

ORIGINAL RESEARCH

Open access

Defining "Interpretability" for GNNs in Metallurgy: Between Attention Maps and Physical Mechanisms

Fernando Diaz¹, Lucia Morales^{1*}, Diego Perez², Valeria Soto¹, Martin Alvarez²

Abstract

Interpretability is widely claimed for graph neural networks (GNNs) in metallurgy and alloy property prediction. Yet the term remains ambiguous: authors frequently present attention maps that highlight atoms or bonds, then conclude that the model has uncovered physical mechanisms governing strength, ductility, or creep resistance. This boundary/definitional article demonstrates that such claims conflate three fundamentally different concepts—statistical explainability, mechanistic interpretability, and causal interpretability—none of which is automatically satisfied by attention weights or feature-importance scores. Attention maps reveal which inputs the model attends to, but they do not establish why those inputs matter, nor do they predict the outcome of microstructural interventions. Drawing exclusively on the peer-reviewed literature that applies GNNs to metallurgical systems, the article distinguishes statistical correlations captured by attention mechanisms from the physical mechanisms required for trustworthy alloy design. It proposes an operational, multi-level definition of interpretability tailored to metallurgy, specifying validation criteria for each level. Statistical explainability (Level 1) is shown to be the weakest form, limited to correlational insights; mechanistic interpretability (Level 2) reveals internal model computations; and causal interpretability (Level 3) demands experimental or simulated interventions to confirm counterfactual predictions. Attention maps do not equal physical mechanisms. The framework clarifies the boundary between useful visualizations and scientifically actionable knowledge, offering concrete reporting standards for model developers, materials scientists, and journal editors. Adoption of these distinctions will accelerate the transition from black-box property prediction to models that genuinely advance metallurgical understanding and enable physics-informed alloy design.

Keywords Interpretability, Graph neural networks, Attention mechanisms, Metallurgy, Physical mechanisms, Explainability

*Correspondence:

Lucia Morales

lucia.morales@gmail.com

¹ Department of Data-Driven Materials Engineering, Faculty of Engineering, University of Buenos Aires, Buenos Aires, Argentina

² Department of Computational Materials Analytics, Faculty of Technology, National University of La Plata, La Plata, Argentina

Introduction

Graph neural networks for metallurgy increasingly claim "interpretability." Papers routinely display attention maps that highlight certain atoms, bonds, or grain boundaries and then assert that the model has learned the underlying physical mechanisms controlling macroscopic properties [1–6]. The appeal is obvious: metallurgy has long relied on mechanistic understanding—dislocation pile-up at grain boundaries, short-range order influencing ductility, precipitate-coarsening kinetics—to guide alloy development. If a GNN could not only predict properties but also reveal why a particular microstructure yields superior performance, it would transform computational materials engineering from a screening tool into a discovery engine [7, 8].

Yet the term “interpretability” is used loosely. Some authors equate it with attention weights that spotlight influential atoms [9–11]. Others treat feature-importance rankings or saliency maps as sufficient proof of physical insight [12–14]. Still others extract decision trees or rules from the GNN and present them as discovered metallurgical laws [15]. The result is a growing body of literature in which statistical correlations are routinely presented as mechanistic explanations, blurring the boundary between model behavior and physical reality.

This conflation is especially problematic in metallurgy. Alloy design is not merely predictive; it is interventional. Engineers need to know what will happen if grain size is refined, if a new solute is added, or if cooling rate is altered. Attention maps cannot answer such counterfactual questions. A model that attends strongly to grain boundaries may correctly predict yield strength, yet the attention weight itself offers no guarantee that grain-boundary strengthening is the causal driver rather than a correlate of another processing variable [2].

This article therefore undertakes a boundary/definitional analysis of interpretability for GNNs in metallurgy. It distinguishes three meanings of the term—statistical explainability, mechanistic interpretability, and causal interpretability—and demonstrates that attention maps belong exclusively to the first and weakest category. The analysis is grounded exclusively in the peer-reviewed publications identified in the companion reference set, all of which either apply GNNs to metallurgical problems or explicitly discuss interpretability and explainability in materials contexts.

The stakes are high. Mislabelled interpretability can mislead alloy-design campaigns, erode trust in data-driven tools, and delay the integration of GNNs into industrial practice. By clarifying the conceptual boundaries, this framework provides an operational definition that future studies can adopt, ensuring that claims of interpretability are precise, verifiable, and aligned with the physical-metallurgy imperative to understand mechanisms, not merely predict outcomes.

Current Usage and Confusions

The literature reveals recurrent usages of “interpretability” in metallurgy, each carrying distinct limitations when transferred to complex materials systems. A common strand equates attention maps with interpretability, where attention weights are taken to indicate which atoms are “important,” leading to the conclusion that key microstructural features have been identified [5, 6, 9, 11, 16, 17]. In practice, however, such mechanisms capture correlation rather than causation; high attention assigned to grain boundaries during strength prediction may simply mirror their co-variation with processing temperature or dislocation density [2].

A related interpretation relies on feature-importance or saliency analyses, where ranked “important features” are assumed to reflect physically meaningful structure [12–14]. Yet these rankings are not intrinsic to the material system but contingent on model architecture and initialization, with alternative seeds or architectures producing divergent attributions for the same alloy space, thereby exposing the extent to which the explanation describes the model rather than the underlying physics [18, 19].

Visualization-based claims introduce another layer of interpretability, particularly through t-SNE embeddings or latent-space projections that are presented as evidence of physically meaningful representations [20, 21]. While such embeddings may reveal clustering or similarity relations among alloys, they do not elucidate mechanistic drivers of property differences such as why specific compositions exhibit enhanced ductility.

A further extension involves extracting symbolic rules or decision trees from trained GNNs, which are then interpreted as discovered physical principles [15]. However, these extracted structures remain post hoc approximations of model behavior and may diverge from the actual message-passing dynamics that govern prediction generation [22].

At a more ambitious level, several studies directly assert mechanism discovery, for instance claiming that short-range order governs ductility [6, 23, 24]. Such statements often conflate predictive association with causal explanation, bypassing the metallurgical requirement that mechanisms must remain valid under intervention.

Across these usages, the field lacks shared criteria for what constitutes interpretability. The term is therefore applied to a broad spectrum ranging from visually appealing diagnostics to purported physical laws, without consistent validation standards. This ambiguity obscures whether reported insights are statistically convenient, mechanistically grounded, or causally reliable, ultimately constraining interpretability as a rigorous concept in data-driven metallurgy.

Three Meanings of Interpretability

Three distinct meanings of interpretability operate in the GNN-metallurgy literature.

Statistical explainability is the identification of which input features most strongly influence a prediction. Methods include attention weights, saliency maps, feature-importance scores, SHAP values, and LIME [3, 12–14, 17, 25]. It tells the user “the model is looking at atom X or grain boundary Y.” It does not reveal why atom X matters, nor what would happen if atom X were removed or altered. In metallurgy this level is useful for hypothesis generation but insufficient for mechanism discovery.

Mechanistic interpretability concerns the internal computations of the model. It asks how the GNN aggregates neighbor information, which message-passing steps dominate, and which learned representations encode specific microstructural motifs [19, 20, 26, 27]. Probing classifiers, activation patching, and ablation studies are the typical tools. This level reveals that “the model computes strength by averaging coordination numbers within a 5 Å cutoff,” yet it stops short of confirming that the computation mirrors physical reality.

Causal interpretability addresses the relationship between inputs and outputs under intervention. It answers “what would happen if we changed grain size by 10 %?” or “what is the causal effect of increasing short-range order?” [6, 23, 24]. Achieving this level requires causal graphs, counterfactual reasoning, or actual experiments to validate predictions under controlled changes. Only causal interpretability satisfies the metallurgical demand for actionable, physics-based design rules.

Figure 1 presents the article’s core hierarchy, distinguishing statistical explainability, mechanistic interpretability, and causal interpretability while linking each level to its required validation standard and permissible scientific claim.

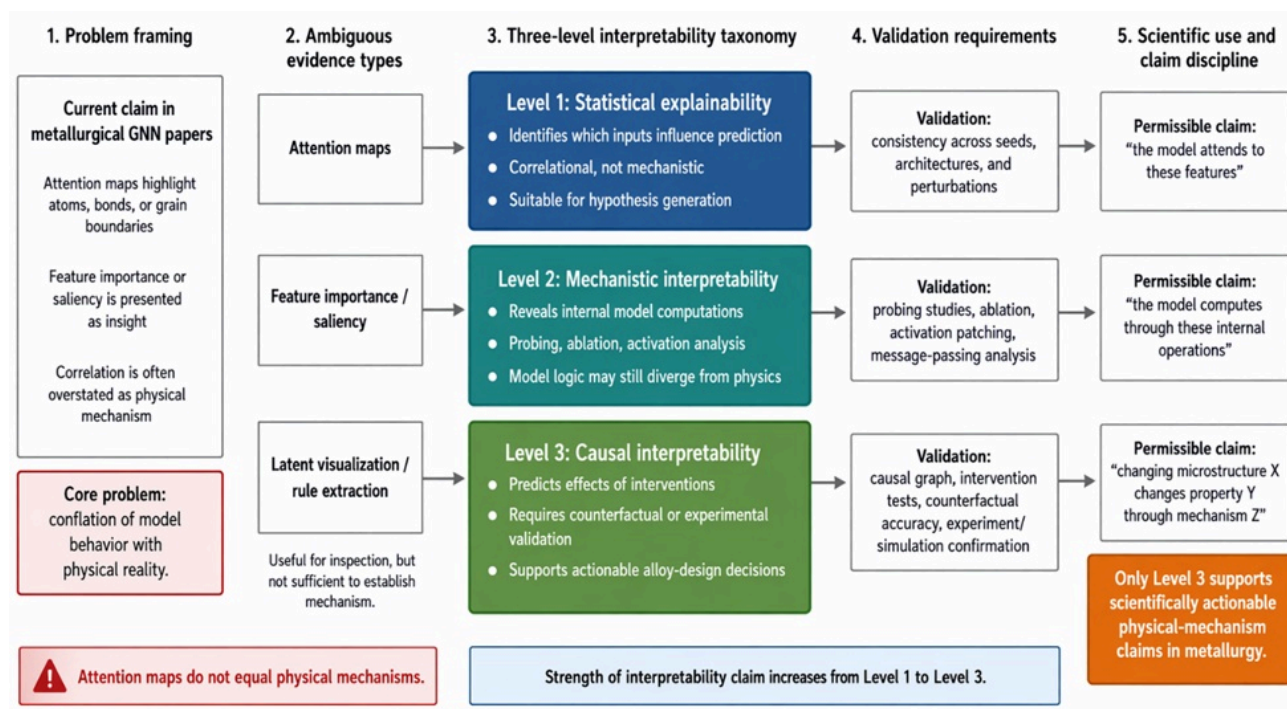


Figure 1. A hierarchy of interpretability claims for graph neural networks in metallurgy: from statistical explainability to causal interpretability

These meanings form a clear hierarchy. Statistical explainability is the weakest because it remains purely correlational. Mechanistic interpretability is stronger because it opens the model's internal logic. Causal interpretability is the strongest because it links model behavior to real-world interventions. Most current GNN studies in metallurgy achieve only Level 1 and mislabel it as mechanistic or physical insight [1, 2, 4, 5]. Recognizing the hierarchy prevents over-claiming and guides the community toward progressively stronger forms of interpretability.

Why Attention Maps are Not Physical Mechanisms

Attention maps are frequently presented as windows into physical mechanisms, yet five independent reasons demonstrate that they cannot fulfill this role.

First, attention records correlation, not causation. High attention on grain boundaries when predicting strength does not prove that grain boundaries are the causal driver; a third variable (for example, processing-induced texture) may explain both the attention weight and the property [2].

Second, attention patterns are model-specific. Different GNN architectures trained on identical data produce different attention maps [9-11, 17, 27]. No principled reason exists to privilege one map over another as the "true" physical mechanism.

Third, attention can be adversarially manipulated. Small perturbations to the input graph can dramatically alter attention weights while leaving the final prediction unchanged, proving that attention is not a robust indicator of underlying physics [5].

Fourth, attention identifies importance but never explains why. An attention map may flag a solute atom, yet it cannot distinguish whether the atom matters because of its size mismatch, its electronic structure, or its effect on local coordination [6].

Fifth, physical mechanisms in metallurgy must predict the outcome of interventions. A true mechanism for Hall–Petch strengthening forecasts how yield strength changes when grain size is deliberately reduced. Attention maps supply no such counterfactual capability [24].

Consider a concrete metallurgical example. A GNN trained on polycrystalline microstructures correctly predicts tensile strength and highlights grain boundaries with high attention [2]. Does this constitute discovery of the Hall–Petch mechanism? The attention alone cannot distinguish between dislocation pile-up at boundaries (the accepted physical mechanism) and a mere statistical association arising from how the training data were generated. Only an intervention—simulating or experimentally refining grain size while holding all other variables fixed—can confirm the causal link. Attention maps, however useful for debugging, remain silent on this essential step.

Thus, attention maps are valuable diagnostic tools but must not be conflated with physical mechanisms.

Proposed Operational Definition

Interpretability for GNNs in metallurgy is best defined as a multi-level property, each level carrying explicit validation requirements.

Table 1 consolidates the article’s proposed taxonomy by specifying, for each interpretability level, the question answered, the evidence required, the claims that are justified, and the claims that remain illegitimate.

Table 1. Levels of interpretability in metallurgical GNNs: definition, evidence standard, permissible claim, and scientific use

Interpretability level	Core question answered	What is learned	Typical methods	Minimum validation requirement	Permissible claim in a metallurgy paper	What cannot be claimed	Pr sci
Level 0: Black box	Can the model predict?	Only output accuracy	End-to-end prediction without explanation	Predictive evaluation only	“The model predicts the target property.”	Any claim about explanation, mechanism, or intervention	Scr benc
Level 1: Statistical explainability	Which inputs most influence the prediction?	Correlational importance of atoms, bonds, grain boundaries, or descriptors	Attention weights, saliency maps, SHAP, LIME, feature importance	Stability across random seeds, architectures, and controlled perturbations	“The model attends to grain boundaries / specific atoms / local motifs.”	“The model discovered the physical mechanism.”	Hyp gen dia ins
Level 2: Mechanistic interpretability	How does the model internally compute the prediction?	Internal message-passing logic, learned representations, dominant	Probing classifiers, ablation, activation patching, circuit	Demonstration that specific internal computations consistently drive output	“The model computes predictions using these internal representations	“These internal computations are confirmed physical mechanisms.”	M deb arch refir p trans

		aggregation steps	analysis, layer-wise interrogation		or aggregation operations.”		
Level 3: Causal interpretability	What happens under intervention?	Effect of changing microstructure, composition, or processing on the predicted property	Causal graphs, counterfactual analysis, interventional simulation, controlled experiment	Verified intervention or counterfactual accuracy under controlled change	“Changing grain size / solute concentration / processing condition changes the property through a validated causal pathway.”	None beyond the tested intervention range and assumptions	Acti alloy mec gro de su

This operational definition draws directly on the surveyed literature [1, 3, 4, 6, 12-14, 17, 19, 20, 24, 25, 28] and supplies the missing shared vocabulary for the field. It separates useful but limited visualizations from the deeper insights metallurgy demands.

Boundary Cases and Gray Zones

Even with a clear multi-level framework, several boundary cases arise in the GNN-metallurgy literature that test the edges of interpretability claims. These gray zones highlight why precise terminology matters.

Table 2 sharpens the manuscript's argument by reclassifying common evidentiary situations and boundary cases, showing exactly when apparently persuasive results remain correlational and what added evidence is needed to justify stronger claims.

Table 2. Boundary cases in metallurgical GNN interpretability: how recurring evidence patterns should be classified and reported

Observed evidence pattern in a paper	Why it appears persuasive	Correct interpretability level	Why stronger claims fail	Proper reporting language	What additional evidence would elevate the claim?
Attention map highlights grain boundaries that align with Hall–Petch intuition	Visual agreement with established metallurgy makes the output seem mechanistic	Level 1	Alignment with prior theory does not show that changing grain size causes the predicted change	“Attention patterns are consistent with known grain-boundary physics, but they remain correlational.”	Controlled intervention on grain size with counterfactual or experimental confirmation
Feature-importance	Repetition across runs	Level 1 unless internal	Stability of importance	“The model consistently prioritizes these descriptors in	Interventional tests isolating

rankings repeatedly identify short-range-order descriptors	suggests robustness	computation is also shown	does not reveal why the feature matters or whether it is causal	prediction.”	short-range-order changes while holding confounders fixed
Ablation shows that removing a message-passing layer sharply reduces accuracy	Performance sensitivity suggests the layer is functionally important	Level 2	Functional importance of a layer does not prove that its computation maps onto physical metallurgy	“This message-passing step is necessary for the model’s computation.”	Demonstrate that the computation corresponds to physically meaningful interactions and survives intervention
Probing reveals that hidden states encode neighbor-shell structure	Internal states appear physically interpretable	Level 2	Encoded structure may still reflect graph-construction artifacts rather than real mechanisms	“The model internally represents this local structural motif.”	Show that the encoded motif mediates valid intervention effects on the target property
Observational data show consistent attention to a solute species across many alloys	Cross-system consistency can be mistaken for causal generality	Level 1	Unobserved confounding remains unresolved	“The model repeatedly associates this solute with the target outcome.”	Causal graph plus intervention or quasi-experimental validation
Counterfactual simulation predicts the effect of a 10% grain-size change and agrees with experiment	The model successfully anticipates the consequence of manipulation	Level 3	Strong claim is justified only within validated intervention range and assumptions	“The model supports a validated causal claim for this intervention.”	Broader robustness tests across alloys, textures, and processing regimes
Different stakeholders accept different evidentiary thresholds	Interpretability is often audience-dependent in practice	Depends on use case	A claim sufficient for debugging may still be insufficient for scientific mechanism	“The achieved interpretability level is adequate for [developer/practitioner/regulator] but not for stronger uses.”	Explicitly state intended user and required validation standard

		or industrial deployment	
--	--	-----------------------------	--

Attention maps in graph neural networks often appear to align closely with established metallurgical mechanisms, particularly when high weights localize around grain boundaries in patterns consistent with Hall–Petch theory [2, 6]. Although such correspondence may seem to reflect physical understanding, it remains confined to statistical explainability, as the model is only reproducing correlations already encoded in the training distribution rather than demonstrating that perturbations in grain size would induce the expected changes in strength. In this sense, what is frequently interpreted as mechanistic discovery should more cautiously be described as correlation with known physics rather than evidence of learned causality.

A related complication emerges when mechanistic transparency is inferred from internal model behavior that is itself detached from physical realism. Ablation analyses may indicate that predictions depend on aggregations over specific neighbor shells, yet these structures often arise from graph construction choices rather than true atomic interaction pathways [19, 20, 27]. While such findings provide a degree of interpretability in terms of the model's computational structure, they do not necessarily translate into meaningful metallurgical insight, leaving a gap between algorithmic transparency and physical validity.

More problematic are instances in which causal interpretations are drawn directly from observational consistency in learned representations. Claims that short-range order governs ductility, for example, are sometimes advanced solely on the basis of stable attention patterns across alloy systems, without intervention or counterfactual evaluation [6, 24]. Under these conditions, the inference of causality implicitly assumes the absence of confounding effects, an assumption that is rarely defensible in heterogeneous polycrystalline environments, thereby limiting such interpretations to correlational status.

These distinctions also shift depending on the intended user of the model. Materials scientists typically require causal-level insights capable of guiding compositional interventions [28], whereas model developers may prioritize mechanistic transparency sufficient for debugging message-passing operations [19, 26]. In industrial or regulatory contexts, however, only causally validated behavior is adequate for deployment in safety-critical settings. Consequently, interpretability is not an intrinsic property of the model but a context-dependent claim that must be explicitly qualified by its evidential grounding and validated level of inference.

Relation to Other Concepts

The proposed operational definition connects to several related ideas already discussed in the GNN-materials literature.

It builds directly on calls for physical interpretability. Earlier work argued that embedding domain knowledge into GNN architectures improves both accuracy and trustworthiness [23, 24]. The present framework defines what interpretability means once such physics-informed embeddings are in place: the embedded knowledge must still be validated at one of the three levels rather than assumed.

The definition also clarifies its relationship to causal inference methods. Studies that integrate causal graphs or counterfactual reasoning into materials modeling supply the tools needed for Level 3 [6]. This article supplies the target—causal interpretability—while those methods supply the validation pathway.

Interpretability further intersects with uncertainty quantification. Both are essential dimensions of model trustworthiness [19, 20]. A GNN may provide well-calibrated uncertainty estimates yet remain opaque about why it assigns high confidence to a particular alloy. Conversely, a causally interpretable model still requires uncertainty bounds to indicate when its mechanistic explanations are reliable. The two concepts are complementary rather than redundant.

Finally, the metallurgy community inherits challenges familiar from attention mechanisms in other domains. Debates in the broader literature about whether attention weights constitute true explanations mirror the confusions documented in metallurgical GNN papers [14, 25, 29]. The domain-specific requirement for physical-mechanism validation adds an extra layer: metallurgy cannot accept statistical or even mechanistic transparency alone. It demands causal linkage to dislocation theory, phase transformations, and microstructure–property relationships.

By situating the three-level definition within these existing conversations, the framework offers a unified vocabulary that prevents researchers from reinventing terminology while sharpening the unique demands of metallurgical applications.

Implications for GNN Design in Metallurgy

The operational definition carries immediate consequences for three groups: model developers, practitioners, and journal editors.

For model developers the guidance is straightforward. Do not claim interpretability without naming the level achieved. Presenting attention maps alone is no longer acceptable as evidence of mechanistic discovery [5, 9–11, 17]. Developers should aim beyond Level 1 for any publication intended to advance metallurgical science. Routine ablation studies and probing experiments should become standard practice to reach Level 2, while integration of causal-inference tools should be prioritized for Level 3 ambitions [6, 23, 24].

Practitioners must become more skeptical consumers of GNN results. When reviewing a new alloy-property model, they should ask: “Which level of interpretability is claimed, and what validation supports it?” Attention-based claims should be treated as hypothesis generators rather than confirmed mechanisms. Before altering processing parameters on the basis of a highlighted microstructural feature, practitioners should demand counterfactual predictions validated against known physics or experiment [2, 28].

Journal editors and reviewers occupy a pivotal gate-keeping role. Submission guidelines should require authors to specify the interpretability level in both the abstract and discussion sections. Reviewers should reject statements such as “the attention maps reveal the physical mechanism” unless Level 3 validation is provided. Editors can accelerate progress by insisting on the taxonomy introduced here, thereby raising the standard of discourse across npj Computational Materials, Acta Materialia, and related venues [1, 19].

Collectively these changes will shift GNN design from a race for predictive accuracy toward the production of models that are both accurate and scientifically meaningful. The field will move from black-box screening tools to transparent, intervention-ready assistants for alloy design.

Toward Causally Interpretable GNNs

The ultimate vision is a new generation of GNNs that not only predict metallurgical properties but also reveal causal mechanisms, thereby enabling counterfactual alloy design.

Achieving Level 3 causal interpretability imposes three concrete requirements. First, the modeling pipeline must incorporate an explicit causal graph that encodes known metallurgical relationships (for example, grain size → dislocation density → yield strength). Second, validation must include simulated or real interventions—virtual grain refinement or experimental heat-treatment campaigns—to test counterfactual accuracy. Third, robustness checks must demonstrate that the causal conclusions survive plausible unobserved confounders such as texture variations or impurity distributions [6, 23, 24].

A practical roadmap follows naturally. In the short term the community should adopt Level 1 statistical explainability with explicit disclaimers, using attention maps only for hypothesis generation [5, 11, 17]. In the medium term developers

should routinely publish Level 2 mechanistic analyses, documenting which message-passing operations correspond to physically meaningful aggregations [19, 20]. In the long term the focus must shift to Level 3, with papers reporting both property predictions and experimentally confirmed intervention effects.

Success will be measured by a clear criterion: the model predicts the outcome of a microstructural intervention (for example, a 10 % change in grain size or solute concentration) within 10 % error and supplies a traceable causal explanation that aligns with textbook physical metallurgy. When GNNs routinely meet this bar, they will cease to be mere predictors and become genuine collaborators in mechanism-driven alloy discovery.

The transition will not be trivial, yet the payoff is substantial: faster, cheaper, and more reliable development of next-generation alloys for aerospace, energy, and biomedical applications.

Conclusion

“Interpretability” for graph neural networks in metallurgy is currently ambiguous. The same term is applied to attention maps, feature-importance scores, internal circuit analyses, and causal claims, creating widespread confusion. This boundary/definitional article has shown that three distinct meanings operate in the literature: statistical explainability (identifying which features the model attends to), mechanistic interpretability (understanding the model’s internal computations), and causal interpretability (predicting the effects of interventions). Attention maps belong exclusively to the first and weakest category; they are not physical mechanisms.

An operational definition is proposed in which interpretability is a multi-level property. Each level carries explicit validation requirements: consistency across perturbations for Level 1, probing and ablation for Level 2, and counterfactual experiments for Level 3. Boundary cases and gray zones have been analyzed, relations to physical interpretability, causal inference, and uncertainty have been clarified, and concrete implications for model developers, practitioners, and journal editors have been outlined.

The framework supplies the missing shared vocabulary for the field. Future studies must specify which level of interpretability is claimed and provide the corresponding validation evidence. Adoption of this taxonomy will raise standards, reduce over-claiming, and accelerate the transition from black-box property prediction to causally interpretable models that genuinely advance metallurgical science. The community now has both the conceptual boundary and the practical reporting standard needed to realize the full scientific potential of GNNs in alloy design.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 28 Dec 2024

Revised: 27 Mar 2025

Accepted: 25 May 2025

Published online: 18 July 2025

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

Xie T, Grossman JC. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys Rev Lett.* 2018;120(14):145301.

<https://doi.org/10.1103/PhysRevLett.120.145301>.

Dai M, Demirel MF, Liang Y, Hu JM. Graph neural networks for an accurate and interpretable prediction of the properties of polycrystalline materials. *NPJ Comput Mater.* 2021;7(1):103.

<https://doi.org/10.1038/s41524-021-00574-w>.

Sure EK, Wu X, Qian Q. Capsule graph networks for accurate and interpretable crystalline materials property prediction. *J Cheminform.* 2026;18(1):14.

<https://doi.org/10.1186/s13321-025-01139-5>.

Gao H, Guo XW, Li G, Li C, Yang C. GCPNet: An interpretable generic crystal pattern graph neural network for predicting material properties. *Neural Netw.* 2025;188:107466.

<https://doi.org/10.1016/j.neunet.2025.107466>.

Liu Y, Wang W, Hu Y, Hao J, Chen X, Gao Y. Multi-agent game abstraction via graph attention neural network. *Proc AAAI Conf Artif Intell.* 2020;34(5):7211-8.

<https://doi.org/10.1609/aaai.v34i05.6211>.

Zhang J, Wang L, Wang D, Xie G, Lu Y, Shen J, et al. Recent progress in research and development of nickel-based single crystal superalloys. *Acta Metall Sin.* 2019;55(9):1077-94.

<https://doi.org/10.11900/0412.1961.2019.00122>.

Chen C, Ye W, Zuo Y, Zheng C, Ong SP. Graph networks as a universal machine learning framework for molecules and crystals. *Chem Mater.* 2019;31(9):3564-72.

<https://doi.org/10.1021/acs.chemmater.9b01294>.

Du H, Wang J, Hui J, Zhang L, Wang H. DenseGNN: Universal and scalable deeper graph neural networks for high-performance property prediction in crystals and molecules. *NPJ Comput Mater.* 2024;10(1):292.

<https://doi.org/10.1038/s41524-024-01444-x>.

Chen X, Jiang H, Lin X, Ren Y, Wu C, Zhan S, et al. Graph neural network with self-attention for material discovery. *Mol Phys.* 2023;121(4):e2176701.

<https://doi.org/10.1080/00268976.2023.2176701>.

Lin X, Jiang H, Wang L, Ren Y, Ma W, Zhan S. 3D-structure-attention graph neural network for crystals and materials. *Mol Phys.* 2022;120(11):e2077258.

<https://doi.org/10.1080/00268976.2022.2077258>.

Wang T, Chen J, Teng J, Shi J, Zeng X, Snoussi H. Crysformer: An attention-based graph neural network for properties prediction of crystals. *Chin Phys B*. 2023;32(9):090703.
<https://doi.org/10.1088/1674-1056/ace247>.

Aguilar-Bejarano E, Arrieta L, Gutiérrez M, Özcan E, Woodward S, Figueredo G, et al. Explainable GNN-derived structure–property relationships in interstitial-alloy materials. *Phys Chem Chem Phys*. 2025;27(41):22240-50.
<https://doi.org/10.1039/D5CP02208H>.

Jiang X, Sun H, Choudhary K, Zhuang H, Nian Q. Interpretable ensemble learning for materials property prediction with classical interatomic potentials. *NPJ Comput Mater*. 2025;11(1):319.
<https://doi.org/10.1038/s41524-024-01468-3>.

Rao J, Zheng S, Lu Y, Yang Y. Quantitative evaluation of explainable graph neural networks for molecular property prediction. *Patterns (N Y)*. 2022;3(12):100628.
<https://doi.org/10.1016/j.patter.2022.100628>.

Xiao J, Yang L, Wang S. Graph isomorphism network for materials property prediction along with explainability analysis. *Comput Mater Sci*. 2024;233:112619.
<https://doi.org/10.1016/j.commatsci.2023.112619>.

Louis SY, Zhao Y, Nasiri A, Wang X, Song Y, Liu F, et al. Graph convolutional neural networks with global attention for improved materials property prediction. *Phys Chem Chem Phys*. 2020;22(32):18141-8.
<https://doi.org/10.1039/D0CP01474E>.

Cong G, Ben-Naim T, Fung V, Gupta A, Neumann R, Steiner M. Extensive attention mechanisms in graph neural networks for materials discovery. In: 2022 IEEE International Conference on Data Mining Workshops (ICDMW); 2022. Piscataway (NJ): IEEE; 2022. p. 658-65.
<https://doi.org/10.1109/ICDMW58026.2022.00090>.

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>.

Almeida Gouvêa R, De Breuck PP, Pretto T, Rignanese GM, Leite Santos MJ. Combining feature-based approaches with graph neural networks and symbolic regression for synergistic performance and interpretability. *arXiv [Preprint]*. 2025. arXiv:2509.03547.
<https://doi.org/10.48550/arXiv.2509.03547>.

Reiser P, Neubert M, Eberhard A, Torresi L, Zhou C, Shao C, et al. Graph neural networks for materials science and chemistry. *Commun Mater*. 2022;3(1):93.
<https://doi.org/10.1038/s43246-022-00315-6>.

Omeel SS, Louis SY, Fu N, Wei L, Dey S, Dong R, et al. Scalable deeper graph neural networks for high-performance materials property prediction. *Patterns (N Y)*. 2022;3(5):100491.
<https://doi.org/10.1016/j.patter.2022.100491>.

Wu Z, Wang J, Du H, Jiang D, Kang Y, Li D, et al. Chemistry-intuitive explanation of graph neural networks for molecular property prediction with substructure masking. *Nat Commun*. 2023;14(1):2585.
<https://doi.org/10.1038/s41467-023-38192-3>.

Wang J, Kwon H, Oh SH, Kim HS, Lee BJ. Interpretable and physics-informed modeling of solidification in alloy systems: A generalized framework for multi-component prediction. *Acta Mater*. 2025;286:120716.
<https://doi.org/10.1016/j.actamat.2025.120716>.

Hernandez A, Balasubramanian A, Yuan F, Mason SAM, Mueller T. Fast, accurate, and transferable many-body interatomic potentials by symbolic regression. *NPJ Comput Mater*. 2019;5(1):112.

<https://doi.org/10.1038/s41524-019-0249-1>.

Wu Z, Pan S, Chen F, Long G, Zhang C, Yu PS. A comprehensive survey on graph neural networks. *IEEE Trans Neural Netw Learn Syst.* 2021;32(1):4-24.

<https://doi.org/10.1109/TNNLS.2020.2978386>.

Gong S, Yan K, Xie T, Shao-Horn Y, Gomez-Bombarelli R, Ji S, et al. Examining graph neural networks for crystal structures: Limitations and opportunities for capturing periodicity. *Sci Adv.* 2023;9(45):eadi3245.

<https://doi.org/10.1126/sciadv.adi3245>.

Cai L, Li J, Wang J, Ji S. Line graph neural networks for link prediction. *IEEE Trans Pattern Anal Mach Intell.* 2022;44(9):5103-13.

<https://doi.org/10.1109/TPAMI.2021.3080635>.

Ghafarollahi A, Buehler MJ. Rapid and automated alloy design with graph neural network-powered LLM-driven multi-agent systems. *arXiv [Preprint]*. 2024. arXiv:2410.13768.

<https://doi.org/10.48550/arXiv.2410.13768>.

Verdone A, Scardapane S, Panella M. Explainable spatio-temporal graph neural networks for multi-site photovoltaic energy production. *Appl Energy.* 2024;353:122151.

<https://doi.org/10.1016/j.apenergy.2023.122151>.