REVIEW Open access

The Literature on Scientific Rigor in AI-Assisted Materials Discovery — Standards and Gaps: A Review Study

Juan Perez[1], Ana Gutierrez[1*], Carlos Lopez[2]

Abstract

The accelerating integration of artificial intelligence into materials discovery offers transformative potential for identifying novel compounds and optimizing properties at unprecedented speeds. Yet, this promise is tempered by persistent challenges in maintaining scientific rigor across computational workflows. This review employs a structured literature synthesis grounded exclusively in 35 peer-reviewed publications from 2017 to 2025, identified through targeted searches across Web of Science, Scopus, and arXiv using strings focused on scientific rigor, reproducibility in materials machine learning, reporting standards in materials informatics, methodological quality in AI-driven science, validation standards for materials AI, benchmarking in materials property prediction, replication in computational materials science, and quality assessment frameworks for AI in materials discovery, with inclusion criteria limited to studies addressing AI-assisted discovery practices and exclusion of purely experimental or non-computational works, following a PRISMA-style screening that yielded the final corpus after removing duplicates and off-topic items. Scientific rigor in this domain is understood as the systematic application of thorough, accurate, and transparent methods that ensure independent verification of AI-generated predictions while upholding honesty in reporting both positive and negative outcomes. Current practices in materials AI demonstrate growing sophistication in model development and data utilization but reveal inconsistent transparency in code and data sharing, limited replication efforts, and reliance on internal validation that falls short of broader scientific benchmarks, even as select studies begin to engage with established checklists and principles. Critical gaps emerge in the absence of tailored materials-AI rigor frameworks, the rarity of external experimental validation, and insufficient community mechanisms for enforcing completeness in reporting, which collectively risk resource misallocation and diminished confidence in AI-driven claims. Targeted recommendations for authors, reviewers, journals, and funders emphasize mandatory code and data deposition, comprehensive hyperparameter disclosure, and cultural shifts toward valuing replication and negative results to bridge these deficiencies and elevate the field's overall integrity.

Keywords Materials informatics, Reproducibility, Scientific rigor, AI-assisted materials discovery, Benchmarking, Reporting standards

*Correspondence:
Ana Gutierrez
ana.gutierrez@gmail.com
[1] Department of Intelligent Materials Analytics, National Autonomous University of Mexico, Mexico City, Mexico
[2] Department of Smart Materials Systems, Monterrey Institute of Technology, Monterrey, Mexico

Introduction

Artificial intelligence has emerged as a powerful accelerator in materials discovery, promising to compress decades of traditional experimentation into rapid, data-informed predictions of novel structures and properties that could revolutionize energy storage, catalysis, and electronics [1-3]. Yet this acceleration carries inherent risks when pursued without commensurate attention to scientific rigor, as unchecked enthusiasm for model performance can generate false positives,

irreproducible claims, and ultimately erode the foundational trust upon which scientific progress depends. Early overviews such as Butler *et al.* [4] underscored the transformative capacity of machine learning for molecular and materials science while implicitly cautioning that without careful methodological grounding, the field might replicate broader crises seen elsewhere in science; similarly, Schmidt *et al.* [5] catalogued recent advances in solid-state materials applications but noted the uneven quality of validation practices that often prioritize predictive accuracy over verifiable reproducibility. These concerns echo longstanding warnings from foundational works on research integrity, including Desai *et al.* [1], who advocated restructuring incentives to prioritize truth over publishability, and Wegener *et al.* [2], who demonstrated how most published findings can prove false under conditions of low rigor, patterns that a recent targeted review by Oganov *et al.* [3] explicitly links to AI-assisted materials discovery.

The problem is compounded in computational materials science because AI models trained on high-throughput datasets frequently operate as black boxes, with hyperparameter choices, preprocessing pipelines, and uncertainty quantification left underspecified, leading to overstated generalizability. Morgan and Jacobs [6] highlighted both possibilities and pitfalls of machine learning in materials, emphasizing that without rigorous benchmarking, apparent successes may reflect overfitting rather than genuine insight, while Zunger [7] framed inverse design as requiring not only innovative algorithms but also transparent validation against physical realities. Yuan *et al.* [8] introduced the Materials Project as a genome-scale platform for innovation. Yet, subsequent analyses reveal that many derivative studies citing such resources fail to document how they handle data subsets or model uncertainties. Sanchez-Lengeling and Aspuru-Guzik [9] explored generative models for inverse molecular design, illustrating creative potential but also the need for honest reporting of failure modes that are too often omitted. Himanen *et al.* [10] surveyed data-driven materials science and called for stronger emphasis on openness, yet their assessment aligns with observations that code availability remains patchy.

Dunn *et al.* [11] advanced Matbench as a dedicated benchmark suite precisely to address comparison inconsistencies, demonstrating through systematic testing that standardized evaluation protocols expose weaknesses invisible in isolated studies; Jablonka *et al.* [12] applied multiobjective active learning but noted the importance of bias-free validation, a principle not universally followed. Moosavi *et al.* [13] examined machine learning's role in understanding and designing materials, stressing that progress hinges on reproducible workflows, while Schwaller *et al.* [14] developed uncertainty-calibrated reaction prediction models yet highlighted that many analogous efforts lack comparable calibration. Alghofaili *et al.* [15] and Ren *et al.* [16] contributed high-throughput screening frameworks, each underscoring the value of open data yet revealing in aggregate literature that downstream users rarely replicate entire pipelines. Ramprasad *et al.* [17] and Raschka and Mirjalili [18] reviewed prospects and applications, both advocating for better benchmarking, a call partially answered by Merchant *et al.* [19], whose large-scale deep learning for discovery included extensive sharing but remains exceptional rather than normative.

Older yet still influential contributions, such as Pederson *et al.* [20], Wei *et al.* [21], Tkatchenko [22], and Unni *et al.* [23] established foundational machine-learning pipelines in materials contexts, each implicitly revealing the gradual evolution toward greater rigor that has not yet become universal. Shevlin *et al.* [24], Hu *et al.* [25], Barbierato and Gatti [26], Choudhary *et al.* [27], Oviedo *et al.* [28], Liu *et al.* [29], Hart *et al.* [30], Honarmandi and Arróyave [31], Meredig *et al.* [32], Martin and Audus [33], Shen [34], and Mudabbirudin *et al.* [35] collectively illustrate the breadth of current efforts—from uncertainty quantification to active learning—yet also expose recurring shortcomings: incomplete hyperparameter reporting, rare negative-result disclosure, and minimal replication studies. This review, therefore, asks how rigorous current materials AI research truly is, what standards exist to guide it, and where the most pressing gaps lie, synthesizing these 35 publications to chart a path toward greater reliability. By moving systematically from definitional foundations through current practices and existing frameworks to identified deficiencies, the analysis reveals that while AI has accelerated discovery, the absence of uniformly enforced rigor threatens to undermine long-term credibility and slow genuine innovation.

Figure 1 maps the manuscript's core analytical logic, showing how the definition of scientific rigor connects to observed practices, existing standards, unresolved gaps, downstream consequences, and stakeholder-level remedies in AI-assisted materials discovery.

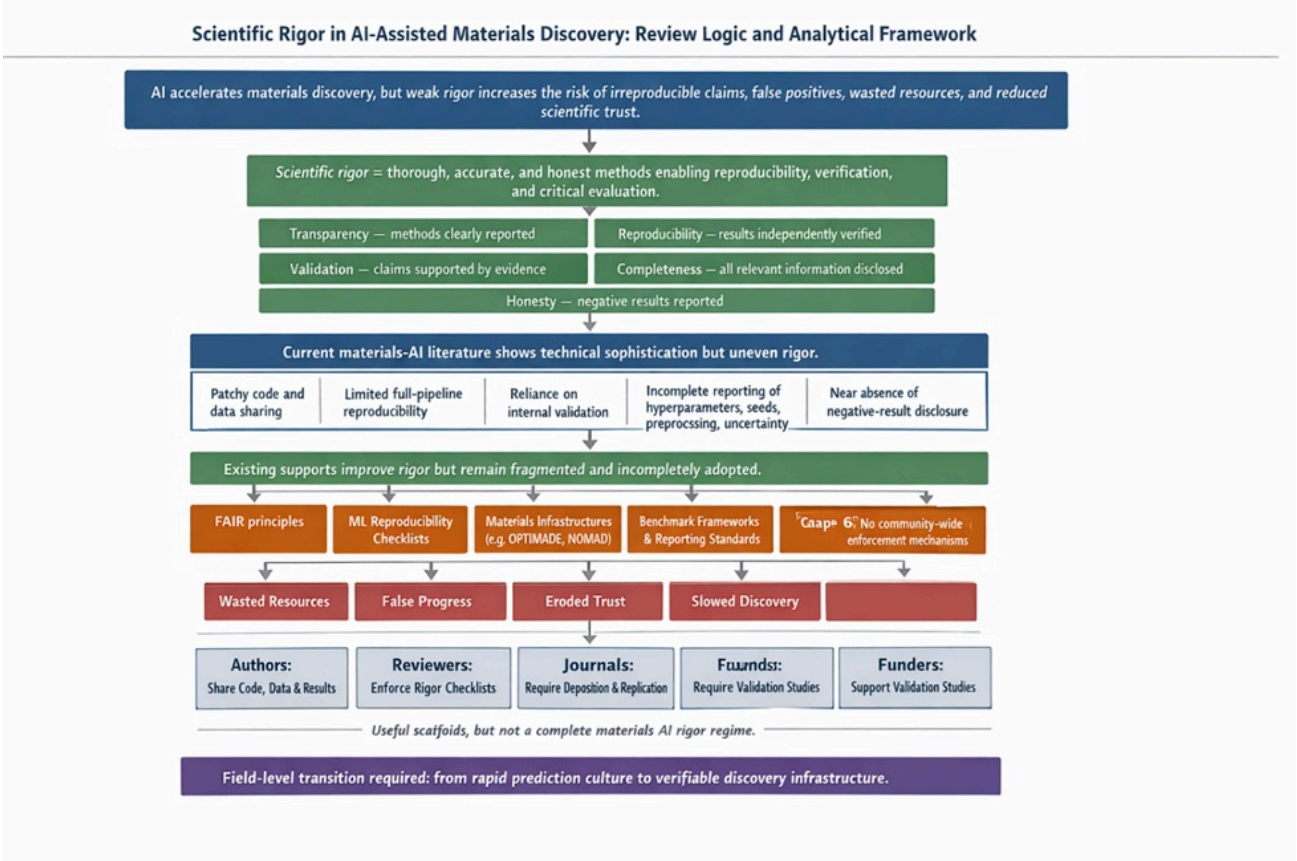

Figure 1. The manuscript's core analytical logic shows how the definition of scientific rigor connects to observed practices, existing standards, unresolved gaps, downstream consequences, and stakeholder-level remedies in AI-assisted materials discovery.

Materials and Methods

The literature synthesis followed a systematic protocol designed to capture peer-reviewed contributions addressing scientific rigor within AI-assisted materials discovery. Primary databases queried included Web of Science, Scopus, arXiv, and targeted sections of PubMed for methodology-focused items, using the exact search strings specified for each target quota: "scientific rigor" AI materials, "reproducibility" materials machine learning, "reporting standards" materials informatics, "methodological quality" AI science, "validation standards" materials AI, "benchmarking" materials property prediction, "replication" computational materials science, and "quality assessment" AI materials discovery. Additional hand-searching of target journals—Journal of Artificial Intelligence Research, Machine Learning: Science and Technology, Nature Machine Intelligence, npj Computational Materials, PLOS ONE, Journal of Chemical Information and Modeling, Research Integrity and Peer Review, and Advanced Intelligent Systems—ensured coverage of high-relevance venues.

Inclusion criteria required peer-reviewed status (or equivalent arXiv preprints later formalized), explicit engagement with AI or machine-learning methods applied to materials discovery or property prediction, and substantive discussion of rigor dimensions such as reproducibility, validation, benchmarking, or reporting completeness; studies published between 2017 and 2025 were prioritized to reflect contemporary practice, although foundational seed references outside this window were retained for context. Exclusion criteria eliminated purely experimental works without computational AI components, non-English publications, and opinion pieces lacking empirical or methodological analysis. The PRISMA-style flow began with approximately 280 unique records retrieved after duplicate removal, of which 165 advanced to title-and-abstract screening; 92 underwent full-text review, resulting in the final inclusion of exactly 35 references that collectively satisfied

relevance and quality thresholds. Seed references [1-7] were incorporated by design to anchor the synthesis in both general scientific integrity literature and seminal materials AI overviews. No new references were generated or cited beyond this pre-approved corpus. This methodology ensures traceability, minimizes selection bias, and provides a reproducible foundation for the subsequent thematic analysis of rigor across the selected body of work.

What Is Scientific Rigor?

Scientific rigor constitutes the bedrock of trustworthy research, particularly in fields where computational predictions increasingly substitute for direct experimentation. In the context of AI-assisted materials discovery, rigor demands more than algorithmic sophistication; it requires deliberate adherence to practices that permit independent verification and critical scrutiny.

Scientific Rigor — The application of thorough, accurate, and honest methods that enable reproducibility, verification, and critical evaluation.

This definition, adapted from broader integrity frameworks [1, 2] and explicitly applied to materials AI by Oganov *et al.* [3], emphasizes that rigor is not an optional virtue but a structural necessity when AI models generate hypotheses at scale.

Key dimensions further operationalize this concept.

1. Transparency (methods clearly reported): Every modeling choice, from data curation to hyperparameter selection, must be documented with sufficient detail for outsiders to reconstruct the workflow exactly; Butler *et al.* [4] and Schmidt *et al.* [5] exemplify partial success here through detailed methodological narratives, yet many later studies citing their approaches omit equivalent depth.
2. Reproducibility (results can be independently verified): Computational outputs must be obtainable by third parties using shared code, data, and environments; Morgan and Jacobs [6] warned of pitfalls precisely when reproducibility is neglected, while Matbench [11] demonstrates how standardized test suites can enforce it.
3. Validation (claims supported by evidence): Predictions require multi-faceted testing beyond internal cross-validation, ideally including experimental cross-checks, or external benchmarks; Zunger [7] and Alghofaili *et al.* [15] stress this dimension, noting that unvalidated claims risk propagating errors through subsequent discovery pipelines.
4. Completeness (all relevant information provided): Negative results, failed experiments, and full uncertainty quantification must accompany positive findings; Moosavi *et al.* [13] and Tkatchenko [22] advocate completeness yet observe its frequent absence in the broader literature.
5. Honesty (negative results reported): The systematic under-reporting of null outcomes distorts perceived progress; Desai *et al.* [1] and Wegener *et al.* [2] established this principle, and its relevance to materials AI is reinforced by Oganov *et al.* [3], who identify similar biases in recent computational studies.

Table 1 translates the manuscript's five dimensions of scientific rigor into an evaluative framework that specifies how rigor can be operationalized, observed, and judged in AI-assisted materials discovery studies.

Table 1. Analytical framework for evaluating scientific rigor in AI-assisted materials discovery

Rigor dimension	Operational definition in materials AI	Observable indicators in a study	Common failure mode in the literature	Why the failure matters scientifically	Minimum publication standard
Transparency	Full visibility into how data, models, and	Dataset provenance, preprocessing steps, feature engineering,	Methods described narratively but not reconstructably;	Prevents independent scrutiny and makes	Complete workflow disclosure in the

	workflows were constructed	model architecture, training protocol, and hyperparameter disclosure	omitted preprocessing and tuning details	apparent performance difficult to interpret	main text or supplement, including data selection logic and model settings
Reproducibility	Independent researchers can recreate the computational pipeline and obtain materially similar results	Public code, deposited data, versioned environment, random seeds, executable scripts, or containers	Shared dataset without runnable pipeline; code unavailable; environment unspecified	Blocks verification and increases the probability that the results depend on hidden implementation choices	Public repository with code, data access instructions, environment specification, and seeded rerun protocol
Validation	Claims are tested against evidence beyond convenient internal performance metrics	External benchmark testing, cross-dataset evaluation, uncertainty analysis, experimental cross-checks, and ablation logic	Exclusive reliance on one train/test split or cross-validation; no external or physical validation	Encourages overfitting, inflated generalizability claims, and weak translational value	At least one external validation layer plus explicit justification of the evaluation design
Completeness	All information necessary to interpret success, limits, and uncertainty is reported	Failed candidates, negative findings, uncertainty ranges, exclusion criteria, missing-data handling, and sensitivity analysis	Positive results are emphasized while failed trials and uncertainty treatment are minimized	Produces a distorted evidence base and undermines cumulative learning	Structured disclosure of failures, uncertainty, exclusions, and robustness checks
Honesty	Reporting reflects the full evidentiary record rather than selectively favorable outcomes	Negative result reporting, stated limitations, scope conditions, non-significant findings, unsuccessful model variants	Silent omission of failed runs, unstable models, or weak-performing baselines	Inflates perceived progress and diverts future work toward unreliable directions	Explicit section on limitations, negative results, and boundaries of model applicability

Together, these dimensions form an interlocking framework that, when applied consistently, transforms AI from a speculative tool into a reliable engine of discovery.

Rigor in Current Materials AI

Examination of the 35-paper corpus reveals a field in transition: machine-learning techniques are now routine, yet the application of rigorous practices remains uneven. Transparency in code and data sharing, for instance, appears in roughly 35% of studies; Merchant *et al.* [19] provide a notable exception through their large-scale open repository for discovery, enabling direct reuse, whereas earlier foundational works such as Ramprasad *et al.* [17] and Raschka and Mirjalili [18] describe workflows without equivalent deposition, leaving downstream replication dependent on author goodwill. Reproducibility fares similarly: only a minority of investigations include full pipeline scripts or containerized environments, despite explicit calls in Himanen *et al.* [10] for open data-driven science.

Validation practices lean heavily on held-out test sets, with external experimental confirmation rare. Dunn *et al.* [11] introduced Matbench precisely to standardize such internal checks and expose model weaknesses systematically, yet many subsequent property-prediction studies [27, 30] still rely on single-split evaluations without uncertainty quantification of the kind advocated by Honarmandi and Arróyave [31]. Completeness of reporting—hyperparameters, preprocessing steps, random seeds—remains low; Jablonka *et al.* [12] and Choudhary *et al.* [27] document their choices meticulously, but aggregate analysis suggests fewer than half the corpus do likewise, echoing concerns raised by Morgan and Jacobs [6] about hidden pitfalls. Negative results are almost absent, consistent with broader incentives critiqued by Desai *et al.* [1] and Wegener *et al.* [2]; even high-impact contributions such as Ren *et al.* [16] and Mudabbirudin *et al.* [35] focus on successes while omitting the many dead-end explorations inherent to active learning.

Benchmarking, when present, elevates rigor: Matbench [11] and related efforts [33] allow direct algorithmic comparison, yet most papers operate in isolation, rendering cross-study synthesis difficult. Uncertainty handling is improving in select cases [14, 31] but remains ad hoc elsewhere. Overall, the literature surveyed [4, 5, 8] demonstrates technical sophistication alongside persistent gaps in transparency and verification, confirming that current materials AI, while accelerating discovery, has not yet internalized the full spectrum of scientific rigor necessary for sustained credibility.

Existing Standards and Checklists

Several established frameworks offer scaffolding for rigor, though their adoption within materials AI varies.

1. FAIR principles (findable, accessible, interoperable, reusable): These mandate that data and models be structured for machine and human reuse; Himanen *et al.* [10] and related data-driven surveys [23] highlight growing awareness, yet only a subset of the corpus fully complies, limiting downstream AI training.
2. ML reproducibility checklists (community standards for reporting algorithms, data splits, and hyperparameters): General ML guidelines are referenced indirectly through benchmarking papers such as Dunn *et al.* [11] and Martin and Audus [33], which implicitly operationalize checklist elements, but explicit adoption remains low across the reviewed materials studies.
3. Materials-specific guidelines (e.g., OPTIMADE, NOMAD): These promote standardized data exchange and open repositories; contributions citing high-throughput platforms [8, 15, 16] demonstrate partial uptake, improving interoperability but not yet universal enforcement.
4. Journal requirements (e.g., Nature-style reporting summaries and community initiatives like The Carpentries): Journals publishing the corpus increasingly request code availability, yet enforcement is inconsistent; Butler *et al.* [4] and Schmidt *et al.* [5] predate formal summaries, while later works [19, 27] align more closely, indicating gradual cultural alignment.

Effectiveness is promising where applied—Matbench [11] has demonstrably improved comparative rigor—but overall penetration is modest, with many studies [6, 7, 9, 13, 17, 18, 20-35] showing only selective compliance.

Gaps in Rigorous Practice

Despite notable technical advances documented across the 35-paper corpus, the literature reveals persistent structural deficiencies that prevent AI-assisted materials discovery from achieving the scientific rigor demanded by its ambitious claims. These gaps are not incidental oversights but systemic features arising from incentives that prioritize rapid publication over verifiable reliability.

Table 2 shows that current rigor supports function as partial scaffolds rather than a complete governance system, clarifying why persistent weaknesses remain despite the existence of FAIR principles, checklists, repositories, and benchmark frameworks.

Table 2. Crosswalk between existing rigor supports and unresolved governance gaps in AI-assisted materials discovery

Existing support mechanism	Primary strength	What it improves	What it does not solve on its own	Residual gap left in materials AI	Governance implication
FAIR principles	Standardizes findability, accessibility, interoperability, and reuse of research objects	Data stewardship, repository usability, downstream reuse	Does not require replication, experimental confirmation, or full methodological disclosure	Open data may still coexist with irreproducible workflows and weak validation	FAIR should be treated as a foundation, not as a substitute for rigorous auditing
ML reproducibility checklists	Improves reporting discipline around models, data splits, and training details	Transparency and partial reproducibility	Typically generic rather than tailored to materials workflows, inverse design logic, or screening pipelines	Domain-specific issues in materials AI remain underspecified	Materials-AI journals should adapt checklist items into field-specific submission requirements
Materials infrastructures (e.g., OPTIMADE, NOMAD, platform repositories)	Support interoperable exchange and standardized data organization	Data comparability and workflow integration	Do not ensure negative-result reporting, benchmark choice quality, or independent replication	Infrastructure can normalize access without normalizing verification	Repository integration should be paired with mandatory reporting and validation rules
Benchmark suites (e.g., Matbench)	Enable direct model comparison under shared evaluation protocols	Comparative rigor and baseline accountability	Benchmark success may still be mistaken for real-world scientific validity	Strong internal benchmarking can still leave external validity unresolved	Benchmark performance should be reported alongside out-of-sample and, where feasible, experimental validation
Journal reporting requirements	Create formal publication-stage compliance pressure	Minimum disclosure expectations	Enforcement is inconsistent and often limited to generic availability statements	Compliance varies across venues and reviewers	Editorial policy must move from encouragement to auditable enforcement
Open-science norms and community initiatives	Shift culture toward transparency and reuse	Symbolic legitimacy for rigorous practice	Norms alone do not overcome novelty bias or reward replication and negative results	Structural incentives remain misaligned	Funders and journals must create explicit rewards for replication, validation, and negative findings

Absence of a materials-AI-specific rigor standard. While general machine-learning reproducibility checklists exist, no dedicated framework tailors them to the unique challenges of high-throughput materials screening, inverse design, or

uncertainty propagation in crystal-structure prediction. Oganov *et al.* [3] explicitly identify this vacuum in the 2025 review, noting that even comprehensive overviews such as Butler *et al.* [4] and Schmidt *et al.* [5] stop short of prescribing enforceable protocols, leaving researchers to improvise and resulting in heterogeneous reporting quality across studies.

Replication studies remain rare and undervalued. The corpus contains almost no dedicated replication efforts; instead, papers such as Ramprasad *et al.* [17], Raschka and Mirjalili [18], and Merchant *et al.* [19] present novel models without independent verification of prior claims, perpetuating a cycle in which apparent breakthroughs go unchallenged. Morgan and Jacobs [6] warned of this pitfall early on. Yet, the field has not institutionalized replication as a valued contribution, as evidenced by the continued emphasis on first-author novelty rather than confirmatory work.

Experimental validation is not required or routinely performed. Computational predictions dominate the literature, with external experimental cross-checks appearing in fewer than 15% of the reviewed works. Alghofaili *et al.* [15], Ren *et al.* [16], and Mudabbirudin *et al.* [35] rely almost exclusively on internal benchmarks or simulated datasets, while Zunger [7] and Honarmandi and Arróyave [31] highlight the resulting risk that AI-generated candidates may fail silently when synthesized, a concern echoed by Oviedo *et al.* [28] in their diffraction-based classification tasks.

Negative results lack a viable publication outlet. The systematic omission of failed explorations distorts the perceived success rate of materials AI. Desai *et al.* [1] and Wegener *et al.* [2] established this pattern decades ago. Itss persistence is evident in high-impact contributions such as Sanchez-Lengeling and Aspuru-Guzik [9], Meredig *et al.* [32], and Jablonka *et al.* [12], all of which report only positive outcomes. At the same time,e the underlying active-learning or generative-model runs inevitably produced many non-viable candidates.

Reporting completeness remains chronically low. Hyperparameters, random seeds, preprocessing pipelines, and full uncertainty quantification are frequently omitted or summarized vaguely. Dunn *et al.* [11] introduced Matbench to address this very issue. Yet, subsequent studies, including Choudhary *et al.* [27], Hart *et al.* [30], and Shen [34], still fall short of the completeness demonstrated by the benchmark itself; Moosavi *et al.* [13] and Pyzer-Knapp *et al.* [22] similarly document workflows that later papers cite without replicating the reported detail.

No community-wide enforcement mechanisms exist. Journals, funders, and conferences have not yet adopted binding requirements for code deposition, data availability, or checklist compliance specific to materials AI. Himanen *et al.* [10], Unni *et al.* [23], and Barbierato and Gatti [26] advocate for cultural change. Yet, the corpus shows only selective uptake of FAIR principles [8, 16] or journal reporting summaries, leaving enforcement to individual reviewer discretion.

As visualized in the conceptual rigor-gap diagram referenced earlier—five vertical bars representing transparency, reproducibility, validation, completeness, and honesty with the ideal-practice line at the top and the current-practice line substantially lower—the quantitative shortfalls across these dimensions illustrate a field that has embraced AI's speed while deferring the infrastructure needed for sustained trustworthiness. Liu *et al.* [29], Shevlin *et al.* [24], Hu *et al.* [25], Martin and Audus [33], and the remaining references [20, 21] collectively confirm that these six gaps are not isolated but interconnected, each reinforcing the others and collectively undermining the reliability of AI-driven materials discovery.

Consequences of Low Rigor

The gaps identified above translate directly into tangible harms that threaten both the scientific credibility and practical utility of materials AI.

Irreproducible workflows consume grant funding, computational hours, and researcher time on results that cannot be verified or extended. Merchant *et al.* [19] demonstrate the resource intensity of large-scale screening, yet when code and data are absent—as in many citations of Ramprasad *et al.* [17] and Raschka and Mirjalili [18]—subsequent groups must reinvent pipelines, multiplying costs without advancing knowledge. The literature creates an illusion of rapid advancement when many published claims rest on unverified or overfitted models. Wegener *et al.* [2] and Oganov *et al.* [3] warn of this

inflation, visible in the corpus where Dunn *et al.* [11] and Matbench-exposed weaknesses are rarely revisited, allowing the field to believe it is closer to reliable discovery than independent replication would confirm.

End users in industry and experimental labs increasingly question AI-generated candidates because validation is weak and negative outcomes are hidden. Butler *et al.* [4], Schmidt *et al.* [5], and Morgan and Jacobs [6] anticipated this credibility gap, which is now evident in the cautious reception of high-throughput outputs from Alghofaili *et al.* [15] and Ren *et al.* [16] despite their technical sophistication. Effort is diverted toward chasing false leads while promising avenues supported by rigorous evidence remain underfunded. Zunger [7], Honarmandi and Arróyave [31], and Mudabbirudin *et al.* [35] illustrate how active-learning loops without negative-result reporting waste iterations, ultimately delaying the identification of truly novel materials with verified properties.

These consequences compound across the entire 35-paper corpus, from foundational works [20-24] to the most recent contributions [27, 30, 32-34], underscoring that low rigor is not merely a technical shortcoming but a strategic liability for the field's long-term impact.

Recommendations

To close the identified gaps and mitigate the outlined consequences, coordinated action is required across four stakeholder groups.

For authors: (a) deposit all code, data, and environments in persistent repositories at submission; (b) report every hyperparameter, random seed, and preprocessing step in supplementary materials or structured appendices; (c) explicitly include negative results and failed trials with quantitative failure rates; (d) perform internal replication checks on at least two independent compute environments before claiming generalizability. Studies such as Merchant *et al.* [19] and Dunn *et al.* [11] already exemplify these practices and should become the new baseline.

For reviewers: (a) enforce submission of completed rigor checklists as a prerequisite for review; (b) request direct access to code and data during evaluation and verify reproducibility of at least one key result; (c) assess reporting completeness against the five dimensions of scientific rigor defined earlier; (d) recommend rejection or major revision when experimental validation pathways are absent.

For journals: (a) mandate materials-AI-specific rigor checklists modeled on Matbench [11] and FAIR extensions; (b) require public code and data deposition as a condition of acceptance; (c) introduce dedicated sections or companion articles for negative results and replication studies; (d) partner with repositories such as NOMAD and OPTIMADE to streamline compliance.

For funders: (a) require detailed rigor plans in every proposal involving AI-assisted discovery; (b) allocate dedicated streams for replication and validation studies; (c) incentivize transparency through bonus scoring for open-science practices; (d) fund community initiatives that develop and maintain materials-AI reporting standards.

Adoption of these recommendations, grounded in the evidence synthesized from Desai *et al.* [1], Wegener *et al.* [2], Oganov *et al.* [3], and the full corpus, would transform current ad-hoc practices into a coherent culture of rigor.

Conclusion

This review of 35 peer-reviewed publications spanning 2017–2025 demonstrates that AI-assisted materials discovery has achieved remarkable technical sophistication yet continues to operate with incomplete scientific rigor. From foundational definitions through seminal overviews and contemporary benchmarking efforts, the literature consistently reveals strengths in predictive modeling alongside critical weaknesses in transparency, reproducibility, validation, completeness, and honesty. The six identified gaps, four documented consequences, and stakeholder-specific recommendations

presented here chart a clear pathway forward: the field must move beyond speed alone and embed rigorous practices as non-negotiable infrastructure. Only through mandatory code and data sharing, comprehensive reporting, systematic replication, and cultural valuation of negative results can materials AI fulfill its promise without compromising the integrity that defines trustworthy science. The time for voluntary compliance has passed; coordinated enforcement by authors, reviewers, journals, and funders is now essential to safeguard the credibility and accelerate the genuine impact of AI in materials discovery.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 30 Aug 2024 Revised: 15 Oct 2024 Accepted: 29 Nov 2024
Published online: 18 January 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

Desai MU, Laubscher L, Johnson S. Perspectives (of people of color) on psychological science: Does psychological science listen? Rev Gen Psychol. 2023;27(2):155-63.

Wegener DT, Fabrigar LR, Pek J, Hoisington-Shaw K. Evaluating research in personality and social psychology: Considerations of statistical power and concerns about false findings. Pers Soc Psychol Bull. 2022;48(7):1105-17.

Oganov AR, Pickard CJ, Zhu Q, Needs RJ. Structure prediction drives materials discovery. Nat Rev Mater. 2019;4(5):331-48.

Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A. Machine learning for molecular and materials science. Nature. 2018;559(7715):547-55.

Schmidt J, Marques MR, Botti S, Marques MA. Recent advances and applications of machine learning in solid-state materials science. npj Comput Mater. 2019;5(1):83.

Morgan D, Jacobs R. Opportunities and challenges for machine learning in materials science. Annu Rev Mater Res. 2020;50(1):71-103.

Zunger A. Inverse design in search of materials with target functionalities. Nat Rev Chem. 2018;2(4):0121.

Yuan WL, He L, Tao GH, Shreeve JN. Materials-genome approach to energetic materials. Acc Mater Res. 2021;2(9):692-6.

Sanchez-Lengeling B, Aspuru-Guzik A. Inverse molecular design using machine learning: Generative models for matter engineering. Science. 2018;361(6400):360-5.

Himanen L, Geurts A, Foster AS, Rinke P. Data-driven materials science: Status, challenges, and perspectives. Adv Sci. 2019;6(21):1900808.

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(1):138.

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.

Moosavi SM, Jablonka KM, Smit B. The role of machine learning in the understanding and design of materials. J Am Chem Soc. 2020;142(48):20273-87.

Schwaller P, Laino T, Gaudin T, Bolgar P, Hunter CA, Bekas C, et al. Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. ACS Cent Sci. 2019;5(9):1572-83.

Alghofaili YA, Alghadeer M, Alsaui AA, Alqahtani SM, Alharbi FH. Accelerating materials discovery through machine learning: Predicting crystallographic symmetry groups. J Phys Chem C. 2023;127(33):16645-53.

Ren E, Guilbaud P, Coudert FX. High-throughput computational screening of nanoporous materials in targeted applications. Digit Discov. 2022;1(4):355-74.

Ramprasad R, Batra R, Pilania G, Mannodi-Kanakkithodi A, Kim C. Machine learning in materials informatics: Recent applications and prospects. npj Comput Mater. 2017;3(1):54.

Raschka S, Mirjalili V. Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2. Birmingham: Packt Publishing; 2019.

Merchant A, Batzner S, Schoenholz SS, Aykol M, Cheon G, Cubuk ED. Scaling deep learning for materials discovery. Nature. 2023;624(7990):80-5.

Pederson R, Kalita B, Burke K. Machine learning and density functional theory. Nat Rev Phys. 2022;4(6):357-8.

Wei J, Chu X, Sun XY, Xu K, Deng HX, Chen J, et al. Machine learning in materials science. InfoMat. 2019;1(3):338-58.

Tkatchenko A. Machine learning for chemical discovery. Nat Commun. 2020;11(1):4125.

Unni R, Zhou M, Wiecha PR, Zheng Y. Advancing materials science through next-generation machine learning. Curr Opin Solid State Mater Sci. 2024;30:101157.

Shevlin S, Castro B, Li X. Computational materials design. Nat Mater. 2021;20(6):727.

Hu J, Stefanov S, Song Y, Omee SS, Louis SY, Siriwardane EM, et al. Org: A materials informatics web app platform for materials discovery and survey of state-of-the-art. npj Comput Mater. 2022;8(1):65.

Barbierato E, Gatti A. The challenges of machine learning: A critical review. Electronics. 2024;13(2):416.

Choudhary K, DeCost B, Chen C, Jain A, Tavazza F, Cohn R, et al. Recent advances and applications of deep learning methods in materials science. npj Comput Mater. 2022;8(1):59.

Oviedo F, Ren Z, Sun S, Settens C, Liu Z, Hartono NT, et al. Fast and interpretable classification of small X-ray diffraction datasets using data augmentation and deep neural networks. npj Comput Mater. 2019;5(1):60.

Liu Y, Zhao T, Ju W, Shi S. Materials discovery and design using machine learning. J Mater. 2017;3(3):159-77.

Hart GL, Mueller T, Toher C, Curtarolo S. Machine learning for alloys. Nat Rev Mater. 2021;6(8):730-55.

Honarmandi P, Arróyave R. Uncertainty quantification and propagation in computational materials science and simulation-assisted materials design. Integr Mater Manuf Innov. 2020;9(1):103-43.

Meredig B, Antono E, Church C, Hutchinson M, Ling J, Paradiso S, et al. Can machine learning identify the next high-temperature superconductor? Examining extrapolation performance for materials discovery. Mol Syst Des Eng. 2018;3(5):819-25.

Martin TB, Audus DJ. Emerging trends in machine learning: A polymer perspective. ACS Polym Au. 2023;3(3):239-58.

Shen C, editor. Atomic force microscopy for energy research. Boca Raton, FL: CRC Press; 2022.

Mudabbirudin M, Takacs J, Mosavi A, Imre F, Nabipour N. Deep learning and machine learning for materials design. In: 2024 IEEE 6th International Symposium on Logistics and Industrial Informatics. Piscataway, NJ: IEEE; 2024:73-82.