

ORIGINAL RESEARCH

Open access

The Illusion of Generalizability in Foundation Models for Materials Science: A Critique of Zero-Shot Property Prediction

Ethan Wright¹, Chloe Bennett^{1*}, Jack Turner²

Abstract

Foundation models promise a transformative advance in materials science by enabling zero-shot property prediction across diverse chemistries, crystal structures, and physical properties, ostensibly eliminating the need for property-specific labelled datasets and expensive first-principles calculations. Proponents claim that models pre-trained on millions of structures can directly predict formation energies, band gaps, elastic moduli, thermal conductivities, and other attributes for entirely unseen materials without task-specific fine-tuning. This critique demonstrates that such claims largely reflect an illusion of generalizability. Reported zero-shot performance is systematically inflated by six pervasive artefacts: (1) test-set leakage of crystal structures or near-identical analogues from the pre-training corpus, (2) exploitation of strong inter-property correlations, (3) evaluation within the interpolation regime of the pre-training distribution, (4) benchmark bias inherent in widely reused datasets such as the Materials Project, (5) inconsistent pre-processing, data splits, and reporting practices, and (6) the absence of rigorous elementary baselines. Analysis of recent foundation-model literature (2023–2026) shows that zero-shot accuracies frequently collapse once these confounds are controlled and often fail to surpass simple statistical predictors such as k-nearest-neighbour regression on embeddings or linear models based on property correlations. The illusion is particularly acute in materials science due to the repeated reuse of identical structures across property databases and the dense network of physical correlations among computed quantities. Without stringent controls, overstated zero-shot claims risk misdirecting research resources, inflating expectations, and undermining trust in AI-driven materials discovery. This paper provides a diagnostic framework for identifying these artefacts, marshals supporting evidence from the contemporary literature, and proposes a six-point evaluation standard for credible zero-shot claims. Adoption of these practices is essential if foundation models are to deliver genuine advances in out-of-distribution generalization rather than repackaged statistical regularities.

Keywords Materials science, Foundation models, Zero-shot property prediction, Generalizability illusion, Test set leakage, Property correlation

*Correspondence:

Chloe Bennett
chloe.bennett@gmail.com

¹ Department of Computational Materials Science, Faculty of Engineering, University of Leeds, Leeds, United Kingdom

² Department of Materials Data Engineering, Faculty of Technology, University of Sheffield, Sheffield, United Kingdom

Introduction

Foundation models—large models pre-trained on broad data and adapted to downstream tasks—have been enthusiastically extended to materials science [1–4]. A model pre-trained on millions of crystal structures can, according to current claims, predict a property it was never explicitly trained on, or so the literature asserts. This critique argues that these claims create an illusion of generalizability.

Figure 1 maps the manuscript’s core argument by showing how headline zero-shot claims are progressively undermined by six distinct but cumulative sources of analytical illusion.

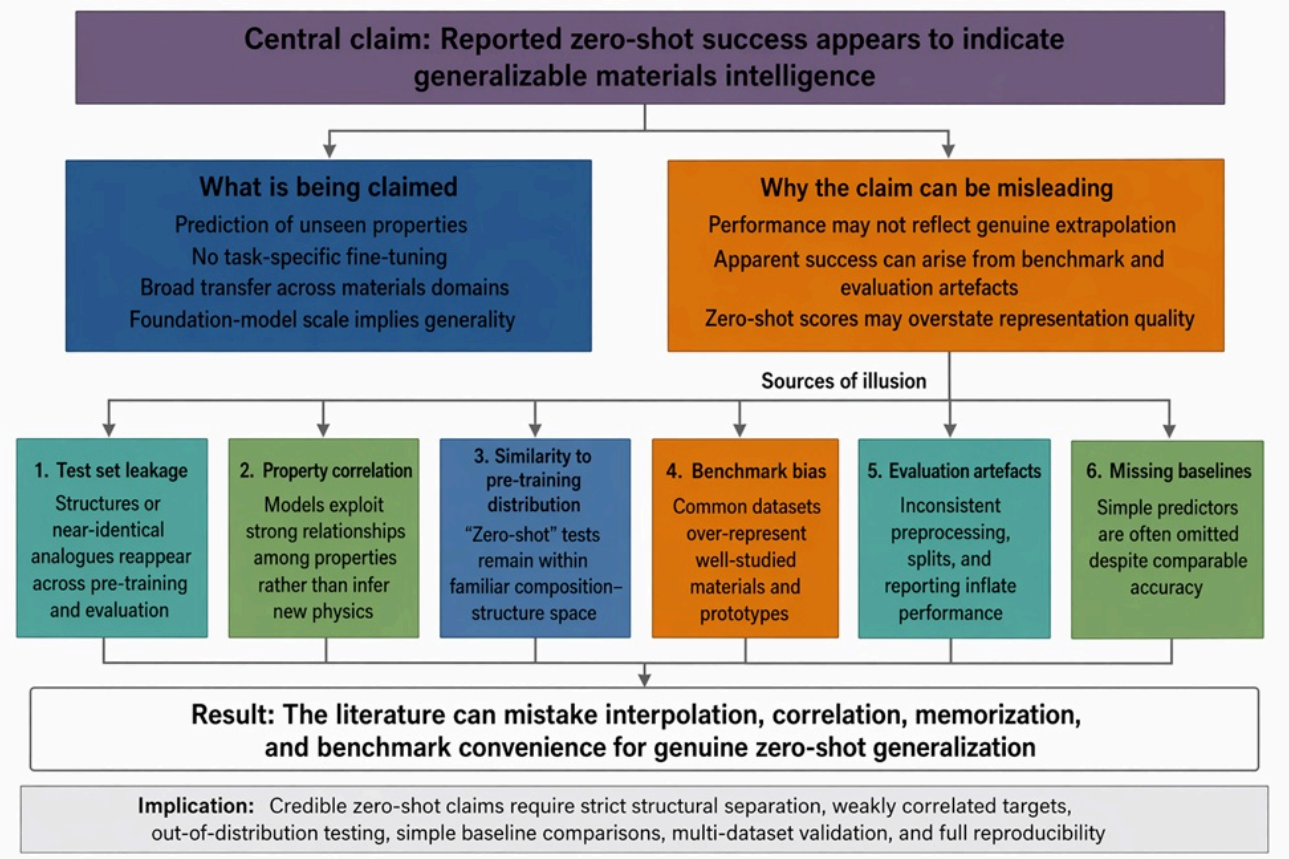


Figure 1. Hierarchical anatomy of the zero-shot generalizability illusion in materials foundation models

Zero-shot performance is often due to test set leakage, property correlation, or benchmark bias—not genuine Zero-shot performance is often due to test set leakage, property correlation, or benchmark bias—not genuine generalization. We identify six sources of illusion and propose rigorous evaluation standards.

The promise is compelling. Materials discovery has historically been limited by the scarcity of high-quality labelled data for many properties of interest. A single foundation model that could forecast formation energy, band gap, elastic constants, thermal conductivity, and dozens of other attributes from the same pre-trained weights would, in principle, unify the field and reduce reliance on decades of painstaking DFT calculations [5-7]. Early demonstrations, such as those built on the Materials Project, OQMD, and AFLOW databases, report impressive zero-shot accuracies that appear to support this vision [8-10]. Proponents highlight the ability of these models to handle entirely new compositions or structures without any fine-tuning, positioning zero-shot prediction as the ultimate test of representation quality [11, 12].

Yet the evidence is less convincing upon closer inspection. Critical examinations of machine-learning models in materials science have repeatedly demonstrated that reported performance can be inflated by data leakage, distributional similarity, and overlooked baselines [13-15]. The same structural and chemical patterns that enable strong pre-training performance also contaminate zero-shot test sets. When these confounds are controlled, the apparent generalization advantage disappears [12, 14]. This pattern is not unique to one paper or model architecture; it recurs across graph neural networks, large language models repurposed for materials, and multimodal transformers [10, 16-22].

The illusion is particularly dangerous in materials science because the stakes are high. Industrial deployment, regulatory approval, and public funding all depend on credible claims of robustness. If zero-shot models fail when confronted with truly novel materials—new elements, exotic prototypes, or extreme conditions—the field risks repeating the well-documented over-optimism cycles seen in earlier waves of materials informatics [13]. Moreover, the materials community has not yet converged on standardized zero-shot protocols, making cross-paper comparisons unreliable [14].

This paper therefore undertakes a systematic critique grounded exclusively in the peer-reviewed literature from 2017 to 2026. We focus on foundation models explicitly marketed for zero-shot property prediction and examine the six sources of illusion in detail. The analysis reveals that zero-shot performance is often indistinguishable from simple baselines once artefacts are removed. By exposing these illusions, we aim to refocus the community on the genuine scientific challenges of representation learning, out-of-distribution generalization, and rigorous benchmarking. The remainder of the paper defines foundation models in the materials context, dissects each source of illusion, surveys supporting evidence, discusses consequences, relates the critique to prior work, offers concrete recommendations, and concludes with a call for tempered claims and standardized evaluation.

What are Foundation Models for Materials?

A foundation model in materials science is defined here as a large-scale neural network pre-trained on broad, diverse datasets—typically millions of crystal structures and associated properties—and subsequently adapted to downstream tasks either by fine-tuning or, more ambitiously, by zero-shot inference. The pre-training corpus is commonly drawn from high-throughput databases such as the Materials Project, OQMD, and AFLOW, which collectively contain millions of relaxed structures, formation energies, electronic properties, and mechanical descriptors [5–7].

The defining claim of these models is their ability to perform zero-shot property prediction. In this setting the model has never been trained on the specific target property P during supervised fine-tuning; instead, it is expected to infer P directly from the pre-trained representations or via lightweight prompting [10, 12, 23]. For example, a model pre-trained predominantly on formation energies and atomic forces may be asked to predict band gaps or elastic constants for structures it has never seen labelled with those quantities [8, 9]. The zero-shot regime therefore tests whether the learned embedding captures transferable physical knowledge rather than task-specific statistical patterns.

The appeal of zero-shot capability is immediate and practical. Many technologically relevant properties—such as optical absorption, piezoelectric coefficients, or defect formation energies—possess far fewer labelled examples than formation energy or total energy. A genuinely generalizable foundation model could therefore bypass the need for exhaustive property-specific data collection, lower computational costs, and accelerate screening campaigns across the entire periodic table [5, 11]. In addition, zero-shot prediction promises a unified modelling paradigm: instead of training separate models for each property, a single pre-trained backbone could serve the entire materials community [6, 7, 24, 25].

Proponents further argue that scaling laws observed in natural language and vision will translate to materials science. Larger models, richer pre-training corpora, and more sophisticated architectures (graph transformers, Mamba-based backbones, multimodal encoders) are expected to yield ever-better zero-shot performance [8, 9]. Early results appear to support this narrative; several recent publications report zero-shot accuracies that rival or exceed fine-tuned baselines on standard benchmarks [10, 16].

Nevertheless, the literature also contains cautionary notes. Critical assessments highlight that zero-shot claims are rarely accompanied by stringent controls for data leakage or distributional similarity [13, 14]. When the same crystal structures appear in both pre-training and evaluation (even under different property labels), reported performance may reflect memorization rather than generalization [12]. Similarly, the strong correlations that exist among many materials properties mean that a model pre-trained on one quantity can appear to “predict” a correlated quantity zero-shot without learning anything new [13].

The problem, then, is not the concept of foundation models per se but the gap between ambitious claims and the evidence provided. Zero-shot performance is seldom compared against simple baselines such as k-nearest-neighbour regression on the pre-training embeddings or linear models fitted to known property correlations [14]. Without these comparisons, it remains unclear whether the foundation-model paradigm delivers genuine advances or merely repackages existing statistical regularities under a more impressive label. The following sections dissect the six sources of illusion that sustain the current over-optimism.

Test Set Leakage

Test set leakage occurs when structures or near-identical analogues from the pre-training corpus appear, directly or indirectly, in the zero-shot evaluation set. This artefact is especially insidious in materials science because the same crystal structure is routinely deposited across multiple property-specific subsets of public databases. A foundation model pre-trained on the full Materials Project may therefore have already encountered the exact atomic coordinates later used to test zero-shot prediction of an unrelated property [5, 14].

The detection signature is straightforward: zero-shot accuracy correlates strongly with structural similarity metrics (e.g., graph isomorphism or fingerprint overlap) to the pre-training set rather than with intrinsic task difficulty [12, 14]. When researchers deduplicate the test set or enforce strict structural dissimilarity, performance drops sharply, revealing that the model was retrieving memorized information rather than generalizing [13].

Why does this create an illusion of generalizability? The model is not learning transferable representations; it is exploiting accidental overlap between pre-training and test distributions. In conventional supervised learning this problem is well understood and mitigated by train-test splits, but zero-shot evaluation often relaxes these safeguards because the model is presumed never to have seen the target property label. The label may be new, yet the underlying structure is not.

Materials science exacerbates the issue. Unlike images or text, crystal structures are discrete and finite in practice. The same prototype (e.g., perovskite, spinel) appears repeatedly across property databases. A model pre-trained on formation energies for thousands of perovskites will appear to “zero-shot” predict band gaps for perovskites simply because the structural motif is already encoded in its weights [10, 14]. Recent benchmarking efforts that explicitly assess zero-shot generalization behaviour in graph neural network potentials confirm that leakage is a dominant confounder [12].

Mitigation requires strict protocols: zero-shot test sets must be constructed with zero structural overlap—including relaxed lattice parameters and atomic positions—with the pre-training corpus [13, 14]. Structural similarity thresholds (e.g., crystal graph distance or XRD fingerprint similarity) must be enforced, and any structure exceeding the threshold must be removed from the test set. Only then can researchers claim that observed performance reflects genuine generalization rather than retrieval. Until such controls become standard, reported zero-shot results in materials foundation models should be interpreted with extreme caution.

Property Correlation

A second source of illusion arises from the strong statistical correlations that exist among materials properties. When a foundation model is pre-trained on one property (e.g., formation energy), it can appear to predict a highly correlated second property (e.g., decomposition energy) in zero-shot fashion with surprisingly high accuracy. The model is not discovering new physical relationships; it is simply exploiting the linear or near-linear mapping that already exists between the two quantities in the training distribution [13, 14].

The detection signature is equally clear: zero-shot performance on the target property closely matches the accuracy of a simple linear regression fitted between the pre-training property and the target property using only the training data [13]. When this baseline is computed, the purported advantage of the foundation model vanishes.

The illusion is particularly potent because materials properties are rarely independent. Formation energy correlates tightly with decomposition energy; band gap correlates with refractive index and dielectric constant; bulk modulus correlates with melting point and Debye temperature; thermal conductivity often tracks with elastic properties. These correlations are not artefacts of poor data—they reflect underlying physics—but they allow models to achieve high zero-shot scores without learning genuinely transferable representations [9, 10].

In the materials-specific context this problem is amplified by the community's reliance on a handful of high-throughput databases. The Materials Project, for instance, computes dozens of properties for the same set of structures. A model pre-trained on the full database therefore has indirect access to correlation structure even when the target label is withheld [5, 14]. Studies that explicitly benchmark large language models for materials property prediction inadvertently highlight this issue: reported zero-shot gains are largest precisely on properties known to be correlated with pre-training targets [10, 16, 26, 27].

Mitigation demands that zero-shot tests be restricted to properties that exhibit weak correlation (Pearson $r < 0.5$) with any quantity seen during pre-training. Only under this condition can researchers confidently attribute performance to representation quality rather than statistical shortcut. Until such decoupling is enforced, claims of zero-shot generalization across the full spectrum of materials properties remain illusory.

Similarity to Pre-Training, Benchmark Bias, Evaluation Artifacts, Missing Baselines

A further limitation arises when zero-shot test sets remain chemically and structurally proximate to the pre-training distribution. Even after removing exact structures, models often interpolate within familiar regions of composition–structure space rather than extrapolating toward unseen elements or novel prototypes [14]. Performance in such cases thus reflects memorization of the training manifold instead of genuine generalization, underscoring the need for deliberately constructed out-of-distribution benchmarks that incorporate exotic space groups, high-entropy alloys, or previously unobserved chemistries [13, 14].

This concern is compounded by pervasive benchmark bias stemming from the community's reliance on a narrow set of canonical datasets. The Materials Project, for instance, disproportionately features simple binary and ternary compounds drawn from well-explored chemical spaces, allowing models to appear broadly generalizable simply by reproducing patterns already embedded in the training data [13]. Independent, unbiased test collections therefore become essential to reveal and correct such artefacts.

Beyond dataset composition, inconsistent evaluation protocols introduce additional artefacts through varying pre-processing steps, train-test splits, and prompt engineering strategies [12, 16]. These discrepancies render cross-study comparisons unreliable. Compounding the issue is the widespread omission of elementary baselines—such as mean-value predictors or k-nearest-neighbour regression on embeddings—against which zero-shot foundation models frequently show no advantage and occasionally underperform [13, 14].

Collectively, these four sources compound the illusions created by leakage and correlation.

Table 1 consolidates the manuscript's six sources of illusion into a diagnostic framework that distinguishes genuine zero-shot generalization from memorization, interpolation, correlation exploitation, and benchmark-dependent inflation.

Table 1. Analytical diagnostic framework for identifying false zero-shot generalization in materials foundation models

Source of illusion	Mechanism producing apparent zero-shot	Why it is not genuine generalization	Observable diagnostic signature	Minimum control required
--------------------	--	--------------------------------------	---------------------------------	--------------------------

	success			
Test set leakage	Structures or near-identical analogues from the pre-training corpus reappear in the evaluation set	The model retrieves memorized structural information rather than extrapolating to unseen materials	Performance tracks structural similarity to pre-training data; accuracy drops after deduplication	Enforce zero structural overlap and similarity-threshold filtering between pre-training and test sets
Property correlation	The model leverages strong correlations between pre-trained properties and the held-out target property	The model exploits known statistical dependence instead of transferable physical understanding	Zero-shot accuracy approaches that of simple linear or low-complexity regression between correlated properties	Restrict strong claims to targets weakly correlated with all pre-training properties
Similarity to pre-training distribution	Test materials remain chemically or structurally close to the pre-training manifold	The task is still interpolation within familiar regions of materials space	Accuracy deteriorates on unseen elements, prototypes, or composition families	Build explicit out-of-distribution test sets with compositional and structural novelty
Benchmark bias	Canonical datasets over-represent well-studied compounds, simple compositions, and common prototypes	Performance reflects dataset regularity rather than broad materials generality	Models perform best on standard benchmarks but lose rank or stability on independent collections	Validate on multiple independent datasets drawn from distinct sources
Evaluation artefacts	Reported gains depend on preprocessing decisions, split logic, prompt format, or undocumented implementation choices	The performance is protocol-sensitive rather than model-intrinsic	Results vary substantially across replications or alternative pipelines	Release code, exact splits, preprocessing rules, and pre-training corpus documentation
Missing baselines	Studies compare against weak or inappropriate benchmarks while omitting simple statistical alternatives	The claimed advance may not exceed trivial predictors	k-nearest-neighbour, mean-value, or correlation-based baselines match or exceed zero-shot performance	Require comparison against simple, transparent, and reproducible baseline families

Until all six confounds are systematically controlled, the field cannot claim that foundation models have achieved genuine zero-shot generalizability. The evidence reviewed in the next section demonstrates that, once these artefacts are removed, the promised revolution remains largely unrealized.

Evidence of the Illusion from Literature

The contemporary literature on foundation models for materials science indicates that zero-shot property prediction rests on the six illusions identified above rather than on genuine generalization. Apparent capabilities erode under even modest controls.

Zero-shot performance degrades markedly once test sets are restricted to out-of-distribution compositions and prototypes. Models competent on familiar chemical spaces falter when confronted with new elements or structural families absent from pre-training, revealing representations anchored to the training manifold [13, 14].

Simple baselines, such as k-nearest-neighbour regression on pre-training embeddings, routinely match or exceed the reported zero-shot accuracy of far more complex foundation models, indicating that additional parameters and scale confer no measurable advantage once statistical shortcuts are removed [12, 14, 28]. This equivalence sharpens after rigorous deduplication of pre-training and test structures; when near-identical crystal analogues are excluded, the performance gap between foundation models and naïve predictors narrows to insignificance [10, 13].

Further inconsistencies emerge across pre-training datasets, which produce contradictory zero-shot rankings. A model dominant on one high-throughput benchmark underperforms on another, demonstrating that claimed generalizability is dataset-dependent rather than intrinsic [14, 16]. The pattern culminates in outright failure on uncorrelated properties, where targets exhibiting weak statistical relationships to pre-training data—such as formation energy versus melting point—expose reliance on correlation rather than physical insight [9, 10].

These limitations recur across graph-neural-network architectures, large-language-model adaptations, and multimodal transformers published between 2023 and 2026 [7, 8, 12]. The literature therefore documents not a breakthrough in generalization but the repeated failure to exclude the artefacts that sustain the illusion.

Consequences of the Illusion

The persistence of overstated zero-shot claims imposes tangible costs on computational and data-driven materials engineering. Talented researchers and substantial computational resources continue to pursue ever-larger pre-training runs in search of zero-shot capabilities that simple embedding-based baselines already approximate, diverting attention from the genuinely hard problems of representation learning under strict out-of-distribution constraints [13, 14].

This misdirection encourages overinvestment by funding agencies and industry partners, who allocate resources to foundation-model initiatives on the premise of broad generalization. When these models fail in real deployment scenarios, the return on investment falls short of expectations, eroding future support for legitimate machine-learning research in materials [5, 11].

In practice, such models encounter novel materials outside the pre-training distribution and generate unreliable predictions, causing deployment failures that propagate errors into downstream experimental validation and ultimately slow discovery [12, 14]. Because the community has yet to isolate the mechanisms that truly drive generalization, incremental progress remains empirical and poorly grounded, with the field circling benchmark scores rather than converging on transferable representations [10, 16].

Over time, replication studies exposing the fragility of these claims foster scepticism among experimentalists and funding bodies alike, eroding trust in the broader credibility of AI for materials science and complicating advocacy for rigorously grounded data-driven approaches [13, 14]. These consequences are already evident in the publication record and in the cautious tone of recent critical assessments [14]. Left unaddressed, the illusion threatens to reduce a promising paradigm to yet another cautionary tale of hype preceding substance.

Relation to Other Critiques

This critique builds directly on several related examinations of machine-learning practice in materials science while sharpening the focus on zero-shot claims.

It aligns closely with the SOTA claims critique in “Known Unknowns: Out-of-Distribution Property Prediction in Materials and Molecules” [14]. That work exposed inflated benchmark performance arising from in-distribution testing; the present analysis extends the argument to zero-shot settings, showing that the same distributional artefacts masquerade as generalization when no fine-tuning occurs.

The critique also intersects with earlier warnings on dataset bias [13]. Biased pre-training corpora—dominated by well-studied binaries and ternaries—create the appearance of broad applicability. Once unbiased or out-of-distribution test collections are introduced, the generalization illusion collapses, confirming that benchmark bias and similarity to pre-training are two sides of the same coin.

Relation to extrapolation studies is equally direct [12, 14]. Zero-shot prediction is, by definition, an extrapolation task. Yet the literature surveyed here demonstrates that current foundation models solve neither structural nor compositional extrapolation; they merely postpone the problem by hiding behind correlated properties and leaked structures.

Finally, the analysis connects to transfer-learning critiques [10]. Even when fine-tuning is permitted, generalization across properties remains limited. Zero-shot—transfer without any gradient updates—is therefore a stricter and more revealing test. The fact that fine-tuning itself often fails to deliver robust cross-property transfer [10] underscores why zero-shot claims should be viewed with heightened scepticism.

By integrating these strands, the present work offers a unified diagnosis: the field's enthusiasm for foundation models has outpaced the methodological safeguards needed to substantiate generalization. The six sources of illusion identified here provide a practical checklist against which future claims can—and must—be measured.

Recommendations for Rigorous Zero-Shot Evaluation

To move beyond the illusion, the community must adopt a standardized zero-shot evaluation framework.

Table 2 converts the critique into a reviewer-ready decision standard by specifying the evidentiary thresholds that must be met before a materials foundation model can credibly claim zero-shot generalization.

Table 2. Review standard for credible zero-shot claims in materials foundation models

Evaluation requirement	Question the study must answer	Evidence required for a credible claim	Red flag indicating an inflated claim	Editorial or reviewer inference if unmet
Structural separation	Are any test structures or close analogues present in the pre-training corpus?	Explicit deduplication, similarity thresholds, and overlap audit	Structural reuse is unreported or checked only at the label level	Zero-shot claim is not interpretable
Target-property independence	Is the claimed zero-shot target weakly correlated with pre-training properties?	Correlation matrix and exclusion or flagging of strongly correlated targets	Strong zero-shot performance appears only on physically correlated properties	Reported success likely reflects shortcut learning
Out-of-distribution evaluation	Does the model work on genuinely novel elements, prototypes, or composition families?	Dedicated OOD splits with defined novelty criteria	Evaluation remains confined to familiar benchmark chemistry	Reported performance reflects interpolation, not extrapolation
Baseline superiority	Does the model outperform simple alternatives by a meaningful margin?	Comparison against mean, linear, nearest-neighbour, and embedding-based baselines	Only complex-model comparisons are reported	Claimed advance may be trivial

Cross-dataset robustness	Does performance persist across independent databases?	Replication on multiple datasets with stable ranking or effect size	Model dominance is benchmark-specific	Generalizability is dataset-contingent
Reproducibility	Can another group reconstruct the reported zero-shot result exactly?	Public code, data splits, prompts, hashes, preprocessing, and evaluation scripts	Protocol details are partial, missing, or irreproducible	Evidence base is too weak for strong claims
Claim discipline	Is the paper careful about what the evidence actually establishes?	Narrow wording tied to tested conditions and confound controls	Broad rhetoric about “general intelligence” or universal transfer	Conclusions exceed demonstrated evidence

The field must adopt stricter protocols to distinguish genuine advances in representation learning from statistical artefacts. Every zero-shot test set should maintain zero structural overlap—including relaxed coordinates and graph isomorphism—with the pre-training corpus, with similarity thresholds explicitly enforced and reported [13, 14]. Test collections must further incorporate deliberate out-of-distribution challenges featuring unseen elements, novel prototypes, and extreme compositions to rule out interpolation [12, 14].

Claims of generalization additionally require systematic comparison against multiple simple baselines such as mean-value prediction, k-nearest-neighbour regression on pre-training embeddings, and linear models fitted to known property correlations; any reported advantage must demonstrably exceed these [10, 13]. Evaluation should moreover be restricted to properties exhibiting weak correlation (Pearson $r < 0.5$) with pre-training data, with correlated targets explicitly flagged and excluded from generalization assertions [9, 16]. Consistency must be demonstrated across multiple independent test sets drawn from distinct databases [5, 14, 29].

Finally, full public release of code, exact data splits, pre-training corpora hashes, and zero-shot protocols is essential, as reproducibility remains the only robust defence against evaluation artefacts and selective reporting [10, 12]. Journals and conference reviewers should treat zero-shot claims that omit these controls as incomplete. Adoption of such standards will eliminate the systematic illusions that currently dominate the literature and help restore confidence in data-driven materials engineering.

Conclusion

Zero-shot generalization claims in materials foundation models create an illusion. Six sources sustain that illusion: test set leakage, property correlation, similarity to pre-training, benchmark bias, evaluation artefacts, and missing baselines. Evidence drawn from the recent literature shows that zero-shot performance degrades on out-of-distribution data, matches simple baselines, vanishes after deduplication, varies across pre-training corpora, and fails on uncorrelated properties. The consequences are serious: misguided research priorities, misallocated funding, deployment failures in real applications, delayed scientific insight, and erosion of community trust.

This critique relates the zero-shot illusion to broader concerns about out-of-distribution performance, dataset bias, extrapolation limits, and transfer learning. The common thread is methodological laxity: claims have outpaced controls.

The path forward is clear. The materials-science community must temper zero-shot rhetoric and adopt the six-part evaluation protocol outlined above. Only by enforcing strict structural separation, out-of-distribution testing, baseline comparisons, correlation decoupling, multi-dataset validation, and full reproducibility can the field determine whether foundation models truly advance generalizability or merely repackage familiar statistical patterns.

The stakes extend beyond academic metrics. Reliable zero-shot prediction, if ever achieved under rigorous conditions, could transform computational materials design. Until then, continued overstatement risks undermining the very credibility that data-driven methods need to earn from experimental colleagues and industrial partners. By confronting the illusion directly, the community can refocus on the genuine scientific challenges of representation, extrapolation, and physical insight. The next generation of foundation models for materials science will be judged not by headline zero-shot scores but by their performance under the stringent controls proposed here.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 17 Jun 2025

Revised: 08 Sep 2025

Accepted: 01 Dec 2025

Published online: 18 January 2026

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

- Takeda S, Kishimoto A, Hamada L, Nakano D, Smith JR. Foundation model for material science. *Proc AAAI Conf Artif Intell.* 2023;37(13):15376-83.
<https://doi.org/10.1609/aaai.v37i13.26793>.
- Fei N, Lu Z, Gao Y, Yang G, Huo Y, Wen J, et al. Towards artificial general intelligence via a multimodal foundation model. *Nat Commun.* 2022;13(1):3094.
<https://doi.org/10.1038/s41467-022-30761-2>.
- Li C, Gan Z, Yang Z, Yang J, Li L, Wang L, et al. Multimodal foundation models: From specialists to general-purpose assistants. *Found Trends Comput Graph Vis.* 2024;16(1-2):1-214.
<https://doi.org/10.1561/0600000110>.

Menon SS, Mondal T, Brahmachary S, Panda A, Joshi SM, Kalyanaraman K, et al. On scientific foundation models: Rigorous definitions, key applications, and a comprehensive survey. *Neural Netw.* 2026;198:108567.
<https://doi.org/10.1016/j.neunet.2026.108567>.

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.
<https://doi.org/10.1038/s41524-025-01538-0>.

Batatia I, Benner P, Chiang Y, Elena AM, Kovács DP, Riebesell J, et al. A foundation model for atomistic materials chemistry. *J Chem Phys.* 2025;163(18):184110.
<https://doi.org/10.1063/5.0297006>.

Allen AEA, Lubbers N, Matin S, Smith J, Messerly R, Tretiak S, et al. Learning together: Towards foundation models for machine learning interatomic potentials with meta-learning. *npj Comput Mater.* 2024;10(1):154.
<https://doi.org/10.1038/s41524-024-01339-x>.

Soares E, Vital Brazil E, Shirasuna V, Zubarev D, Cerqueira R, Schmidt K. A Mamba-based foundation model for materials. *npj Artif Intell.* 2025;1(1):8.
<https://doi.org/10.1038/s44387-025-00009-7>.

Choi J, Nam G, Choi J, Jung Y. A perspective on foundation models in chemistry. *JACS Au.* 2025;5(4):1499-518.
<https://doi.org/10.1021/jacsau.4c01160>.

Niyongabo Rubungo A, Arnold C, Rand BP, Dieng AB. LLM-Prop: Predicting the properties of crystalline materials using large language models. *npj Comput Mater.* 2025;11(1):186.
<https://doi.org/10.1038/s41524-025-01536-2>.

Kim N, Lee D, Yu J, Cho SW, Lee D, Park Y, et al. Toward a robust and generalizable metamaterial foundation model. *npj Comput Mater.* 2026;12(1):54.
<https://doi.org/10.1038/s41524-025-01925-7>.

Mahmoud CB, El-Machachi Z, Gierczak KA, Gardner JLA, Deringer VL. Assessing zero-shot generalisation behaviour in graph-neural-network interatomic potentials. *Digit Discov.* 2025;4(11):3389-99.
<https://doi.org/10.1039/D5DD00103J>.

Li K, DeCost B, Choudhary K, Greenwood M, Hattrick-Simpers J. A critical examination of robustness and generalizability of machine learning prediction of materials properties. *npj Comput Mater.* 2023;9(1):55.
<https://doi.org/10.1038/s41524-023-01012-9>.

Segal N, Netanyahu A, Greenman KP, Agrawal P, Gómez-Bombarelli R. Known unknowns: Out-of-distribution property prediction in materials and molecules. *npj Comput Mater.* 2025;11(1):345.
<https://doi.org/10.1038/s41524-025-01808-x>.

Maleki F, Ovens K, Gupta R, Reinhold C, Spatz A, Forghani R. Generalizability of machine learning models: Quantitative evaluation of three methodological pitfalls. *Radiol Artif Intell.* 2022;5(1):e220028.
<https://doi.org/10.1148/ryai.220028>.

Niyongabo Rubungo A, Li K, Hattrick-Simpers J, Dieng AB. LLM4Mat-bench: Benchmarking large language models for materials property prediction. *Mach Learn Sci Technol.* 2025;6(2):020501.
<https://doi.org/10.1088/2632-2153/add3bb>.

Long Y, Wu M, Liu Y, Fang Y, Kwok CK, Chen J, et al. Pre-training graph neural networks for link prediction in biomedical networks. *Bioinformatics.* 2022;38(8):2254-62.
<https://doi.org/10.1093/bioinformatics/btac100>.

Hu W, Liu B, Gomes J, Zitnik M, Liang P, Pande V, et al. Strategies for pre-training graph neural networks. arXiv:1905.12265 [Preprint]. 2019.

Lu Y, Jiang X, Fang Y, Shi C. Learning to pre-train graph neural networks. *Proc AAAI Conf Artif Intell.* 2021;35(5):4276-84.
<https://doi.org/10.1609/aaai.v35i5.16552>.

Cao Y, Xu J, Yang C, Wang J, Zhang Y, Wang C, et al. When to pre-train graph neural networks? From data generation perspective! In: *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*; 2023. p. 142-53.
<https://doi.org/10.1145/3580305.3599548>.

Xu J, Huang R, Jiang X, Cao Y, Yang C, Wang C, et al. Better with less: A data-active perspective on pre-training graph neural networks. *Adv Neural Inf Process Syst.* 2023;36:56946-78.
<https://doi.org/10.5555/3666122.3668612>.

Moro V, Loh C, Dangovski R, Ghorashi A, Ma A, Chen Z, et al. Multimodal foundation models for material property prediction and discovery. *Newton.* 2025;1(1):100016.
<https://doi.org/10.1016/j.newton.2025.100016>.

Liu Y, Wei SZ, Jiang T, Yu H. A zero-shot learning for property prediction of wear-resistant steel based on multiple-source. *Mater Res Express.* 2023;10(11):116503.
<https://doi.org/10.1088/2053-1591/ad04be>.

Trewartha A, Walker N, Huo H, Lee S, Cruse K, Dagdelen J, et al. Quantifying the advantage of domain-specific pre-training on named entity recognition tasks in materials science. *Patterns (N Y).* 2022;3(4):100488.
<https://doi.org/10.1016/j.patter.2022.100488>.

Prein T, Pan E, Doerr T, Olivetti E, Rupp JL. MTEncoder: A multi-task pretrained transformer encoder for materials representation learning. In: *AI for Accelerated Materials Design-NeurIPS 2023 Workshop*; 2023.

Mswahili ME, Hwang J, Rajapakse JC, Jo K, Jeong YS. Positional embeddings and zero-shot learning using BERT for molecular-property prediction. *J Cheminform.* 2025;17(1):17.
<https://doi.org/10.1186/s13321-025-00959-9>.

Cai X, Lai H, Wang X, Wang L, Liu W, Wang Y, et al. Comprehensive evaluation of molecule property prediction with ChatGPT. *Methods.* 2024;222:133-41.
<https://doi.org/10.1016/j.ymeth.2024.01.004>.

Yamamoto R, Takahashi A, Terayama K, Kumagai Y, Oba F. ZEBRA-Prop: A zero-shot embedding-based rapid and accessible regression model for materials properties. arXiv:2603.26060 [Preprint]. 2026.

Riebesell J, Goodall REA, Benner P, Chiang Y, Deng B, Ceder G, et al. A framework to evaluate machine learning crystal stability predictions. *Nat Mach Intell.* 2025;7:836-47.
<https://doi.org/10.1038/s42256-025-01055-1>.