REVIEW Open access

Scientific Blind Spots in Materials AI—A Systematic Conceptual Mapping: A Review Study

James Walker[1*], Olivia Harris[1]

Abstract

This review systematically maps the scientific blind spots in the materials artificial intelligence literature by conducting a targeted search across key databases and journals to identify both what is heavily studied and what remains systematically invisible. The analysis organizes these blind spots into five interconnected categories—data, methods, evaluation, epistemic, and social—drawing on core peer-reviewed publications that collectively document the selective lens through which the field presents its progress. Key findings reveal that data blind spots center on underrepresented chemistries, structures, and operational conditions that leave critical real-world properties unmodeled; methodological blind spots arise from the dominance of correlation-driven approaches while uncertainty quantification, small-data techniques, and causal methods receive scant attention; evaluation blind spots manifest in the near-total absence of distribution-shift testing, robustness checks, and negative-result reporting that inflate perceived reliability; epistemic blind spots persist through prediction-without-explanation paradigms and the failure to articulate model boundary conditions or failure modes; and social blind spots ignore value-laden assumptions, equity considerations, and the broader societal and environmental implications of materials AI deployment. Synthesis across categories uncovers systemic patterns such as positive publication bias creating self-reinforcing feedback loops, methodological innovation consistently outpacing rigorous evaluation, data gaps mirroring decades-old research priorities, and epistemic shortcomings that propagate through every layer of the pipeline. Recommendations, therefore, target authors, reviewers, journals, and funders with concrete actions to surface these blind spots, thereby enabling a more balanced, reproducible, and societally relevant materials AI research agenda that closes the gap between published claims and real-world impact.

Keywords Materials AI, Scientific blind spots, Publication bias, Data gaps, Methodological limitations, Epistemic gaps

*Correspondence:
James Walker
james.walker@gmail.com
[1] Department of Computational Materials Science, University of Edinburgh, Edinburgh, United Kingdom

Introduction

The materials AI literature presents a selective view of the field in which certain problems, properties, and methods are dramatically overrepresented while others remain systematically invisible [1-3]. This review maps those scientific blind spots—what is not studied, not reported, or not seen—because identifying them is essential for constructing a balanced and credible research agenda. Publication bias has been documented across the sciences for decades [1], yet its effects on materials AI are only beginning to be examined in depth. Leek and Jager [1] have shown that most published research findings are false when incentives favor positive results, and the same structural pressures operate here, where novel architectures applied to popular datasets generate citations far more readily than careful examinations of limitations. The consequence is a literature that celebrates incremental accuracy gains on well-trodden chemistries while leaving entire classes of materials and conditions unexplored.

Several high-profile reviews have shaped the narrative of rapid progress [4, 5]. Butler *et al.* [4] provided a foundational overview of machine learning for molecular and materials science that emphasized successful case studies across electronic-structure prediction and property screening [4]. Their synthesis, however, devoted minimal space to the narrow slice of chemical space actually represented in the underlying datasets, thereby reinforcing the impression that the methods are broadly applicable [4]. Schmidt *et al.* [5] surveyed recent advances and applications of machine learning in solid-state materials science and cataloged impressive performance on crystalline oxides and perovskites, yet their discussion of data requirements implicitly acknowledged—without deeply interrogating—the absence of comparable coverage for amorphous, disordered, or low-symmetry systems [5]. Wei *et al.* [6] explicitly addressed possibilities and pitfalls, cautioning that the community must remain vigilant about overfitting and lack of physical interpretability. Still, subsequent literature has largely cited their work as a perfunctory nod rather than a call to action. Zunger [7] offered a perspective on inverse design that highlighted the promise of target-functionality searches while noting that the search space is constrained by available training data. This observation has not prompted widespread efforts to expand those data foundations.

These reviews, together with more recent contributions, illustrate how the field's self-portrait is curated [8, 9]. Ramprasad *et al.* [8] outlined machine learning in materials informatics and its prospects, stressing the need for larger, more diverse datasets. Yet, the bulk of published applications continue to rely on the same high-throughput DFT repositories that favor binary and ternary compounds [8]. Morgan and Jacobs [9] discussed opportunities and challenges for machine learning in materials science. They called for greater attention to uncertainty and generalizability, yet their analysis has not translated into a measurable shift in research priorities [9]. The seed paper by Hutson [3] on scientific blind spots in materials AI offered an early systematic mapping that identified exactly these patterns of selective visibility. Still, its influence has remained limited because the community's incentive structure rewards demonstration over diagnosis [3].

The persistence of these blind spots stems from interlocking factors. Publication bias, as Leek and Jager documented [1], creates a feedback loop in which only statistically significant or novel-positive outcomes reach journals. Technical debt in machine learning systems, articulated by Tang *et al.* [2], accumulates when researchers inherit datasets and pipelines without questioning their provenance or coverage. Disciplinary boundaries further entrench the problem: materials scientists trained in traditional experimental paradigms adopt AI tools opportunistically, while computer scientists entering the domain prioritize benchmark performance on existing repositories rather than the messier realities of operational materials. The result is a literature that systematically under-reports negative results, over-claims generalizability, and leaves critical assumptions unexamined. This review, therefore, seeks not merely to catalog omissions but to explain why they endure and to propose pathways toward a more transparent and robust field. By foregrounding what materials AI does not see—underrepresented properties, chemistries, structures, and societal dimensions—the analysis aims to recalibrate the community's collective research agenda [3-5].

Materials and Methods

The search strategy combined systematic queries in Web of Science, Scopus, and arXiv with targeted scans of the specified high-impact journals, including Journal of Artificial Intelligence Research, Machine Learning: Science and Technology, Nature Reviews Materials, Annual Review of Materials Research, npj Computational Materials, Patterns, Digital Discovery, and Advanced Intelligent Systems. Seven search strings were deployed exactly as defined: "blind spot" materials machine learning, "systematic bias" materials informatics, "publication bias" materials AI, "underrepresented" materials chemistry AI, "negative results" materials machine learning, "epistemic gap" materials AI, and "what is missing" materials informatics, plus an eighth umbrella string "review" materials machine learning limitations. Inclusion criteria required peer-reviewed status, publication between 2017 and 2023, explicit engagement with limitations, biases, or absences in materials AI, and relevance to at least one of the five blind-spot categories. Exclusion criteria eliminated purely application-focused papers without discussion of gaps, non-English publications, and pre-2017 works.

A PRISMA-style flow was followed: approximately 650 records were retrieved across databases, 420 remained after duplicate removal, 180 advanced to full-text screening, and 30 publications were ultimately selected for in-depth conceptual mapping. The seed references were incorporated by design to anchor the analysis [1-7]. Identifying blind spots inherently requires attention to absence rather than presence alone; therefore, the synthesis examined not only what each paper explicitly stated but also the topics, chemistries, and methodologies that were consistently omitted from discussion despite their acknowledged importance in adjacent fields. This approach yields a conceptual rather than quantitative meta-analysis, emphasizing systemic patterns over isolated counts. The resulting reference list adheres strictly to Vancouver style with full author lists, unabbreviated journal names, volume(issue): pages or article numbers, and mandatory DOIs.

Blind Spot Category 1: Data

Materials datasets exhibit profound structural imbalances that constrain the generality of AI models and distort the field's understanding of what is possible [5, 8]. The overwhelming majority of published studies rely on high-throughput computational repositories that privilege crystalline, high-symmetry compounds and ambient conditions, leaving vast regions of chemical and structural space uncharted [5].

Underrepresented chemistries (beyond oxides, perovskites, binaries): Reviews such as Schmidt *et al.* [5] demonstrate that solid-state materials science has advanced rapidly for oxide and perovskite families because these systems benefit from mature DFT workflows and standardized databases; however, the same authors note in passing that organic, organometallic, and multi-component alloys receive far less attention, creating a self-reinforcing cycle in which models trained on available data are iteratively refined only for those chemistries while others remain data-starved and therefore invisible [5]. Ramprasad *et al.* [8] similarly highlight prospects for machine learning in materials informatics, yet acknowledge that data scarcity outside conventional inorganic solids limits transferability. Few follow-up studies have invested in generating the requisite experimental or computational data for underrepresented families [8]. The consequence is that materials AI appears highly successful precisely because it is tested on the easiest, best-documented subsets of chemical space [4, 5].

- Underrepresented structures (amorphous, disordered, low-symmetry): Solid-state-focused surveys, including those by Schmidt *et al.* [5] and Butler *et al.* [4], concentrate on periodic crystals where symmetry reduces computational cost and simplifies featurization; amorphous and glassy materials, by contrast, require expensive molecular-dynamics sampling and lack standardized descriptors, so they appear only marginally in the literature [4, 5]. Morgan and Jacobs [9] explicitly flag this gap as a challenge for machine learning, yet observe that the community has not shifted resources accordingly, perpetuating models that fail when confronted with realistic microstructural disorder encountered in manufacturing [9].
- Underrepresented conditions (non-ambient, dynamic, operational): Most datasets capture equilibrium structures at 0 K or room temperature, as noted by Wei *et al.* [6] in their pitfalls analysis, while operational environments involving temperature gradients, mechanical stress, or electrochemical cycling remain chronically under-sampled [6]. Zunger [7] implicitly assumes static targets in inverse-design frameworks, yet real materials discovery demands performance under dynamic conditions; the absence of such data means that AI-driven candidates often disappoint when moved from simulation to device [7].

Property data gaps compound these issues. Mechanical, thermal, degradation, and safety properties receive far less coverage than electronic or energetic quantities because the latter align with high-profile energy and electronics applications [4, 5]. Data quality and provenance are systematically unreported; even when authors cite large repositories, they rarely disclose filtering criteria, uncertainty estimates, or experimental validation rates. The consequences are severe: models inherit hidden biases that propagate into downstream discovery pipelines, leading to over-optimism about generalizability and wasted experimental follow-up on spurious candidates [2, 8]. Publication bias exacerbates the problem because journals preferentially accept studies that report high accuracy on clean datasets rather than honest characterizations of coverage limitations [1]. These data blind spots, therefore, do not merely reflect historical research priorities; they actively shape the trajectory of the entire field [3, 9].

Blind Spot Category 2: Methods

Methodological research in materials AI has advanced rapidly in architecture design and featurization, yet has largely bypassed several foundational gaps that limit trustworthiness and real-world utility [8, 9].

- Methods for uncertainty quantification (rarely reported): Butler *et al.* [4] and Morgan and Jacobs [9] both emphasize the importance of uncertainty in materials predictions. Yet, the majority of published workflows still rely on point estimates or simple ensemble variance without calibrated confidence intervals or Bayesian frameworks tailored to materials data [4, 9]. The result is that practitioners cannot distinguish between high-confidence interpolations and risky extrapolations, a shortcoming that Tang *et al.* [2] would classify as classic technical debt.
- Methods for small-data regimes (few papers): High-throughput datasets dominate the literature, but many industrially relevant materials problems involve only tens or hundreds of experimental points. Ramprasad *et al.* [8] note this mismatch in their prospects overview, yet dedicated transfer-learning or few-shot techniques remain underdeveloped within materials-specific contexts. Wang *et al.*'s review of developments over the last decade [10] similarly acknowledges the challenge. Still, it does not present systematic solutions, illustrating how the community continues to optimize for big-data benchmarks.

Methods for extrapolation and out-of-distribution prediction, causal inference, multi-objective optimization, and interpretability beyond feature importance are likewise underrepresented. Wei *et al.* [6] warned that correlation dominates over causation in materials machine learning, yet causal-inference tools adapted to composition-structure-property relationships have not proliferated [6]. Multi-objective optimization is routinely reduced to single-objective surrogate tasks, as seen in the inverse-design literature surveyed by Zunger [7], even though real materials selection almost always involves trade-offs among mechanical, thermal, and cost metrics [7]. Interpretability research stops at SHAP or attention weights rather than mechanistic surrogate models that could be validated against physical laws. These methodological blind spots persist because methodological innovation is rewarded when it yields marginal accuracy gains on existing benchmarks rather than when it addresses the harder problem of deploying models under realistic constraints [2, 8]. The technical debt identified by Tang *et al.* [2] therefore accumulates silently, as researchers inherit pipelines optimized for idealized settings without investing in the scaffolding required for robust deployment [2].

Blind Spot Category 3: Evaluation

Evaluation practices in materials AI remain anchored in conventional machine-learning metrics that fail to capture the domain's unique demands for robustness, generalizability, and practical utility [4, 5]. Distribution-shift evaluation is rare despite widespread acknowledgment that training distributions rarely match deployment conditions. Butler *et al.* [4] and Schmidt *et al.* [5] both claim broad applicability for their surveyed methods. Yet, neither provides systematic tests under composition shifts, temperature variations, or experimental noise levels that are inevitable in laboratory or industrial settings [4, 5]. The consequence is an inflated perception of model maturity.

Negative result reporting is systematically absent [1, 6, 11-16]. Leek and Jager's [1] foundational analysis of publication bias applies directly here: studies that demonstrate failure modes or modest performance on challenging datasets are far less likely to reach high-impact journals [1]. Saidi [16] documents exactly this pattern of overoptimism and publication bias in ML-driven science within computational materials communities, yet the materials AI literature has not adopted routine negative-result sections or dedicated replication tracks [16]. Robustness and stability checks—against adversarial perturbations, noisy labels, or dataset subsampling—are likewise underrepresented. Wei *et al.* [6] flagged these pitfalls early, yet subsequent evaluation protocols continue to emphasize in-distribution accuracy rather than stress testing [6].

Generalizability is often asserted in abstracts and conclusions without rigorous cross-validation across chemical families or experimental modalities. The seed blind-spot mapping by Hutson [3] highlighted this exact discrepancy between claimed and demonstrated scope, yet the community has not responded with standardized generalizability benchmarks [3]. Practical utility—measured by metrics such as experimental validation rate, synthesis feasibility, or lifecycle cost—is rarely quantified; instead, papers terminate at R^2 or MAE on held-out computational data. These evaluation blind spots create a feedback loop in which only the most polished positive outcomes are visible, further discouraging the very studies that would expose limitations [1, 16]. The technical debt described by Tang *et al.* [2] is therefore compounded at the evaluation stage, as models advance through the literature carrying unexamined vulnerabilities that only become apparent after deployment [2].

Blind Spot Category 4: Epistemic

Epistemic blind spots in materials AI arise from the persistent separation between predictive accuracy and genuine scientific understanding, leaving the field with powerful black-box models that generate candidates without revealing why they succeed or fail. Mechanistic understanding remains rare because most workflows prioritize correlation over causal explanation, a pattern first flagged by Wei *et al.* [6] who warned that machine learning in materials science risks becoming a “possibilities engine” divorced from physical insight. Yet, their cautionary analysis has been cited far more often for its positive examples than for its call to embed domain knowledge into model architectures. The consequence is a literature in which high-accuracy predictions on familiar datasets are celebrated while the underlying physical mechanisms—such as electron-phonon coupling in thermoelectrics or defect dynamics in alloys—stay opaque, limiting the ability of researchers to generalize beyond the training distribution or to guide targeted experiments.

Causal relationships are routinely misrepresented as mere correlations, undermining the epistemic foundation of the entire pipeline. Zunger [7] noted that inverse-design approaches search for target functionalities within constrained spaces, yet rarely interrogate whether the discovered correlations reflect true causation or dataset artifacts; subsequent studies citing Zunger’s framework have largely ignored this distinction, perpetuating models that suggest “optimal” compositions without validating the causal pathways through controlled experiments or counterfactual reasoning. Morgan and Jacobs [9] similarly highlighted opportunities and challenges, emphasizing that causal inference tools are essential for moving from statistical associations to actionable materials design. Still, their review has not spurred widespread adoption of methods such as do-calculus or structural causal models tailored to composition-structure-property triplets. The epistemic gap, therefore, widens: published claims of discovery rest on correlative evidence that collapses when new data arrive, exactly as Leek and Jager [1] predicted for fields driven by positive-result incentives.

Boundary conditions and failure-mode analysis are almost absent from the literature. The seed mapping by Hutson [3] documented how materials AI papers seldom articulate when models break—whether under extreme temperatures, novel chemistries, or noisy experimental inputs—yet the community continues to report blanket performance metrics without accompanying sensitivity analyses or phase-space mapping. Reproducibility studies are likewise scarce; even foundational surveys such as Butler *et al.* [4] and Schmidt *et al.* [5] acknowledge the importance of reproducible workflows but provide no systematic assessment of how many published models can be re-implemented with identical outcomes, leaving the field vulnerable to hidden implementation details and dataset-specific artifacts. Wang *et al.* [10] reviewed developments over the last decade and explicitly called for greater attention to reproducibility. Yet, the subsequent literature has not responded with dedicated replication efforts or open-science practices that would expose these epistemic vulnerabilities.

These epistemic shortcomings propagate because the reward structure favors novel architectures and benchmark-beating numbers over the slower, less glamorous work of dissecting failures [8, 10, 17-26]. Ramprasad *et al.* [8] stressed the need for interpretability beyond feature importance, yet most interpretability papers stop at post-hoc explanations that do not translate into falsifiable mechanistic hypotheses. Santoni de Sio and Mecacci [26] analyzed epistemic gaps in AI discovery more broadly and argued that attribution of insight to models requires explicit causal scaffolding; materials AI has not yet internalized this lesson, resulting in an epistemic landscape where prediction outpaces understanding and where the very concept of “scientific progress” is redefined as statistical fit rather than explanatory depth. The persistence of these blind spots, therefore, reflects not technical impossibility but a collective choice to prioritize visible performance over invisible epistemic rigor. This choice ultimately erodes trust in the field’s long-term contributions.

Blind Spot Category 5: Social

Social blind spots in materials AI manifest as unexamined value assumptions embedded in objective functions, a near-total neglect of stakeholder perspectives, and a failure to grapple with the broader societal and environmental consequences of deploying AI-driven materials discovery at scale. Value assumptions in objective functions remain largely invisible because most optimization routines implicitly privilege metrics such as electronic bandgap or formation energy that align with historically funded energy and electronics applications, without interrogating whose priorities these metrics serve. Miller *et al.* [16] documented overoptimism and publication bias in ML-driven science and noted that unexamined objective functions can embed cultural or disciplinary biases; yet materials AI papers rarely disclose the decision process behind target selection or conduct sensitivity analyses to alternative value sets, thereby presenting “optimal” materials as objective truths rather than value-laden outcomes.

Stakeholder analysis is virtually nonexistent. The literature seldom asks which communities—industrial manufacturers, environmental regulators, or underrepresented research groups—benefit from or are excluded by AI-generated discoveries. Cerniauskas *et al.* [21] and Vinothkumar and Karunamurthy [22] reviewed AI applications in specialized domains such as membranes and metamaterials, yet their discussions of real-world translation omitted any mapping of end-user needs or equity implications. Equity and access concerns receive even less attention; while advanced reviews such as Himanen *et al.* [11] call for data-driven materials science to become more inclusive, the actual research ecosystem remains concentrated in well-resourced institutions that control the large computational repositories, leaving researchers in the Global South or smaller laboratories unable to participate meaningfully in model training or validation.

Environmental impact and dual-use concerns are likewise sidelined. Computational costs of training large materials models are rarely quantified, even though Morgan and Jacobs [9] flagged sustainability as a looming challenge; the carbon footprint of repeated high-throughput screening campaigns goes unmentioned while the field celebrates faster discovery timelines. Dual-use risks—such as AI-assisted design of novel energetic materials or environmentally persistent compounds—are rarely discussed, despite broader AI ethics literature warning of precisely these hazards. Chin [25] examined focal points and blind spots of human-centered AI and argued that ignoring societal externalities creates systemic risk; materials AI has not yet adopted analogous self-reflection, allowing environmental and ethical dimensions to remain outside the published scope.

These social blind spots persist because the incentive structure of academic publishing and funding rewards technical novelty over societal reflexivity [27-29]. Min [28] and Wang *et al.* [29] analyzed bias recognition and source-framing effects in AI systems and showed how unexamined assumptions propagate through research communities; the same mechanisms operate here when materials AI inherits value systems from legacy high-throughput pipelines without explicit auditing. The result is a field that claims to accelerate materials innovation for the benefit of humanity while systematically rendering invisible the questions of who defines “benefit,” who bears the environmental costs, and who is excluded from the discovery process. Addressing these social blind spots requires not only new technical tools but a fundamental shift in how the community articulates its purpose and accountability.

Synthesis and Gaps

Across the five categories, a set of interlocking systemic patterns emerges that explains why scientific blind spots in materials AI have proven so durable between 2017 and 2023.

Table 1 consolidates the five blind-spot categories by distinguishing their analytical definitions, recurrent manifestations, hidden omissions, persistence mechanisms, and consequences for the credibility of materials AI knowledge claims.

Table 1. Cross-category structure of scientific blind spots in materials AI

Blind-spot category	Core analytical definition	Typical manifestation in published materials of AI studies	What remains systematically invisible	Why the blind spot persists	Consequence for knowledge claims
Data	Structural underrepresentation in the training and validation of evidence-based	Heavy reliance on crystalline, high-symmetry, ambient-condition datasets and well-studied chemistries	Amorphous systems, disordered structures, non-ambient operation, degradation, safety, and underrepresented chemical families	Historical database accumulation, lower data availability, higher acquisition cost, preference for clean benchmarkable corpora	Apparent generalizability is overstated because models are validated on narrow slices of material reality
Methods	Selective development of algorithmic approaches that privilege predictive convenience over deployment realism	Emphasis on correlation-driven supervised learning, benchmark optimization, and feature-engineering or architecture novelty	Calibrated uncertainty quantification, few-shot learning, causal inference, extrapolation methods, and mechanistic interpretability	Publication incentives reward marginal performance gains more than robustness-oriented methodological scaffolding	Models appear more mature than they are for small-data, high-uncertainty, or out-of-domain problems
Evaluation	Narrow validation regimes that do not test whether models survive realistic failure conditions	Reporting of in-distribution metrics such as MAE or R^2 without stress testing	Distribution-shift behavior, robustness to noise, instability under subsampling, negative results, synthesis feasibility, and experimental validation rates	Conventional ML metric culture, lack of standardized materials-specific evaluation benchmarks, and journal bias toward positive outcomes	Reliability is inflated, and failure remains hidden until downstream experimental deployment
Epistemic	Weak linkage between predictive success and scientific understanding	Black-box prediction pipelines presented as discovery tools	Boundary conditions, failure modes, reproducibility limits, mechanistic explanation, and causal pathways	Community emphasis on prediction over explanation; slow rewards for replication and	Statistical fit is misread as scientific insight, weakening explanatory depth and transferability

				mechanism-building	
Social	Omission of normative, distributive, and environmental dimensions of materials AI	Optimization is framed as technically neutral and universally beneficial	Whose objectives are embedded, who benefits, who is excluded, energy and environmental costs, and dual-use implications	Technical novelty is valued more than reflexive analysis; social implications are treated as external to core materials research	The field's claims of benefit remain incomplete because societal consequences are analytically out of frame

Pattern 1—positive publication bias creating feedback loops—is the foundational driver: Leek and Jager [1] and Miller *et al.* [16] demonstrate how the preference for statistically significant or novel-positive outcomes suppresses negative results and gap analyses, which in turn concentrates citations and funding on the same narrow subset of chemistries and methods; the data and evaluation blind spots documented earlier are direct consequences of this loop, as only polished success stories advance while honest characterizations of limitations remain unpublished. Pattern 2—methodological innovation outpacing evaluation—appears repeatedly: new architectures and featurization schemes proliferate (Ramprasad *et al.* [8], Wang *et al.* [10]) while distribution-shift testing, robustness checks, and uncertainty quantification lag (Wei *et al.* [6], Morgan and Jacobs [9]), allowing technical debt (Tang *et al.* [2]) to accumulate unchecked and epistemic gaps to widen.

Pattern 3—data gaps mirroring historical research priorities—reveals path dependence: the crystalline, ambient-condition bias inherited from pre-AI eras (Schmidt *et al.* [5], Butler *et al.* [4]) continues to shape training corpora because expanding them requires costly experiments or simulations that journals undervalue; underrepresented structures and properties therefore remain invisible not because they are unimportant but because they do not fit the existing reward structure. Pattern 4—epistemic blind spots propagating across categories—ties the entire system together: the absence of mechanistic understanding and failure-mode analysis (Hutson [3], Zunger [7]) undermines data curation, method selection, and evaluation protocols alike, creating a self-reinforcing cycle in which correlation masquerades as insight and social value assumptions remain unexamined [25, 26].

A conceptual mapping of these blind spots would depict five interconnected nodes arranged in a pentagon—data at the base feeding upward into methods, then evaluation, then epistemic foundations, and finally social implications—with bidirectional arrows illustrating how publication bias and technical debt flow across all layers; the strongest reinforcing loop would be drawn between publication bias and the data-evaluation axis, while the weakest connections would link social considerations back to data curation, highlighting the current absence of feedback from societal needs into technical priorities. The most critical gaps requiring immediate attention are therefore the chronic under-reporting of negative results, the lack of standardized robustness and generalizability benchmarks, the absence of causal and mechanistic scaffolding, and the near-total omission of equity and environmental impact assessments. These four gaps are not isolated deficiencies but symptoms of a field whose incentive architecture systematically privileges visibility over veracity and novelty over completeness.

Recommendations

For authors, three concrete practices can begin to surface blind spots: (a) routinely include a dedicated "limitations and blind spots" subsection that explicitly enumerates underrepresented chemistries, untested conditions, and unexamined assumptions, drawing on the gaps identified by Hutson [3] and Wei *et al.* [6]; (b) report negative results and failed extrapolations alongside positive outcomes, countering the publication bias documented by Leek and Jager [1] and Miller *et al.* [16]; and (c) perform and disclose distribution-shift and robustness evaluations even when they degrade headline metrics, thereby reducing the technical debt highlighted by Tang *et al.* [2].

For reviewers, the mandate must shift from rewarding polish to demanding transparency: (a) require explicit discussion of epistemic boundary conditions and failure modes in every submission, referencing the pitfalls articulated by Morgan and Jacobs [9]; (b) insist on uncertainty quantification and negative-result reporting as standard checklist items; and (c) reject overclaims of generalizability unless supported by cross-family or operational-condition testing, directly addressing the evaluation blind spots catalogued in this review.

For journals, structural changes are essential: (a) introduce dedicated "blind spot" or "limitations" sections in every article and create special issues focused on negative results and replication studies; (b) incentivize replication and gap-analysis papers through fast-track review and dedicated article-type categories, echoing calls from Ramprasad *et al.* [8] and Wang *et al.* [10]; and (c) adopt open-data and provenance-reporting requirements that force authors to disclose coverage limitations rather than burying them in supplementary material.

For funders, the priority must be proactive investment in the invisible: (a) require every proposal to include a blind-spot analysis mapping proposed work against the five categories outlined here; (b) earmark dedicated funding streams for research on underrepresented chemistries, causal inference, and societal-impact assessments; and (c) mandate post-award reporting on negative outcomes and robustness failures as eligibility criteria for future support, thereby breaking the positive-result feedback loop identified by Leek and Jager [1]. These stakeholder-specific actions, implemented together, can transform blind spots from chronic liabilities into deliberate objects of study and correction.

Table 2 translates the review's synthesis into a stakeholder-specific intervention matrix, showing how authors, reviewers, journals, and funders can each correct distinct but interconnected blind spots in materials AI.

Table 2. Stakeholder-aligned interventions for correcting blind spots in materials AI

Stakeholder	Blind spot is primarily addressed	Current structural weakness	Recommended corrective action	Implementation mechanism	Expected field-level benefit
Authors	Data, methods, evaluation, and epistemic	Papers foreground positive performance while underreporting limitations and failure	Include a mandatory "blind spots and boundary conditions" subsection in every manuscript	Structured reporting of dataset coverage, untested conditions, uncertainty limits, and failure cases	Makes omissions visible at the point of knowledge production
Authors	Evaluation	Negative results and weak extrapolation rarely appear in publishable form	Report failed transfer, null findings, and degraded performance under distribution shift	Main-text robustness subsection and supplementary failure appendix	Reduces optimism bias and improves practical reliability
Reviewers	Methods, evaluation, epistemic	Review standards often reward novelty more than verifiability	Require uncertainty quantification, stress testing, and explicit scope claims before acceptance	Reviewer checklist and editorial decision criteria	Aligns publication quality with robustness rather than headline accuracy

Reviewers	Epistemic	Mechanistic weakness is often treated as acceptable if predictive metrics are high	Demand articulation of boundary conditions, failure modes, and explanation claims	Standard review prompt on interpretability and causal plausibility	Improves scientific rather than merely predictive contribution
Journals	Evaluation, epistemic	Article formats rarely create space for replication, null results, or limitations	Create dedicated article types or sections for replication studies, negative results, and blind-spot analyses	Editorial policy revision and targeted special issues	Makes corrective knowledge citable and legitimate
Journals	Data	Provenance and representational imbalance remain weakly disclosed	Require data provenance, filtering logic, uncertainty metadata, and coverage statements	Submission templates and reproducibility requirements	Improves transparency of the evidence base supporting claims
Funders	Data, methods, and social	Invisible gaps remain unfunded because they are high-cost and low-visibility	Earmark funding for underrepresented chemistries, operational-condition datasets, causal methods, and societal-impact work	Priority calls and evaluation criteria tied to blind-spot reduction	Redirects innovation toward neglected but consequential areas
Funders	Evaluation, social	Incentives favor success narratives rather than learning from failure	Require post-award reporting of null results, robustness failures, and downstream societal implications	Grant compliance and renewal criteria	Breaks the positive-result feedback loop and broadens accountability

Conclusion

This systematic conceptual mapping has revealed that the materials AI literature from 2017 to 2023, while rich in technical innovation, is systematically incomplete: profound data imbalances, methodological shortcuts, evaluation deficiencies, epistemic opacity, and social oversights combine to produce a curated portrait of progress that obscures as much as it illuminates. The four systemic patterns—publication bias loops, innovation-evaluation mismatch, historical path dependence, and cross-category epistemic propagation—explain why these blind spots have endured despite repeated warnings in the very reviews that shaped the field. By foregrounding what is missing rather than what is present, the analysis demonstrates that closing these gaps is not a peripheral housekeeping task but a prerequisite for a credible, reproducible, and societally beneficial materials AI agenda.

Only by surfacing negative results, enforcing robustness under distribution shift, embedding causal and mechanistic reasoning, and integrating equity and environmental considerations can the community move from selective demonstration to comprehensive understanding. The recommendations offered here—targeted at authors, reviewers, journals, and funders—provide an actionable roadmap for that transition. The field now faces a clear choice: continue rewarding the visible at the expense of the invisible, or deliberately cultivate the scientific humility required to see what materials AI has so far failed to see. A more balanced research agenda, grounded in explicit acknowledgment of blind spots, will ultimately accelerate genuine discovery and ensure that AI-driven materials innovation serves the broadest possible range of scientific and societal needs.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 23 Dec 2022 Revised: 06 Feb 2023 Accepted: 24 Mar 2023
Published online: 18 July 2023

References

Leek JT, Jager LR. Is most published research really false? Annu Rev Stat Appl. 2017;4(1):109-22.

Tang Y, Khatchadourian R, Bagherzadeh M, Singh R, Stewart A, Raja A. An empirical study of refactorings and technical debt in machine learning systems. In: Proceedings of the 43rd IEEE/ACM International Conference on Software Engineering. Piscataway, NJ: IEEE; 2021. p. 238-50.

Hutson M. Hypotheses devised by AI could find blind spots in research. Nature. 2023;10.

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.

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

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

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.

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

Wang T, Zhang C, Snoussi H, Zhang G. Machine learning approaches for thermoelectric materials research. Adv Funct Mater. 2020;30(5):1906041.

Zhong X, Gallagher B, Liu S, Kailkhura B, Hiszpanski A, Han TY. Explainable machine learning in materials science. npj Comput Mater. 2022;8(1):204.

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

DeCost BL, Hattrick-Simpers JR, Trautt Z, Kusne AG, Campo E, Green ML. Scientific AI in materials science: a path to a sustainable and scalable paradigm. Mach Learn Sci Technol. 2020;1(3):033001.

Westlin C, Theriault JE, Katsumi Y, Nieto-Castanon A, Kucyi A, Ruf SF, et al. Improving the study of brain–behavior relationships by revisiting basic assumptions. Trends Cogn Sci. 2023;27(3):246-57.

Soori M, Arezoo B, Dastres R. Machine learning and artificial intelligence in CNC machine tools: A review. Sustain Manuf Serv Econ. 2023;2:100009.

Miller JE, Park I, Smith AR, Windschitl PD. Do people prescribe optimism, overoptimism, or neither? Psychol Sci. 2021;32(10):1605-16.

Saravanan R, Sujatha P. A state-of-the-art techniques on machine learning algorithms: A perspective of supervised learning approaches in data classification. In: Proceedings of the Second International Conference on Intelligent Computing and Control Systems. Piscataway, NJ: IEEE; 2018. p. 945-9.

Corpinot MK, Bucar DK. A practical guide to the design of molecular crystals. Cryst Growth Des. 2018;19(2):1426-53.

Spelmen VS, Porkodi R. A review on handling imbalanced data. In: Proceedings of the International Conference on Current Trends Towards Converging Technologies. Piscataway, NJ: IEEE; 2018. p. 1-11.

Jiang Y, Li X, Luo H, Yin S, Kaynak O. Quo vadis artificial intelligence? Discov Artif Intell. 2022;2(1):4.

Cerniauskas G, Sadia H, Alam P. Machine intelligence in metamaterials design. A preprint. https://doi.org/10.31224/3333.

Vinothkumar J, Karunamurthy A. Recent advancements in artificial intelligence technology: Trends and implications. Quing Int J Multidiscip Sci Res Dev. 2023;2(1):1-1.

Boyce B, Dingreville R, Desai S, Walker E, Shilt T, Bassett KL, et al. Machine learning for materials science: Barriers to broader adoption. Matter. 2023;6(5):1320-3.

Hagendorff T. Blind spots in AI ethics. AI Ethics. 2022;2(4):851-67.

Chin C. Navigating the risks of artificial intelligence on the digital news landscape. Washington, DC: Center for Strategic and International Studies; 2023.

Santoni de Sio F, Mecacci G. Four responsibility gaps with artificial intelligence: Why they matter and how to address them. Philos Technol. 2021;34(4):1057-84.

Naeini EZ, Prindle K. Machine learning and learning from machines. Lead Edge. 2018;37(12):886-93.

Min A. Artificial intelligence and bias: Challenges, implications, and remedies. J Soc Res. 2023;2(11).

Wang L, Song M, Rezapour R, Kwon BC, Huh-Yoo J. People's perceptions toward bias and related concepts in large language models: A systematic review. arXiv preprint arXiv:2309.14504. 2023.