critically assess what these benchmarks capture effectively—and, equally important, what they fail to measure—focusing on issues of data quality, generalization, interpretability, and ethical robustness. Third, we propose pathways toward more holistic benchmarking practices that integrate active learning, explainable artificial intelligence (XAI), and fairness-aware evaluation criteria. Drawing on peer-reviewed literature published, this review highlights the urgent need for benchmarking paradigms that better reflect the interdisciplinary, data-heterogeneous, and application-driven nature of MAI [13, 14].

Benchmarking frameworks in MAI: Foundations and evolution

Benchmarking in MAI involves using standardized datasets, task definitions, and evaluation protocols to systematically compare the performance of ML models on core materials science problems, including property prediction, structure optimization, phase classification, and stability assessment [15]. Early benchmarking efforts in the field were typically narrow in scope, focusing on single properties such as formation energy, band gap, or elastic constants. While valuable, these benchmarks provided limited insight into model transferability or robustness across diverse materials classes and application contexts [16].

More recent benchmarking frameworks have evolved toward multi-task, multi-property, and multi-fidelity settings, reflecting the increasing complexity of MAI workflows. The Matbench benchmark suite exemplifies this shift by offering a curated collection of 13 tasks derived from experimentally and computationally relevant datasets, spanning mechanical, thermal, and electronic properties [14]. By standardizing train–test splits and evaluation metrics, Matbench enables reproducible comparison of a wide range of ML models, from traditional descriptor-based regressors to state-of-the-art graph neural networks (GNNs). Results reported on Matbench have consistently shown that GNN-based architectures outperform feature-engineered approaches on large datasets, emphasizing the critical role of explicit structural representations in learning materials–property relationships [17].

Complementing Matbench, the JARVIS-Leaderboard provides a community-driven benchmarking ecosystem that spans multiple methodological domains, including ML models, electronic structure calculations, and interatomic force fields [1]. A defining feature of the JARVIS-Leaderboard is its strong emphasis on reproducibility and transparency: contributors are required to submit not only model predictions but also executable code, detailed metadata, and links to peer-reviewed digital object identifiers (DOIs) [3]. This approach addresses a long-standing limitation of earlier benchmarks, in which reported performance was difficult to reproduce or verify independently. Across these platforms, commonly reported metrics include MAE, RMSE, classification accuracy, and F1 scores, providing quantitative assessments of predictive performance [18]. Notably, the integration of active learning strategies within the JARVIS framework has demonstrated

substantial gains in data efficiency, with reports of up to seven-fold reductions in the number of required calculations for effective materials optimization [19].

Table 1 synthesizes the dominant evaluation dimensions captured by current MAI benchmarking platforms and contrasts them with critical epistemic, ethical, and practical dimensions that remain largely unmeasured.

Table 1. What current MAI benchmarks measure vs. what they miss

Benchmarking dimension	Commonly measured in MAI benchmarks	Typically missed or under-measured	Implications for MAI practice
Predictive performance	MAE, RMSE, accuracy, F1 score on fixed train–test splits	Sensitivity to distribution shift, extrapolative failure modes	Overstates real-world reliability, especially under experimental or industrial deployment
Computational efficiency	Training time, inference speed, scalability	Energy cost, carbon footprint, and hardware accessibility	Favors resource-intensive models, obscuring sustainability trade-offs
Dataset consistency	Standardized datasets (e.g., DFT-derived repositories)	Experimental noise, synthesis constraints, negative results	Limits transferability from computational to experimental settings
Reproducibility	Fixed splits, partial code availability	Environment dependence, stochastic variability, workflow provenance	Weakens independent verification and cumulative science
Interpretability	Optional post-hoc feature attribution	Explanation fidelity, physical plausibility, stability	Restricts scientific insight and mechanistic understanding
Bias and fairness	Aggregate performance over dominant data regimes	Performance disparity across chemistries, elements, or sustainability classes	Reinforces historical research biases
Ethical robustness	Rarely formalized	Resource equity, environmental impact, and downstream societal effects	Disconnects benchmarking from responsible innovation goals

Nevertheless, while existing benchmarks excel at evaluating predictive accuracy under controlled conditions, they often fall short in capturing broader aspects of model utility and real-world relevance. For example, many Matbench tasks rely heavily on density functional theory (DFT)–derived datasets, which, although internally consistent, may not adequately represent experimental noise, synthesis constraints, or measurement variability [20]. This reliance raises concerns about the generalizability of benchmarked models when deployed in experimental or industrial settings. Consequently, there is a growing recognition of the need for next-generation benchmarks that incorporate heterogeneous data sources—including experimental measurements, negative results, and uncertainty annotations—to more rigorously assess model robustness and practical applicability [21].

Data bias and fairness in MAI datasets

Data bias constitutes a critical yet insufficiently quantified challenge in the benchmarking of materials artificial intelligence (MAI) models. Most widely used MAI datasets—such as those derived from the Materials Project, Open Quantum Materials Database (OQMD), and similar high-throughput repositories—are inherently skewed toward chemically simple systems, frequently studied elements, and thermodynamically stable phases [9]. This imbalance arises from both computational convenience and historical research priorities, resulting in systematic underrepresentation of rare

elements, complex chemistries, metastable phases, and synthesis-constrained materials. As a consequence, ML models trained on these datasets often exhibit strong performance on well-represented material classes but perform poorly when extrapolated to underexplored regions of the materials space [22].

Such biases do not remain confined to datasets; they are amplified by the learning process itself. Models optimized for aggregate accuracy metrics tend to prioritize dominant data regimes, reinforcing existing inequities in materials discovery and limiting the exploration of unconventional or high-risk–high-reward candidates. For example, although bias mitigation strategies—such as synthetic data generation, reweighting schemes, and transfer learning—have been proposed to address class imbalance, their effectiveness within MAI pipelines has rarely been evaluated under standardized benchmarking conditions [9]. Without rigorous assessment, these techniques risk introducing artificial correlations or obscuring physically meaningful trends, further complicating model interpretation and deployment.

Fairness in MAI extends beyond statistical balance to encompass broader ethical and societal considerations. In practical applications, MAI models may implicitly favor materials optimized for specific industrial infrastructures, geographic regions, or resource availability, thereby marginalizing alternatives that are more sustainable, abundant, or economically accessible [23]. While platforms such as the JARVIS-Leaderboard have begun to incorporate diversity-related metrics, systematic and comparative evaluations of fairness across MAI benchmarks remain sparse [1]. Recent studies have demonstrated that biased training data can lead to discriminatory outcomes in alloy and catalyst design, where models disproportionately recommend high-performance yet resource-intensive or environmentally burdensome materials, sidelining lower-cost or greener options [24].

Active learning offers a promising pathway to mitigate these effects by explicitly prioritizing diversity-aware sampling strategies. By selecting new data points that maximize coverage of underrepresented regions in composition or structure space, active learning has been shown to reduce bias and improve predictive robustness in small-sample regression tasks relevant to MAI [25]. However, such strategies are not yet routinely embedded into benchmarking frameworks, limiting their broader adoption.

A major gap in current MAI benchmarks is the absence of standardized methodologies for quantifying and reporting bias. Metrics originally developed in general AI research—such as demographic parity, equalized odds, or performance disparity across predefined subgroups—could be adapted to materials contexts to assess equitable model behavior across compositional, structural, or application-specific categories [17]. Integrating such metrics into MAI benchmarks would represent an important step toward more responsible and inclusive materials discovery.

Reproducibility challenges in MAI

Reproducibility is a foundational principle of scientific research, yet it remains a persistent challenge in MAI benchmarking. Many published MAI studies do not provide complete access to source code, detailed documentation of dependencies, or precise records of software versions and computational environments, making independent replication difficult or impossible [3]. Even minor discrepancies in library versions, data preprocessing steps, or training protocols can lead to substantial variation in reported performance, undermining confidence in benchmark results.

These issues are particularly acute in MAI workflows that combine multiple computational stages, such as density functional theory (DFT) calculations, feature extraction, ML model training, and post hoc validation. A notable case study examining a widely used MAI tool revealed substantial obstacles related to code modularity, software versioning, and undocumented assumptions, highlighting the need for structured, sequential workflows and explicit cross-referencing between manuscripts and code repositories [3]. Such challenges are exacerbated by the stochastic nature of many ML algorithms, where random initialization, data shuffling, and hyperparameter optimization can significantly influence outcomes [26].

Benchmarking initiatives such as Matbench have taken important steps toward addressing these concerns by mandating open-source code submission and standardized train–test splits [14]. Similarly, the JARVIS-Leaderboard enforces peer-reviewed contributions, metadata reporting, and continuous integration testing, thereby promoting transparency and long-term maintainability [1]. Despite these advances, reproducibility gaps persist, particularly for complex, multi-fidelity pipelines that integrate computational predictions with experimental validation [27]. In such cases, differences in experimental protocols, measurement uncertainty, and data curation practices further complicate reproducibility.

What remains largely missing from current MAI benchmarks is a unified set of reproducibility guidelines tailored to the unique challenges of materials research. Recent recommendations emphasize the inclusion of detailed checklists covering data provenance, preprocessing steps, computational environments, random seed control, and statistical significance testing [28]. Embedding these requirements directly into benchmarking frameworks would not only improve reproducibility but also elevate the credibility and long-term impact of MAI studies.

Interpretability and explainable AI in MAI

The increasing reliance on complex ML architectures—such as deep neural networks and graph-based models—has intensified concerns about interpretability in MAI. While these models often achieve superior predictive accuracy, their "black-box" nature limits trust and adoption, particularly among domain scientists who seek mechanistic understanding rather than purely empirical correlations [6]. Interpretability is especially critical in materials science, where model insights can inform theory development, guide experimental design, and reveal previously unknown physical relationships.

Explainable artificial intelligence (XAI) techniques have emerged as powerful tools to address this challenge. Methods such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) have been applied to MAI models to quantify feature importance and uncover relationships between compositional or structural descriptors and target properties [7]. For example, the XElemNet framework integrates XAI with deep learning to attribute predicted material properties to specific elemental contributions, yielding explanations that are consistent with established chemical intuition and physical principles [8].

Despite these successes, interpretability remains peripheral in most MAI benchmarks. While platforms like Matbench occasionally include XAI-based analyses, these evaluations are neither standardized nor required [14]. This omission limits the ability to compare models not only on predictive accuracy but also on their capacity to generate scientifically meaningful explanations. In applied materials design, XAI has demonstrated value in elucidating atomic-scale interactions in complex systems such as multicomponent alloys, thereby supporting rational design strategies [29].

A particularly underexplored area is the integration of XAI into active learning workflows. In principle, model explanations could guide data acquisition by identifying regions of high uncertainty or poorly understood mechanisms, leading to more efficient and interpretable learning cycles [30]. However, current benchmarks rarely assess such synergistic approaches.

Future MAI benchmarking efforts should therefore move beyond accuracy-centric evaluation and explicitly incorporate interpretability metrics, including measures of explanation faithfulness, stability, and plausibility [6, 7]. By doing so, benchmarks can better align model development with the dual goals of predictive performance and scientific insight, ultimately fostering greater trust and adoption of MAI methodologies.

Ethical considerations in MAI

Ethical considerations in materials artificial intelligence extend beyond technical performance to encompass fairness, accountability, transparency, and broader societal impact [11]. As MAI models increasingly influence research priorities and downstream industrial decision-making, ethical shortcomings at the benchmarking stage can propagate into real-world materials development pipelines. In particular, biased MAI models may inadvertently reinforce existing inequities in materials access by systematically favoring rare, geopolitically sensitive, or environmentally intensive elements over more

abundant and sustainable alternatives [9]. Such biases are not merely technical artifacts but have tangible implications for resource distribution, supply chain resilience, and environmental justice.

Although fairness-oriented benchmarks have begun to emerge in the broader AI community, their adoption within MAI remains limited [31]. Current MAI benchmarks rarely evaluate whether predictive performance is equitably distributed across material classes, elemental groups, or application domains. As a result, models that perform exceptionally well on narrowly defined, high-value datasets may still exhibit ethically problematic behavior when deployed in contexts where sustainability, cost, or accessibility are primary concerns.

Accountability represents another underdeveloped dimension of ethical MAI benchmarking. Responsible AI practice requires transparent disclosure of how AI models are trained, validated, and applied, including explicit acknowledgment of limitations, uncertainties, and potential negative impacts [32]. This need is particularly acute in the context of generative AI approaches to materials discovery, which raise concerns regarding data provenance, intellectual property rights, and the inadvertent generation of non-physical or irreproducible material candidates [12]. Yet, existing benchmarks seldom require documentation of ethical risks or downstream consequences associated with model deployment.

A notable omission in current MAI benchmarking practices is stakeholder engagement. Ethical evaluation is often confined to technical metrics defined by model developers, without input from communities potentially affected by AI-driven materials innovation, such as manufacturing sectors, environmental regulators, or regions dependent on specific material resources [13]. Incorporating participatory approaches—where diverse stakeholders help define ethical priorities and evaluation criteria—could substantially enhance the societal relevance and legitimacy of MAI benchmarks.

To advance ethical benchmarking, MAI frameworks could draw on normative ethical theories, such as Rawlsian principles of justice, which emphasize fairness and equitable distribution of benefits [28]. Operationalizing such principles may involve explicitly rewarding models that balance performance with sustainability, accessibility, and long-term societal value. Embedding these considerations into benchmark design would mark a critical step toward aligning MAI development with responsible innovation goals.

Active learning and optimization in MAI

Active learning (AL) has emerged as a powerful strategy for addressing data scarcity and high labeling costs in MAI by iteratively selecting the most informative samples for computation or experimentation [25]. In contrast to passive learning paradigms, AL enables models to prioritize data points that maximize expected information gain, thereby improving predictive performance with significantly fewer labeled samples. Benchmark studies in small-sample regression have consistently shown that uncertainty-based acquisition strategies outperform random sampling, particularly in regimes relevant to materials discovery where data generation is expensive [25].

In practical materials optimization workflows, AL is increasingly combined with automated machine learning (AutoML) and high-throughput experimentation to accelerate discovery cycles. Notable successes include the application of AL-driven optimization in electrospray processing and catalyst design, where iterative model–experiment loops have rapidly converged on optimal synthesis conditions [33]. These advances highlight AL's potential to transform MAI from a purely predictive tool into a closed-loop discovery engine.

Benchmarking frameworks such as Matbench and the JARVIS-Leaderboard have begun to support evaluations of AL strategies, enabling comparative analysis of acquisition functions and model architectures [1, 14]. However, significant challenges remain, particularly in benchmarking AL for multi-objective optimization problems common in materials science, where trade-offs between properties such as performance, stability, cost, and sustainability must be considered [15]. Current benchmarks often reduce these problems to single-objective formulations, limiting their realism and practical relevance.

An important aspect largely missing from existing AL benchmarks is the explicit consideration of ethical and environmental constraints. For example, AL strategies could be evaluated not only on their efficiency in improving predictive accuracy but also on their ability to minimize environmental impact, reduce reliance on hazardous synthesis routes, or favor sustainable material choices [23]. Incorporating such constraints into AL benchmarking would better align methodological innovation with responsible materials development.

Case studies: Matbench and JARVIS-Leaderboard

Matbench and the JARVIS-Leaderboard represent two of the most influential benchmarking platforms in MAI, offering complementary perspectives on model evaluation and comparison. Matbench has become a standard reference for property prediction tasks, providing well-curated datasets and standardized evaluation protocols that have revealed the strong performance of graph neural networks on structure-sensitive problems [14]. Its emphasis on reproducibility and fair model comparison has significantly raised the methodological rigor of MAI research.

The JARVIS-Leaderboard extends this benchmarking philosophy by encompassing a broader range of computational approaches, including ML models, electronic structure methods, and interatomic potentials, within a unified evaluation ecosystem [1]. By enforcing requirements for code availability, metadata reporting, and peer-reviewed contributions, JARVIS promotes transparency and long-term reproducibility across the MAI community.

Despite their strengths, these case studies also illustrate the limitations of current benchmarking paradigms. Both platforms primarily emphasize predictive accuracy and computational efficiency, while providing limited mechanisms for auditing data bias, evaluating fairness, or assessing ethical implications [9]. As such, they exemplify the broader trend in MAI benchmarking: strong performance measurement in narrow technical dimensions, coupled with insufficient attention to societal and ethical considerations.

Collectively, these case studies underscore the need for next-generation MAI benchmarks that integrate accuracy, robustness, interpretability, fairness, and ethical responsibility into a unified evaluation framework. Addressing these gaps will be essential for ensuring that MAI advances not only scientific discovery but also broader societal goals.

Results and Discussion

The benchmarking practices in materials artificial intelligence (MAI) reveal a landscape in which quantitative metrics dominate evaluations, while critical qualitative dimensions remain underexplored [1, 5]. Frameworks such as Matbench and JARVIS-Leaderboard have significantly advanced the field by providing standardized datasets and protocols that facilitate direct comparisons of ML models [1, 14]. These tools effectively measure predictive accuracy, computational efficiency, and scalability, enabling researchers to identify superior algorithms for tasks like property prediction and structure generation [15, 16]. For instance, integrating active learning into these benchmarks has demonstrated substantial reductions in data requirements, thereby addressing the high cost of materials experiments [25, 32]. However, the emphasis on error metrics such as MAE and RMSE often masks underlying issues in model reliability, including sensitivity to data perturbations or domain shifts [18,19].

A key discrepancy lies in how data bias and fairness are handled. While benchmarks quantify performance on curated datasets, they rarely incorporate audits for representational bias, which can skew predictions toward well-studied materials like silicon-based semiconductors at the expense of emerging biomaterials or sustainable composites [9, 17]. This omission perpetuates a cycle in which AI models reinforce existing research priorities, potentially delaying innovation in underrepresented areas [22, 23]. Synthetic data generation offers a promising avenue for bias mitigation, but its benchmarking in MAI is nascent, with limited evaluations of how augmented datasets affect long-term model fairness [9, 30]. Furthermore, fairness metrics derived from general AI, such as equal opportunity in classification tasks, could be incorporated into MAI benchmarks to ensure equitable outcomes across material classes [13, 17].

Reproducibility emerges as another area where benchmarks measure procedural compliance but miss deeper systemic challenges [3, 20]. Platforms like JARVIS-Leaderboard mandate code sharing, which improves immediate replicability [1]. Yet, variability in computational environments, such as differences in DFT software versions or hardware configurations, often leads to non-reproducible results [21, 23]. This is compounded by the lack of standardized reporting guidelines in MAI, unlike those in cheminformatics or biomedical AI [20, 24]. For example, studies on computational pathology highlight the value of comprehensive checklists for algorithm reusability, a practice that MAI could adopt to enhance trust [22]. What is missed here is the integration of statistical robustness tests, such as bootstrapping or cross-validation across multiple seeds, to quantify reproducibility under uncertainty [26, 28].

Figure 1 conceptually summarizes how current MAI benchmarking practices emphasize narrow performance metrics while systematically overlooking epistemic, ethical, and societal dimensions.

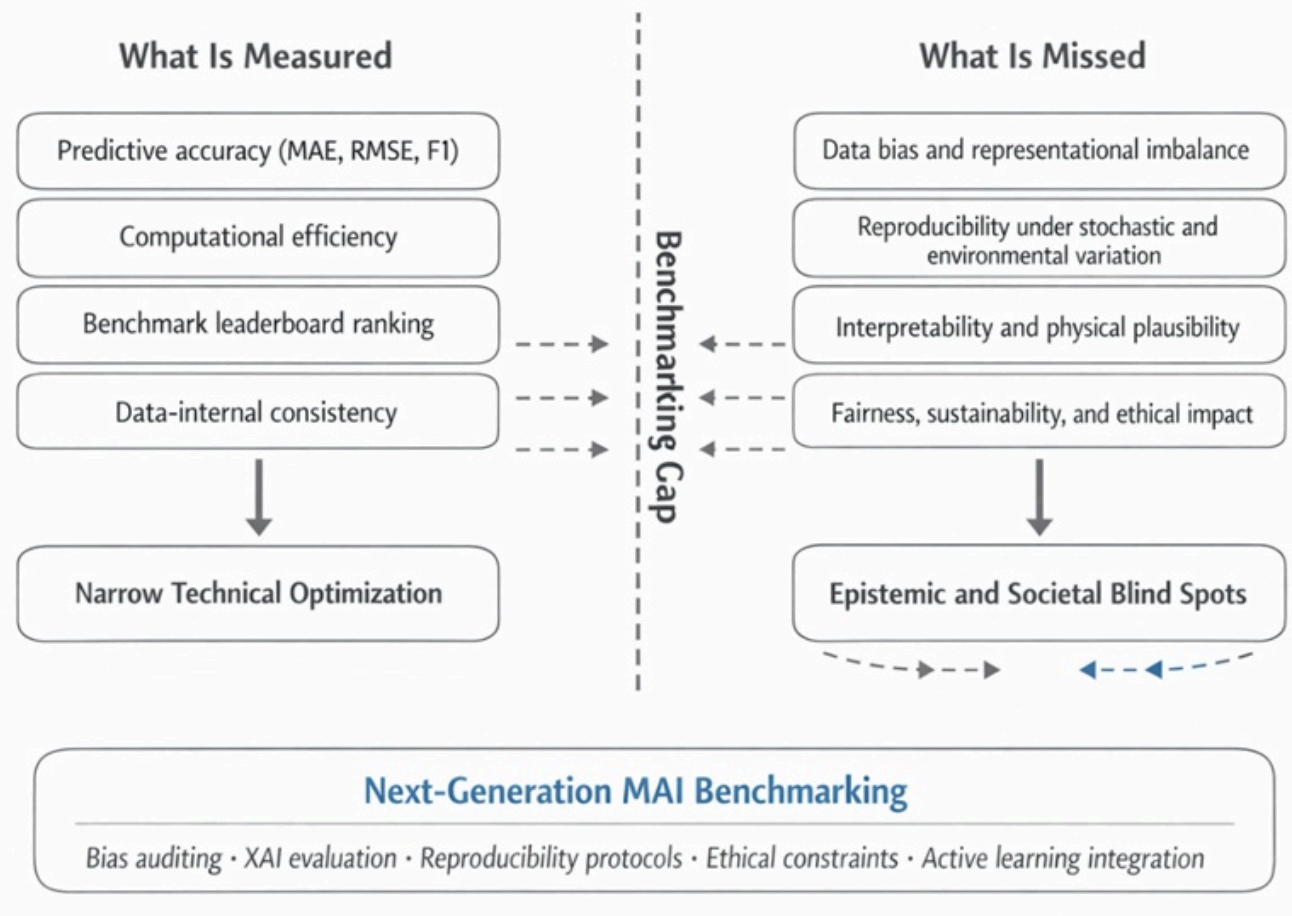

Figure 1. Conceptual framework of benchmarking in materials AI: what is measured and what is missed

Interpretability and explainable AI (XAI) represent a frontier where benchmarking practices are evolving but remain incomplete [6, 7]. Current evaluations often include post-hoc explanations that reveal how models weigh features such as atomic radii or electronic configurations [8, 31]. However, benchmarks seldom assess the fidelity of these explanations to physical principles, leading to potential misinterpretations in high-stakes applications like alloy design [25, 29]. Human-AI interaction studies underscore that interpretability is subjective, with domain experts valuing plausible explanations over purely mathematical ones [25]. Integrating XAI into active learning cycles could bridge this gap, allowing benchmarks to measure not just accuracy but also explanatory utility [30, 32].

Ethical considerations in MAI benchmarking are perhaps the most overlooked, despite growing awareness in broader AI domains [11, 12]. While frameworks measure technical performance, they rarely assess societal impacts, such as the environmental footprint of AI-driven material synthesis or equitable access to the technologies discovered [27, 28]. Principles from Rawlsian ethics advocate for benchmarks that prioritize the least advantaged stakeholders, yet MAI practices lack such mechanisms [28]. For instance, AI models in materials science could inadvertently favor proprietary datasets, exacerbating global disparities in research capabilities [4, 10]. Addressing this requires interdisciplinary collaboration and the incorporation of ethical audits into benchmark designs to ensure alignment with values such as transparency and accountability [12, 13].

Active learning and optimization strategies highlight both strengths and gaps in MAI benchmarks [15, 25]. These methods are well-measured for efficiency gains in data-scarce scenarios, as seen in graph-based relaxation tasks [32]. However, benchmarks miss evaluations under ethical constraints, such as optimizing for low-toxicity materials while maintaining performance [23, 27]. Case studies from Matbench and JARVIS illustrate that while predictive tasks are robustly evaluated, multi-objective optimizations that involve trade-offs among accuracy, cost, and ethics are underexplored [1, 14].

Overall, the discussion reveals that MAI benchmarks excel at measuring technical prowess but fall short in holistic assessments that incorporate bias, reproducibility, interpretability, and ethics. This imbalance risks overhyping AI capabilities while ignoring potential harms, underscoring the need for evolved benchmarking paradigms [2, 4].

Conclusion

In conclusion, benchmarking practices in materials artificial intelligence have made commendable strides in quantifying model performance and fostering community-driven evaluations. What is effectively measured includes predictive metrics and efficiency in frameworks like Matbench and JARVIS-Leaderboard, which have accelerated materials discovery. However, significant oversights persist in areas such as data bias, reproducibility, interpretability, and ethical fairness, undermining the trustworthiness and societal value of MAI.

Future directions should prioritize the development of integrated benchmarks that embed bias audits, XAI evaluations, and ethical checklists. For instance, expanding active learning benchmarks to include multi-fidelity data and fairness constraints could enhance real-world applicability. Collaborative efforts, such as updating JARVIS with stakeholder-inclusive metrics, would promote equitable AI in materials science. Additionally, adopting reporting standards from adjacent fields could improve reproducibility. Ultimately, by addressing what is missed, MAI can evolve into a more robust, inclusive, and impactful discipline.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 03 Aug 2025 Revised: 23 Aug 2025 Accepted: 11 Oct 2025
Published online: 18 January 2026

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. npj Comput Mater. 2024;10:93.
https://doi.org/10.1038/s41524-024-01259-w.

Katsura Y, Akiyama M, Morito H, Fujioka M, Sugahara T. Systematic searches for new inorganic materials assisted by materials informatics. Sci Technol Adv Mater. 2024;26(1):2428154.
https://doi.org/10.1080/14686996.2024.2428154.

Persaud D, Ward L, Hattrick-Simpers J. Reproducibility in materials informatics: Lessons from ‘a general-purpose machine learning framework for predicting properties of inorganic materials’. Digit Discov. 2024;3(2):281-6.
https://doi.org/10.1039/D3DD00199G.

Vergara D, Lampropoulos G, Fernández-Arias P, Antón-Sancho Á. Artificial intelligence reinventing materials engineering: A bibliometric review. Appl Sci. 2024;14(18):8143.
https://doi.org/10.3390/app14188143.

Jain A. Machine learning in materials research: Developments over the last decade and challenges for the future. Curr Opin Solid State Mater Sci. 2024;33:101189.
https://doi.org/10.1016/j.cossms.2024.101189.

Zhong X, Gallagher B, Liu S, Kailkhura B, Hiszpanski A, Han TY-J. Explainable machine learning in materials science. npj Comput Mater. 2022;8:204.
https://doi.org/10.1038/s41524-022-00884-7.

Oviedo F, Lavista Ferres J, Buonassisi T, Butler KT. Interpretable and explainable machine learning for materials science and chemistry. Acc Mater Res. 2022;3(6):597-607.
https://doi.org/10.1021/accountsmr.1c00244.

Wang K, Gupta V, Lee CS, Mao Y, Kilic MNT, Li Y, et al. XElemNet: Towards explainable AI for deep neural networks in materials science. Sci Rep. 2024;14:25178.
https://doi.org/10.1038/s41598-024-76535-2.

Hameed MAS, Qureshi AM, Kaushik A. Bias mitigation via synthetic data generation: A review. Electronics. 2024;13(19):3909. https://doi.org/10.3390/electronics13193909.

Chikware AB, Roman NV, Davids EL. Improving health informatics competencies: A scoping review of the components of health informatics academic programs. Sage Open. 2024;14(4). https://doi.org/10.1177/21582440241293259.

Chen Z, Chen C, Yang G, He X, Chi X, Zeng Z, et al. Research integrity in the era of artificial intelligence: Challenges and responses. Medicine. 2024;103(27):e38811. https://doi.org/10.1097/MD.0000000000038811.

Al-kfairy M, Mustafa D, Kshetri N, Insiew M, Alfandi O. Ethical challenges and solutions of generative AI: An interdisciplinary perspective. Informatics. 2024;11(3):58. https://doi.org/10.3390/informatics11030058.

Neveditsin N, MacKinnon K, Al-Dabbagh M, Neveditsina A. Toward fairness, accountability, transparency, and ethics in AI for social media and health care: Scoping review. JMIR Med Inform. 2024;12:e50048. https://doi.org/10.2196/50048.

Dunn A, Wang Q, Ganose A, Dopp D, Jain A. Benchmarking materials property prediction methods: The Matbench test set and Automatminer reference algorithm. npj Comput Mater. 2020;6:138. https://doi.org/10.1038/s41524-020-00406-3.

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.

Lu Y, Wang H, Zhang L, Yu N, Shi S, Su H. Unleashing the power of AI in science-key considerations for materials data preparation. Sci Data. 2024;11:1039. https://doi.org/10.1038/s41597-024-03821-z.

Ferrara E. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. Sci. 2024;6(1):3. https://doi.org/10.3390/sci6010003.

Yesil MR, Talli I, Pelloso M, Cosma C, Pangrazzi E, Plebani M, et al. Impact of analytical bias on machine learning models for sepsis prediction using laboratory data. Clin Chem Lab Med. 2025;63(10):2022-30. https://doi.org/10.1515/cclm-2025-0491.

Butler KT, Choudhary K, Csanyi G, Ganose AM, Kalinin SV, Morgan D. Setting standards for data driven materials science. npj Comput Mater. 2024;10:231. https://doi.org/10.1038/s41524-024-01411-6.

Hoyt CT, Zdrazil B, Guha R, Jeliazkova N, Martinez-Mayorga K, Nittinger E. Improving reproducibility and reusability in the Journal of Cheminformatics. J Cheminform. 2023;15:62. https://doi.org/10.1186/s13321-023-00730-y.

Fokkens L, Botelho MC, Sturm B, et al. Recommendations to enhance rigor and reproducibility in biomedical research. Gigascience. 2020;9(6):giaa056. https://doi.org/10.1093/gigascience/giaa056.

Wagner SJ, Matek C, Shetab Boushehri S, Boxberg M, Lamm L, Sadafi A, et al. Built to last? Reproducibility and reusability of deep learning algorithms in computational pathology. Mod Pathol. 2024;37(1):100350. https://doi.org/10.1016/j.modpat.2023.100350.

Lu J. After computational reproducibility: Scientific reproducibility and trustworthy AI. Harv Data Sci Rev. 2024;6(1). https://doi.org/10.1162/99608f92.ea5e6f9a.

Kolbinger FR, Veldhuizen GP, Zhu J, Truhn D, Kather JN, et al. Reporting guidelines in medical artificial intelligence: A systematic review and meta-analysis. Commun Med. 2024;4:71. https://doi.org/10.1038/s43856-024-00492-0.

Chu E, Roy D, Andreas J. Are visual explanations useful? A case study in model-in-the-loop prediction. arXiv preprint arXiv:2007.12248. https://doi.org/10.48550/arXiv.2007.12248.

Mersha M, Lam K, Wood J, AlShami AK, Kalita J. Explainable artificial intelligence: A survey of needs, techniques, applications, and future direction. Neurocomputing. 2024;599:128111. https://doi.org/10.1016/j.neucom.2024.128111.

Giarmoleo FV, Ferrero I, Rocchi M, Pellegrini MM. What ethics can say on artificial intelligence: Insights from a systematic literature review. Bus Soc Rev. 2024;129(2):258-92. https://doi.org/10.1111/basr.12336.

Westerstrand S. Reconstructing AI ethics principles: Rawlsian ethics of artificial intelligence. Sci Eng Ethics. 2024;30(5):46. https://doi.org/10.1007/s11948-024-00507-y.

Do V, Akehurst S, Boursier JM, et al. Ethical principles in machine learning and artificial intelligence: Cases from the field and possible ways forward. Humanit Soc Sci Commun. 2020;7:9. https://doi.org/10.1057/s41599-020-0501-9.

Li Y, Li J, Li B, Yuan T, Gong X. On the data quality and imbalance in machine learning-based design and manufacturing—A systematic review. Engineering. 2024. https://doi.org/10.1016/j.eng.2024.06.014.

Sarker IH. Explainable artificial intelligence (XAI)—From theory to methods and applications. IEEE Access. 2024;12:87791-815. https://doi.org/10.1109/ACCESS.2024.3416184.

Chen Y, Li Z, McComb C, Sun Y, Goodall JL, Gong X. Generalization of graph-based active learning relaxation strategies across materials. Mach Learn Sci Technol. 2024;5:025007.

Wang F, Harker A, Edirisinghe M, Parhizkar M. Tackling data scarcity challenge through active learning in materials processing with electrospray. Adv Intell Syst. 2024;6(7):2300798. https://doi.org/10.1002/aisy.202300798.