ORIGINAL RESEARCH **Open access**

A Conceptual Framework for Detecting Scientific Illusions in Generative Materials Outputs

Alejandro Torres[1*], Miguel Fernandez[1]

Abstract

Scientific illusions represent an undetected yet pervasive failure mode in generative materials AI, where model outputs appear scientifically credible and satisfy superficial visual or computational checks but are fundamentally invalid, thereby consuming experimental resources, misleading research trajectories, and eroding community trust in AI-driven discovery pipelines. A scientific illusion is formally defined as any generative model output that meets surface-level plausibility constraints—such as reasonable bond lengths or predicted low formation energy—while violating deeper physical, chemical, or thermodynamic principles that render the material unrealizable or non-existent in nature. The four primary types of scientific illusions specific to materials generation—structural, property, stability, and novelty—arise through well-characterized mechanisms including spurious correlation learning, mode averaging across disparate training distributions, boundary artifacts in latent representations, and systematic training bias toward plausible but unphysical regions, as synthesized from recent surveys and critical reviews in the field. This paper proposes a five-component conceptual framework for pre-validation detection that operates entirely at the conceptual and computational screening level, enabling researchers to flag illusions before any laboratory commitment. Adoption of this framework promises to transform generative materials practice by shifting evaluation paradigms from post-hoc experimental triage to proactive illusion-aware design, ultimately accelerating credible discovery while safeguarding scientific integrity.

Keywords Scientific illusions, Generative materials AI, Hallucination detection, Spurious correlations in materials modeling, Validation of AI-generated materials, Conceptual framework for illusion detection

*Correspondence:
Alejandro Torres
alejandro.torres@gmail.com
[1] Department of AI-Based Materials Systems, University of Chile, Santiago, Chile

Introduction

Generative models in materials science have rapidly advanced the capacity to propose entirely novel crystal structures, chemical compositions, and functional properties that would be intractable to enumerate manually. Yet this creative power comes with a hidden cost: many of the generated candidates look entirely plausible on first inspection—displaying chemically reasonable coordination environments, aesthetically balanced unit cells, and thermodynamically attractive predicted energies—yet they are scientifically wrong at a fundamental level. These outputs, which we term scientific illusions, consume finite experimental resources when synthesized in the laboratory only to fail basic validation tests; they mislead subsequent modeling efforts by propagating invalid assumptions into downstream simulations; and they erode collective trust in AI-assisted discovery when high-profile “breakthroughs” later prove illusory. The problem is not merely one of occasional error but of a systematic, under-recognized failure mode that current evaluation protocols are ill-equipped to address [1-5].

As Gill and Kell have argued in their critical review, the integration of artificial intelligence into materials discovery has accelerated hypothesis generation far beyond the rate at which traditional validation can keep pace. Generative models, in particular, excel at interpolating and extrapolating within learned manifolds. Yet, the very smoothness that makes them appear powerful also masks the production of outputs that satisfy local statistical regularities while violating global physical laws. Huang *et al.*, in their comprehensive survey of generative models for materials discovery, similarly highlight how diffusion-based and variational autoencoder architectures routinely output structures that pass rudimentary geometric filters yet collapse under phonon analysis or dynamical stability checks. Even foundational machine-learning-for-materials contributions, such as those by Butler *et al.* [4] and Schmidt *et al.* [5], implicitly acknowledge that model predictions can appear credible without being realizable. However, the literature has until recently lacked a dedicated conceptual vocabulary for this phenomenon.

The present work, therefore, articulates “scientific illusions” as a distinct category of generative failure that is neither the generic hallucination studied in large language models nor the simple numerical error common to regression tasks. Scientific illusions occupy a middle ground: they are epistemically seductive because they align with human-trained intuitions of what a “good” material should look like, yet they are ontologically invalid. This paper provides a conceptual framework for detecting such illusions before experimental validation, moving the field from reactive correction to proactive filtering. By grounding the discussion exclusively in existing peer-reviewed literature on generative materials workflows, spurious correlation detection, and validation challenges, the framework offers a structured, citation-supported pathway for researchers to interrogate model outputs at the conceptual level. The sections that follow first define the core construct, delineate its four canonical types, unpack the generative mechanisms that produce illusions, and finally present a five-component detection pipeline that can be applied uniformly across diffusion, autoregressive, and graph-based generative architectures.

Defining Scientific Illusions

Scientific illusions must be distinguished from related but non-identical failure modes if the materials community is to develop targeted detection strategies. Definition 1: A scientific illusion is a generative model output that appears scientifically plausible (i.e., satisfies surface-level geometric, compositional, or energetic constraints typically used in rapid screening) but is fundamentally incorrect because it violates deeper physical principles, cannot be realized in any thermodynamically accessible state, or misrepresents the underlying material system in ways that render it experimentally inaccessible [6-10].

This definition deliberately separates scientific illusions from the broader category of hallucinations familiar in large language model literature. Hallucinations, as discussed by Sun *et al.* in their classification of distorted AI-generated content, typically involve the fabrication of non-existent facts or citations that are easily falsified by external knowledge bases; in contrast, scientific illusions in materials generation are internally consistent within the model’s latent space yet externally inconsistent with established physics. They are also distinct from generic errors—simple mispredictions of a scalar property that can be corrected by retraining—because illusions persist even when the model is otherwise well-calibrated on benchmark datasets. Finally, scientific illusions differ from noise: random invalid outputs are usually flagged immediately by basic sanity checks, whereas illusions survive those checks precisely because they mimic the statistical signature of real materials [11-15].

The epistemological danger lies in the “looks-right-but-is-wrong” character. A generated perovskite-like structure may exhibit ideal octahedral tilting angles and reasonable Goldschmidt tolerance factors yet possess impossible B-site cation ordering that violates Pauling’s rules at the bond-valence level. Because the violation is subtle, it evades casual inspection and even many automated filters currently deployed in high-throughput screening pipelines. As Zunger’s seminal work on inverse design underscores, the search for target functionalities frequently yields candidates that satisfy the optimizer’s objective function while failing unspoken physical constraints—an observation that pre-dates modern generative architectures yet anticipates the illusion problem. Similarly, Gómez-Bombarelli *et al.* [6] demonstrated that

continuous latent representations of molecules can produce chemically plausible yet synthetically unrealizable structures; the same representational smoothness now powers materials-scale generators and simultaneously amplifies illusion risk. By formalizing scientific illusions under Definition 1, this framework equips the community with a precise term that captures the seductive plausibility unique to generative materials outputs.

Types of Scientific Illusions

Type 1: Structural illusion

A structural illusion arises when a generated atomic configuration satisfies superficial geometric criteria—reasonable bond lengths, coordination environments, and symmetry assignments—yet fails to meet deeper crystallographic or chemical bonding constraints required for physical realizability. The resulting structure can appear entirely plausible in static representations, giving the impression of stability while harboring latent inconsistencies that would trigger rapid rearrangement or decomposition under realistic conditions. This mismatch between visual plausibility and physical feasibility is particularly problematic because it propagates through subsequent stages of analysis. When such geometries are passed into density-functional theory workflows, the computations proceed as if the structure were valid, producing electronic or energetic outputs that carry no meaningful correspondence to real materials. The illusion, therefore, operates not only at the point of generation but across the entire discovery pipeline, embedding error into downstream predictions and potentially into the published record [16-19].

Type 2: Property illusion

A distinct yet related failure emerges when predicted functional properties appear credible in isolation but lack any physically grounded connection to the generated structure. In this case, the system produces outputs that align with desirable performance metrics—such as a specific band gap or piezoelectric response—while the underlying atomic arrangement is incapable of supporting those properties through any known mechanism. This disjunction often reflects a decoupling between generative and predictive components within the modeling pipeline, where structural generation and property estimation are learned on separate datasets and integrated only at the level of output evaluation. Under such conditions, the coherence between structure and function becomes contingent rather than enforced. The implications are especially significant in application-oriented contexts, where seemingly promising candidates may guide experimental efforts toward synthesis targets that cannot, in practice, realize the predicted performance, resulting in prolonged cycles of unproductive validation [17-21].

Type 3: Stability illusion

Another form of misrepresentation occurs when thermodynamic indicators suggest stability while the structure remains dynamically unstable. Generative models frequently optimize for scalar energy measures, such as formation energy or distance to the convex hull, without incorporating constraints related to lattice dynamics or vibrational modes. As a consequence, a candidate may appear energetically favorable within static calculations yet exhibit instabilities that would become evident through phonon analysis or finite-temperature simulations. The absence of such checks allows structures with imaginary frequencies or mechanically unstable configurations to pass as viable discoveries. This disconnect introduces substantial risk, as experimental efforts may be directed toward synthesizing phases that cannot persist under ambient conditions, leading to outcomes such as phase separation or amorphization that invalidate the original prediction [22-24].

Type 4: Novelty illusion

A further complication arises in the assessment of novelty, where generated candidates are presented as unprecedented despite being structurally or conceptually derivative. Models trained on extensive chemical datasets can produce compositions that differ nominally from known materials while preserving their underlying topology or bonding motifs. These outputs may evade direct duplication checks due to slight variations in stoichiometry, yet they offer little genuine

departure from established compounds. The illusion of novelty is reinforced when the absence of an exact match in the training data is interpreted as evidence of discovery, rather than prompting a more careful comparison at the level of structural archetypes. This phenomenon carries particular implications for publication and intellectual-property practices, where claims of innovation may be overstated, thereby obscuring the boundary between incremental variation and substantive advancement.

Each type shares the core illusion signature—surface plausibility masking deeper invalidity—yet demands distinct detection lenses, a point elaborated in later sections.

Table 1 differentiates the four scientific illusion types by linking their superficial plausibility signatures to hidden sources of invalidity, diagnostic criteria, and downstream research risks.

Table 1. Analytical matrix linking scientific illusion types to hidden invalidity, detection criteria, and downstream research risk

Scientific illusion type	Surface-level plausibility signature	Deeper source of invalidity	Primary detection criteria	Immediate downstream risk if undetected
Structural illusion	Reasonable bond lengths, acceptable coordination appearance, visually coherent symmetry, and apparently realistic unit cell	Violation of crystallographic rules, forbidden Wyckoff occupation, impossible bonding logic, and chemically incompatible coordination environment	Bond-length violations; coordination-number anomalies; and symmetry/Wyckoff incompatibility	Invalid structure enters DFT or simulation workflows, producing meaningless electronic or mechanical predictions
Property illusion	Attractive target property values, such as a plausible band gap, piezoelectric coefficient, conductivity, or magnetic response	Structure–property decoupling; predicted property lacks causal support from atomic arrangement or bonding configuration	Property inconsistent with structure; violation of known physical bounds; incompatibility with known scaling relations	Application-driven discovery campaigns pursue materials incapable of delivering claimed functionality
Stability illusion	Favorable formation energy, low convex-hull distance, and apparent thermodynamic promise	Dynamic or mechanical instability despite acceptable static energetics; saddle-point rather than realizable minimum	Negative phonon indicators; violation of Born stability criteria; inconsistent phase-diagram placement	Experimental effort is invested in phases that decompose, separate, or amorphize under realistic conditions
Novelty illusion	Output appears new because stoichiometry or labeling differs from known entries	Candidate is database-redundant, trivially substituted, or topologically derivative rather than genuinely novel	Exact/near-exact database match; trivial prototype variation; absence of justified departure from literature precedents	Publication and intellectual-property claims are inflated while genuine innovation is obscured
Cross-type common signature	Output survives superficial screening	Surface plausibility masks ontological invalidity and	Requires a multi-stage rather than a single-metric evaluation	Time, computational budget, laboratory resources, and

	and appears credible to human experts	creates epistemic seduction		community trust are all degraded

Mechanisms of Illusion Generation

Spurious correlation

Generative models can internalize non-causal statistical associations present in training data, such as linking a particular elemental combination with low formation energy solely because those elements co-occur in stable compounds for reasons unrelated to the target property. When the model extrapolates beyond the training distribution, the spurious link produces structures that inherit the correlation without the underlying causality. As Zhang *et al.* demonstrate in their work on robustness to spurious correlations, this mechanism is especially pronounced in high-dimensional composition spaces where dimensionality reduction techniques inadvertently amplify coincidental patterns [25-27].

Mode averaging

Many generative architectures operate by sampling from a smoothed posterior that averages across multiple valid modes in the training distribution. The resulting output can occupy an intermediate, non-physical state—akin to averaging two distinct crystal lattices to produce a hybrid geometry that satisfies neither. This mechanism is documented in several studies of diffusion models for materials, where the denoising process converges to a mean that lies outside any realizable basin. The illusion is particularly convincing because the averaged structure retains fragments of real motifs, lending it visual credibility [25-29].

Boundary artifacts

Latent representations learned by variational or diffusion models possess finite support; near the boundaries of the encoded chemical space, sampling can generate structures whose fractional coordinates or lattice parameters fall into physically impossible regimes (e.g., negative volumes or overlapping atoms). These boundary artifacts survive initial relaxation steps because the model's prior favors plausible-looking geometries, yet they collapse under stricter geometric or symmetry enforcement. Recent analyses of graph-based generators highlight how edge-case sampling systematically produces such artifacts.

Training bias

When training corpora lack sufficient negative examples—structures that are chemically intuitive yet provably unstable—the model never learns the boundaries of impossibility. Consequently, it freely generates candidates in unphysical regions while still assigning them high plausibility scores. This bias is exacerbated by the community's historical emphasis on reporting only successful syntheses, creating an incomplete counterfactual landscape. Merchant *et al.* and Sanchez-Lengeling *et al.* both note that the absence of "failed" examples in public datasets directly contributes to over-confident generation of illusory candidates.

Collectively, these four mechanisms explain why illusions are not rare anomalies but predictable by-products of current generative paradigms, setting the stage for systematic detection.

A Framework for Detection

The proposed conceptual framework for detecting scientific illusions comprises five sequential, mutually reinforcing components that can be applied to any generative output without requiring new experimental data.

Component 1: Plausibility screening

This initial gate applies fast, rule-based filters for obvious geometric and compositional violations—bond-length ranges, coordination-number statistics, charge-balance neutrality, and packing-efficiency thresholds—derived from established crystal-chemistry heuristics. Outputs failing this screen are immediately flagged as low-level illusions; those passing proceed to deeper analysis.

Consistency checking

The second component verifies whether the candidate satisfies fundamental physical laws: conservation of mass and charge, space-group symmetry compatibility, and adherence to thermodynamic inequalities (e.g., positive definite elastic tensors). Conceptual consistency is assessed by mapping the structure onto known symmetry databases and checking for disallowed Wyckoff site occupations or forbidden point-group operations.

Stability prediction

Leveraging lightweight, pre-trained surrogate models for phonon dispersion or elastic-constant estimation, this component flags candidates whose static energy minimum is contradicted by dynamical instability indicators. Because the framework remains conceptual, the emphasis is on the logical integration of fast stability oracles rather than any specific implementation.

Counterfactual test

The framework introduces a perturbation-based diagnostic: small, physically motivated changes (e.g., isotopic substitution, lattice expansion, or elemental doping within allowed ranges) are applied conceptually to the candidate, and the model's response is examined for expected physical behavior. An illusory structure typically exhibits discontinuous or non-monotonic changes that contradict established structure–property gradients, thereby exposing its fragility.

Literature cross-check

The final component performs a conceptual mapping of the generated composition and topology against the corpus of known materials, asking whether the candidate contradicts established knowledge without providing a physically justified rationale for the discrepancy.

Figure 1 presents the conceptual architecture linking illusion-generating mechanisms and illusion types to the five-component pre-validation detection framework and its divergent outcomes.

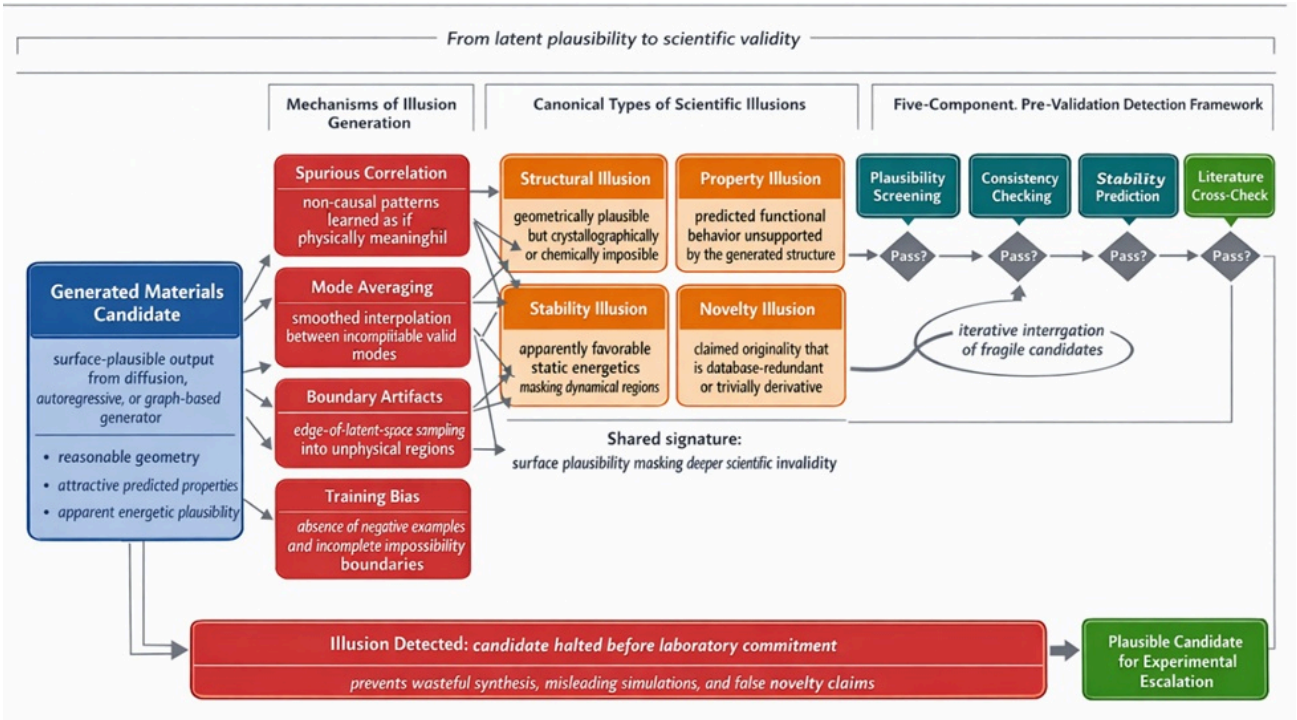

Figure 1. Conceptual architecture of scientific illusion detection in generative materials AI.

The detection framework can be visualized conceptually as a linear pipeline diagram: a generative output enters from the left and flows sequentially through five rectangular stages labeled plausibility screening, consistency checking, stability prediction, counterfactual test, and literature cross-check. Between each stage sits a diamond-shaped decision gate (Pass/Fail); failing any gate routes the candidate to an “illusion detected” terminal box on the right, while passing all gates leads to a “plausible candidate” box. Arrows indicate iterative feedback loops between stability prediction and counterfactual test, underscoring that the components are not strictly independent but can be revisited for deeper interrogation. This pipeline structure ensures that detection is both modular and cumulative, allowing researchers to isolate the precise failure point for any given illusion.

The five components together form a self-contained, pre-validation conceptual scaffold that directly addresses the mechanisms outlined earlier and the illusion types delineated previously.

Table 2 shows how each illusion-generation mechanism is most effectively intercepted by specific components of the five-stage detection framework, clarifying why cumulative screening is necessary.

Table 2. Conceptual alignment between illusion-generation mechanisms and the five detection framework components

Illusion-generation mechanism	How the mechanism produces scientifically seductive outputs	Most vulnerable illusion types	Detection framework component with the highest leverage	Logic of interception within the framework
Spurious correlation	The model learns non-causal associations that mimic physical regularities	Property illusion; novelty illusion; stability illusion	Consistency checking and counterfactual test	Consistency checks expose law-level contradictions, while perturbation response reveals that the model has learned

	without embodying underlying mechanisms			statistical shortcuts rather than causal structure
Mode averaging	Sampling between distinct valid modes generates intermediate structures that look plausible but belong to no realizable basin	Structural illusion; property illusion; stability illusion	Plausibility screening and stability prediction	Screening catches geometrically distorted hybrids, while stability analysis reveals that averaged outputs do not correspond to dynamically viable states
Boundary artifacts	Edge-of-latent-space sampling pushes coordinates, lattice parameters, or topological relations into physically impossible regimes	Structural illusion; stability illusion	Plausibility screening and consistency checking	Early heuristics detect impossible distances or packing, and consistency checks identify broken symmetry or forbidden structural assignments
Training bias	The absence of negative examples leaves the model poorly calibrated at the boundary between plausible and impossible materials	Novelty illusion; stability illusion; property illusion	Literature cross-check and counterfactual test	Cross-checking reveals derivative or unjustified outputs, while counterfactual interrogation shows fragile behavior inconsistent with physically meaningful candidates
Mechanism interaction effect	Multiple mechanisms may jointly produce a candidate that survives single-filter evaluation	All four illusion types	Full five-component pipeline	No single gate is sufficient because illusion robustness often results from overlapping generative distortions that must be intercepted cumulatively

Detection Criteria

Detection criteria operationalize the five-component framework by supplying concrete, conceptually grounded checks tailored to each illusion type. These criteria function as diagnostic lenses rather than empirical thresholds, allowing researchers to interrogate generative outputs at the level of physical and chemical consistency before any laboratory or high-fidelity computational commitment.

For structural illusions, the criteria are: Criterion 1: Bond-length violations. Any generated structure must be examined for interatomic distances that fall outside chemically reasonable ranges established by decades of crystallographic data; a candidate displaying, for instance, a metal–oxygen distance compressed below the sum of ionic radii minus a physically justified tolerance immediately signals an illusion because such compression cannot persist without catastrophic lattice collapse. Criterion 2: Coordination-number anomalies. The local environment around each atom is cross-checked against valence and packing rules; an atom assigned a coordination number incompatible with its oxidation state or electron count—such as a tetrahedral silicon forced into five-fold coordination—reveals the structure as a statistical artifact rather than a viable material. Criterion 3: Symmetry violations. The output geometry is mapped onto allowed space groups and Wyckoff positions; any occupation that breaks translational or point-group symmetry without a compensating distortion (for example, a cubic lattice containing atoms at positions forbidden by the space-group operations) constitutes prima facie evidence of a structural illusion. These three criteria, applied sequentially after Plausibility Screening, catch the majority of geometrically seductive yet crystallographically impossible outputs.

For property illusions, the criteria focus on structure–property decoupling: Criterion 1: Property inconsistent with structure. The assigned functional property (band gap, thermal conductivity, magnetization) is evaluated against the explicit atomic

arrangement; if the predicted value contradicts the orbital overlap or phonon spectrum implied by the geometry, the output is flagged because the generator has learned an independent mapping that ignores causal physics. Violation of known physical bounds. Properties must respect universal constraints such as positive-definiteness of elastic tensors or non-negativity of densities of states; a candidate claiming negative compressibility or imaginary dielectric response without justification fails this criterion and exposes the illusion. Criterion 3: Incompatibility with scaling relations. Established empirical or theoretical scaling laws—such as the inverse relationship between band gap and lattice parameter in oxides—are applied conceptually; any deviation without an accompanying structural rationale indicates the property value is an artifact of mode averaging rather than a genuine prediction.

Stability illusions are diagnosed through dynamical indicators: Criterion 1: Low static energy contradicted by negative phonons. Even when formation energy appears favorable on the convex hull, the presence of imaginary vibrational modes (conceptually checked via fast surrogate models) demonstrates that the structure sits at a saddle point rather than a minimum, rendering it dynamically unstable. Criterion 2: Violation of mechanical stability criteria. Elastic constants must satisfy Born stability conditions (positive eigenvalues of the stiffness matrix); a candidate whose computed moduli yield negative values under infinitesimal strain fails this test and is classified as a stability illusion. Criterion 3: Inconsistent phase-diagram placement. The candidate's composition and energy are compared conceptually against known phase boundaries; placement inside a two-phase region without a stabilizing entropy term or pressure dependence signals that the apparent stability is illusory.

Novelty illusions require database-aware scrutiny: Criterion 1: Exact or near-exact match to known database entries. The generated stoichiometry and topology are conceptually hashed and compared against established repositories; identity or trivial substitution (for example, isovalent doping at a symmetry-equivalent site) reveals the output as non-novel despite surface differences in labeling. Criterion 2: Trivial variant of documented prototypes. If the structure is a simple permutation of a known archetype (perovskite, spinel, layered double hydroxide) without introducing new topological motifs or electronic features, it fails the novelty criterion. Criterion 3: Absence of justified deviation from literature precedents. Any claimed novelty must be accompanied by a physically motivated explanation for departing from prior art; lack of such justification—coupled with high structural similarity—flags the output as a novelty illusion.

When applied together within the five-component pipeline, these criteria transform detection from an ad-hoc exercise into a repeatable conceptual protocol, directly countering the mechanisms of spurious correlation, mode averaging, boundary artifacts, and training bias identified earlier.

Relation to Existing Concepts

Scientific illusions share surface similarities with several well-documented phenomena in artificial intelligence, yet possess distinct epistemological and materials-specific characteristics that justify their separate treatment. The relation to hallucinations, for instance, is one of partial overlap: while hallucinations in large language models involve the fabrication of non-existent facts or citations, as classified by Sun *et al.* [18], scientific illusions in generative materials outputs maintain internal consistency within the model's learned manifold while violating external physical reality. This distinction matters because hallucination-mitigation strategies focused on factual grounding (retrieval-augmented generation) cannot be transplanted wholesale to materials generators, where the "fact" in question is a physical law rather than a textual assertion.

Spurious correlations, extensively analyzed by Zhang *et al.* [19] and Zhou *et al.* [20], constitute a primary generative mechanism rather than the illusion itself; the illusion emerges only when the spurious pattern survives plausibility screening and produces a candidate that appears valid to domain experts. Overfitting, a concept familiar from Butler *et al.* [4] and Schmidt *et al.* [5], describes the model's excessive fidelity to training data but does not capture the seductive plausibility that allows illusory outputs to evade even well-calibrated validation pipelines. Reward hacking, discussed implicitly in the context of inverse design by Zunger [7] and Sanchez-Lengeling *et al.* [28], occurs when the generator

optimizes for proxy objectives such as low predicted energy or high similarity to training examples, thereby producing candidates that satisfy the reward function yet fail deeper physical tests.

The framework, therefore, positions scientific illusions as a higher-order failure mode that subsumes elements of hallucinations, spurious correlations, overfitting, and reward hacking while remaining irreducible to any single concept. By distinguishing illusions through Definition 1 and the four-type taxonomy, the present work equips the community with a vocabulary that bridges the general AI literature—represented by Merchant *et al.* [27] and Pyzer-Knapp *et al.* [10]—with materials-specific epistemology. This conceptual clarification prevents the misapplication of generic debiasing techniques and instead calls for illusion-specific diagnostics that operate at the intersection of representation learning and physical law enforcement.

Implications for Materials AI Practice

Adoption of the scientific-illusion detection framework necessitates targeted changes across three stakeholder groups. For authors the implications are threefold: first, every generative study must report the outcome of the five-component pipeline and the specific detection criteria applied to each proposed candidate; second, manuscripts should include explicit counterfactual test results demonstrating that accepted candidates respond to perturbations in physically expected ways; third, validation protocols must incorporate multiple independent stability oracles rather than relying on a single energy model. These practices, aligned with the data-integrity emphasis of Reeves-McLaren *et al.*, elevate the evidentiary standard without imposing new experimental burdens.

Reviewers are likewise called to evolve their evaluation lens: they should routinely request illusion-detection summaries and probe any plausible-looking output that lacks documented passage through all five framework components; questions about symmetry violations, dynamical stability, and database cross-checks become standard rather than optional. Such scrutiny, building on the ethical considerations articulated by Tortora [25] and the scientist's guide of Odah *et al.* [16], protects the literature from the propagation of seductive but invalid claims.

For the broader community, the framework implies the development of shared resources: illusion-specific benchmarks that compile known historical cases of structural, property, stability, and novelty illusions; open-source toolkits implementing the conceptual criteria as modular filters compatible with common generative architectures; and curated case-study repositories that document both successful detections and instructive failures. These initiatives, echoing the call for robust validation in Gomes [1] and Fuhr and Sumpter [2], will foster an illusion-aware culture in which generative models are evaluated not only on creativity and diversity but on epistemic reliability. Collectively, these practice shifts shift the field from post-hoc experimental triage toward proactive filtering, conserving resources, and restoring confidence in AI-accelerated materials discovery.

Conclusion

This paper has articulated scientific illusions as a distinct and previously under-recognized failure mode in generative materials AI—outputs that satisfy surface plausibility yet violate deeper physical principles—and has furnished a complete conceptual framework for their detection. From the formal definition 1 through the four illusion types, four generative mechanisms, five-component detection pipeline, and type-specific criteria, the framework offers a modular, citation-grounded scaffold that researchers can apply uniformly across diffusion, autoregressive, and graph-based generators. By distinguishing illusions from hallucinations, spurious correlations, overfitting, and reward hacking, the work clarifies their unique epistemological status while relating them productively to existing literature.

The practical implications outlined for authors, reviewers, and the community chart a clear pathway toward illusion-aware generative materials practice. Ultimately, the framework calls for a new standard: illusion detection must become an integral, transparent step in every generative workflow, ensuring that only candidates passing all five components

advance to experimental validation. Such vigilance will prevent the wasteful expenditure of laboratory resources on seductive but unrealizable structures, safeguard the integrity of published discoveries, and accelerate the genuine contribution of artificial intelligence to materials science. The field now possesses both the vocabulary and the conceptual tools to move beyond the era of undetected illusions toward one of reliable, epistemically robust discovery.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 09 Mar 2024 Revised: 01 Jun 2024 Accepted: 23 Jun 2024
Published online: 18 July 2024

References

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

Fuhr AS, Sumpter BG. Deep generative models for materials discovery and machine learning-accelerated innovation. Front Mater. 2022;9:865270.
https://doi.org/10.3389/fmats.2022.865270.

Gomez-Villa A, Martín A, Vazquez-Corral J, Bertalmío M, Malo J. On the synthesis of visual illusions using deep generative models. J Vis. 2022;22(8):2.

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.

Otyepka M, Pykal M, Otyepka M. Advancing materials discovery through artificial intelligence. Appl Mater Today. 2025;47:102981.

Roy M, Lambard G. Quantum-aware generative AI for materials discovery: A framework for robust exploration beyond DFT biases. arXiv preprint arXiv:2512.12288. 2025 Dec 13.

Pyzer-Knapp EO, Manica M, Staar P, Morin L, Ruch P, Laino T, et al. Foundation models for materials discovery–current state and future directions. npj Comput Mater. 2025;11(1):61.

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

Aboutalebi SH. Ensuring data integrity in AI-driven materials science: Why f-sum rules and Kramers-Kronig relations matter. Nanoscale Adv Mater. 2025;2(1):10-5.

Takahara I, Mizoguchi T, Liu B. Accelerated inorganic materials design with generative AI agents. Cell Rep Phys Sci. 2025;6(12):103019.

Kochkov D, Smith JA, Alieva A, Wang Q, Brenner MP, Hoyer S. Machine learning–accelerated computational fluid dynamics. Proc Natl Acad Sci U S A. 2021;118(21):e2101784118.

Pan Y, Hou H, Pei X, Zhao Y. Feature purify: An examination of spurious correlations in high-entropy alloys. Mater Des. 2024;239:112785.

Odah M. Artificial intelligence meets drug discovery: A systematic review on AI-powered target identification and molecular design. Basel: Preprints.org; 2025 Mar 12. https://doi.org/10.20944/preprints202503.0912.v1.

Belkina M, Daniel S, Nikolic S, Haque R, Lyden S, Neal P, et al. Implementing generative AI (GenAI) in higher education: A systematic review of case studies. Comput Educ Artif Intell. 2025;8:100407.

Sun Y, Sheng D, Zhou Z, Wu Y. AI hallucination: Towards a comprehensive classification of distorted information in artificial intelligence-generated content. Humanit Soc Sci Commun. 2024;11(1):1-4.

Zhang LH, Ranganath R. Robustness to spurious correlations improves semantic out-of-distribution detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press; 2023;37(12):15305-12.

Zhou X, Wei F, Duan L, Yao A, Li W. The devil is in the spurious correlations: Boosting moment retrieval with dynamic learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE; 2025:20981-90.

Mao C, Cha A, Gupta A, Wang H, Yang J, Vondrick C. Generative interventions for causal learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE; 2021:3947-56.

Nam J, Mo S, Lee J, Shin J. Breaking the spurious causality of conditional generation via fairness intervention with corrective sampling. arXiv preprint arXiv:2212.02090. 2022 Dec 5.

Ghiurău D, Popescu DE. Distinguishing reality from AI: Approaches for detecting synthetic content. Computers. 2024;14(1):1.

Nightingale SJ, Farid H. AI-synthesized faces are indistinguishable from real faces and more trustworthy. Proc Natl Acad Sci U S A. 2022;119(8):e2120481119.

Tortora L. Beyond discrimination: Generative AI applications and ethical challenges in forensic psychiatry. Front Psychiatry. 2024;15:1346059.

Gray KL, Davis JP, Bunce C, Noyes E, Ritchie KL. Training human super-recognizers' detection and discrimination of AI-generated faces. R Soc Open Sci. 2025;12(11):250921.

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

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

Isayev O, Oses C, Toher C, Gossett E, Curtarolo S, Tropsha A. Universal fragment descriptors for predicting properties of inorganic crystals. Nat Commun. 2017;8(1):15679.