

ORIGINAL RESEARCH

Open access

Toward Community-Agreed Metrics for Generative Materials Models: A Position on Inverse Design Reporting

Sven Larsson^{1*}, Erik Johansson¹, Anna Nilsson²

Abstract

Generative models for materials inverse design, including variational autoencoders, generative adversarial networks, and diffusion models, have rapidly emerged as powerful tools for proposing crystal structures conditioned on target properties. Yet, despite significant methodological progress, evaluation practices remain fragmented and inconsistent, limiting comparability and obscuring genuine scientific advancement. This review introduces a unified five-dimensional taxonomy—validity, novelty, diversity, property optimization, and stability—and uses it to diagnose systematic distortions in current reporting. Across the literature, metrics are defined heterogeneously, diversity is frequently omitted, stability is selectively assessed, and baseline comparisons are often absent. These inconsistencies inflate reported performance and prevent reliable benchmarking. In response, a minimal reporting standard is proposed, requiring transparent multi-metric evaluation, explicit dataset and split disclosure, baseline-controlled property assessment, distribution-level reporting, and computational-cost transparency. Establishing such community-agreed protocols is essential for transforming generative materials research into a reproducible, comparable, and cumulatively advancing field.

Keywords Inverse design, Generative materials models, Evaluation metrics, Validity, Novelty, Diversity

*Correspondence:

Sven Larsson

sven.larsson@gmail.com

¹ Department of Materials Data Science, Faculty of Engineering, Karolinska Institute, Stockholm, Sweden

² Department of Computational Materials Engineering, Faculty of Technology, Lund University, Lund, Sweden

Introduction

Generative models for inverse design—VAEs, GANs, and diffusion models—promise to revolutionize materials discovery by generating novel crystals with target properties [1–13]. Rather than screening existing libraries, these models directly propose candidate structures conditioned on desired electronic, mechanical, or thermal characteristics [14–17]. Early graph-based forward models, such as the crystal graph convolutional neural networks introduced by Xie and Grossman and the universal graph networks of Chen *et al.*, demonstrated that machine-learned surrogates could predict material properties with near-DFT accuracy [18, 19]. These predictive frameworks naturally evolved into generative counterparts, enabling inverse workflows that map property targets back to atomic configurations.

Yet a critical problem has emerged: evaluation metrics and reporting practices are wildly inconsistent. One paper claims success based on 90 % validity, defined as structures that are physically plausible after basic geometric checks [6]. Another reports success solely through property-conditioning accuracy, measuring how closely generated structures match a target band gap or formation energy [8]. A third emphasizes novelty, claiming that 95 % of outputs lie outside the training distribution [20, 21]. Rarely do papers report validity, novelty, diversity, property optimization, and stability together, making objective comparisons across studies impossible [22–24].

The consequences extend beyond academic inconvenience. Without standardized metrics, it is unclear whether a reported “high-performing” generative model truly advances discovery or merely reproduces known chemistry with minor perturbations [25, 26]. Inconsistencies also hinder the creation of community benchmarks. While datasets such as the Materials Project, Crystallography Open Database, and PerovSketch are widely cited, their usage varies in composition splits, preprocessing, and held-out sets, undermining reproducibility [4, 27, 28].

This review provides a systematic analysis of evaluation practices for generative materials models published [29, 30]. We draw exclusively from the curated reference set to examine how validity, novelty, diversity, property optimization, and stability are currently operationalized. We synthesize reporting patterns, expose inconsistencies, articulate desiderata for community-agreed metrics, and pinpoint gaps that must be closed to enable fair, reproducible comparisons.

Figure 1 maps the manuscript’s central analytical argument by tracing how fragmented metric practices generate comparison failure, obscure scientific progress, and motivate a standardized multi-metric reporting architecture for generative materials inverse design.

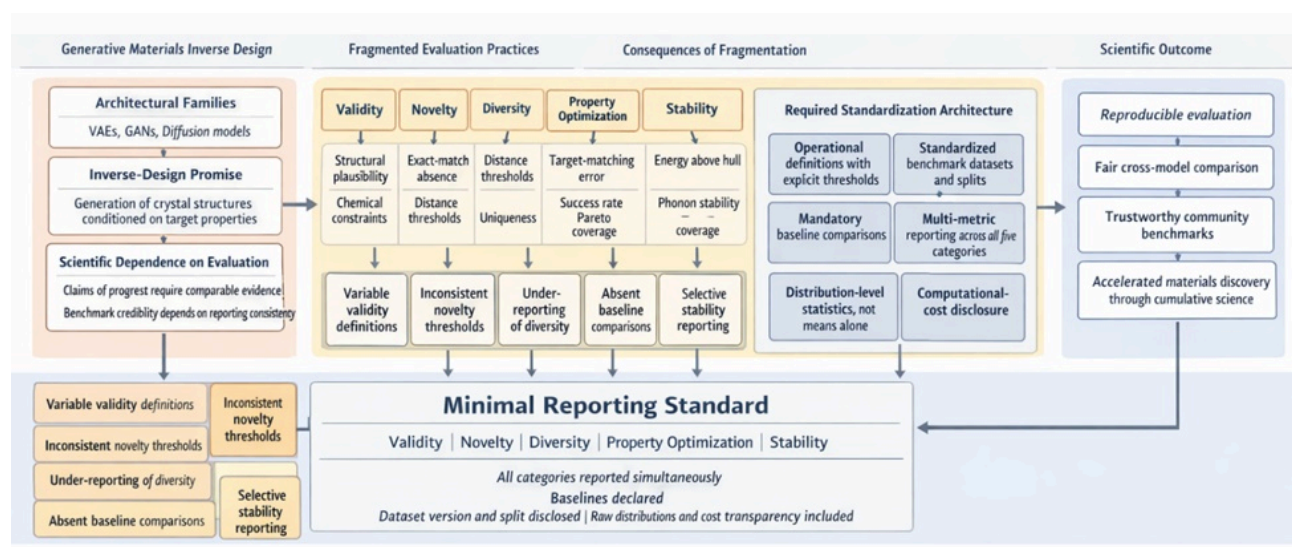


Figure 1. Analytical Architecture of Evaluation Fragmentation and Standardization in Generative Materials Inverse Design

Taxonomy of Evaluation Metrics

A coherent taxonomy is necessary to render the fragmented landscape of generative-materials evaluation analytically tractable. The literature converges on five interrelated dimensions—validity, novelty, diversity, property optimization, and stability—each capturing a distinct constraint governing whether generated structures can meaningfully enter downstream discovery pipelines. Validity establishes a foundational filter by quantifying the proportion of physically plausible outputs, operationalized through structural criteria such as the absence of atomic overlap and bond lengths constrained within 20 % of empirical ranges, chemical constraints including charge neutrality and consistency with electronegativity and oxidation states, and symmetry compatibility with crystallographic space groups [5, 6, 10]. Empirical evidence illustrates the tightening effect of these layered constraints: Kim *et al.* report that while 92 % of GAN-generated crystals satisfy geometric criteria after post-processing, enforcement of oxidation-state consistency reduces this proportion to 68 %, revealing the sensitivity of validity assessments to chemically informed screening [5].

Beyond mere plausibility, the generative process must demonstrate epistemic contribution through novelty, typically defined as the absence of generated structures within the training distribution. This requirement is operationalized through exact-match exclusion, intra-batch uniqueness, and distance-based thresholds in descriptor space [20–22, 25,

26]. However, high nominal novelty can mask latent redundancy; Betzalel *et al.* show that models achieving 95 % exact-match novelty still produce structurally similar outputs when assessed through continuous similarity measures, indicating that discrete criteria alone fail to capture meaningful innovation [21]. A related implication concerns diversity, which reflects the extent to which generated structures span chemical and structural design spaces. Measures based on pairwise distances in graph or fingerprint representations, coverage of known prototypes, and representation across space groups provide complementary perspectives on distributional breadth [4, 9, 23, 28]. In practice, insufficient diversity manifests as dense clustering around limited structural motifs, constraining exploratory capacity despite nominal validity and novelty.

This shift toward distributional evaluation introduces the question of functional relevance, addressed through property optimization metrics that quantify alignment with target performance criteria. Such alignment is typically evaluated through mean absolute error relative to conditioning targets, coverage of Pareto-efficient solutions in multi-objective settings, and the proportion of candidates exceeding predefined thresholds [8, 10, 31, 32]. Yet reported gains often lack contextual grounding; Ma *et al.* demonstrate diffusion-based conditioning on band gaps with mean absolute errors of 0.15 eV, although the absence of baseline comparisons complicates interpretation of model efficacy [8]. Beyond target attainment, the transition from computational candidate to experimentally viable material hinges on stability, which imposes thermodynamic and dynamical feasibility constraints. Formation energy relative to the convex hull, phonon spectra devoid of imaginary modes, and resistance to decomposition collectively operationalize this requirement [9, 10, 31, 33]. The disparity between generative output and physically stable configurations remains substantial, as evidenced by Bastek *et al.*, who observe that although 85 % of diffusion-generated mechanical metamaterials satisfy geometric validity, only 35 % remain within 0.1 eV/atom of the convex hull following relaxation [9].

Taken together, this taxonomy encodes a sequential filtering logic intrinsic to materials discovery: plausibility precedes differentiation, which in turn conditions exploratory breadth, targeted functionality, and ultimately physical realizability. Partial reporting across these dimensions obscures this progression and yields systematically incomplete evaluations [20-22]. The discussion that follows interrogates the extent of such inconsistencies and outlines the methodological standardization required to align evaluative practice with the epistemic demands of AI-driven materials science.

Table 1 consolidates the manuscript's analytical core by linking each evaluation category to its operational sub-metrics, its characteristic reporting distortion, and the precise standardization requirement needed to restore comparability.

Table 1. Crosswalk between Metric Categories, Operational Sub-Metrics, Recurring Reporting Distortions, and Standardization Requirements

Metric category	Core scientific question	Representative sub-metrics	Typical distortion in current literature	Why distortion is analytically serious	Required standardization response
Validity	Is the generated structure physically and chemically plausible at the point of generation?	Structural validity; chemical validity; symmetry validity; post-relaxation validity	Studies collapse distinct checks into a single “validity” score, or report only geometric plausibility while omitting oxidation-state, charge-neutrality, symmetry, or	A reported validity value becomes non-equivalent across papers because the underlying test ladder differs in depth and strictness	Require tiered reporting of geometric, chemical, symmetry, and post-relaxation validity with explicit numerical tolerances

			relaxed-structure criteria		
Novelty	Is the generated structure genuinely absent from prior data rather than a memorized or near-duplicate output?	Exact-match novelty; uniqueness; nearest-neighbor distance; descriptor-threshold novelty	Papers use incompatible novelty definitions, including composition-only checks, exact-match exclusion, or loose similarity filters without threshold disclosure	Apparent innovation can be overstated when near-duplicates are counted as novel	Require simultaneous reporting of exact-match novelty and distance-based novelty, with descriptor space and threshold explicitly declared
Diversity	Does the generated set explore broad chemical and structural regions rather than collapse around a narrow motif family?	Pairwise descriptor distances; prototype coverage; space-group coverage; compositional spread	Diversity is often omitted entirely or asserted rhetorically without quantitative evidence	High validity and novelty can coexist with severe mode collapse, producing little real discovery value	Require descriptor-space distributions, prototype coverage, and coverage-sensitive statistics rather than verbal claims of diversity
Property optimization	Do generated structures actually satisfy conditioning targets better than trivial alternatives?	MAE to target; success rate; threshold attainment; Pareto-front coverage	Property-target performance is reported without baseline controls such as random sampling or nearest-neighbor interpolation	Reported success may reflect dataset priors or easy targets rather than genuine generative capability	Require comparison against random-sampling, interpolation, and property-free baselines for every target metric
Stability	Are generated structures thermodynamically or dynamically plausible enough to matter for downstream realization?	Energy above hull; phonon stability; decomposition energy; relaxed-candidate survival rate	Stability is frequently evaluated only on filtered top candidates or after extensive post hoc pruning	Best-case subset reporting masks the true quality distribution of raw outputs and exaggerates practical usefulness	Require raw-generation stability statistics before filtering, together with the exact stability test and cutoff used

Current Reporting Practices and their Inconsistencies

Despite a well-defined conceptual taxonomy, reporting practices in generative materials research remain fragmented, producing systematic inconsistencies that undermine cross-study comparability. A central point of divergence lies in how validity is operationalized: some studies restrict the notion to geometric plausibility, emphasizing non-overlapping atoms and reasonable bond lengths [5, 6], whereas others impose additional chemical constraints such as charge neutrality and oxidation-state consistency [10], and a smaller subset extends the criterion to post-relaxation stability under DFT [31]. This heterogeneity introduces substantial ambiguity, such that nominally similar validity scores may correspond to

fundamentally different screening regimes. A related instability emerges in the treatment of novelty, where definitions range from strict exclusion of training-set duplicates [4, 28] to more permissive structural similarity thresholds or composition-level comparisons [20, 21, 25], often leading to inflated claims of innovation when looser criteria are adopted. Beyond these definitional inconsistencies, diversity—although conceptually central to exploratory capacity—is frequently omitted or weakly substantiated; even studies reporting high validity and novelty can exhibit pronounced mode collapse, with generated structures concentrated around a narrow set of prototypes [6, 8], sometimes spanning fewer than 20 space groups despite substantially richer training distributions [9].

This unevenness extends to the evaluation of functional performance, where property optimization is typically reported through isolated error metrics without reference to baseline strategies, leaving unclear whether generative models meaningfully outperform trivial alternatives such as random sampling or interpolation [8, 10]. Interpretability is further compromised by selective stability analysis, as many studies restrict thermodynamic evaluation to a small subset of top-performing candidates, thereby overstating overall model quality while neglecting the broader distribution of generated outputs [9, 31]. Compounding these issues, the absence of standardized datasets and evaluation splits—ranging from heterogeneous Materials Project subsets to proprietary or custom-generated data—precludes reproducible benchmarking and renders aggregate comparison infeasible [4, 27, 28]. These patterns are not isolated anomalies but recur systematically across the surveyed literature, yielding a body of work rich in methodological variation yet limited in cumulative insight.

Desiderata for Community-Agreed Metrics

Addressing these structural limitations requires a shift toward rigorously defined and collectively adopted evaluation standards. Central to this effort is the need for operational clarity, whereby each metric is anchored in explicit, reproducible criteria: validity must be decomposed into geometric, chemical, and symmetry constraints; novelty must integrate both exact-match exclusion and distance-based thresholds within a specified descriptor space; and diversity must be tied to clearly defined representations and aggregation procedures [20–23, 34]. This emphasis on definitional precision naturally extends to the establishment of shared benchmark datasets, where fixed training corpora—such as standardized subsets of the Materials Project with predefined splits—and common evaluation tasks enable meaningful comparison across models [4, 27, 28].

Equally important is the systematic inclusion of baseline comparisons, ensuring that reported performance gains are contextualized against simple yet informative reference strategies, including random sampling, interpolation, and unconstrained structure generation [8, 10]. Within this framework, comprehensive reporting across all core dimensions—validity, novelty, diversity, property optimization, and stability—becomes a minimal requirement rather than an optional extension [20–22, 26]. This broadened scope must be accompanied by distributional transparency, where summary statistics are supplemented with full error distributions, structural distance profiles, and stability landscapes to reveal underlying trade-offs and failure modes [9, 31]. Practical considerations further necessitate explicit disclosure of computational cost, including training resources and generation efficiency, to ground methodological advances in realistic deployment constraints [4, 8]. Taken together, these requirements reposition evaluation as a reproducible and comparative enterprise, enabling systematic progress rather than isolated demonstration.

Gaps between Current Practice and Desiderata

Notwithstanding increasing recognition of these standards, a substantive gap persists between current practice and the conditions required for cumulative scientific advancement. Validity definitions remain largely ad hoc, with no agreed-upon hierarchy spanning geometric plausibility, chemical consistency, and post-relaxation stability, thereby sustaining incompatibility across reported results [5, 6, 10]. Novelty assessment similarly lacks depth, as reliance on exact-match criteria obscures the distributional proximity of generated structures to known compounds, limiting insight into true generative innovation [20, 21, 25]. The omission of diversity metrics continues to constrain interpretability, particularly in

studies that assert broad exploration without quantifying structural coverage or representation across crystallographic classes [6, 8, 9].

Parallel limitations are evident in property optimization, where the absence of baseline comparisons leaves the marginal contribution of generative models unresolved [8, 10], and in benchmarking practices, where heterogeneous data splits inhibit reproducibility and prevent the emergence of standardized leaderboards [4, 27, 28]. Stability, although critical for experimental relevance, is frequently relegated to post hoc filtering, with insufficient attention to the intrinsic thermodynamic quality of the full generated distribution [9, 10, 31]. These gaps, taken together, do more than introduce technical noise; they fundamentally constrain the field's ability to transition from isolated methodological advances to a coherent and accumulative science, necessitating coordinated intervention across authorship, peer review, and benchmark design

Table 2 translates the review from diagnosis to prescription by showing how each identified gap produces a specific epistemic consequence and by specifying the minimum reporting obligation required to make future studies benchmarkable.

Table 2. Structured Gap Matrix Linking Present Reporting Practice, Scientific Consequence, and Minimal Reporting Obligations for Generative Materials Studies

Gap in current practice	What is currently missing	Immediate scientific consequence	Higher-order field consequence	Minimum reporting obligation for future studies
Ad hoc validity definitions	No common ladder separating geometric, chemical, symmetry, and relaxed-structure validity	Validity values from different papers cannot be compared directly	Benchmark tables aggregate non-equivalent claims under one label	Report each validity layer separately with explicit thresholds and post-processing rules
Under-specified novelty	Exact-match checks dominate, while distance-to-training-set distributions are usually absent	Novelty claims fail to distinguish memorization from genuine extrapolative generation	The field overestimates innovation and under-detects mode-local exploration	Report both exact-match novelty and nearest-neighbor or descriptor-distance distributions relative to the training set
Routine omission of diversity	Few studies quantify structural spread, prototype coverage, or space-group breadth	Generated libraries may appear productive while remaining narrowly concentrated	Community progress favors models that optimize headline metrics at the expense of exploration	Report at least one descriptor-space diversity statistic and one discrete coverage statistic such as prototype or space-group coverage
Lack of baselines for property optimization	MAE, success rate, or target attainment are reported without random or interpolation baselines	It is unclear whether the model improves on trivial candidate selection strategies	Apparent conditioning progress may reflect dataset bias rather than modeling advance	Report performance against random sampling, nearest-neighbor interpolation, and other low-complexity baselines
No standardized benchmark split regime	Different papers use incompatible datasets,	Cross-paper performance differences cannot be	Leaderboards, meta-analyses, and cumulative	Declare benchmark version, data source, split protocol, preprocessing

	preprocessing pipelines, and train/test splits	attributed confidently to model quality	comparison remain impossible	choices, and held-out evaluation regime in a fixed comparable format
Selective stability reporting	Stability is often reported only for curated top subsets after filtering or relaxation	Bulk-generation feasibility remains unknown	Models appear experimentally promising even when most raw outputs are unusable	Report the full raw-generation stability distribution before any filtering, ranking, or candidate selection
Limited distributional transparency	Means are favored over histograms, quantiles, and tail behavior	Trade-offs and failure modes remain hidden	Reviewers and readers cannot diagnose whether performance is robust or cherry-picked	Report distributions for novelty distances, diversity metrics, property errors, and stability values in addition to summary statistics
Missing computational cost disclosure	Training GPU-hours, inference time, and hardware context are inconsistently reported	Practical reproducibility and deployment realism cannot be assessed	The field risks rewarding expensive models whose marginal gains are not decision-relevant	Report training cost, inference cost per generated structure, hardware specification, and evaluation-time cost for stability checks

Comparison with Other Fields

Generative modeling has reached greater maturity in adjacent disciplines, offering clear lessons for materials inverse design. In molecular generative models for drug discovery, evaluation standards have coalesced around a core set of metrics that directly parallel the taxonomy introduced here. Studies routinely report validity (both geometric and chemical), uniqueness within the generated batch, novelty relative to the training set, and diversity via the Fréchet ChemNet Distance (FCD) [20–22, 25, 26, 29]. Preuer *et al.* demonstrated that FCD provides a distribution-level measure of how well generated molecules occupy realistic chemical space, while Handa *et al.* emphasized the necessity of baseline comparisons against random sampling to avoid over-optimistic claims [20, 22]. These practices emerged after early GAN- and VAE-based papers revealed the same fragmentation now visible in materials science; the community responded by publishing standardized benchmarks and open-source evaluation suites.

Computer-vision generative models offer a second instructive parallel. Metrics such as the Fréchet Inception Distance (FID) and precision/recall curves for diversity have become de-facto standards because they quantify both fidelity to the training distribution and coverage of the target manifold [20, 23, 34, 35]. While image-based scores cannot be transplanted directly to periodic atomic structures, the conceptual framework—joint evaluation of quality, novelty, and diversity—translates cleanly.

Materials inverse design nevertheless faces domain-specific demands absent from molecular or image generation. Periodic boundary conditions, crystallographic symmetry enforcement, and compositional constraints introduce hardness not encountered when generating SMILES strings or pixel arrays [5, 6, 10]. Most critically, thermodynamic and dynamic stability must be verified through DFT relaxation or phonon calculations—computational steps that are orders of magnitude more expensive than the validity checks used in drug discovery [9, 31]. Property conditioning is also more stringent: a generated crystal must simultaneously satisfy multiple interdependent targets (band gap, elastic moduli, thermal conductivity) under the constraint of charge neutrality and space-group symmetry, whereas molecular models often optimize a single scalar bioactivity score [8, 10, 16, 32].

Consequently, the materials field cannot simply import existing protocols; it must extend them. The five-category taxonomy and multi-metric reporting proposed in this review represent exactly such an extension—borrowing the rigor of molecular benchmarks while embedding the stability and periodicity requirements unique to crystalline matter [4, 11, 27, 28]. Without this synthesis, cross-field comparisons remain impossible and progress in generative materials engineering will continue to lag behind its molecular counterpart.

Proposed Minimal Reporting Standard

To eliminate the inconsistencies identified in Section 3 and narrow the gaps outlined in Section 5, generative materials studies must converge on a minimal reporting standard grounded in full transparency across all evaluative dimensions. This framework requires explicit disclosure of validity, novelty, diversity, property optimization, and stability, each accompanied by clearly specified metrics, thresholds, and methodological assumptions. Within validity, geometric plausibility must be quantified through the proportion of structures free from atomic overlap under explicitly stated bond-length tolerances, while chemical consistency must be reported through satisfaction of charge neutrality and oxidation-state constraints. Extending beyond feasibility, novelty must capture both exclusion from the training distribution and distributional distance to nearest neighbors within a declared descriptor space, thereby distinguishing superficial variation from substantive generative departure [25, 26]. A related requirement concerns diversity, which must be operationalized through descriptor-dependent measures such as pairwise distances or structural coverage, with accompanying distributions that reveal concentration or dispersion within the generated ensemble.

This emphasis on distributional characterization carries through to functional evaluation, where property optimization must be contextualized against baseline strategies, including random sampling and nearest-neighbor interpolation, to establish whether observed performance reflects genuine model capability rather than trivial inheritance from the training manifold [8, 10]. Stability assessment, in turn, must reflect the intrinsic thermodynamic profile of the entire generated set, requiring disclosure of formation-energy distributions prior to any post hoc filtering and, where applicable, phonon-based criteria [9, 31]. Beyond metric-specific reporting, reproducibility demands precise documentation of dataset composition, including the version and partitioning of benchmark corpora such as Materials Project subsets or PerovSketch, alongside explicit train/test splits [4, 27, 28]. Computational considerations must also be made visible, with generation efficiency, training cost, and hardware configuration reported to situate methodological advances within practical constraints. The expectation of open evaluation code and hyperparameter disclosure further reinforces the transition from opaque experimentation to verifiable scientific practice.

Adoption of such a standard reconfigures evaluation from an author-dependent narrative into a shared engineering protocol, enabling consistent comparison across studies and facilitating cumulative progress. Journals such as *npj Computational Materials*, *Digital Discovery*, and *Nature Machine Intelligence* are well positioned to institutionalize these requirements through structured submission checklists, while benchmark developers can support compliance through automated pipelines that extract all relevant metrics directly from generated structures. The resulting ecosystem would replace isolated headline figures with multidimensional performance profiles, allowing leaderboards to reflect the full spectrum of generative capability and thereby supporting more reliable model assessment [20–23].

Challenges and Open Questions

Even under a standardized reporting regime, several conceptual and practical challenges remain unresolved. A fundamental difficulty arises in defining validity for crystalline systems, where periodic boundary conditions, symmetry constraints, and long-range electrostatics introduce complexities absent from molecular representations such as SMILES [5, 6, 10]. Although a hierarchical notion of validity spanning geometric, chemical, symmetry, and post-relaxation criteria is conceptually appealing, a community-endorsed formulation has yet to emerge. This ambiguity is compounded by the computational burden of rigorous evaluation, as large-scale DFT relaxation and phonon analysis remain prohibitively expensive for high-throughput assessment, often requiring extensive computational resources [9, 31]. While surrogate

models offer a promising alternative, their integration introduces an additional layer of uncertainty that must itself be systematically benchmarked within the same multi-metric framework.

A related tension concerns the interplay among evaluation dimensions, where improvements in one metric may systematically degrade another. In practice, high validity frequently coincides with reduced novelty due to generative collapse toward the training distribution, obscuring whether models are truly exploring new regions of materials space or merely refining known configurations [20, 21, 23]. Such trade-offs remain difficult to interpret without joint, distributional reporting across all categories. Ambiguity also arises in property-conditioned generation when target specifications extend beyond physically realizable regimes, as in extreme band-gap requests; without explicit acknowledgment of feasibility constraints, reported success rates risk misrepresenting model performance [8, 10, 16]. Finally, the treatment of stability remains incomplete, as conventional reliance on convex-hull proximity at zero temperature neglects dynamic and environmental factors that govern real-world synthesis, including temperature- and pressure-dependent effects [9, 31, 33]. These unresolved issues indicate that while metric standardization is necessary for coherence, it must be complemented by continued methodological development addressing the underlying scientific complexities.

Recommendations for the Community

Achieving meaningful standardization requires coordinated intervention across the research ecosystem. Editorial and review processes must prioritize comprehensive evaluation by mandating multi-metric reporting and baseline comparisons as prerequisites for publication, thereby discouraging reliance on isolated performance indicators and aligning author incentives with methodological rigor [4, 27, 28]. In parallel, benchmark developers play a critical role in establishing shared infrastructure, including fixed dataset partitions, reference implementations of all core metrics, and automated evaluation pipelines that generate unified performance reports, enabling consistent comparison across studies.

For authors, adherence to the proposed framework entails not only reporting all relevant metrics but also embracing transparency in methodological disclosure, including open access to generated structures, evaluation code, and full distributional analyses rather than summary statistics alone. Baseline comparisons should be treated as integral to evaluation rather than supplementary, ensuring that claims of generative capability are grounded in demonstrable improvement over trivial alternatives [8, 10, 20–22, 29]. Through sustained alignment across these stakeholders, the field can transition from a collection of fragmented contributions to a coherent and reproducible body of knowledge, accelerating the identification of genuinely effective generative approaches and strengthening the connection between computational design and experimental realization.

Conclusion

Generative materials models have advanced rapidly, but their evaluation remains too inconsistent to support rigorous comparison or cumulative progress. This review demonstrates that fragmentation across validity, novelty, diversity, property optimization, and stability systematically undermines claims of model performance. When metrics are inconsistently defined, selectively reported, or detached from baseline comparisons, apparent improvements become difficult to interpret and impossible to benchmark reliably.

Progress now depends less on proposing new architectures and more on enforcing shared evaluative standards. A minimal reporting framework—grounded in explicit metric definitions, standardized datasets, baseline comparisons, distributional transparency, and computational-cost disclosure—provides the necessary foundation for reproducible and comparable research. Without such alignment, the field risks continued fragmentation; with it, generative inverse design can evolve into a coherent scientific discipline capable of delivering robust and experimentally relevant discoveries.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 19 Apr 2023

Revised: 31 Jul 2023

Accepted: 19 Nov 2023

Published online: 18 January 2024

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

References

- Harshvardhan GM, Gourisaria MK, Pandey M, Rautaray SS. A comprehensive survey and analysis of generative models in machine learning. *Comput Sci Rev.* 2020;38:100285.
<https://doi.org/10.1016/j.cosrev.2020.100285>.
- Ruthotto L, Haber E. An introduction to deep generative modeling. *GAMM Mitt.* 2021;44(2).
<https://doi.org/10.1002/gamm.202100008>.
- Oussidi A, Elhassouny A. Deep generative models: Survey. In: *Int Conf Intell Syst Comput Vis.* Piscataway (NJ): IEEE; 2018. p. 1-8.
<https://doi.org/10.1109/ISACV.2018.8354080>.
- Chen L, Zhang W, Nie Z, Li S, Pan F. Generative models for inverse design of inorganic solid materials. *J Mater Inform.* 2021;1(1):4.
<https://doi.org/10.20517/jmi.2021.07>.
- Kim S, Noh J, Gu GH, Aspuru-Guzik A, Jung Y. Generative adversarial networks for crystal structure prediction. *ACS Cent Sci.* 2020;6(8):1412-20.
<https://doi.org/10.1021/acscentsci.0c00426>.

Long T, Fortunato NM, Opahle I, Zhang Y, Samathrakris I, Shen C, et al. Constrained crystals deep convolutional generative adversarial network for the inverse design of crystal structures. *NPJ Comput Mater.* 2021;7(1):66.
<https://doi.org/10.1038/s41524-021-00526-4>.

Noh J, Gu GH, Kim S, Jung Y. Machine-enabled inverse design of inorganic solid materials: Promises and challenges. *Chem Sci.* 2020;11(19):4871-81.
<https://doi.org/10.1039/D0SC00594K>.

Ma W, Cheng F, Xu Y, Wen Q, Liu Y. Probabilistic representation and inverse design of metamaterials based on a deep generative model with semi-supervised learning strategy. *Adv Mater.* 2019;31(35):1901111.
<https://doi.org/10.1002/adma.201901111>.

Bastek JH, Kochmann DM. Inverse design of nonlinear mechanical metamaterials via video denoising diffusion models. *Nat Mach Intell.* 2023;5(12):1466-75.
<https://doi.org/10.1038/s42256-023-00762-x>.

Zeni C, Pinsler R, Zügner D, Fowler A, Horton M, Fu X, et al. MatterGen: A generative model for inorganic materials design. *arXiv.* 2023;2312.03687.
<https://doi.org/10.48550/arXiv.2312.03687>.

Lu S, Zhou Q, Chen X, Song Z, Wang J. Inverse design with deep generative models: Next step in materials discovery. *Natl Sci Rev.* 2022;9(8).
<https://doi.org/10.1093/nsr/nwac111>.

Zhang Q, Tao M, Chen Y. gDDIM: Generalized denoising diffusion implicit models. *arXiv.* 2022;2206.05564.
<https://doi.org/10.48550/arXiv.2206.05564>.

Lalena JN, Cleary DA, Duparc OB. Principles of inorganic materials design. 3rd ed. Hoboken: John Wiley & Sons; 2020.

Liu Z, Zhu D, Rodrigues SP, Lee KT, Cai W. Generative model for the inverse design of metasurfaces. *Nano Lett.* 2018;18(10):6570-6.
<https://doi.org/10.1021/acs.nanolett.8b03171>.

Sánchez-Lengeling B, Aspuru-Guzik A. Inverse molecular design using machine learning: Generative models for matter engineering. *Science.* 2018;361(6400):360-5.
<https://doi.org/10.1126/science.aat2663>.

Zhong C, Zhang J, Lu X, Zhang K, Liu J, Hu K, et al. Deep generative model for inverse design of high-temperature superconductor compositions with predicted $T_c > 77$ K. *ACS Appl Mater Interfaces.* 2023;15(25):30029-38.
<https://doi.org/10.1021/acsami.3c00593>.

Cai H, Srinivasan S, Czaplewski DA, Martinson AB, Gosztola DJ, Stan L, et al. Inverse design of metasurfaces with non-local interactions. *NPJ Comput Mater.* 2020;6(1):116.
<https://doi.org/10.1038/s41524-020-00369-5>.

Xie T, Grossman JC. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys Rev Lett.* 2018;120(14):145301.
<https://doi.org/10.1103/PhysRevLett.120.145301>.

Chen C, Ye W, Zuo Y, Zheng C, Ong SP. Graph networks as a universal machine learning framework for molecules and crystals. *Chem Mater.* 2019;31(9):3564-72.
<https://doi.org/10.1021/acs.chemmater.9b01294>.

Preuer K, Renz P, Unterthiner T, Hochreiter S, Klambauer G. Fréchet ChemNet distance: A metric for generative models for molecules in drug discovery. *J Chem Inf Model.* 2018;58(9):1736-41.

<https://doi.org/10.1021/acs.jcim.8b00234>.

Betzalel E, Penso C, Navon A, Fetaya E. A study on the evaluation of generative models. *arXiv*. 2022;2206.10935.
<https://doi.org/10.48550/arXiv.2206.10935>.

Handa K, Thomas MC, Kageyama M, Iijima T, Bender A. On the difficulty of validating molecular generative models realistically: A case study on public and proprietary data. *J Cheminform*. 2023;15(1):112.
<https://doi.org/10.1186/s13321-023-00781-1>.

Xu Q, Huang G, Yuan Y, Guo C, Sun Y, Wu F, et al. An empirical study on evaluation metrics of generative adversarial networks. *arXiv*. 2018;1806.07755.
<https://doi.org/10.48550/arXiv.1806.07755>.

Naeem MF, Oh SJ, Uh Y, Choi Y, Yoo J. Reliable fidelity and diversity metrics for generative models. In: *Proc Mach Learn Res*. 2020;119:7176-85.

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).
<https://doi.org/10.1002/wcms.1608>.

Dollar O, Joshi N, Beck DA, Pfaendtner J. Attention-based generative models for de novo molecular design. *Chem Sci*. 2021;12(24):8362-72.
<https://doi.org/10.1039/D1SC01050F>.

Ren Z, Tian SIP, Noh J, Oviedo F, Xing G, Li J, et al. An invertible crystallographic representation for general inverse design of inorganic crystals with targeted properties. *Matter*. 2022;5(1):314-35.
<https://doi.org/10.1016/j.matt.2021.11.032>.

Yao Z, Sánchez-Lengeling B, Bobbitt NS, Bucior BJ, Kumar SGH, Collins SP, et al. Inverse design of nanoporous crystalline reticular materials with deep generative models. *Nat Mach Intell*. 2021;3(1):76-86.
<https://doi.org/10.1038/s42256-020-00271-1>.

Bian Y, Xie XQ. Generative chemistry: Drug discovery with deep learning generative models. *J Mol Model*. 2021;27(3):71.
<https://doi.org/10.1007/s00894-021-04674-8>.

Mehmood R, Bashir R, Giri KJ. Deep generative models: A review. *Indian J Sci Technol*. 2023;16(7):460-7.
<https://doi.org/10.17485/IJST/v16i7.2296>.

Vlassis NN, Sun W. Denoising diffusion algorithm for inverse design of microstructures with fine-tuned nonlinear material properties. *Comput Methods Appl Mech Eng*. 2023;413:116126.
<https://doi.org/10.1016/j.cma.2023.116126>.

Wang J, Li R, He C, Chen H, Cheng R, Zhai C, et al. An inverse design method for supercritical airfoil based on conditional generative models. *Chin J Aeronaut*. 2022;35(3):62-74.
<https://doi.org/10.1016/j.cja.2021.03.006>.

Medina E, Farrell PE, Bertoldi K, Rycroft CH. Navigating the landscape of nonlinear mechanical metamaterials for advanced programmability. *Phys Rev B*. 2020;101(6):064101.
<https://doi.org/10.1103/PhysRevB.101.064101>.

Maiorca A, Bohy H, Yoon Y, Dutoit T. Objective evaluation metric for motion generative models: Validating Fréchet Motion Distance on foot skating and over-smoothing artifacts. In: *Proc ACM SIGGRAPH Conf Motion Interact Games*. 2023. p. 1-11.
<https://doi.org/10.1145/3623264.3624443>.

Duan Y, Hong Y, Niu L, Zhang L. Few-shot defect image generation via defect-aware feature manipulation. *Proc AAAI Conf Artif Intell.* 2023;37(1):571-8.
<https://doi.org/10.1609/aaai.v37i1.25132>.