

ORIGINAL RESEARCH

Open access

Why GNNs Fail for Metastable Phase Formation Energies: A Failure Mode Analysis

Kevin O'Brien^{1*}, Liam Murphy¹, Sean Doyle²

Abstract

Graph neural networks (GNNs) routinely achieve low errors in formation energy prediction, yet they systematically fail for metastable phases lying above the convex hull. These kinetically accessible materials, critical for batteries, catalysis, and thin-film applications, are severely underrepresented in training data dominated by stable ground states. This failure-mode analysis identifies four interlocking issues: (1) convex hull bias that pulls predictions downward toward stable energies, (2) energy range compression that collapses the predicted dynamic range, (3) local environment extrapolation failures caused by strained or unusual atomic geometries unseen in relaxed training data, and (4) error asymmetry producing consistent underprediction for metastable structures. We describe the origins, signatures, and discovery consequences of each mode, propose simple diagnostic tests for practitioners, and outline targeted mitigation strategies including metastable data augmentation, hull-aware regularization, and physics-informed extrapolation. Current GNN performance on convex-hull benchmarks does not guarantee reliability for metastable discovery. Addressing these failure modes is essential for trustworthy machine-learning-assisted design of functional materials that exist because of their position above the convex hull.

Keywords Graph neural networks, Data-driven materials engineering, Metastable phases, Formation energy prediction, Convex hull bias, Machine learning failure modes

*Correspondence:

Kevin O'Brien

kevin.obrien@gmail.com

¹ Department of Computational Materials Analytics, Faculty of Engineering, University College Dublin, Dublin, Ireland

² Department of Materials Data Systems, Faculty of Technology, Trinity College Dublin, Dublin, Ireland

Introduction

Formation energy is the most predicted property in materials machine learning [1-3]. It serves as the fundamental indicator of thermodynamic stability and the cornerstone for constructing phase diagrams, assessing synthesizability, and guiding high-throughput screening campaigns. Graph neural networks now deliver impressive accuracy on standard benchmarks, often reporting mean absolute errors below 0.05 eV per atom on held-out test sets [4-6]. These results have fueled optimism that GNNs can accelerate materials discovery across energy storage, catalysis, and electronics [7-10].

Yet a closer inspection reveals a critical limitation. The test sets on which these accuracies are measured consist almost exclusively of stable or near-stable compounds—materials whose formation energies place them on or near the convex hull [11, 12]. Metastable phases, which possess formation energies above the hull, are rarely present in the training or evaluation data. This omission is not accidental. Large databases such as the Materials Project were constructed with an emphasis on thermodynamic ground states because those structures define the lower envelope of the energy landscape and are the easiest to compute reliably with density-functional theory [13]. As a result, GNNs learn the statistical patterns

of low-energy, relaxed configurations and have little exposure to the higher-energy, often distorted structures that characterize metastable phases.

Metastable phases are far from exotic curiosities. They are routinely accessed through non-equilibrium synthesis routes such as rapid quenching, vapor deposition, or electrochemical cycling, and they frequently exhibit the very properties that make a material technologically useful [14–17]. In battery electrodes, many intercalation compounds at intermediate lithium concentrations are metastable yet reversible under operating conditions. In catalysis, metastable surface reconstructions often provide the active sites with optimal binding energies. High-entropy alloys and thin films produced by physical vapor deposition similarly stabilize metastable configurations that cannot be reached by conventional equilibrium methods. The practical importance of these phases is therefore undeniable.

Despite their relevance, the literature on GNN-based formation energy prediction has largely overlooked the extrapolation challenge posed by metastable structures. Models are trained and validated within the narrow energy window of convex-hull compounds, where local atomic environments are relaxed and formation energies cluster near zero. When the same models encounter structures lying 0.2–2 eV per atom above the hull, their predictions degrade sharply. Recent studies that have begun to probe performance beyond the convex hull confirm that extrapolation errors are not random but exhibit systematic patterns traceable to the training distribution [11, 18, 19]. This paper conducts a dedicated failure mode analysis to explain why GNNs fail for metastable phase formation energies.

Figure 1 provides a hierarchical overview of how convex-hull-dominated training data induce four linked GNN failure modes in metastable formation-energy prediction, the diagnostic signatures by which they can be identified, and the mitigation principles required to restore discovery reliability.

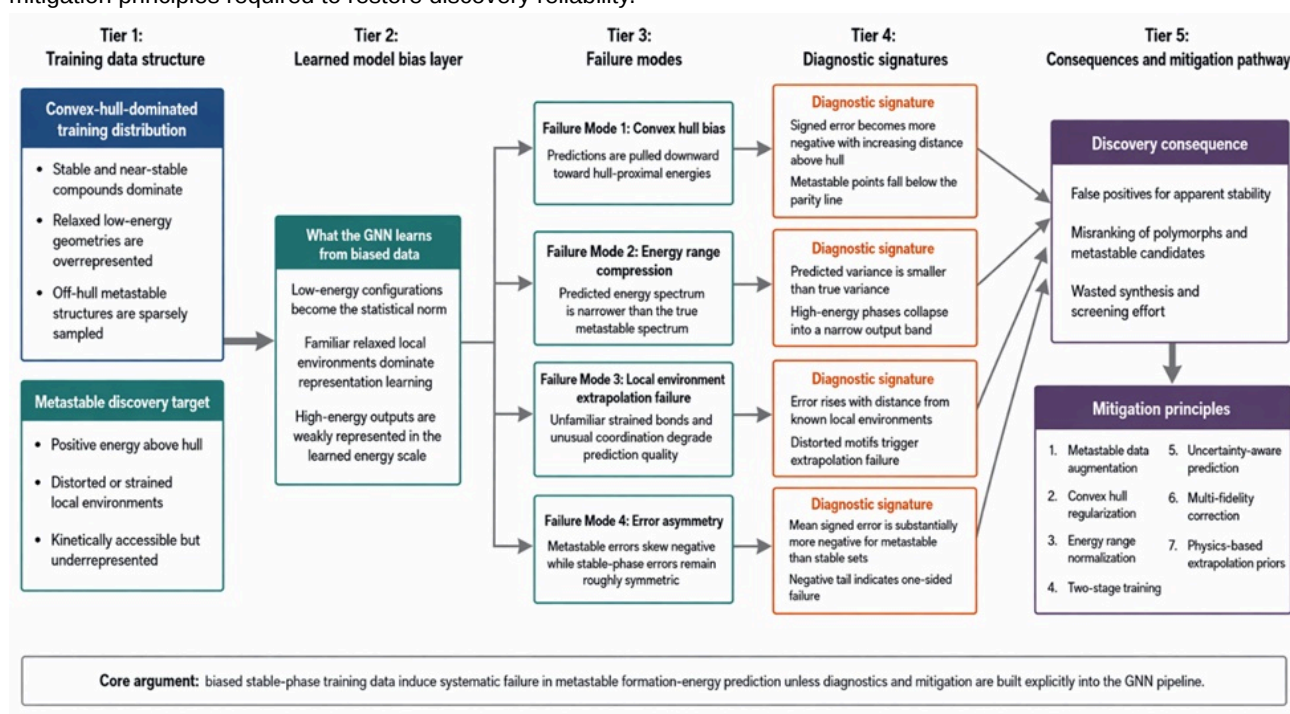


Figure 1. Conceptual failure architecture of graph neural networks for metastable formation energy prediction

The analysis is strictly conceptual. No new simulations, datasets, or numerical benchmarks are introduced. Instead, we synthesize insights from the computational materials literature to isolate the mechanisms of failure, define their signatures, and outline detection principles and mitigation strategies. The goal is to equip both model developers and materials practitioners with a clear understanding of the limitations inherent in current GNN architectures and the

conceptual changes required to overcome them. By focusing on metastable phases, this work addresses a blind spot that currently limits the reliability of machine learning in one of the most promising frontiers of materials engineering.

What are Metastable Phases?

A metastable phase is a material that is not the global thermodynamic ground state at a given composition and temperature but is separated from lower-energy states by kinetic barriers that prevent rapid transformation. On a convex-hull diagram—formed by plotting formation energy against composition—the stable compounds sit on the lower convex envelope. Any structure whose formation energy lies above this envelope is metastable. The vertical distance from the hull, often denoted conceptually as ΔE_{hull} , quantifies how far above the ground state the phase resides. Small positive values (for example, 0.05–0.2 eV per atom) indicate marginally metastable phases that can be stabilized by modest kinetic control; larger values correspond to highly unstable configurations that are difficult to synthesize but may still appear under extreme non-equilibrium conditions.

Metastable phases matter because real materials synthesis is rarely an equilibrium process. Kinetic pathways frequently trap atoms in local energy minima that are higher than the global minimum [14–17]. In lithium-ion batteries, layered oxides such as Li_xCoO_2 at intermediate lithium contents adopt metastable structures during charge–discharge cycling; these structures enable reversible intercalation even though they are not the lowest-energy arrangement at that composition. In heterogeneous catalysis, metastable surface terminations or reconstructed facets often expose under-coordinated sites that bind reaction intermediates more effectively than the thermodynamically stable surface. High-entropy alloys derive much of their strength and corrosion resistance from the large number of metastable local configurations that frustrate the formation of ordered, low-energy phases. Thin-film deposition techniques such as sputtering or molecular-beam epitaxy routinely produce metastable polymorphs that cannot be obtained by bulk equilibrium cooling. In each case, the functional material exists only because synthesis conditions allow access to regions of the energy landscape above the convex hull.

Materials databases reflect the historical emphasis on thermodynamic stability. The Materials Project and similar repositories were designed to compute and store the lowest-energy structures at each composition so that convex hulls could be constructed accurately for phase-diagram prediction [12, 13]. High-throughput workflows therefore prioritize relaxed, low-energy configurations and apply strict energy-above-hull filters (typically $\Delta E_{\text{hull}} < 0.1$ eV per atom) before including a structure in the final dataset. Metastable candidates with higher energies are either discarded or underrepresented. The training data supplied to GNNs therefore contain few, if any, examples of structures with formation energies more than a few tenths of an electronvolt above the hull. This systematic underrepresentation is not a flaw in data collection but a deliberate design choice that optimizes databases for stable-phase thermodynamics while leaving the metastable regime sparsely sampled.

The consequence for machine learning is straightforward. GNNs learn to associate realistic atomic graphs with low formation energies because that is the dominant pattern in the training distribution. When presented with a metastable structure—whose graph may contain elongated bonds, unusual coordination polyhedra, or higher internal strain—the model has no statistical precedent for assigning the correspondingly higher energy. The literature on data bias in materials property prediction confirms that such distributional skew produces models that generalize poorly outside the convex-hull region [11, 12, 20]. Understanding metastable phases therefore requires more than a thermodynamic definition; it demands recognition that the very data ecosystem on which GNNs are built has been engineered to exclude the higher-energy structures that practitioners most need to evaluate.

Convex Hull Bias

Convex hull bias arises when a GNN internalizes the statistical predominance of low-energy, hull-proximal compounds and treats any deviation above the hull as anomalous. During training, the overwhelming majority of formation energies cluster tightly around the convex hull ($\Delta E_{\text{hull}} \approx 0$). The model therefore learns that “reasonable” atomic configurations

correspond to low formation energies. When it encounters a metastable structure whose true formation energy lies above the hull, the network's learned bias pulls the prediction downward toward the familiar low-energy regime. The result is systematic underestimation: the predicted formation energy is lower—sometimes substantially lower—than the true value [11, 12, 18].

This bias is not merely a regression artifact. It reflects the model's implicit assumption that the training distribution defines the entire plausible energy landscape. Because the training set contains almost no examples with large positive ΔE_{hull} , the GNN has no incentive to reserve capacity for high-energy predictions. Message-passing layers that aggregate local environments reinforce this effect: the features learned from relaxed, stable geometries are reused even when the input graph describes a strained metastable arrangement, and the final readout layer maps those features back onto the narrow energy range seen during training. The literature on graph-based learning beyond stable materials explicitly documents this extrapolation failure, showing that errors grow monotonically with distance above the hull [11, 19].

The practical consequence is severe for materials discovery. A model that underestimates the formation energy of a metastable phase may classify it as stable (predicted $\Delta E_{\text{hull}} < 0.1$ eV per atom) when in reality it lies well above the hull. Researchers relying on the prediction may then attempt synthesis, only to discover that the targeted phase is thermodynamically inaccessible under the intended conditions. The wasted experimental effort compounds across screening campaigns that evaluate thousands of candidates. In extreme cases, the bias can invert the relative stability ordering between polymorphs, leading to incorrect phase-diagram predictions and misguided design strategies [14, 15, 18].

Detection of convex hull bias follows a simple conceptual signature. When predicted formation energies are plotted against true values for a set of metastable test structures, the points lie systematically below the diagonal line of perfect agreement. The magnitude of underprediction increases with ΔE_{hull} : structures close to the hull show modest errors, while those far above exhibit larger deviations. This pattern is absent or far weaker on stable test sets, confirming that the bias is specific to the off-hull regime. A representative illustration is the common TiO₂ system. The stable rutile polymorph sits on the convex hull, while the metastable anatase phase lies approximately 0.1 eV per atom above it. A GNN trained predominantly on rutile-like local environments predicts an anatase formation energy that is pulled downward, closer to the rutile value than DFT calculations warrant. Similar behavior has been observed in other oxide systems and in metastable intermetallics [15, 16].

Convex hull bias is therefore the foundational failure mode. It originates directly from the composition of training databases, propagates through the inductive biases of graph message passing, and produces downstream errors that compromise the reliability of GNN-driven discovery pipelines. Any attempt to mitigate later failure modes must first address this root distributional skew.

Energy Range Compression, Local Environment Extrapolation

Energy range compression, occurs when the GNN's output distribution is narrower than the true distribution of formation energies in the metastable regime. Training data span a limited interval—typically from roughly -3 eV per atom for strongly bound compounds to near zero for hull-proximal structures. The model therefore learns to produce predictions within this compressed band. When presented with metastable structures whose true energies may reach $+1$ to $+5$ eV per atom relative to the hull, the network still outputs values near the lower end of its learned range. The variance of predicted energies is substantially smaller than the variance of true energies, and the highest-energy structures are systematically underpredicted [19–21].

This compression is a direct consequence of regression toward the training mean. Without exposure to high-energy examples, the network has no mechanism to expand its dynamic range. Even architectures equipped with sophisticated

attention or global pooling layers remain constrained by the statistical envelope of the training labels [22]. The result is a model that cannot distinguish among metastable phases of differing energetic instability; all are mapped to roughly the same narrow energy window. Screening workflows that rely on predicted formation energies to rank candidates therefore lose discriminatory power precisely in the region where kinetic accessibility varies most strongly.

Failure mode 3, local environment extrapolation, stems from the mismatch between the atomic neighborhoods present in stable training structures and those appearing in metastable phases. Stable compounds relax under thermodynamic driving forces to optimal bond lengths, ideal coordination numbers, and minimal strain. Metastable structures, by contrast, often contain elongated or compressed bonds, under- or over-coordinated atoms, and polyhedral distortions that are energetically costly yet kinetically trapped. GNNs encode local environments through message-passing iterations that learn filters tuned to the relaxed geometries of the training set. When these filters encounter unfamiliar local motifs, the propagated features become unreliable and the final energy prediction degrades [11, 19, 20].

Detection of local environment extrapolation relies on a conceptual similarity metric between the local graph around each atom in a metastable structure and the nearest analogous environment in the training set. Errors rise sharply as this distance increases, confirming that the model is operating outside its learned feature space. A concrete illustration is a strained perovskite oxide in which octahedral tilting angles deviate markedly from the values observed in stable cubic or tetragonal perovskites. The GNN, having never encountered such large tilt magnitudes during training, cannot correctly assess the associated strain energy contribution, leading to underpredicted formation energies [23, 24].

Energy range compression and local environment extrapolation reinforce each other. Compressed outputs hide the true spread of metastable energies, while poor local extrapolation ensures that the model cannot even recognize the structural features responsible for elevated energies. Together they render GNN predictions untrustworthy for any screening task that extends beyond the convex hull.

Error Asymmetry

Error asymmetry is the tendency of GNNs to produce errors that are systematically negative for metastable phases while remaining roughly symmetric for stable phases. On stable test sets the distribution of signed errors (predicted minus true formation energy) is centered near zero, with over- and under-predictions balancing out. On metastable test sets the same distribution shifts to a negative mean, indicating consistent underprediction. The magnitude of this negative bias can exceed 0.2–0.3 eV per atom, far larger than the mean absolute error typically reported for stable compounds [18, 20, 21].

The origin of asymmetry lies in the interplay between convex hull bias and the regression properties of neural networks. Because the training distribution is skewed toward low energies, the model's loss landscape favors predictions that remain within that low-energy basin. When a metastable input is processed, residual pull from the dominant training pattern shifts the output downward. Stable structures, by contrast, lie within the densely sampled region where the model has learned balanced corrections, so positive and negative deviations cancel. The literature on training bias in materials property prediction confirms that such one-sided error patterns emerge precisely when the test distribution lies outside the training support [12, 13].

The consequences for materials discovery are direct and costly. Underprediction transforms genuinely metastable phases into apparently stable ones, generating false positives in virtual screening. Synthesis efforts are then directed toward phases that are thermodynamically inaccessible, consuming experimental resources and delaying genuine discoveries. In generative-model pipelines that propose new structures and rely on GNNs for rapid energy evaluation, asymmetric errors can bias the search toward unrealistically low-energy candidates that are in fact metastable or unstable. The cumulative effect is a loss of trust in machine-learning predictions exactly when they are most needed—outside the convex-hull regime.

Detection of error asymmetry requires only the computation of mean signed error on separate stable and metastable partitions. If the mean signed error on the metastable set is more negative than on the stable set by a statistically meaningful margin (for example, >0.1 eV per atom), asymmetry is present. The full error distribution further reveals a longer negative tail for metastable structures, confirming the directional bias. Conceptual comparison of mean absolute error (which remains moderate) versus mean signed error (which is large and negative) highlights that the model is not merely imprecise but systematically wrong in one direction [18, 20].

Error asymmetry is therefore the observable manifestation of the preceding failure modes. Convex hull bias supplies the directional pull, energy range compression limits the upward dynamic range, and local environment extrapolation prevents accurate compensation for structural novelty. The resulting negative skew is the signature that practitioners should examine first when evaluating GNN reliability for metastable discovery.

Table 1 consolidates the four failure modes into a single analytical structure by linking each mode to its distributional origin, model mechanism, observable signature, discovery consequence, and primary mitigation lever.

Table 1. Theoretical consolidation of the four failure modes in metastable formation-energy prediction

Failure mode	Immediate source in training distribution	Model-level mechanism	Observable prediction signature	Discovery-level consequence	Primary mitigation lever
Convex hull bias	Stable and near-stable structures dominate the label distribution; off-hull examples are rare	The model internalizes hull-proximal energies as the default outcome and pulls unfamiliar inputs toward the low-energy regime	Signed errors become increasingly negative with greater distance above the hull; metastable points systematically fall below parity	Metastable phases are misclassified as stable, creating false positives for synthesis targeting	Deliberate inclusion of off-hull training structures plus convex-hull-aware loss regularization
Energy range compression	Training labels occupy a narrow energy interval concentrated near stable compounds	The regressor learns a restricted dynamic range and cannot expand outputs into the true metastable spectrum	Predicted variance is substantially smaller than true variance; highly unstable phases are squeezed into a narrow band	Candidate ranking loses resolution in the metastable regime, obscuring meaningful differences in kinetic accessibility	Energy-range normalization and explicit exposure to broader off-hull label support
Local environment extrapolation failure	Stable training data mostly contain relaxed bonds, common coordination states, and low-strain motifs	Message-passing filters are tuned to familiar local environments and perform unreliably on distorted or unusual atomic neighborhoods	Error increases with graph-feature distance from the nearest training environment; strained motifs trigger disproportionate underprediction	Structurally novel metastable candidates are penalized precisely where discovery demands generalization	Metastable structural augmentation, physics-guided descriptors, and extrapolation-aware architecture design

Error asymmetry	Distributional skew toward low energies makes balanced correction possible in stable regions but not in metastable ones	The combined effect of convex hull bias, compressed range, and poor extrapolation produces one-sided negative error	Mean signed error is near zero for stable phases but strongly negative for metastable phases; a long negative tail appears	Screening pipelines become systematically overoptimistic about metastable stability and waste experimental resources	Partitioned evaluation, uncertainty-aware prediction, and metastable-specific calibration
-----------------	---	---	--	--	---

Detection Principles

Detection of the four failure modes relies on simple, interpretable tests that practitioners can apply to any GNN without retraining or new experiments. Model behavior on stable versus metastable partitions of existing data isolates each failure.

Convex hull bias is confirmed when signed prediction error becomes increasingly negative with growing conceptual distance above the hull for every test structure, the slope of this trend quantifying the strength of the model's pull toward the low-energy regime [11, 12, 18]. Energy range compression is present when the ratio of predicted variance to true variance of formation energies on the metastable set falls below roughly 0.7, as the model squeezes the entire metastable energy spectrum into the narrow band learned from stable training examples [19-21].

A related implication arises through local environment extrapolation failure. Prediction error correlates positively with the graph-feature distance to the nearest local environment encountered during training for each atom in a metastable structure, directly exposing the mismatch between relaxed stable neighborhoods and the strained motifs typical of metastable phases [11, 19, 23].

This shift also introduces error asymmetry, which is active when the metastable mean signed error is more negative than the stable mean by more than 0.1 eV per atom; the full distribution of signed errors further reveals a pronounced negative tail for metastable structures while the stable distribution remains centered near zero [12, 18, 20].

Beyond this immediate concern, the false positive rate signals unreliability for discovery pipelines when the fraction of true metastable phases incorrectly classified as stable exceeds 10%, with predicted distance above hull below 0.1 eV per atom, because the model will routinely suggest syntheses that cannot succeed [14-16].

Such diagnostics require no additional data generation, only careful partitioning of existing structures into hull-proximal and off-hull subsets. Applied routinely, they allow model developers to certify whether a GNN suffers from any of the four failure modes before it is deployed for metastable screening [11, 18].

Table 2 converts the manuscript's diagnostic principles into a decision framework that links each test to an interpretation threshold, an operational risk judgment, and a specific corrective response.

Table 2. Diagnostic-to-decision framework for evaluating and deploying GNNs in the metastable regime

Diagnostic principle	What is compared	Decision rule / threshold from manuscript logic	What the result means conceptually	Immediate deployment implication	Most direct corrective action

Convex hull distance test	Signed prediction error versus distance above hull	Increasingly negative slope with higher distance above hull	The model is being pulled back toward the stable-data manifold	Do not trust off-hull stability calls without validation	Add off-hull training data and hull-aware regularization
Energy range test	Variance of predicted energies versus variance of true energies on metastable structures	Predicted variance / true variance < 0.7	The model has learned a compressed metastable energy spectrum	Candidate ranking is unreliable even when average error appears acceptable	Rescale labels to the broader metastable range and retrain with wider support
Local environment similarity test	Prediction error versus distance from nearest training local environment	Clear positive error–distance relationship	Structural novelty is exceeding the learned feature space	Novel motifs should be routed to higher-fidelity evaluation	Expand training environments and incorporate physics-based local priors
Error asymmetry test	Mean signed error for stable partition versus metastable partition	Metastable mean signed error more negative than stable by >0.1 eV per atom	Failure is directional rather than random	Report separate performance by regime; avoid single aggregated benchmark claims	Calibrate on metastable data and add uncertainty estimation
False positive rate test	Fraction of true metastable phases predicted as stable	>10% false positives	The model is operationally unsafe for discovery screening	Synthesis recommendations based on predicted stability are not trustworthy	Tighten decision thresholds and require secondary validation before down-selection
Integrated deployment judgment	Results across all five tests taken together	Two or more failed tests indicate metastable-unreliable deployment status	The model is fit for stable interpolation but not metastable discovery	Restrict use to hull-proximal screening until corrected	Apply staged mitigation rather than incremental benchmarking

Mitigation Principles

Mitigation begins by incorporating metastable structures directly into training, where additional off-hull examples generated through random structural perturbations, high-temperature molecular dynamics snapshots, or targeted high-energy sampling—followed by accurate reference calculations—disrupt convex-hull dominance and instill the proper energy scale for higher-lying configurations [12, 19, 25].

Convex hull regularization complements this by introducing a penalty term in the loss function that explicitly discourages predictions below the hull, enforcing the physical constraint that prevents the model from collapsing metastable inputs toward ground-state behavior and thereby countering the inherent downward bias [11, 18].

Energy range normalization further addresses the compression mechanism by rescaling formation energy labels across the full span from stable minima to representative metastable values, compelling the network to capture the broader dynamic range essential for reliable extrapolation rather than defaulting to the narrow stable regime [20, 21, 25].

A related refinement employs two-stage training, first establishing core chemical intuition on the stable-dominated set before fine-tuning on a curated metastable subset under elastic weight consolidation to expand competence without catastrophic forgetting [19, 25].

Beyond this, uncertainty-aware prediction equips the GNN with an auxiliary head that flags high epistemic uncertainty on metastable inputs, enabling deferral to reference calculations and safeguarding against erroneous extrapolations [12, 18].

Multi-fidelity learning enhances generalization by training a correction network on sparse high-fidelity labels atop abundant low-fidelity metastable structures, allowing the model to internalize off-hull energy differences that transfer effectively to novel candidates [19, 20].

Finally, physics-based extrapolation integrates strain-energy or coordination priors into the architecture or post-processing, supplying the missing physical intuition for distorted local geometries that deviate from training distributions [11, 23].

These strategies require only modest adjustments to data pipelines and loss functions yet fundamentally transform GNNs from convex-hull specialists into reliable instruments for credible metastable prediction [14, 15, 25].

Relation to Other Failure Modes

Convex hull bias and its companions are not isolated quirks but specific manifestations of broader challenges in data-driven materials engineering. The root cause is dataset bias toward stable compounds, a well-documented limitation of databases that prioritize thermodynamic ground states [12, 13]. The four failure modes described here are the direct, observable consequences of that bias when applied to formation energy prediction.

The problem also represents a hard instance of the extrapolation challenge. Metastable prediction demands simultaneous extrapolation in both energy and structure space: the model must output values outside the training label range while processing atomic graphs that contain unseen local motifs [11, 19, 25]. Most GNN literature focuses on interpolation within the convex-hull manifold; off-hull extrapolation exposes the limits of message-passing architectures more sharply than any in-distribution test.

Active learning offers a natural connection. By iteratively querying the most uncertain or most off-hull candidates, active learning can deliberately expand the training distribution into metastable regions, reducing all four failure modes at their source [12, 18]. Generative models further amplify the issue: they frequently propose metastable or high-energy structures, yet downstream GNN energy evaluators must correctly rank them [26, 27]. Without mitigation of convex hull bias, energy range compression, and error asymmetry, generative pipelines will favor unrealistic low-energy candidates and discard promising metastable discoveries [14–16]. Recognizing these interrelations makes clear that addressing metastable failure modes is not a niche fix but a necessary step toward trustworthy machine learning across the entire materials discovery workflow.

Implications for Metastable Materials Discovery

For practitioners screening candidates for batteries, catalysis, or thin films, the implications are immediate. Never trust GNN formation energies for structures suspected to lie above the hull without independent validation. Use the detection principles to flag risky predictions and reserve reference calculations for high-uncertainty or high- ΔE_{hull} cases. This disciplined workflow prevents wasted synthesis efforts on false positives [16–18].

For model developers the message is equally clear. Every new GNN architecture must be evaluated on a dedicated metastable test set, not solely on standard convex-hull benchmarks. Report error versus distance above hull, variance compression ratios, and mean signed errors separately for stable and metastable partitions. Adopt at least convex hull regularization and metastable data augmentation as baseline practices before claiming broad applicability [11, 12, 25].

For benchmark designers the path forward is equally concrete. Construct and publish metastable-specific test sets drawn from diverse chemical spaces and energy ranges. Require submissions to include false positive rate for stability classification and error-versus- ΔE_{hull} curves. These expanded benchmarks will drive the community toward genuinely robust models rather than convex-hull specialists [13, 18]. Collectively these changes shift metastable materials discovery from an unreliable extrapolation exercise to a predictable, validated process, unlocking the kinetic phases that conventional equilibrium-focused pipelines have overlooked.

Conclusion

GNNs for formation energy prediction fail systematically when applied to metastable phases. The failure originates in the convex-hull-dominated training distributions of current materials databases and manifests through four interlocking modes: convex hull bias that pulls predictions downward, energy range compression that collapses the predicted spectrum, local environment extrapolation that mishandles strained geometries, and error asymmetry that produces consistent negative bias for off-hull structures. Each mode is detectable through straightforward tests—convex hull distance plots, variance ratios, local similarity correlations, signed-error comparisons, and false positive rates—that require no new computations.

Mitigation is equally straightforward in principle: enrich training data with metastable examples, impose convex hull regularization, normalize energy ranges, adopt two-stage learning, incorporate uncertainty quantification, leverage multi-fidelity corrections, and embed physics priors for extrapolation. These conceptual adjustments do not demand revolutionary architectures; they require only deliberate expansion of the training distribution and modest regularization. When implemented, they will enable GNNs to support reliable discovery of kinetically accessible materials in batteries, catalysis, high-entropy alloys, and thin films.

The broader lesson is that performance on stable test sets is an insufficient guarantee of utility. True progress in data-driven materials engineering demands explicit attention to the off-convex-hull regime where most functional phases actually reside. By diagnosing these failure modes, defining their detection signatures, and outlining principled mitigations, this analysis supplies the conceptual foundation for the next generation of metastable-aware GNNs. The field can now move beyond optimistic in-distribution benchmarks toward models that genuinely accelerate the discovery of materials that exist because of, rather than despite, their position above the convex hull.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 23 Jun 2025

Revised: 18 Sep 2025

Accepted: 29 Nov 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

- Yang TX, Dou P. Prediction of formation energy for oxides in ODS steels by machine learning. *Mater Des.* 2024;248:113503.
<https://doi.org/10.1016/j.matdes.2024.113503>.
- Kiyohara S, Shibui C, Bae S, Kumagai Y. Machine-learning prediction of charged-defect formation energies from crystal structures. *Phys Rev Lett.* 2025;135(24):246101.
<https://doi.org/10.1103/h66h-y5k6>.
- Rengaraj V, Jost S, Bethke F, Plessl C, Mirhosseini H, Walther A, et al. A two-step machine learning method for predicting the formation energy of ternary compounds. *Computation.* 2023;11(5):95.
<https://doi.org/10.3390/computation11050095>.
- Pandey S, Qu J, Stevanović V, St John P, Gorai P. Predicting energy and stability of known and hypothetical crystals using graph neural network. *Patterns (N Y).* 2021;2(11):100361.
<https://doi.org/10.1016/j.patter.2021.100361>.
- Choudhary K, DeCost B. Atomistic line graph neural network for improved materials property predictions. *NPJ Comput Mater.* 2021;7(1):185.
<https://doi.org/10.1038/s41524-021-00650-1>.
- Davariashtiyani A, Kadkhodaei S. Formation energy prediction of crystalline compounds using deep convolutional network learning on voxel image representation. *Commun Mater.* 2023;4(1):105.
<https://doi.org/10.1038/s43246-023-00433-9>.
- Meng K, Long R. A universal machine learning framework driven by artificial intelligence for ion battery cathode material design. *JACS Au.* 2025;5(8):3833-45.
<https://doi.org/10.1021/jacsau.5c00526>.
- Dong W, Wang Q, Ren L, Wei J, Shen S, Chen J. Graph neural networks reshaping the paradigm of electrocatalyst design for green hydrogen production. *Catal.* 2025;1(1):12.
<https://doi.org/10.1007/s44422-025-00013-7>.
- Rahman MH, Gollapalli P, Manganaris P, Yadav SK, Pilania G, DeCost B, et al. Accelerating defect predictions in semiconductors using graph neural networks. *APL Mach Learn.* 2024;2(1):016122.
<https://doi.org/10.1063/5.0176333>.

Mortazavi B. Recent advances in machine learning-assisted multiscale design of energy materials. *Adv Energy Mater.* 2025;15(9):2403876.
<https://doi.org/10.1002/aenm.202403876>.

Ekström Kelvinius F, Armiento R, Lindsten F. Graph-based machine learning beyond stable materials and relaxed crystal structures. *Phys Rev Mater.* 2022;6(3):033801.
<https://doi.org/10.1103/PhysRevMaterials.6.033801>.

Kumagai M, Ando Y, Tanaka A, Tsuda K, Katsura Y, Kurosaki K. Effects of data bias on machine-learning-based material discovery using experimental property data. *Sci Technol Adv Mater Methods.* 2022;2(1):302-9.
<https://doi.org/10.1080/27660400.2022.2109447>.

Schmidt J, Cerqueira TFT, Romero AH, Loew A, Jäger F, Wang HC, et al. Improving machine-learning models in materials science through large datasets. *Mater Today Phys.* 2024;48:101560.
<https://doi.org/10.1016/j.mtphys.2024.101560>.

Srinivasan S, Batra R, Luo D, Loeffler T, Manna S, Chan H, et al. Machine learning the metastable phase diagram of materials. *arXiv [Preprint]*. 2020. arXiv:2004.08753.
<https://doi.org/10.48550/arXiv.2004.08753>.

Feng J, Dong Z, Ji Y, Li Y. Accelerating the discovery of metastable IrO₂ for the oxygen evolution reaction by the self-learning-input graph neural network. *JACS Au.* 2023;3(4):1131-40.
<https://doi.org/10.1021/jacsau.2c00709>.

Liu C, Tamaki H, Yokoyama T, Wakasugi K, Yotsuhashi S, Kusaba M, et al. Shotgun crystal structure prediction using machine-learned formation energies. *NPJ Comput Mater.* 2024;10(1):298.
<https://doi.org/10.1038/s41524-024-01471-8>.

Lupo Pasini ML, Jung GS, Irlé S. Graph neural networks predict energetic and mechanical properties for models of solid solution metal alloy phases. *Comput Mater Sci.* 2023;224:112141.
<https://doi.org/10.1016/j.commatsci.2023.112141>.

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(6):836-47.
<https://doi.org/10.1038/s42256-025-01055-1>.

Hong C, Choi JM, Jeong W, Kang S, Ju S, Lee K, et al. Training machine-learning potentials for crystal structure prediction using disordered structures. *Phys Rev B.* 2020;102(22):224104.
<https://doi.org/10.1103/PhysRevB.102.224104>.

Krautsou AV, Humonen IS, Lazarev VD, Eremin RA, Budennyy SA. Impact of crystal structure symmetry in training datasets on GNN-based energy assessments for chemically disordered CsPbI₃. *Sci Rep.* 2025;15(1):8856.
<https://doi.org/10.1038/s41598-025-92669-3>.

Sheng Z, Zhu H, Shao B, He Y, Liu Z, Wang S, et al. Accelerated discovery of energy materials via graph neural network. *Inorganics.* 2025;13(12):395.
<https://doi.org/10.3390/inorganics13120395>.

Zhou W, Qu A, Cooper KW, Fortin N, Shahbaba B. A model-agnostic graph neural network for integrating local and global information. *J Am Stat Assoc.* 2025;120(550):1225-38.
<https://doi.org/10.1080/01621459.2024.2404668>.

Wen M, Horton MK, Munro JM, Huck P, Persson KA. An equivariant graph neural network for the elasticity tensors of all seven crystal systems. *Digit Discov.* 2024;3(5):869-82.

<https://doi.org/10.1039/D3DD00233K>.

Torlao V, Fajardo EA. Formation energy prediction of material crystal structures using deep learning. *Mater Res Express.* 2025;12(12):125501.

Sugiura T, Mizoguchi T. Achieving robust extrapolation in materials property prediction via decoupled transfer learning. *arXiv [Preprint]*. 2026. arXiv:2602.18054.

<https://doi.org/10.48550/arXiv.2602.18054>.

Cheng G, Gong XG, Yin WJ. Crystal structure prediction by combining graph network and optimization algorithm. *Nat Commun.* 2022;13(1):1492.

<https://doi.org/10.1038/s41467-022-29241-4>.

Wang J, Gao H, Han Y, Ding C, Pan S, Wang Y, et al. MAGUS: machine learning and graph theory assisted universal structure searcher. *Natl Sci Rev.* 2023;10(7):nwad128.