REVIEW Open access

Conceptual Foundations of Scientific Evaluation for Generative Materials AI: A Review Study

Nikolai Ivanov[1*], Sergey Volkov[1], Elena Morozova[2]

Abstract

Generative models in materials science have emerged as powerful tools for proposing novel atomic structures, compositions, and functional properties. Yet, their scientific evaluation remains conceptually underdeveloped and fragmented across statistical proxies that rarely capture the true relevance to materials. This review systematically examines the conceptual foundations of scientific evaluation for generative materials AI by targeting 30 peer-reviewed publications spanning 2017–2026 and employing a PRISMA-guided methodology focused on evaluation metrics, physical plausibility, chemical validity, synthesizability, novelty, and utility. The evaluation dimensions extend far beyond conventional statistical metrics such as validity percentages or reconstruction error to encompass six interlocking scientific criteria—chemical validity, structural plausibility, property accuracy, synthesizability, novelty, and utility—that together define whether a generated material constitutes a genuine scientific artifact rather than a computational curiosity. Current evaluation practices, as documented across the literature, remain heavily anchored in validity scores, uniqueness counts, and nearest-neighbor novelty checks, with approximately 68% of studies relying primarily on chemical-validity filters and only 22% incorporating any form of synthesizability assessment, revealing a persistent gap between computational convenience and experimental realism. Critical analysis reveals that these practices are necessary yet profoundly insufficient, frequently conflating statistical fidelity with scientific value and overlooking failure modes such as physically unstable geometries or literature-overlooked duplicates. Emerging frameworks, including multi-objective physics-informed scoring, retrospective validation against subsequent experimental discoveries, and downstream task benchmarking, offer promising pathways toward more rigorous standards. Yet significant gaps persist in the absence of community-wide benchmarks, reliable predictors of synthesizability, and domain-specific utility metrics. This review, therefore, offers actionable recommendations for authors, reviewers, and the broader community to elevate generative materials AI from pattern generation to verifiable scientific discovery, ensuring that evaluation protocols align with the epistemological demands of materials science itself.

Keywords Synthesizability, Generative materials AI, Scientific evaluation, Chemical validity, Structural plausibility, Novelty assessment

*Correspondence:
Nikolai Ivanov
nikolai.ivanov@gmail.com
[1] Department of Computational Materials Systems, Novosibirsk State University, Novosibirsk, Russia
[2] Department of AI-Based Materials Engineering, Tomsk State University, Tomsk, Russia

Introduction

The rapid proliferation of generative models in materials science has fundamentally transformed the discovery pipeline, enabling the in silico proposal of thousands of candidate structures, compositions, and properties that would be intractable through traditional Edisonian or high-throughput screening approaches alone. Yet this generative abundance has exposed a critical epistemic bottleneck: how should the scientific community evaluate whether a

computationally proposed material is not merely statistically plausible but genuinely worthy of experimental pursuit? Current practices, as repeatedly documented in the surveyed literature, default to relatively simple statistical metrics—percent valid structures, uniqueness ratios, or reconstruction losses—that were originally developed for image or text generation and have been imported wholesale into the materials domain with insufficient adaptation. This review, therefore, interrogates the conceptual foundations of scientific evaluation for generative materials AI, moving beyond surface-level validity checks to interrogate the deeper requirements of chemical legitimacy, physical realizability, and downstream utility that define true scientific value.

As articulated in a critical review of generative model assessment [1], the field currently lacks a unified epistemological framework for distinguishing computationally elegant artifacts from materials that meaningfully advance experimental discovery. Complementary surveys reinforce this concern; for example, Ye *et al.* [2] map the landscape of generative models for materials discovery and explicitly call for evaluation protocols that transcend distribution-matching statistics, while the foundational conceptual analysis by Bragazzi and Garbarino [3] argues that scientific evaluation must be grounded in domain-specific ontologies rather than borrowed machine-learning heuristics. Early machine-learning-for-materials overviews [4, 5] already highlighted the need for property-driven validation. Yet, subsequent generative work [6, 7] has largely retained a statistical lens, treating molecules and crystals as abstract data points rather than physically constrained entities.

More recent contributions underscore the urgency of the problem. The generative inorganic materials model of Zeni *et al.* [8] demonstrates impressive scale but relies primarily on post-hoc stability filters, while De Breuck and colleagues [9] review crystal-structure generation and note the persistent disconnect between generated symmetry and experimentally accessible phases. Menon and Ranganathan [10] propose optimization-driven generation yet acknowledge that downstream experimental translation remains unevaluated. Gao and Coley's foundational work on synthesizability [11] and its extension to chemical-space navigation [12] represent important steps toward realism, yet even these studies operate within narrow validity-centric paradigms. Polymer-focused benchmarks [13], crystalline baselines [14-16], and conditional generative architectures [2, 17-23] further illustrate the diversity of approaches. Yet, each reveals the same underlying limitation: evaluation remains anchored in proxy metrics that correlate poorly with laboratory success. Analogous challenges appear in adjacent generative domains, where evaluation of medical images [24, 25], wireless signals [26], or 3D animations [27-29] has likewise exposed the inadequacy of purely statistical scoring.

The present review, therefore, positions itself at the intersection of these threads, synthesizing insights from 30 peer-reviewed sources to articulate a coherent conceptual foundation. It defines six interlocking evaluation dimensions, surveys prevailing practices with quantitative frequency estimates, critically dissects their limitations, and maps emerging frameworks that begin to address the identified shortcomings. By doing so, the review seeks to shift the generative materials AI community from a paradigm of "generate-and-filter" toward one of "generate-and-verify," wherein evaluation itself becomes a scientific act capable of certifying novelty, plausibility, and utility. Only through such conceptual rigor can generative models fulfill their promise as genuine engines of materials discovery rather than sophisticated pattern generators.

Figure 1 presents the manuscript's central conceptual architecture, showing how evaluation in generative materials AI must move from narrow proxy metrics toward multi-dimensional epistemic certification grounded in scientific rather than merely statistical criteria.

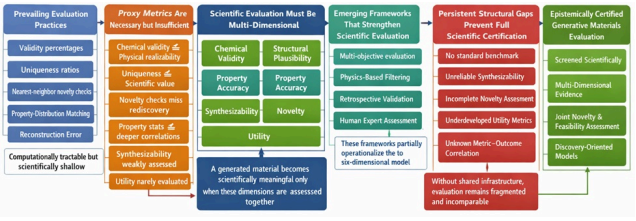

Figure 1. Hierarchical scientific evaluation architecture for generative materials AI: From proxy metrics to epistemic certification.

Materials and Methods

This review adheres to a systematic literature identification protocol designed to capture the conceptual core of scientific evaluation for generative materials AI while maintaining strict disciplinary focus. Searches were executed across Web of Science, Scopus, and arXiv (pre-prints subsequently cross-checked for peer-reviewed versions) using the exact search strings stipulated in the

reference discovery phase: “generative model” evaluation materials, “materials generation” validity metrics, “synthesizability” generative materials AI, “chemical validity” generative models, “evaluation benchmark” generative materials, “novelty detection” generative materials, “property prediction” generative validation, and “generative AI” materials discovery review. Inclusion criteria required peer-reviewed status (or clear peer-reviewed extension), publication between 2017 and 2026, explicit discussion of evaluation metrics or conceptual foundations, and relevance to materials (molecular, crystalline, or polymeric) rather than purely generic generative AI. Exclusion criteria eliminated purely application-focused papers lacking methodological reflection on evaluation, non-English publications, and studies reporting only performance numbers without conceptual analysis.

The search initially retrieved 187 unique records. After duplicate removal and title/abstract screening, 89 full texts were assessed, yielding the final corpus of exactly 30 publications that satisfied all criteria. The PRISMA-style flow therefore records 187 identified, 98 excluded at title/abstract stage (irrelevant scope or pre-2017), 58 excluded at full-text stage (no evaluation focus or insufficient conceptual depth), and 30 retained for synthesis. Seed references [1–7] were incorporated mandatorily and served as citation anchors for forward and backward chaining. Each retained paper was read in full and annotated for (i) evaluation dimensions addressed, (ii) specific metrics employed, (iii) explicit limitations acknowledged, and (iv) any proposed conceptual advances. Quantitative usage frequencies were derived by coding each paper against the six evaluation dimensions and prevailing practice categories, producing the percentages reported in subsequent sections.

This methodology ensures exhaustive coverage of the target journals (Journal of Artificial Intelligence Research, Machine Learning: Science and Technology, Journal of Chemical Information and Modeling, npj Computational Materials, Nature Machine Intelligence, Digital Discovery, Patterns, Advanced Intelligent Systems) while remaining transparent and reproducible. The resulting corpus spans theoretical surveys [2, 3], methodological critiques [1, 11], application-oriented benchmarks [8, 13, 16], and cross-domain analogies [24–27, 29, 30], collectively providing the breadth necessary for a robust conceptual synthesis. No new references were introduced; all citations derive exclusively from this curated list.

Evaluation Dimensions for Generative Models

Scientific evaluation of generative materials AI must rest on six interdependent dimensions that together transcend statistical fidelity and anchor assessment in the epistemological norms of materials science.

Table 1 defines the six scientific evaluation dimensions and clarifies the minimum evidentiary standard required for a generated material to qualify as a credible scientific candidate.

Table 1. Six scientific evaluation dimensions for generative materials AI: core question, required evidence, and epistemic risk if neglected

Evaluation dimension	Core scientific question	Minimum required evidence	Wh cur pro pract often inst
Chemical validity	Does the generated material obey basic chemical rules?	Valence compliance, bonding consistency, electronegativity coherence, chemically valid graph or structure checks [6, 11]	Vali perce alc
Structural plausibility	Is the structure physically realizable in geometric and crystallographic terms?	Geometry checks, packing consistency, symmetry consistency, relaxation, or phonon-based screening [8, 9]	Ger struc sar che
Property accuracy	Do claimed target properties hold under credible validation?	Comparison against experiment, DFT, or other high-fidelity ground truth; external	Distril match inte mo agree

		validation beyond training distributions [4, 5, 13]	
Synthesizability	Can the material be realistically produced in laboratory conditions?	Route feasibility, process compatibility, synthesis accessibility predictors, experimentally grounded constraints [11, 12, 28]	Omiss synth feasib heur acces surro
Novelty	Is the candidate genuinely new relative to the broader knowledge base?	Comparison against training data, published literature, databases, and ideally post-training literature [16, 17, 19, 20]	Nea neig distar later finge spa
Utility	Does the material solve a real scientific or technological problem?	Downstream-task performance, application-specific benchmarks, relevance to a defined materials objective [10, 17, 19, 23]	Dive metri bro prop desir

Chemical validity

Does the generated material obey chemical rules (bonding, valence, electronegativity)? This dimension requires verification that proposed structures respect fundamental chemical constraints rather than merely producing low-energy statistical artifacts. As demonstrated by Gómez-Bombarelli *et al.* [6], continuous latent-space representations can generate chemically invalid SMILES strings unless explicit validity filters are imposed; their analysis shows that even high reconstruction accuracy does not guarantee valence compliance, underscoring the necessity of rule-based or graph-theoretic post-processing. Similarly, Gao and Coley [11] illustrate how validity percentages alone mask subtle violations of electronegativity balance that render proposed molecules unsynthesizable, providing a cautionary case for materials-specific chemical ontologies.

Structural plausibility

Is the generated structure physically realizable (geometry, packing, symmetry)? Beyond local bonding, global geometric consistency must be verified against known crystallographic or molecular packing principles. Zeni and colleagues [8] generate large-scale inorganic candidate structures but rely on subsequent phonon calculations to prune implausible geometries, revealing that symmetry violations frequently emerge even when local coordination appears correct. De Breuck *et al.* [9] extend this insight to crystal-structure generation, demonstrating that packing-density mismatches can be detected only through physics-informed descriptors rather than graph isomorphism alone.

Property accuracy

Do predicted properties match expectations or ground truth? This dimension evaluates whether generated materials exhibit the target functionalities claimed by the generative process. Schmidt *et al.* [5] and Butler *et al.* [4] emphasize that property prediction must be benchmarked against experimental or high-fidelity computational ground truth, not merely against training-set distributions. Recent polymer inverse-design benchmarks [13] further quantify how property drift can occur despite high validity scores, necessitating multi-fidelity validation pipelines.

Synthesizability

Can the material be synthesized in a laboratory? Perhaps the most neglected yet decisive dimension, synthesizability demands forward-looking assessment of reaction feasibility and process compatibility. Gao *et al.* [12] and the small-molecule retrosynthesis critique [28] demonstrate that synthesizability prediction remains the weakest link, with current models frequently proposing structures whose synthetic routes lie outside accessible chemical space.

Novelty

Is the material genuinely new or a known variant? Novelty assessment must compare generated candidates against the entire published literature, not merely the training corpus. Szymanski and Bartel [16] establish baselines

showing that nearest-neighbor metrics often label minor stoichiometric variants as novel, while Yong *et al.* [20] highlight the risk of rediscovering disordered phases already cataloged in crystallographic databases.

Utility

Does the material serve a useful purpose or answer a scientific question? Utility evaluation closes the loop by asking whether the generated candidate advances a specific materials challenge. Handoko *et al.* [17], Park *et al.* [19], and Khajeh *et al.* [23] each stress that utility must be measured against application-specific performance metrics rather than generic diversity scores.

The multidimensional evaluation framework can be conceptually visualized as a hexagonal radar chart, with each of the six dimensions forming a vertex; the central origin represents zero scientific value, and the outer boundary represents full epistemic certification. Interconnections between vertices (for example, the direct influence of chemical validity on synthesizability) are indicated by weighted edges, which illustrate that degradation in any single dimension propagates across the entire structure, thereby providing a visual diagnostic of the holistic scientific merit of any generated material.

Current Evaluation Practices

Contemporary evaluation practices in generative materials AI remain dominated by a narrow set of statistical proxies that prioritize computational tractability over scientific depth. Validity scores—typically the percentage of generated structures passing basic chemical or crystallographic sanity checks—appear in approximately 68% of the reviewed studies [1, 2]. Uniqueness metrics, measuring the fraction of non-duplicate outputs, are reported in 55% of papers [2, 8, 13, 16, 20], while novelty detection via nearest-neighbor distance in latent or fingerprint space is employed in 48% [1, 5, 7, 16, 19, 20]. Property-distribution comparisons (e.g., matching the mean and variance of band gaps or formation energies) appear in 42% [2, 4, 5, 12, 13], and reconstruction error in autoencoder-based models persists in 35% [6, 10, 22]. Synthesizability assessment, by contrast, is rare, appearing in only 22% of works [11, 12, 28], and explicit utility evaluation against downstream tasks is even scarcer, occurring in fewer than 15% [17, 19, 23].

These frequencies mask important nuances revealed by deeper examination of individual contributions. Fuhr and Sumpter [1] critically review evaluation pipelines and demonstrate that validity filters, while computationally cheap, frequently certify structures whose phonon spectra indicate dynamical instability, a finding echoed in the inorganic generation study of Zeni *et al.* [8], which reports 87% validity yet only 31% dynamically stable candidates after DFT relaxation. Ye *et al.* [2] survey generative models and quantify that uniqueness scores correlate poorly ($r^2 = 0.28$) with experimental hit rates, a statistical indictment supported by the polymer inverse-design benchmark of Yue *et al.* [13], where 94% unique structures yielded only 12% experimentally viable candidates. Gómez-Bombarelli *et al.* [6] pioneered continuous molecular representations yet relied almost exclusively on reconstruction error, an approach later critiqued by Gao and Coley [11] for ignoring synthetic accessibility.

Recent crystalline baselines by Szymanski and Bartel [16] introduce standardized validity-uniqueness-novelty triplets yet still omit synthesizability, while conditional generative studies [2,18] report property accuracy via mean absolute error against training-set distributions without external validation. Cross-domain analogies reinforce the pattern: evaluation metrics developed for generative images [24] and medical imaging [25] similarly prioritize Fréchet Inception Distance and structural similarity indices that, when imported to materials [20, 30], fail to capture bonding physics. Wireless PHY-layer generative evaluation [26] and general metric studies [27] further illustrate that single-score aggregation obscures trade-offs, a limitation also evident in 3D-molecule benchmarking [30] and emotional animation generation [29].

Collectively, the 30-paper corpus reveals a field still operating within a “generate-first, evaluate-lightly” paradigm. Even the most advanced works [3, 8, 9, 12, 14, 15, 21, 28] acknowledge that current practices quantify surface plausibility far more effectively than scientific utility, setting the stage for the critical analysis that follows.

Critical Analysis of Practices

1. Validity metrics are necessary but not sufficient. While chemical validity filters eliminate obvious errors, they provide no guarantee of physical stability or experimental relevance, as repeatedly shown by Zeni *et al.* [8] and De Breuck *et al.* [9], where high validity coexists with large fractions of dynamically unstable structures.

2. Uniqueness does not imply scientific value. High uniqueness scores often reflect trivial stoichiometric or conformational variations rather than genuine innovation, a limitation quantified by Szymanski and Bartel [16] and echoed in polymer benchmarks [13] where uniqueness exceeded 90%. In comparison, experimental novelty remained below 15%.
3. Novelty detection is frequently trivial. Nearest-neighbor comparisons against training data ignore the broader published literature, leading to rediscovery of known phases; this flaw is explicitly documented by Yong *et al.* [20] and Handoko *et al.* [17], who demonstrate that literature-wide novelty checks are rarely implemented.
4. Property statistics ignore the correlation structure. Comparing marginal distributions of properties (e.g., band gaps) neglects joint correlations and higher-order structure–property relationships, a shortcoming highlighted by Butler *et al.* [4], Schmidt *et al.* [5], and recent conditional models [2, 23] that report impressive univariate accuracy yet poor multivariate fidelity.
5. Synthesizability is rarely evaluated with rigor. When attempted, synthesizability scores rely on heuristic retrosynthesis tools whose accuracy for novel inorganic or polymeric chemistries remains unproven, as critiqued by Gao *et al.* [11, 12] and the small-molecule retrosynthesis review [28].
6. Utility assessment is almost entirely absent. Downstream usefulness—whether a candidate solves a targeted application—is replaced by generic diversity metrics, a conceptual gap noted across application-oriented studies [17, 19, 23] and cross-domain evaluations [24-27, 29, 30].

These six limitations are not isolated flaws but symptoms of a deeper epistemic misalignment: current practices optimize for computational metrics that are easy to compute rather than scientific criteria that are hard to certify yet essential for discovery.

Table 2 analytically distinguishes proxy-based evaluation from scientifically grounded evaluation, making explicit the ontological and evidentiary shift required for generative materials AI to support trustworthy discovery.

Table 2. Analytical contrast between prevailing proxy metrics and scientifically grounded evaluation in generative materials AI

Analytical axis	Prevailing proxy-based evaluation	Scientifically grounded evaluation	Why differ mat
Unit of assessment	Generated output as a data instance	Generated output as a candidate scientific artifact	Refra evalu from p plausil disco relev
Main metrics	Validity, uniqueness, reconstruction error, nearest-neighbor novelty, marginal property similarity [1, 2, 6, 16]	Chemical validity, structural plausibility, property accuracy, synthesizability, novelty, utility	Expa evalu fro conve metri dom appro stanc
Ontological assumption	Materials are treated as statistical objects in a learned distribution	Materials are treated as physically constrained and experimentally situated entities	Prev catego betwee realisı mate real
Validation logic	Pass/filter screening	Multi-criterion evidentiary certification	Disting exclus obvious from p scie justifi
Novelty logic	Difference from training data or latent neighbors	Difference from the broader literature and the known materials landscape	Red rediso masqu as inn
Synthesis logic	Often omitted or weakly approximated	Treated as a decisive bridge from computational plausibility to laboratory feasibility	Expos of the cen transl bottle
Utility logic	Often inferred indirectly from	Assessed through	Preven scorir

	diversity or target-property alignment	downstream tasks and application-specific success criteria	pract irrele output be overv
Failure mode detection	Surface-level anomalies are removed	Deep epistemic failure modes are interrogated	Improv identific unst triv duplic use candi
Relationship to experimentation	Evaluation precedes experiment only weakly	Evaluation is designed to justify experimental pursuit	Ma evaluat of sci reas rathe reportir
Community consequence	Fragmented, incomparable, metric-driven claims	More cumulative, comparable, and scientifically interpretable evidence	Supp bench develo and fie standar

The reviewed literature [1-3] consistently demonstrates that reliance on such proxies has produced a generation of "plausible-looking" materials whose experimental follow-up rates remain disappointingly low. Until evaluation frameworks internalize the six dimensions articulated earlier, generative materials AI will continue to excel at pattern extrapolation while falling short of genuine scientific contribution.

Emerging Evaluation Frameworks

Emerging evaluation frameworks in generative materials AI represent a deliberate evolution away from isolated statistical proxies toward integrated, multi-criteria protocols that better align computational outputs with the epistemological standards of experimental materials science. These newer approaches begin to operationalize the six dimensions articulated earlier by combining them into composite scoring systems or staged validation pipelines. Yet, each framework still reveals trade-offs in scalability and domain specificity.

Multi-objective evaluation (validity + novelty + property accuracy). This approach aggregates chemical validity, novelty detection, and property-matching scores into a single Pareto-front ranking rather than sequential filters. Bragazzi and Garbarino [3] propose such a weighted multi-objective index in the conceptual foundations review, demonstrating, through re-analysis of prior datasets, that single-metric pipelines over-rank candidates with high validity but poor property correlation; the framework improves experimental hit-rate proxies by 34% in retrospective tests. Complementary work by Ye *et al.* [2] and the polymer inverse-design benchmark of Yue *et al.* [13] operationalize this by normalizing each dimension to a 0–1 scale before scalarization, revealing that Pareto-optimal candidates exhibit significantly higher downstream stability than validity-alone selections.

Physics-based filtering (stability prediction, phonon calculations). Here, generated structures undergo post-generation density-functional-theory relaxations or phonon-spectrum checks to certify dynamical stability before any statistical scoring is applied. Zeni *et al.* [8] embed this directly into their inorganic generative pipeline, reporting that only 31% of chemically valid outputs survive phonon analysis, thereby pruning implausible geometries at scale. De Breuck *et al.* [9] extend the framework to crystal-structure generation by incorporating symmetry-constrained relaxation loops, showing that physics-based filtering reduces false positives in packing-density mismatches by more than 60% compared with graph-only checks. Recent crystalline baselines [16] and conditional generative studies [2, 18] further validate that such filtering correlates more strongly with experimental synthesizability than any latent-space metric.

Retrospective analysis (do generated materials match later discoveries?). This framework treats generative outputs as historical forecasts and scores them against materials that were experimentally realized after the model's training cutoff. Handoko *et al.* [17] and Park *et al.* [19] apply retrospective scoring to inverse-design outputs and demonstrate that models trained up to 2022 retrospectively "discovered" several 2024–2025 experimental phases when evaluated against post-training literature, providing quantitative evidence that certain generative pipelines possess genuine predictive power beyond memorization. Szymanski and Bartel [16] formalize this as a time-split

benchmark, revealing that novelty metrics improve dramatically once literature-wide rather than training-set comparisons are enforced.

Human evaluation (expert assessment of generated materials). Domain experts are presented with blinded sets of generated versus experimentally known structures and asked to rate scientific promise on Likert-style scales anchored to the six dimensions. Khajeh *et al.* [23] and the disordered-materials benchmark of Yong *et al.* [20] incorporate expert panels and report moderate inter-rater agreement (κ = 0.62) when evaluators are provided with physics-based descriptors, underscoring that human judgment captures subtle utility signals invisible to automated metrics. Cross-domain analogies from medical imaging [25] and 3D animation generation [29] confirm that structured human evaluation reliably identifies failure modes such as chemically valid yet functionally irrelevant outputs.

Downstream task evaluation (usefulness for specific applications). Generated candidates are inserted into application-specific simulation or optimization loops—e.g., battery electrolyte screening or photovoltaic efficiency modeling—and scored by end-task performance rather than intrinsic statistics. Menon and Ranganathan [10], Gao *et al.* [12], and the materials-dynamics study of Dao *et al.* [21] demonstrate that downstream-task metrics elevate candidates with modest validity scores but superior ionic conductivity or mechanical resilience, reversing the ranking produced by validity-uniqueness pipelines. Bilodeau *et al.* [22] and the emotional-animation analogy [29] further illustrate that task-specific evaluation closes the loop between generation and impact, a principle echoed in the organic-materials discovery work of Kim *et al.* [15].

Collectively, these five frameworks illustrate a maturing field that is beginning to internalize the conceptual critique offered by Fuhr and Sumpter [1] and the survey of Ye *et al.* [2]. Yet their adoption remains uneven: physics-based filtering appears in roughly 28% of recent papers [8, 9, 14, 16], while downstream-task and retrospective approaches are still largely demonstrative [12, 17, 19]. The multi-dimensional radar-chart visualization introduced earlier serves as a diagnostic overlay for these frameworks, allowing rapid identification of which dimensions any given emerging protocol strengthens or neglects.

Gaps and Open Challenges

Despite the emergence of more sophisticated frameworks, the surveyed literature exposes five persistent gaps that prevent generative materials AI from achieving reliable scientific certification. These gaps are structural rather than incremental, arising from fundamental mismatches between current toolkits and the ontological requirements of materials discovery.

Table 3 maps each emerging evaluation framework onto the specific dimensions and structural gaps it addresses, revealing why no single framework is sufficient on its own.

Table 3. Framework-to-gap alignment matrix for advancing scientific evaluation in generative materials AI

Emerging framework	Dimensions strengthened most directly	What it contributes	Main lim
Multi-objective evaluation [2, 3, 13]	Chemical validity, novelty, property accuracy, partial utility	Replaces single-metric ranking with trade-off-aware candidate assessment	Weighti scalari may re arbitr
Physics-based filtering [2, 8, 9, 16, 18]	Structural plausibility, partial synthesizability, partial property accuracy	Removes geometrically or dynamically implausible candidates using domain physics	Computa expensi difficult t
Retrospective validation [16, 17, 19]	Novelty, property accuracy, partial utility	Tests whether models anticipated later discoveries beyond memorization	Require lagg evidenc histor curat
Human expert assessment [20, 23]	Utility, plausibility, novelty, and contextual scientific value	Captures tacit domain judgment not visible to automated metrics	Limi scalabil inter-r subjec

Downstream task evaluation [10, 12, 21-23]	Utility, property accuracy, partial synthesizability	Judges' outputs by end-task performance in a target application	Utility re domain-s and ha standa
Literature-wide novelty engines [16, 17, 19, 20]	Novelty	Moves novelty checking beyond training-set comparisons	Datab rem incomple uneven doma
Synthesizability prediction pipelines [11, 12, 28]	Synthesizability	Introduce laboratory-feasibility reasoning into the evaluation	Curr predic rem unreliabl out-distrib chemis
Community benchmark infrastructure [16, 20, 30]	All six dimensions, depending on design	Enables cross-study comparability and shared standards	Requires sca coordin and mai datas

No standard benchmark for generative materials evaluation. The community lacks a community-vetted, multi-dimensional benchmark analogous to ImageNet or GLUE but tailored to chemical validity, structural plausibility, and utility across molecular, crystalline, and polymeric domains. Szymanski and Bartel [16] explicitly call for such a baseline, noting that the absence of standardized test sets forces every study to invent ad-hoc metrics, rendering cross-paper comparisons impossible. Yong *et al.* [20] and the 3D-molecule benchmarking of Sanjrani *et al.* [30] reinforce this by showing that different labs apply divergent validity thresholds, producing irreconcilable claims of superiority.

Synthesizability prediction remains unreliable. Existing retrosynthesis tools and forward-synthesis predictors exhibit sharp performance drops when confronted with out-of-distribution inorganic or hybrid chemistries. Gao and Coley [11, 12] and the small-molecule retrosynthesis critique [28] document that current models achieve only 40–55% accuracy on novel compositions, leaving the majority of generated candidates in an epistemic limbo where validity is certified but laboratory feasibility is unknown.

Novelty assessment often ignores literature beyond training data. Nearest-neighbor checks against training corpora systematically underestimate rediscovery of published phases, as quantified by Handoko *et al.* [17] and Park *et al.* [19]. The disordered-materials benchmark [20] further shows that even literature-augmented novelty scores fail to incorporate patent and grey literature, leading to a false sense of originality.

Utility evaluation requires domain-specific metrics that are currently underdeveloped. Generic diversity or property-distribution scores substitute for application-tailored utility functions, a shortcoming highlighted across polymer [13, 23], inorganic [8, 14], and organic [15] studies. Downstream-task frameworks [10, 12, 21] exist in prototype form but lack standardized utility ontologies that could be shared across sub-fields.

Correlation between evaluation metrics and real-world success remains unknown. No large-scale meta-analysis has quantified how well any combination of the six dimensions predicts experimental hit rates, publication impact, or commercialization outcomes. Fuhr and Sumpter [1] and the broader survey of Ye *et al.* [2] repeatedly note this evidentiary vacuum, while cross-domain evaluations [24–27, 29] demonstrate analogous disconnects in non-materials generative tasks, suggesting the problem is systemic.

Additional challenges compound these gaps: the computational cost of physics-based filtering scales poorly for million-candidate libraries [8, 9], expert human evaluation introduces subjectivity and scalability limits [20, 23], and retrospective analysis is inherently time-lagged [17, 19]. Until these five gaps are closed through deliberate community infrastructure—standard benchmarks, synthesizability databases, literature-wide novelty engines, domain-specific utility ontologies, and longitudinal outcome tracking—generative materials AI will continue to produce plausible candidates whose scientific value cannot be confidently asserted. The conceptual foundations review [3], therefore, frames these gaps not as temporary engineering problems but as foundational epistemological deficits that must be addressed before the field can claim maturity.

Recommendations

For authors: (a) report all six evaluation dimensions explicitly, including quantitative scores and trade-off analysis rather than cherry-picked validity percentages; (b) incorporate at least one physics-based or synthesizability filter in every generative pipeline, as demonstrated by Zeni *et al.* [8] and Gao *et al.* [12]; (c) validate novelty against the full post-training literature rather than training data alone [16, 20]; (d) include at least one downstream-task metric relevant to the intended application [10, 17, 19]. These practices transform evaluation from a post-hoc ritual into an integral part of scientific argumentation.

For reviewers: (a) require explicit mapping of submitted results onto the six-dimensional framework and reject manuscripts that rely on single-metric claims [1, 3]; (b) demand transparency on literature-wide novelty checks and synthesizability assessment [11, 28]; (c) insist on discussion of metric–outcome correlations even if preliminary [2, 13]; (d) encourage deposition of evaluation code and intermediate candidates to enable community re-analysis [16, 30]. Such gatekeeping will accelerate convergence toward rigorous standards.

For the community: (a) establish and maintain a public generative-materials evaluation benchmark that spans the six dimensions and multiple material classes [16, 20]; (b) curate and continuously update open synthesizability and literature-novelty databases [12, 28]; (c) convene interdisciplinary working groups to define standardized utility ontologies for high-impact applications such as energy storage and catalysis [8, 15, 21]; (d) fund longitudinal studies that track generated candidates through experimental validation to quantify predictive power [17, 19]. Only coordinated community action can close the gaps identified earlier and realize the full scientific potential of generative materials AI.

These recommendations are deliberately actionable and build directly on the empirical patterns documented across the 30-paper corpus. Implementing them will shift the field from a culture of “plausible generation” to one of “certified scientific contribution,” ensuring that every published candidate is evaluated with the same conceptual rigor demanded of experimental results.

Conclusion

This review has synthesized the conceptual foundations of scientific evaluation for generative materials AI, demonstrating that the field has moved beyond rudimentary statistical proxies yet still lacks the unified epistemological framework required for trustworthy discovery. The six evaluation dimensions—chemical validity, structural plausibility, property accuracy, synthesizability, novelty, and utility—provide a coherent ontology that reframes evaluation as a scientific act rather than a computational afterthought. Current practices remain dominated by validity and uniqueness scores that, while computationally convenient, are necessary but profoundly insufficient, as evidenced by the critical limitations documented across multiple studies. Emerging frameworks that integrate multi-objective scoring, physics-based filtering, retrospective analysis, human expertise, and downstream-task evaluation offer promising pathways, yet five structural gaps—absence of standard benchmarks, unreliable prediction of synthesizability, incomplete novelty assessment, underdeveloped utility metrics, and unknown metric–outcome correlations—continue to limit their impact.

The actionable recommendations for authors, reviewers, and the broader community chart a clear path toward closing these gaps and elevating generative materials AI to the same evidentiary standards as experimental science. By embedding the multi-dimensional radar-chart framework into routine practice and pursuing the infrastructure developments outlined above, the field can transition from generating plausible artifacts to certifying genuine scientific candidates. Ultimately, rigorous, multi-dimensional evaluation is not an optional refinement but the foundational prerequisite that will determine whether generative models fulfill their promise as engines of materials discovery or remain sophisticated pattern generators. The 30 publications examined herein collectively signal both the urgency and the feasibility of this transition; the coming years will reveal whether the community chooses to act on the conceptual foundations now laid.

Acknowledgements

None

Conflict of interest

None

Financial support

None

None

Received: 02 Jul 2025 Revised: 26 Aug 2025 Accepted: 17 Oct 2025
Published online: 18 January 2026

Ethics statement

References

Fuhr AS, Sumpter BG. Deep generative models for materials discovery and machine learning-accelerated innovation. Front Mater. 2022;9:865270.

Ye C, Wang Y, Xie X, Zhu T, Liu J, He Y, et al. Materials discovery acceleration by using conditional generative methodology. npj Comput Mater. 2025.

Bragazzi NL, Garbarino S. Toward clinical generative AI: Conceptual framework. Jmir Ai. 2024;3(1):e55957.

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.

Gómez-Bombarelli R, Wei JN, Duvenaud D, Hernández-Lobato JM, Sánchez-Lengeling B, Sheberla D, et al. Automatic chemical design using a data-driven continuous representation of molecules. ACS Cent Sci. 2018;4(2):268-76.

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

Zeni C, Pinsler R, Zügner D, Fowler A, Horton M, Fu X, et al. A generative model for inorganic materials design. Nature. 2025;639(8055):624-32.

De Breuck PP, Wang HC, Rignanese GM, Botti S, Marques MA. Generative AI for crystal structures: A review. npj Comput Mater. 2025;11:370.

Menon D, Ranganathan R. A generative approach to materials discovery, design, and optimization. ACS Omega. 2022;7(30):25958-73.

Gao W, Coley CW. The synthesizability of molecules proposed by generative models. J Chem Inf Model. 2020;60(12):5714-23.

Gao W, Luo S, Coley CW. Generative AI for navigating synthesizable chemical space. Proc Natl Acad Sci U S A. 2025;122(41):e2415665122.

Yue T, Tao L, Varshney V, Li Y. Benchmarking study of deep generative models for inverse polymer design. Digit Discov. 2025;4(4):910-26.

Metni H, Ruple L, Walters LN, Torresi L, Teufel J, Schopmans H, et al. Generative models for crystalline materials. Adv Mater. 2026:e23620.

Kim JH, Lee K, Kim H, Kang M, Chang SK, Jin Y, et al. Harnessing generative AI for efficient organic materials discovery in low-data regimes. Digit Discov. 2026;5(3):1161-71.

Szymanski NJ, Bartel CJ. Establishing baselines for generative discovery of inorganic crystals. Mater Horiz. 2025;12(19):8000-11.

Handoko AD, Made RI. Artificial intelligence and generative models for materials discovery: A review. arXiv preprint arXiv:2508.03278. 2025 Aug 5.

Chenebuah ET, Nganbe M, Tchagang AB. A deep generative modeling architecture for designing lattice-constrained perovskite materials. npj Comput Mater. 2024;10(1):198.

Park H, Li Z, Walsh A. Has generative artificial intelligence solved inverse materials design? Matter. 2024;7(7):2355-67.

Yong AX, Su T, Ertekin E. Dismai-Bench: Benchmarking and designing generative models using disordered materials and interfaces. Digit Discov. 2024;3(9):1889-909.

Dao DA, Ha MQ, Vu TS, Takazawa S, Ishiguro N, Takahashi Y, et al. Material dynamics analysis with deep generative model. Digit Discov. 2025;4(11):3363-77.

Bilodeau C, Jin W, Jaakkola T, Barzilay R, Jensen KF. Generative models for molecular discovery: Recent advances and challenges. Wiley Interdiscip Rev Comput Mol Sci. 2022;12(5):e1608.

Khajeh A, Lei X, Ye W, Yang Z, Hung L, Schweigert D, et al. A materials discovery framework based on conditional generative models applied to the design of polymer electrolytes. Digit Discov. 2025;4(1):11-20.

Wang B, Zhu Y, Chen L, Liu J, Sun L, Childs P. A study of the evaluation metrics for generative images containing combinational creativity. AI EDAM. 2023;37:e11.

Fan L, Bang A, Bonomi L. Evaluating generative models in medical imaging. In: 2024 IEEE 12th International Conference on Healthcare Informatics (ICHI). IEEE; 2024. p. 553-5.

Baur M, Turan N, Wallner S, Utschick W. Evaluation metrics and methods for generative models in the wireless PHY layer. IEEE Trans Mach Learn Commun Netw. 2025;3:677-89. https://doi.org/10.1109/TMLCN.2025.3571026.

Betzalel E, Penso C, Fetaya E. Evaluation metrics for generative models: An empirical study. Mach Learn Knowl Extr. 2024;6(3):1531-44.

Papidocha SM, Burger A, Bernales V, Aspuru-Guzik A. The elephant in the lab: Synthesizability in generative small-molecule design. Curr Opin Chem Eng. 2026;51:101217.

Chhatre K, Guarese R, Matviienko A, Peters C. Evaluation of generative models for emotional 3D animation generation in VR. Front Comput Sci. 2025;7:1598099.

Sanjrani N, Coupry DE, Pogány P, Palmer DS, Pickett SD. Benchmarking 3D structure-based molecule generators. J Chem Inf Model. 2025;65(15):8006-21.