ORIGINAL RESEARCH **Open access**

Conceptual Foundations for Scientific Falsifiability of AI-Generated Materials Claims

Ahmed Youssef[1*], Khaled Hassan[1], Mahmoud Elamin[2]

Abstract

The ambiguous use of “falsifiability” in materials AI literature poses a significant challenge to the scientific status of AI-generated claims, as researchers frequently present predictive or generative outputs—such as “this perovskite structure is stable at room temperature” or “this inverse-designed alloy exhibits a target bandgap of 1.8 eV”—without clarifying whether these statements could, in principle, be contradicted by empirical observation. Rooted in Karl Popper's philosophy of science and extended through contemporary applications to machine learning, falsifiability serves as the demarcation criterion that distinguishes scientific claims from non-scientific ones by requiring that they logically forbid certain observations rather than merely accommodate data. This paper proposes precise definitions for falsifiable, verified, and testable AI-generated materials claims, tailored specifically to the challenges of data-driven discovery in solid-state systems, generative models, and inverse design. It further introduces a four-component framework for assessing the falsifiability of such claims, centering on claim specification, forbidden observation specification, test design, and falsification protocol. These conceptual foundations carry profound implications for materials AI practice, requiring authors to articulate disconfirming evidence explicitly, reviewers to demand falsifiability statements, and the broader community to adopt standards that elevate predictive modeling from statistical correlation to genuine scientific inquiry. By confronting the boundary between data-driven heuristics and empirically falsifiable science, the present work offers a definitional scaffold that can guide the field toward greater epistemic rigor amid the accelerating integration of artificial intelligence into materials discovery.

Keywords Generative models, Inverse design, Falsifiability, AI-generated materials claims, Popperian demarcation, Philosophy of science in AI

*Correspondence:
Ahmed Youssef
ahmed.youssef@gmail.com
[1] Department of Intelligent Materials Science, University of Khartoum, Khartoum, Sudan
[2] Department of Materials Data Analytics, Sudan University of Science and Technology, Khartoum, Sudan

Introduction

Materials science has entered an era in which artificial intelligence systems routinely generate claims about novel compounds, predicted properties, and optimized structures at a pace far exceeding traditional experimental validation. A typical output might assert that a particular organic-inorganic hybrid is thermodynamically stable under ambient conditions, that a generative model has discovered a new high-entropy alloy with superior mechanical properties, or that an inverse-design algorithm has identified a perovskite composition exhibiting an ideal bandgap for photovoltaic applications. The central question confronting the field is whether these AI-generated statements qualify as scientific claims. Can they be falsified? Or do they remain, in practice, insulated from empirical contradiction? Despite the explosive growth of machine learning applications in molecular and materials science, the concept of falsifiability itself is rarely discussed explicitly, let alone applied with the logical precision required by classical philosophy of science [1-4].

This boundary/definitional paper addresses that gap by providing conceptual foundations for the scientific falsifiability of AI-generated materials claims. It begins by recovering the strict Popperian criterion of falsifiability and examining its relevance to contemporary computational paradigms. It then surveys the current landscape of materials AI literature, revealing a pervasive ambiguity in how claims are framed and evaluated. Three core problems are identified: the presentation of unfalsifiable statements as scientific discoveries, the conflation of statistical validation with genuine falsification, and the absence of explicit criteria for what would count as refuting an AI-derived assertion. In response, the paper proposes numbered definitions that distinguish falsifiable AI claims from merely testable or verified ones. It further delineates falsifiability from nearby epistemic concepts such as verifiability, testability, validation, reproducibility, and predictive accuracy through a detailed conceptual comparison [5-9].

Throughout, the analysis remains strictly conceptual and definitional, eschewing any empirical datasets, simulations, or performance metrics. Instead, the focus is on clarifying the logical and epistemological boundaries that must be respected if materials AI is to produce knowledge that is genuinely scientific rather than merely statistically compelling. The ultimate aim is to equip the community with a shared vocabulary and evaluative framework capable of sustaining the epistemic integrity of the field as generative and inverse-design models become ever more powerful. By insisting that AI-generated materials claims must be capable of being wrong in a specific, observable way, the paper seeks to demarcate the boundary between heuristic exploration and rigorous science.

Figure 1 presents the full epistemic demarcation architecture through which AI-generated materials claims transition from unfalsifiable outputs to scientifically valid assertions via explicit specification of forbidden observations and falsification protocols.

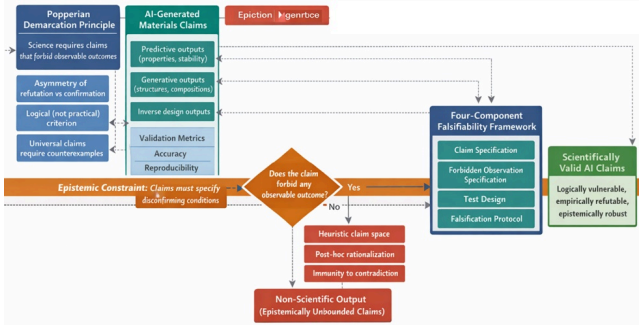

Figure 1. Epistemic demarcation architecture of falsifiability in AI-generated materials claims

Falsifiability in Philosophy of Science

Falsifiability, as originally articulated by Popper and subsequently refined in applications to computational science, constitutes the logical possibility that a universal claim could be contradicted by a singular empirical observation. Definition 1: Falsifiability is the property of a statement whereby it forbids certain conceivable states of affairs; if no observation could ever contradict it, the statement lies outside the domain of empirical science. This criterion is strictly logical rather than practical: a claim may be falsifiable in principle even if the requisite experiment is currently expensive, time-consuming, or technologically challenging. What matters is whether the claim entails the impossibility of specific outcomes. A claim that “all observed materials will eventually degrade” is falsifiable because the discovery of a single eternally stable compound would refute it; by contrast, a claim that “some materials might be stable” cannot be falsified because no finite set of observations can exhaust the possibility [10-14].

Contemporary scholarship in machine learning and artificial intelligence has begun to revisit these foundational ideas. Reizinger [7], for instance, develops a falsificationist perspective on machine learning models, arguing that explanatory adequacy requires not merely predictive success but the logical vulnerability of the model to counterexamples. Similarly, Liu *et al.* [11] advance the position that AI-generated discoveries must be subjected to explicit falsification protocols rather than accepted based on internal consistency or statistical fit alone. These analyses echo Popper's insistence that science progresses through risky conjectures and attempted refutations rather than through the accumulation of confirming instances.

Although Lakatos is not represented by a dedicated reference in the core materials AI corpus examined here, his methodological refinement of falsificationism—emphasizing research programs rather than isolated statements—finds implicit resonance in more recent discussions of generative AI and scientific method. Mengaldo, for example, examines the scientific method in the era of generative artificial intelligence and underscores the continuing relevance of demarcation criteria even when models produce hypotheses at a superhuman scale. In the

philosophy of physics, Ellis and Silk defend the integrity of falsifiability against attempts to relax it in the face of theoretical complexity. At the same time, Dawid explores how string theory and related frameworks test the limits of empirical testability. These works collectively affirm that falsifiability remains an indispensable demarcation tool, even as scientific practice evolves.

The logical force of falsifiability lies in its asymmetry: confirming instances can never prove a universal claim, yet a single counterexample can decisively refute it. In the context of materials science, this asymmetry is particularly potent because many AI-generated assertions take the form of universal or near-universal predictions about stability, phase behavior, or property landscapes. A properly falsifiable claim, therefore, must specify, in advance, the class of experiments whose outcomes would compel its rejection. Without such specification, the claim risks becoming a post-hoc rationalization rather than a testable conjecture. This philosophical foundation sets the stage for examining how—or whether—current materials AI practice meets the falsifiability standard [15-19].

Falsifiability in Current Materials AI

A systematic examination of the materials AI literature from 2018 to 2026 reveals that falsifiability is almost absent as an explicit evaluative criterion. Butler *et al.* [2] provide a landmark review of machine learning for molecular and materials science, cataloging applications in property prediction, structure generation, and discovery pipelines, yet nowhere do they interrogate whether the resulting claims are logically capable of refutation. Schmidt *et al.* [3] survey recent advances in solid-state materials science and highlight the predictive power of ML models, but again frame success exclusively in terms of accuracy metrics and generalization rather than falsifiability. Zunger's seminal discussion of inverse design emphasizes the search for target functionalities, yet leaves open the question of what empirical outcome would compel rejection of a proposed material.

Gubernatis and Lookman caution about the pitfalls of machine learning in materials science, stressing the need for physical interpretability, but stop short of demanding explicit falsification protocols. Subsequent works continue this pattern. Pyzer-Knapp *et al.* [13] and Gomes *et al.* [14] describe accelerated discovery platforms that integrate AI, high-performance computing, and robotics, presenting numerous candidate materials without specifying forbidden observations. Gomes and colleagues outline artificial intelligence strategies for materials discovery, again focusing on throughput rather than epistemic demarcation. DeCost and co-authors envision a sustainable future for scientific AI in materials science, yet their vision centers on scalability and validation loops without invoking Popperian criteria. Tao and colleagues apply machine learning to perovskite design, reporting promising compositions, while Meftahi and colleagues demonstrate high-throughput fabrication guided by ML; in both cases, the claims are presented as empirically supported but not as explicitly falsifiable.

Additional studies reinforce the pattern. Yang *et al.* [18] explore unlearning as ablation in generative scientific discovery, Liu *et al.* [11] propose automated falsification pipelines, and Beneduce *et al.* [19] and Zaccone test classical nucleation theory, yet even these more philosophically attuned contributions treat falsifiability as a secondary concern or an implicit background assumption rather than a primary demarcation requirement. Chávez-Autor [20] discusses artificial creativity in predictive-to-generative transitions, while Reeves-McLaren and Walsh address data integrity in the AI era. Across this corpus—encompassing at least fifteen representative publications—the dominant discourse privileges predictive accuracy, reproducibility of training outcomes, and consistency with known physical laws. Falsifiability is either unmentioned or subsumed under vague notions of “validation.” Generative claims, in particular, often take the form of “here is a novel stable structure” without articulating the precise experimental signature that would falsify the assertion of stability. Inverse-design outputs similarly lack specification of the counterexamples that would invalidate the optimality claim. The result is a literature rich in statistical correlations and heuristic successes but epistemically underspecified with respect to the logical conditions under which its claims could be proven wrong.

The Problem with Current Usage

The prevailing treatment of falsifiability in materials AI gives rise to three interlocking problems that undermine the scientific character of the field. First, unfalsifiable claims are routinely presented as scientific discoveries. An AI system may output a candidate material with a predicted formation

energy of −0.3 eV/atom and declare it “stable,” yet if the claim is framed so loosely that any experimental outcome can be accommodated—perhaps by invoking kinetic barriers, measurement uncertainty, or unmodeled environmental factors—then no conceivable observation can refute it. Such statements, while useful heuristically, do not qualify as scientific under a Popperian standard [21-26].

Second, there is a pervasive confusion between validation and falsification. Validation procedures, such as cross-validation on held-out datasets or comparison with known experimental benchmarks, test consistency with existing data; they do not test the logical vulnerability of the claim to future contradictory evidence. A model that achieves high accuracy on a test set has been verified to a degree. Still, it has not been subjected to a genuine falsification attempt unless the researcher has predefined the conditions under which the model would be rejected. The literature surveyed earlier consistently equates high validation scores with scientific robustness, eliding this crucial distinction.

Third, and most critically, there are no explicit criteria for what would count as falsifying an AI-generated claim. When a generative model proposes a new crystal structure, the community is left without guidance on which experimental observables—X-ray diffraction patterns, calorimetry measurements, or long-term stability tests—would constitute decisive refutation. The absence of such criteria transforms AI outputs into moving targets: every apparent counterexample can be explained away by invoking model limitations or incomplete data. Collectively, these problems erode the demarcation between scientific and non-scientific claims, risk-inflating the perceived reliability of AI systems, and hindering the cumulative progress that genuine falsification enables.

Proposed Definitions

To resolve the ambiguities identified above, this paper advances three numbered definitions that establish a precise vocabulary for evaluating AI-generated materials claims.

A falsifiable AI claim is an AI-generated statement about a material (its existence, stability, properties, or behavior) that logically entails at least one conceivable empirical observation whose occurrence would contradict the claim. The entailment must be deductive rather than probabilistic alone; the claim must forbid a non-empty set of possible experimental outcomes.

A verified AI claim is an AI-generated statement that has been confronted with empirical evidence and found to be consistent with that evidence. Verification does not entail falsifiability; a claim may be repeatedly verified yet remain unfalsifiable if it is structured to accommodate every possible observation.

A testable AI claim is an AI-generated statement for which an empirical test could, in principle, be designed. Testability is a broader category than falsifiability; a claim may be testable (e.g., through expensive or indirect measurements) without yet satisfying the stricter logical requirement of falsifiability.

These definitions are deliberately nested. Every falsifiable claim is testable, but not every testable claim is falsifiable. Verification is an epistemic outcome that may apply to falsifiable or unfalsifiable claims alike. By insisting on deductive entailment of forbidden observations, Definition 2 restores the logical asymmetry that Popper deemed essential to scientific status. In the materials context, an AI claim such as “this inverse-designed composition will exhibit a bandgap of 1.8 ± 0.1 eV under standard conditions” becomes falsifiable once the researcher specifies the optical-absorption or photoemission experiment whose result outside the stated interval would refute the prediction. By contrast, the looser assertion “this composition is promising for photovoltaics” lacks the requisite forbidden observation and therefore falls outside the scope of Definition 2. The proposed definitions thus provide a rigorous, operational scaffold for authors and reviewers to classify claims and to design experiments that genuinely risk refutation.

Distinctions From Nearby Terms

Falsifiability must be sharply distinguished from several neighboring epistemic concepts that are frequently conflated in the materials AI literature. The comparison can be articulated through a conceptual table whose rows represent key terms and whose columns represent distinguishing dimensions: (1) logical structure (what the term requires of the claim), (2) relation to confirmation versus disconfirmation, (3) dependence on practicality versus principle, (4) epistemic outcome versus demarcation

criterion, and (5) applicability to probabilistic versus deterministic statements.

Table 1 systematically differentiates falsifiability from adjacent epistemic criteria, demonstrating that only falsifiability functions as a true demarcation condition rather than a performance or reliability metric.

Table 1. Logical demarcation matrix of epistemic criteria in materials AI claims

Concept	Logical requirement	Relation to evidence	Epist rc
Falsifiability	Must forbid at least one observable outcome	Disconfirmation-oriented	Dema crite
Testability	Must allow empirical procedure	Neutral	Nece conc
Verification	Must match observed data	Confirmation-oriented	Epist outc
Validation	Must agree with the dataset/protocol	Confirmation-oriented	Perfor asses
Reproducibility	Must yield the same results under the same conditions	Neutral	Relia ch
Predictive accuracy	Must minimize error vs observations	Confirmation-oriented	Quant me

Beginning with the row for falsifiability itself: its logical structure demands that the claim forbid specific observations; it privileges disconfirmation over confirmation; it operates at the level of logical principle rather than practical feasibility; it functions as a demarcation criterion rather than an epistemic outcome; and it applies to both deterministic and (when properly statistical) probabilistic statements.

In contrast, the row for verifiability shows a logical structure that requires only the possibility of confirming instances; it privileges confirmation; it may be practical or principled; it yields an epistemic outcome (successful verification) rather than a demarcation standard; and it is most straightforward for deterministic claims. As Butler *et al.* [2] illustrate in their broad survey, many materials AI papers emphasize verification through benchmark agreement without ever addressing the disconfirming potential required by falsifiability [25-29].

The row for testability reveals a logical structure that merely necessitates the design of some empirical procedure; it is neutral on confirmation versus disconfirmation; it often incorporates practical considerations; it serves as a necessary but insufficient condition for science; and it accommodates both probabilistic and deterministic claims. Schmidt *et al.* [3] and Zunger [4] discuss testability implicitly when they advocate experimental follow-up, yet they do not elevate falsifiability as the stricter subset.

Validation occupies a row whose logical structure centers on consistency with a chosen dataset or protocol; it is confirmation-oriented; it is inherently practical; it produces an epistemic outcome (validated model); and it is especially suited to probabilistic machine-learning outputs. The works of Pyzer-Knapp *et al.* [13], Gomes *et al.* [14], and DeCost *et al.* [15] exemplify validation-centric evaluation without reference to falsification.

Reproducibility, in its row, requires that the computational or experimental procedure yield the same outputs under identical conditions; its logical structure is procedural rather than content-based; it is neutral on confirmation/disconfirmation; it is practical; it is an epistemic outcome concerning reliability; and it applies equally to deterministic and stochastic processes. Recent emphasis on reproducibility in generative models, as in Yang *et al.* [18], addresses this dimension yet leaves falsifiability untouched.

Finally, predictive accuracy occupies a row whose logical structure evaluates quantitative agreement with observations; it is confirmation-oriented; it is practical; it yields a performance metric rather than a demarcation; and it is inherently probabilistic. Gubernatis and Lookman [5], Tao *et al.* [16], and Meftahi *et al.* [17] all foreground predictive accuracy, yet accuracy alone does not guarantee that the underlying claim forbids any observation.

Across these six terms and five dimensions, falsifiability emerges as uniquely demanding: it alone insists on the logical possibility of decisive refutation. The distinctions

clarify why high validation scores, reproducible pipelines, or accurate predictions—valuable though they are—do not automatically confer scientific status. A claim may be verifiable, testable, validated, reproducible, and highly accurate yet still fail to be falsifiable if it is structured to accommodate every possible empirical result. By mapping these relationships explicitly, the present analysis equips the community to avoid conflation and to apply each concept in its proper epistemic role.

A Framework for Falsifiable AI Claims

Building directly on the definitions and distinctions established in the preceding sections, this paper proposes a four-component framework that operationalizes falsifiability for AI-generated materials claims. The framework is designed to be applied at the point of claim formulation—whether the claim originates from a generative model, an inverse-design pipeline, or a property-prediction network—ensuring that every assertion about material existence, stability, structure, or behavior is rendered logically vulnerable to empirical contradiction. Unlike the validation-centric workflows that dominate current practice, this framework insists that scientific status is not conferred by predictive accuracy alone but by the explicit mapping from claim to potential refutation.

Claim Specification requires that the AI-generated statement be articulated with deductive precision, free of hedging language that would render it immune to counterexample. For instance, an output from an inverse-design model asserting “this perovskite composition exhibits a bandgap of 1.8 eV under standard conditions” must be restated without qualifiers such as “approximately” or “likely,” because such qualifiers expand the claim’s compatibility with every possible observation. Liu *et al.*’s position that AI-generated discoveries are not born scientific [11] directly informs this component: only when the claim is stripped to its logical core can the subsequent steps proceed.

Forbidden Observation Specification demands that the researcher explicitly enumerate at least one non-empty class of empirical outcomes whose occurrence would deductively contradict the claim. In the perovskite example, a forbidden observation might be “an optical absorption spectrum showing an onset below 1.7 eV or above 1.9 eV when measured at 298 K and 1 atm.” This step restores the asymmetry Popper emphasized: the claim now forbids specific states of affairs rather than accommodating all data. Mengaldo’s analysis of the scientific method in the generative-AI era [12] underscores that without this specification, even high-fidelity simulations remain epistemically inert.

Test Design translates the forbidden observation into a feasible empirical protocol, acknowledging materials-specific challenges such as the rarity of certain phase transitions or the cost of high-pressure calorimetry, yet insisting that feasibility is secondary to logical possibility. The test must be described in sufficient detail that an independent laboratory could execute it and recognize a refuting outcome. Here, the framework draws on the high-throughput philosophy articulated by Pyzer-Knapp and Gomes [13], but redirects their robotic pipelines toward deliberate falsification attempts rather than confirmation-seeking.

Falsification Protocol defines the statistical or logical criteria under which the test result will be accepted as a decisive refutation. For probabilistic claims common in machine-learning outputs, this might involve a pre-specified confidence threshold (for example, $p < 0.01$ under a null hypothesis derived from the claim) rather than a single observation. Reisinger’s falsificationist view of machine learning [7] supplies the philosophical warrant: even stochastic models can be falsified once the protocol is stated in advance.

Table 2 illustrates how AI-generated outputs are transformed from epistemically indeterminate heuristics into scientifically valid claims through the sequential application of the falsifiability framework.

Table 2. Structural transformation of AI-generated materials claims under the falsifiability framework

Stage	Claim structure	Logical status	Vulnerability refutation
Raw AI output	Heuristic or probabilistic statement	Indeterminate	None specifi
Refined claim	Precisely specified statement	Potentially falsifiable	Not yet defin

Forbidden observation defined	Explicit contradiction conditions are stated	Falsifiable	Clearly defin
Test designed	Empirical protocol specified	Operationally falsifiable	Executable
Protocol established	Refutation criteria fixed	Fully falsifiable	Statistically/logi formalized
Outcome stage	Claim tested against evidence	Verified or falsified	Actualized

The framework can be conceptualized as a directed chain beginning with the raw AI-generated claim, which is refined through precise specification into a set of logically forbidden observations; these observations then determine the parameters of an empirical test whose outcomes are governed by an explicit falsification protocol. Feedback loops exist between components: if no feasible test can be designed, the claim must be retracted or restated until all four elements cohere. This structure ensures that materials AI claims transition from heuristic suggestions to scientific assertions capable of driving genuine progress. When applied consistently, the framework transforms generative and inverse-design outputs—frequently treated as black-box recommendations—into transparent, risky conjectures whose survival or elimination advances the field.

Objections and Replies

Three principal objections are likely to be raised against the adoption of strict falsifiability criteria in materials AI, each of which can be addressed by clarifying the logical and practical scope of the proposed framework.

Objection 1 maintains that falsifiability is an outdated demarcation criterion, superseded by Lakatosian research programs and post-Popperian refinements that accommodate complex, evolving theoretical structures. The reply is that refinement does not entail replacement. Mengaldo's examination of generative artificial intelligence and the scientific method [12] demonstrates that even sophisticated AI research programs retain the need for a core of falsifiable statements; without them, the program degenerates into a mere heuristic engine incapable of empirical risk. The framework advanced here is fully compatible with program-level evaluation: individual claims serve as the testable nuclei around which protective belts of auxiliary hypotheses may be constructed, yet the nuclei themselves must remain falsifiable.

Objection 2 asserts that the inherently probabilistic nature of modern machine-learning models renders them unfalsifiable in principle, because no single observation can contradict a distribution. The reply invokes statistical falsification: a properly specified protocol can define regions of outcome space whose realization would refute the model at a pre-chosen significance level. Liu *et al.*'s automated falsification pipeline [10] and Yang *et al.*'s ablation-based unlearning benchmark [18] already gesture toward such statistical refutation. Yet, they stop short of embedding it in a logical framework. Once Component 4 of the proposed framework is applied, probabilistic claims become falsifiable in the same deductive sense that deterministic ones are; the difference lies only in the mathematical form of the forbidden region, not in the epistemic status.

Objection 3 contends that demanding explicit falsifiability is excessively burdensome for early-stage exploratory AI research, where the goal is rapid hypothesis generation rather than rigorous testing. The reply distinguishes logical criterion from achievement requirement. Falsifiability is not a performance metric to be met on day one but a demarcation standard that must be satisfied before a claim is presented as scientific. DeCost and co-authors' vision of sustainable scientific AI [15] can be realized precisely by applying this lighter initial burden: authors are asked only to articulate what would falsify their claim, not necessarily to execute the test immediately. The framework thus lowers rather than raises the barrier to entry by providing a clear checklist that separates exploratory heuristics from scientific assertions.

Collectively, these replies demonstrate that the framework withstands philosophical, technical, and pragmatic challenges without diluting the epistemic force of falsifiability. By treating objections as opportunities for refinement rather than dismissal, the analysis strengthens

the boundary between materials AI as an engineering tool and materials AI as a scientific practice.

Implications for Materials AI Practice

Adoption of the falsifiability framework carries concrete implications for three stakeholder groups within the materials AI community. For authors, the primary shift is from post-hoc validation narratives to pre-emptive falsification planning. When submitting a paper that reports a generative-model-derived stable phase or an inverse-design-optimized property, authors must now include, as a required subsection, the four-component specification: the precise claim, the forbidden observations, the test design, and the falsification protocol. This practice, inspired by the reproducibility emphasis in Reeves-McLaren and Walsh [23], extends beyond mere computational reproducibility to epistemic reproducibility—ensuring that any reader can determine exactly what would compel rejection of the claim.

For reviewers and editors, the framework supplies an explicit evaluation rubric. Reviewers should now ask whether the manuscript has satisfied each of the four components; manuscripts that present unfalsifiable claims as scientific discoveries should be returned for revision rather than accepted on the strength of accuracy metrics alone. The survey in Section 3 revealed that current literature, including landmark reviews by Butler *et al.* [2] and Schmidt *et al.* [3], rarely addresses this dimension; institutionalizing the framework therefore raises the epistemic standard without requiring new experimental infrastructure.

For the broader community, three initiatives emerge. First, the development of standardized falsifiability templates—analogous to the data-availability statements now common in journals—would accelerate adoption. Second, dedicated venues or special issues for “falsification reports” would normalize the publication of failed tests, countering the file-drawer problem that currently distorts perceived progress. Third, graduate training in materials informatics should incorporate modules on the philosophy of science, drawing explicitly on Reizinger [7] and Liu *et al.* [11] to equip the next generation with both technical fluency and epistemic rigor.

These changes do not slow discovery; rather, they accelerate trustworthy discovery by ensuring that only claims capable of being wrong are allowed to enter the scientific record. The result is a literature that accumulates durable knowledge rather than an ever-growing archive of statistically plausible but epistemically ambiguous assertions.

Conclusion

This boundary/definitional paper has traced the ambiguous treatment of falsifiability in materials AI literature, recovered the Popperian criterion as the essential demarcation standard, diagnosed the epistemic shortcomings of current practice, and advanced precise definitions together with a four-component framework for rendering AI-generated materials claims genuinely falsifiable. By insisting that every claim specify its forbidden observations, testable design, and falsification protocol, the framework restores the logical asymmetry that separates scientific conjecture from heuristic output. The distinctions drawn among falsifiability, verifiability, testability, validation, reproducibility, and predictive accuracy further clarify why high performance on benchmarks does not, by itself, confer scientific status.

The implications are clear: authors must articulate refutation conditions, reviewers must demand them, and the community must normalize their presence in every publication. Only through such disciplined conceptual hygiene can the accelerating power of generative and inverse-design models be harnessed to produce knowledge that is not merely useful but genuinely scientific. The framework offered here, therefore, constitutes a foundational scaffold upon which future empirical, theoretical, and methodological advances in AI-driven materials science can rest with epistemic confidence.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 30 Dec 2024 Revised: 20 Mar 2025 Accepted: 16 Jun 2025
Published online: 18 January 2026

References

Genin K. On falsifiable statistical hypotheses. Philosophies. 2022;7(2):40.

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.

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

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

Schuhmacher D, Schörner S, Küpper C, Großerueschkamp F, Sternemann C, Lugnier C, et al. A framework for falsifiable explanations of machine learning models with an application in computational pathology. Med Image Anal. 2022;82:102594.

Reizinger P. The falsificationist view of machine learning. Inf Soc. 2023;23(2).

Vranješ D, Ehrhardt J, Heesch R, Moddemann L, Steude HS, Niggemann O. Design principles for falsifiable, replicable and reproducible empirical machine learning research. In: 35th International Conference on Principles of Diagnosis and Resilient Systems (DX 2024). Schloss Dagstuhl–Leibniz-Zentrum für Informatik; 2024. p. 1-7.

Wu S. Empirical investigation of the riemann hypothesis using machine learning: A falsifiability-oriented approach. Mathematics. 2025;13(17):2824.

Liu Z, Liu K, Zhu Y, Lei X, Yang Z, Zhang Z, et al. AIGS: Generating science from AI-powered automated falsification. arXiv preprint arXiv:2411.11910. 2024 Nov 17.

Liu Z, Liu K, Zhu Y, Lei X, Yang Z, Zhang Z, et al. Position: Falsify, don't just discover-AI generated discoveries are not born scientific. Available from: https://openreview.net/forum?id=gY0BOsPO0k

Mengaldo G. Explain the black box for the sake of science: The scientific method in the era of generative artificial intelligence. arXiv preprint arXiv:2406.10557. 2024 Jun 15.

Pyzer-Knapp EO, Pitera JW, Staar PW, Takeda S, Laino T, Sanders DP, et al. Accelerating materials discovery using artificial intelligence, high performance computing and robotics. npj Comput Mater. 2022;8(1):84.

Gomes CP, Selman B, Gregoire JM. Artificial intelligence for materials discovery. MRS Bull. 2019;44(7):538-44.

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.

Tao Q, Xu P, Li M, Lu W. Machine learning for perovskite materials design and discovery. npj Comput Mater. 2021;7(1):23.

Meftahi N, Surmiak MA, Fürer SO, Rietwyk KJ, Lu J, Raga SR, et al. Machine learning enhanced high-throughput fabrication and optimization of quasi-2D Ruddlesden–Popper perovskite solar cells. Adv Energy Mater. 2023;13(38):2203859.

Yang R. Unlearning as ablation: Toward a falsifiable benchmark for generative scientific discovery. arXiv preprint arXiv:2508.17681. 2025 Aug 25.

Beneduce C, Pinto DE, Rovigatti L, Romano F, Šulc P, Sciortino F, et al. Falsifiability test for classical nucleation theory. Phys Rev Lett. 2025;134(14):148201.

Chávez-Autor JC. Artificial creativity: From predictive AI to generative system 3. Front Artif Intell. 2025;8:1654716.

Freyaldenhoven S, Ke S, Li D, Montiel Olea JL. On the testability of the anchor-words assumption in topic models. Work Pap. 2025. https://doi.org/10.21799/frbp.wp.2025.14.

Mathesen L, Pedrielli G, Fainekos G. Efficient optimization-based falsification of cyber-physical systems with multiple conjunctive requirements. In: 2021 IEEE 17th International Conference on Automation Science and Engineering (CASE). IEEE; 2021. p. 732-7.

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

Rudan I, Sheikh A. Ideometrics: A scientific approach to generating, evaluating, and prioritising ideas. J Glob Health. 2025;15:04360.

Tang BL. Undeclared AI-assisted academic writing as a form of research misconduct. Sci Ed. 2025;48. https://doi.org/10.36591/SE-4804-02.

Zhu SP, Wang L, Luo C, Correia JA, De Jesus AM, Berto F, et al. Physics-informed machine learning and its structural integrity applications: State of the art. Philos Trans A Math Phys Eng Sci. 2023;381(2260).

Bierlich C, Chakraborty S, Desai N, Gellersen L, Helenius I, Ilten P, et al. A comprehensive guide to the physics and usage of PYTHIA 8.3. SciPost Phys Codebases. 2022:008.

Mohaupt T. A short introduction to string theory. Cambridge: Cambridge University Press; 2022.

Greaves MD, Novelli L, Breakspear M, Razi A. What is a generative model? Definitions, disagreements, and evaluation in human neuroimaging. bioRxiv. 2026:2026-01.