

ORIGINAL RESEARCH

Open access

A Critical Examination of "State-of-the-Art" Claims in Materials GNNs: Benchmark Artifacts and Unstated Limitations

Ahmed Benali^{1*}, Karim Boudiaf², Samir Touati¹

Abstract

"State-of-the-art" (SOTA) claims in materials graph neural networks are typically grounded in marginal improvements on standard benchmarks, yet such gains often reflect properties of the evaluation protocol rather than advances in generalization. This analysis demonstrates that widely used practices—particularly random train–test splits, inconsistent baselines, and selective reporting—systematically inflate reported performance. At the same time, critical constraints, including interpolation-dominated evaluation, computational cost, limited extrapolation, and dataset bias, remain largely unacknowledged. Drawing on the 2017–2023 literature, the study shows that these artifacts and omissions distort how progress is measured and interpreted. By reframing SOTA as a context-dependent rather than universal claim, it proposes a shift toward multi-split, multi-metric, and reproducible evaluation that more accurately reflects real-world materials discovery conditions.

Keywords Materials graph neural networks, State-of-the-art claims, Benchmark artifacts, Unstated limitations, Generalization in materials property prediction, Reproducibility issues in materials machine learning

*Correspondence:

Ahmed Benali

ahmed.benali@gmail.com

¹ Department of Computational Materials Systems, Faculty of Engineering, University of Algiers, Algiers, Algeria

² Department of Data-Driven Materials Analytics, Faculty of Science and Technology, University of Tunis El Manar, Tunis, Tunisia

Introduction

Every week, a new paper claims "state-of-the-art" results on crystal property prediction. But what does SOTA actually mean? Often, it means: "Our model achieves a lower MAE on a specific test set than previous models reported." This is problematic. Benchmark artifacts—data leakage, unfair comparisons, test-set overfitting—inflate reported performance [1]. Unstated limitations—test set similarity to training, computational cost, extrapolation failure—are never discussed. This critique examines SOTA claims in materials GNNs, identifies common artifacts and unstated limitations, and proposes rigorous evaluation standards.

The adoption of GNNs in materials science accelerated dramatically after Xie and Grossman introduced crystal graph convolutional neural networks [2] that demonstrated accurate and interpretable prediction of material properties [3]. Subsequent works rapidly positioned GNNs as a universal framework for molecules and crystals [4], established deep learning architectures tailored to atomic systems [5, 6], and introduced E(3)-equivariant networks that improved data efficiency for interatomic potentials [7]. Performance assessments further reinforced the narrative of steady progress [8], while recent scaling studies have pushed the frontier toward large-scale discovery [9]. Comprehensive reviews have catalogued the expanding toolbox of GNN architectures for materials science and chemistry [10–12], and specialized

models have targeted polycrystalline systems [13], polycrystalline property prediction [13], and charge-informed modelling [14]. At first glance, the field appears to be on an unbroken upward trajectory.

Yet beneath the steady stream of SOTA announcements lies a more troubling pattern. Papers routinely cite prior GNNs such as CGCNN [2], SchNet [5], or NequIP [7] as baselines, declare a new architecture superior on one or two benchmark tasks, and conclude that the proposed model is now state-of-the-art. The implied promise—that the new method will accelerate materials discovery—is rarely qualified by the conditions under which superiority was demonstrated. As benchmarking studies have begun to reveal [15], the evaluation protocols themselves contain structural flaws that reward memorization over physical insight. Reproducibility studies and limitation-focused analyses further suggest that reported gains may not survive changes in train-test partitioning or chemical composition [16, 17].

The consequences extend beyond academic leaderboard competition. Materials engineers and experimentalists increasingly rely on published GNN predictions to prioritize synthesis targets or interpret measurements. If SOTA claims are systematically overstated, the field risks directing finite experimental resources toward models that fail when confronted with real deployment scenarios: new element combinations, disordered structures, or properties outside the narrow training distribution. Moreover, the pressure to claim SOTA incentivizes researchers to optimize for benchmark artifacts rather than address fundamental challenges such as long-range interactions, uncertainty quantification, or scalability [18].

This article therefore adopts a critical stance grounded exclusively in the 2017–2023 literature. It dissects the rhetorical structure of SOTA claims, catalogues the benchmark artifacts that inflate them, surfaces the limitations that remain unstated, and categorizes the resulting misleading assertions into a typology. The goal is not to diminish the genuine contributions of GNN research but to clarify what “state-of-the-art” can and cannot mean under current evaluation practices. Only by acknowledging these artifacts and limitations can the community adopt evaluation standards that reflect the true demands of computational and data-driven materials engineering.

Figure 1 maps how benchmark artifacts and unstated limitations interact to convert narrow benchmark advantages into inflated state-of-the-art claims and downstream distortions in materials GNN evaluation.

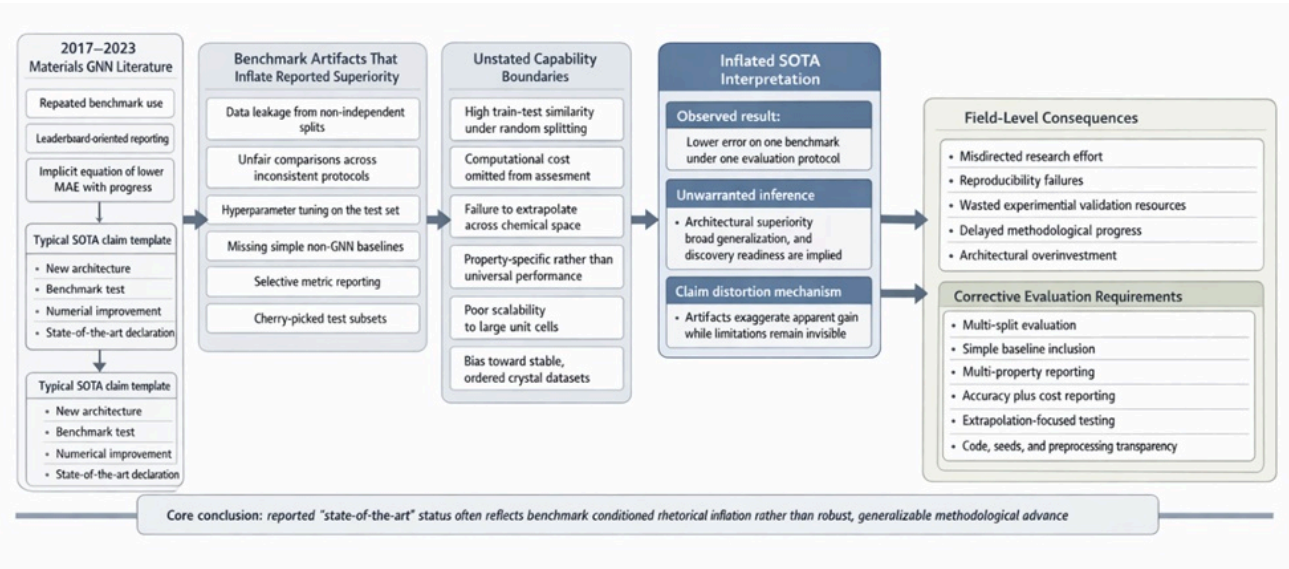


Figure 1. Analytical Architecture of Artifact-Driven “State-of-the-Art” Claims in Materials GNN Evaluation

The Anatomy of a SOTA Claim

A typical SOTA claim follows a predictable four-step pattern. Authors introduce a new architecture, usually framed as an extension of message passing, equivariant layers, or attention mechanisms [19, 20]. They then evaluate the model on a standard benchmark, most often formation-energy or band-gap prediction in the Materials Project [21]. Next, they report a marginally lower MAE than prior models and present that numerical advantage as evidence of superior performance. On this basis, the model is declared state of the art. Variants of this pattern recur from early crystal GNNs [2] to universal graph networks [4], scalable deeper architectures [22], and pretrained charge-informed models [14].

What this structure omits is any substantive treatment of statistical significance, computational cost, generalization scope, or reproducibility. An improvement of 0.001 eV/atom in MAE may appear meaningful yet remain statistically indistinguishable once variation from random seeds or train–test splits is taken into account. Computational cost is rarely evaluated with the same seriousness; a model that is an order of magnitude slower may still be labeled SOTA on accuracy alone, even though high-throughput screening depends critically on throughput and efficiency [8, 9, 23]. Claims about generalization are likewise left unstated but assumed: performance on a single benchmark is taken to imply broader superiority across materials-property prediction, despite evidence that model behavior is strongly property dependent [17]. Reproducibility is weakened further when train–test splits, hyperparameter ranges, random seeds, and preprocessing pipelines are omitted from the main text or displaced to supplementary material without integration into the central argument [24].

The resulting claim is therefore much broader in implication than in substance. What is actually established is narrow and conditional: the model attains lower error on a specific test set under a particular evaluation protocol. What is implicitly suggested, however, is that the model is generally superior for materials-property prediction, a conclusion that is seldom warranted and often sits uneasily with the literature being cited. Comparative studies of machine-learning interatomic potentials, for example, show that no single architecture dominates across all properties or computational regimes [8]. Even so, many SOTA-oriented papers suppress this qualification.

Review articles have documented the evolution of the field and its unresolved challenges [10, 25], yet they often reproduce the same rhetorical logic by cataloguing successive benchmark improvements without questioning the evaluation frameworks that produced them. The consequence is a literature in which numerical gains on isolated benchmarks are treated as evidence of scientific advance, while robustness, cost-efficiency, and practical utility remain secondary. In this light, the SOTA claim says less about intrinsic model superiority than about the benchmark and evaluation design through which superiority was defined, choices that are rarely justified and even less often diversified.

Benchmark Artifacts

Benchmark artifacts systematically distort reported performance in materials GNN literature. Five primary artifacts recur across the 2017–2023 corpus.

Table 1 specifies how each recurrent benchmark artifact undermines a distinct dimension of evidentiary validity, thereby showing why a lower benchmark error alone cannot justify a robust state-of-the-art claim.

Table 1. Benchmark Artifact–Validity Mapping for Evaluating the Credibility of State-of-the-Art Claims in Materials GNNs

Benchmark artifact	Immediate evaluation distortion	Primary validity dimension compromised	How the artifact can falsely support a SOTA claim	What a rigorous paper must report instead
Data leakage from non-independent train-test splits	Test samples remain structurally or compositionally too	External validity	The model appears to generalize, but performance mainly reflects memorization of recurring structural motifs,	Split logic based on composition, prototype, time, or chemistry novelty; quantitative

	similar to training samples		prototypes, or composition families	degradation across split regimes
Unfair comparison across inconsistent preprocessing and metrics	Reported numbers are not directly comparable across papers	Comparative validity	A new model appears superior because the benchmark conditions differ from prior work rather than because the architecture is intrinsically better	Same dataset version, same preprocessing, same target definition, same metric suite, same split protocol across all baselines
Hyperparameter tuning on the test set	The test set becomes partially incorporated into model selection	Inferential validity	The final reported score looks like independent evidence but is already optimized against the evaluation target	Explicit validation protocol, pre-declared tuning budget, fixed test set used once only for final assessment
Missing simple non-GNN baselines	Practical value-added of architectural complexity is obscured	Incremental validity	The model is described as SOTA relative to earlier deep models even when simple composition-based or tabular baselines may perform competitively	Inclusion of linear, kernel, tree-based, and composition-only baselines under matched conditions
Selective reporting of the best metric only	Weaknesses on other properties or metrics remain invisible	Construct validity	“State-of-the-art” is inferred from one favorable indicator while broader task competence is unknown or worse	Multi-metric reporting, including accuracy, calibration, failure cases, variance across seeds, and property coverage
Cherry-picked test subsets	The evaluated task becomes easier than the implied deployment scenario	Ecological validity	The paper claims broad superiority on an unrepresentative high-symmetry or stable subset that excludes precisely the difficult regimes relevant to discovery	Transparent test-set composition, inclusion of hard regimes, and explicit statement of the operational domain of validity

Data leakage arises when random splits of the Materials Project place structurally or compositionally related instances across training and test partitions, enabling models to internalize recurring motifs rather than infer transferable physical structure [15]. This effect becomes particularly problematic given evidence that standard GNN architectures struggle to encode periodicity under non-independent sampling, as shown by Gong *et al.* [17], yet such leakage is rarely acknowledged in SOTA claims. A related distortion emerges through inconsistent evaluation protocols: variation in splits, preprocessing, and metrics undermines comparability, since results derived from random partitions, composition-based splits, or differing normalization schemes cannot be meaningfully aligned, even though prior values are often treated as fixed baselines [15]. This instability is further amplified when hyperparameters are implicitly tuned against the test set, effectively collapsing the distinction between validation and evaluation and inflating reported performance, a concern reinforced by reproducibility analyses demonstrating sensitivity to fixed configurations [16]. Beyond protocol design, selective benchmarking introduces additional bias, as comparisons are typically restricted to prior GNNs while omitting simpler baselines whose performance can be competitive in controlled studies [15, 26]. This selective framing is compounded by partial reporting practices that foreground a single favorable metric while neglecting degraded performance across other properties or regimes [22], alongside evaluations confined to favorable subsets of the data distribution, thereby masking failure modes in more challenging settings [17]. The cumulative effect is a self-reinforcing

cycle in which incremental numerical gains reflect optimization over benchmark artifacts rather than substantive advances in representation or generalization.

Unstated Limitations

Even when reported improvements appear compelling, their interpretation is constrained by structural limitations that are seldom made explicit. The reliance on random splits ensures that test data remain closely aligned with the training distribution, so benchmark success primarily reflects interpolation rather than deployment-relevant generalization [15, 17]. This issue is compounded by the omission of computational cost, as marginal accuracy gains may coincide with substantial increases in training or inference time, a trade-off that is particularly consequential in high-throughput screening contexts where efficiency is decisive [8, 9]. A deeper limitation concerns extrapolation: models trained within a fixed chemical manifold frequently fail when exposed to new elements, structural prototypes, or metastable configurations, undermining their utility in discovery-driven settings [17]. Performance heterogeneity across properties further complicates interpretation, since architectures optimized for formation energy often underperform on electronic or mechanical targets, despite being presented as broadly superior [4, 14, 27]. Scalability introduces an additional constraint, as benchmarks dominated by small unit cells obscure degradation effects that arise in larger systems due to memory limitations and representation smoothing [10, 13, 28]. Finally, dataset bias toward stable, ordered, high-symmetry structures restricts the validity of conclusions, as generalization to disordered or low-symmetry materials remains largely untested [17]. These constraints delineate the operational envelope of current models, and their omission systematically inflates perceived readiness for real-world deployment.

A Typology of Misleading SOTA Claims

Within this evaluative landscape, recurring patterns of overstated claims emerge that reflect structural weaknesses in benchmarking practice. One prominent form involves marginal numerical improvements presented as decisive advances, despite falling within the range of statistical variability and often requiring disproportionate architectural complexity [22]. A related pattern appears when comparisons exclude non-neural baselines, allowing incremental gains over prior GNNs to be framed as substantive progress even when simpler models perform competitively [15]. The problem intensifies when evaluation is restricted to a single metric, typically formation energy, while performance on other properties is omitted or comparatively weaker [14]. Claims of generalization frequently rest on random-split benchmarks that probe interpolation rather than true distributional shift, thereby overstating robustness [15, 17]. This distortion is reinforced by the systematic omission of computational cost, where accuracy gains are emphasized without accounting for substantial increases in resource requirements [9]. Reproducibility gaps further undermine credibility, as the absence of code, hyperparameters, or seed control renders reported improvements difficult to verify [16]. These patterns do not arise independently; rather, they exploit the same underlying vulnerabilities in data partitioning, evaluation design, and reporting practice, sustaining a literature in which benchmark-driven gains are routinely conflated with genuine methodological progress.

Table 2 converts the manuscript's critique into a decision framework that distinguishes defensible, conditional, weak, and misleading forms of state-of-the-art claims in materials GNN research.

Table 2. Decision Framework for Classifying the Strength of State-of-the-Art Claims in Materials GNN Studies

Claim classification	Minimum evidentiary condition	Typical empirical profile	What the authors are justified in claiming	What the authors are not justified in claiming	Recommended wording standard
Defensible conditional	Matched baselines, multiple split	Improvement persists across	Superiority under explicitly defined	Universal superiority across	“Under the reported evaluation regime,

SOTA claim	strategies, uncertainty reporting, computational profile, reproducible code and seeds	random and extrapolative splits; cost is disclosed; gains are not confined to a single property	benchmark conditions and task boundaries	materials tasks or discovery scenarios	model X outperforms included baselines for property P within the tested domain.”
Provisionally strong claim	Most core controls present, but one dimension remains incomplete	Improvement is replicated across several settings, but either cost, scalability, or one baseline family is missing	Strong evidence of task-specific improvement with partial generalizability	Field-wide state-of-the-art status	“Results indicate strong task-specific performance, though broader generalization requires further evaluation.”
Weak benchmark-bound claim	Single split, narrow metric suite, incomplete baseline set	Numerical gain is modest and context-dependent	A lower score on the reported benchmark only	Any general statement about methodological advance	“Model X achieves a lower error on this benchmark configuration.”
Artifact-sensitive claim	Evaluation design contains one or more major artifacts	Performance advantage disappears or may disappear under stricter protocol control	At most, a benchmark-specific observation requiring validation	State-of-the-art, generalization, robustness, or deployment readiness	“Observed gains should be interpreted cautiously because evaluation conditions may inflate performance.”
Misleading SOTA claim	Missing baselines, leakage-prone split, selective reporting, absent reproducibility details	Headline score is foregrounded while major validity threats remain unaddressed	No robust superiority claim is warranted	Any strong comparative or translational conclusion	Avoid SOTA language entirely until evaluation is redesigned.
Deployment-relevant superiority claim	All conditions for defensible SOTA claim plus extrapolation tests, cost disclosure, larger-cell scalability, and hard-domain evaluation	Model remains competitive under chemically novel, larger, or less idealized regimes	Evidence of practical superiority for a defined deployment scenario	Universal readiness for all materials discovery tasks	“Model X shows superior performance for the specified deployment-relevant scenario tested here.”

Case Studies of Problematic SOTA Claims

Representative examples from the 2017–2023 literature illustrate how benchmark artifacts and unexamined assumptions translate into misleading SOTA assertions. A recurring issue emerges in studies relying on random train–test splits for

Materials Project formation-energy prediction, where apparent gains reported in [2, 4, 22] reflect structural overlap between partitions rather than improved generalization; benchmarking analyses show that performance degrades substantially under composition- or prototype-based splits that better approximate discovery settings [15, 17], indicating that the claimed superiority is induced by the evaluation protocol itself. A related distortion appears when architectural comparisons omit relevant baselines, as non-equivariant GNNs are often evaluated only against earlier message-passing models such as CGCNN [2] or SchNet [5], while exclusion of E(3)-equivariant alternatives obscures the role of symmetry-aware inductive bias; once included, these baselines frequently negate the reported advantage, particularly for rotation-sensitive properties [7, 14]. Beyond architectural framing, scaling studies introduce a subtler misalignment by reporting marginal accuracy gains achieved through substantially increased computational cost, yet omitting runtime and resource metrics that determine practical utility in high-throughput contexts [8, 9]. This tension between numerical improvement and usability is mirrored in pretrained “universal” models that achieve strong results on formation energy while underperforming on other properties, revealing that single-task SOTA claims do not translate into general predictive competence across materials domains [4, 8, 14]. These cases reflect a consistent pattern in which narrow evaluation choices produce results that appear robust but fail under broader or more realistic conditions.

Consequences for the Field

The persistence of artifact-driven evaluation produces structural distortions that extend beyond individual studies and shape the trajectory of the field. One immediate effect is the redirection of research effort toward incremental benchmark optimization, diverting attention from fundamental challenges such as extrapolation, long-range interactions, and uncertainty quantification, thereby reinforcing a feedback loop centered on leaderboard performance rather than scientific insight [15, 17]. This dynamic contributes to a deepening reproducibility problem, as incomplete reporting of splits, hyperparameters, and initialization conditions renders many results difficult to verify, and corrections for leakage or baseline selection often eliminate the claimed advantage [16]. The downstream impact is particularly acute in experimental workflows, where reliance on overstated models leads to failed predictions outside the training distribution, consuming scarce resources in synthesis and validation [17]. At a broader level, the inability to distinguish genuine progress from artifact-driven improvement delays the exploration of alternative paradigms, including physics-informed or multi-fidelity approaches, while reinforcing concentration around a narrow set of architectures [9, 22, 29]. Over time, these dynamics erode confidence in materials machine learning as a reliable tool for discovery, shifting the literature toward competitive benchmarking rather than cumulative knowledge development.

Relation to Other Critiques

This analysis extends prior critiques by situating them within a unified account of how evaluation practices shape perceived progress. It builds on the random-split concerns formalized in Matbench benchmarking work [15], demonstrating that leakage is not confined to a single dataset but permeates a broad range of SOTA claims, particularly when combined with inconsistent baselines and selective reporting. A related connection arises with the periodicity limitations identified by Gong et al. [17], whose findings on symmetry and structural generalization are here integrated into a wider framework that includes computational cost, dataset bias, and property dependence, all of which constrain the validity of SOTA assertions. While reproducibility studies have emphasized transparency in code and configuration [16], the present argument highlights that methodological rigor at the implementation level is insufficient when evaluation protocols themselves embed systematic bias. This perspective also deepens extrapolation-focused critiques [17] by clarifying that interpolation-dominated benchmarks inherently overstate model readiness for discovery, as performance under random splits cannot be interpreted as evidence of robustness in chemically distinct regimes. By synthesizing these strands, the discussion reframes isolated methodological concerns as manifestations of a broader evaluative structure that governs how progress is measured and communicated.

Recommendations for Rigorous SOTA Evaluation

To replace artifact-driven SOTA claims with defensible statements of progress, concrete changes are required from authors, reviewers, and benchmark designers.

For authors: (a) report performance across multiple train-test splits (random, composition, prototype, and time-based) and explicitly quantify degradation when moving from interpolation to extrapolation regimes [15, 17]; (b) include simple non-GNN baselines such as composition-only random forests or linear models to establish practical added value [15, 26]; (c) publish full computational profiles—training time, inference latency, memory usage, and scaling with system size—alongside accuracy numbers [8, 9, 23]; (d) evaluate and report at least three physically relevant properties rather than a single headline metric [4, 14, 27]; (e) release complete code, exact hyperparameter settings, random seeds, and preprocessing scripts to enable independent verification [16, 24].

For reviewers: (a) require explicit discussion of potential benchmark artifacts and how the claimed SOTA survives changes in splitting protocol or baseline inclusion; (b) reject comparisons that omit simple baselines or use inconsistent preprocessing; (c) demand multi-split results before accepting any universality claim; (d) treat single-metric, single-split declarations with skepticism and request quantitative uncertainty estimates that incorporate split variability.

For benchmark curators and leaderboard maintainers: (a) design leaderboards that surface accuracy, speed, memory, and extrapolation performance side-by-side rather than ranking on one scalar; (b) enforce submission of results on at least three distinct splitting strategies; (c) incorporate dedicated extrapolation tasks (new elements, disordered alloys, large cells) as first-class evaluation axes.

Adoption of these practices would transform SOTA from a rhetorical device into a transparent, multi-dimensional summary of capability boundaries. The field would then be able to distinguish genuine methodological advances from benchmark gaming and allocate effort toward challenges that matter for materials discovery.

Conclusion

SOTA claims in materials GNNs frequently conflate benchmark-specific gains with genuine methodological progress. When evaluation artifacts and implicit constraints are accounted for, many reported improvements lose their broader significance. A more rigorous standard—grounded in transparent protocols, diverse evaluation regimes, and explicit acknowledgment of limitations—is essential for aligning reported performance with real-world applicability.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 06 May 2023

Revised: 22 Aug 2023

Accepted: 08 Nov 2023

Published online: 18 January 2024

Rights and permissions

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

References

- Shchur O, Mumme M, Bojchevski A, Günnemann S. Pitfalls of graph neural network evaluation. *arXiv*. 2018;1811.05868.
<https://doi.org/10.48550/arXiv.1811.05868>.
- 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>.
- Ying R, Bourgeois D, You J, Zitnik M, Leskovec J. GNNExplainer: Generating explanations for graph neural networks. In: *Adv Neural Inf Process Syst*. Red Hook (NY): Curran Associates; 2019;32.
- 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>.
- Schütt KT, Sauceda HE, Kindermans PJ, Tkatchenko A, Müller KR. SchNet: A deep learning architecture for molecules and materials. *J Chem Phys*. 2018;148(24):241722.
<https://doi.org/10.1063/1.5019779>.
- Kocer E, Ko TW, Behler J. Neural network potentials: A concise overview of methods. *Annu Rev Phys Chem*. 2022;73:163-86.
<https://doi.org/10.1146/annurev-physchem-082720-034254>.
- Batzner S, Musaelian A, Sun L, Geiger M, Mailoa JP, Kornbluth M, et al. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat Commun*. 2022;13(1):2453.
<https://doi.org/10.1038/s41467-022-29939-5>.
- Zuo Y, Chen C, Li X, Deng Z, Chen Y, Behler J, et al. Performance and cost assessment of machine learning interatomic potentials. *J Phys Chem A*. 2020;124(4):731-45.
<https://doi.org/10.1021/acs.jpca.9b08723>.
- Merchant A, Batzner S, Schoenholz SS, Aykol M, Cheon G, Cubuk ED. Scaling deep learning for materials discovery. *Nature*. 2023;624(7990):80-5.
<https://doi.org/10.1038/s41586-023-06735-9>.
- 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>.
- Zhou J, Cui G, Hu S, Zhang Z, Yang C, Liu Z, et al. Graph neural networks: A review of methods and applications. *AI Open*. 2020;1:57-81.

<https://doi.org/10.1016/j.aiopen.2021.01.001>.

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>.

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>.

Deng B, Zhong P, Jun KJ, Riebesell J, Han K, Bartel CJ, et al. CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nat Mach Intell.* 2023;5(9):1031-41.

<https://doi.org/10.1038/s42256-023-00716-3>.

Dunn A, Wang Q, Ganose A, Dopp D, Jain A. Benchmarking materials property prediction methods: The Matbench test set and Automatminer reference algorithm. *NPJ Comput Mater.* 2020;6(1):138.

<https://doi.org/10.1038/s41524-020-00406-3>.

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>.

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>.

Qin S, Jiang S, Li J, Balaprakash P, Van Lehn RC, Zavala VM. Capturing molecular interactions in graph neural networks: A case study in multi-component phase equilibrium. *Digit Discov.* 2023;2(1):138-51.

<https://doi.org/10.1039/D2DD00045H>.

Fan W, Ma Y, Li Q, He Y, Zhao E, Tang J, et al. Graph neural networks for social recommendation. In: *World Wide Web Conf.* New York: Association for Computing Machinery; 2019. p. 417-26.

<https://doi.org/10.1145/3308558.3313488>.

Qian C, Xiong Y, Chen X. Directed graph attention neural network utilizing 3D coordinates for molecular property prediction. *Comput Mater Sci.* 2021;200:110761.

<https://doi.org/10.1016/j.commatsci.2021.110761>.

Pandey S, Qu J, Stevanović V, St John PC, 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>.

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>.

Yang H. AliGraph: A comprehensive graph neural network platform. In: *Proc ACM SIGKDD Int Conf Knowl Discov Data Min.* New York: Association for Computing Machinery; 2019. p. 3165-6.

<https://doi.org/10.1145/3292500.3340404>.

Zhao T, Liu Y, Neves L, Woodford O, Jiang M, Shah N. Data augmentation for graph neural networks. *Proc AAAI Conf Artif Intell.* 2021;35(12):11015-23.

<https://doi.org/10.1609/aaai.v35i12.17315>.

Gupta A, Matta P, Pant B. Graph neural network: Current state of art, challenges and applications. *Mater Today Proc.* 2021;46:10927-32.
<https://doi.org/10.1016/j.matpr.2021.01.950>.

Cong G, Fung V. Improving materials property predictions for graph neural networks with minimal feature engineering. *Mach Learn Sci Technol.* 2023;4(3):035030.
<https://doi.org/10.1088/2632-2153/acefab>.

Wen M, Blau SM, Spotte-Smith EW, Dwaraknath S, Persson KA. BonDNet: A graph neural network for the prediction of bond dissociation energies for charged molecules. *Chem Sci.* 2021;12(5):1858-68.
<https://doi.org/10.1039/D0SC05251E>.

Rusch TK, Bronstein MM, Mishra S. A survey on oversmoothing in graph neural networks. *arXiv.* 2023;2303.10993.
<https://doi.org/10.48550/arXiv.2303.10993>.

Morgan JP, Paiement A, Klinke C. Domain-informed graph neural networks: A quantum chemistry case study. *Neural Netw.* 2023;165:938-52.
<https://doi.org/10.1016/j.neunet.2023.06.030>.