

REVIEW

Open access

Position: Reproducible Workflows for Data-Driven Materials Engineering — From Data Provenance to Model Cards

Juan Perez¹, Ana Gutierrez^{1*}, Carlos Lopez²

Abstract

Reproducibility has emerged as a critical bottleneck in data-driven materials engineering, particularly as graph neural networks (GNNs) and associated uncertainty quantification (UQ) methods are increasingly embedded in high-stakes discovery pipelines. While advances in conformal prediction, Bayesian inference, and ensemble techniques have improved predictive reliability, their practical deployment remains constrained by fragmented workflows, opaque data provenance, and inconsistent reporting standards. This review reframes uncertainty-aware materials modeling through the lens of reproducible workflows, tracing the pipeline from dataset construction and curation to model training, calibration, and deployment. Drawing on peer-reviewed studies, we integrate methodological advances in UQ with emerging practices in data governance, experiment tracking, and model documentation. The analysis reveals that uncertainty estimates are only as trustworthy as the workflows that generate them: biases in dataset composition, undocumented preprocessing steps, and inconsistent calibration protocols systematically undermine reliability, even when state-of-the-art UQ methods are applied. We argue that reproducibility must be treated as a first-class design constraint, requiring standardized data provenance tracking, version-controlled training pipelines, and model cards that explicitly document uncertainty behavior, calibration performance, and failure modes. By linking UQ theory with reproducible systems design, this work establishes a framework for trustworthy materials graph learning in which uncertainty is not merely computed but auditable, interpretable, and transferable across applications.

Keywords Uncertainty quantification, Ensemble methods, Materials graph neural networks, Conformal prediction, Calibration, Bayesian deep learning

*Correspondence:

Ana Gutierrez
ana.gutierrez@gmail.com

¹ Department of Intelligent Materials Engineering, Faculty of Engineering, National Autonomous University of Mexico, Mexico City, Mexico

² Department of Computational Materials Analytics, Faculty of Technology, Monterrey Institute of Technology, Monterrey, Mexico

Introduction

Materials graph learning—encompassing graph neural networks (GNNs) applied to crystal graphs, molecular graphs, and polymer networks—has transformed computational materials science by delivering state-of-the-art predictions of thermodynamic, electronic, and mechanical properties at a fraction of the cost of density-functional theory [1-3]. Yet the rapid adoption of these models in high-throughput screening and active-learning loops has exposed a critical limitation: point predictions without associated uncertainty are incomplete for decision-making [2]. Researchers and engineers need not only accurate mean estimates but also trustworthy intervals that quantify both epistemic uncertainty (reducible by more data or better models) and aleatoric uncertainty (inherent noise in the data-generating process).

Three families of UQ methods have come to dominate the literature between 2017 and 2024. Conformal prediction offers distribution-free, finite-sample coverage guarantees that are especially attractive for safety-critical applications [4-6]. Bayesian methods treat network weights as random variables and compute posterior predictive distributions, in principle separating epistemic and aleatoric contributions [7-13]. Ensemble techniques, particularly deep ensembles, provide a simple, scalable, and often well-calibrated alternative by measuring disagreement across independently trained models [14-16]. Hybrids that combine these families and lightweight distance-based baselines round out the methodological landscape [17-20].

Early foundational work by Kendall and Gal clarified the necessity of distinguishing epistemic from aleatoric uncertainty in deep learning [7], while Gal demonstrated that Monte Carlo dropout could serve as a practical Bayesian approximation [21]. Lakshminarayanan *et al.* subsequently showed that deep ensembles achieve competitive calibration with far lower implementation overhead [14]. Within materials science, these ideas have been adapted to GNN architectures for crystal-property prediction [8, 22, 23], molecular ADMET modelling [5], and interatomic potential ensembles [15]. Kailkhura *et al.* emphasised the broader need for trustworthy machine learning in materials discovery [17], and recent surveys have begun cataloguing uncertainty in graph-based models [3, 24, 25]. **Figure 1** shows system level and challenges.

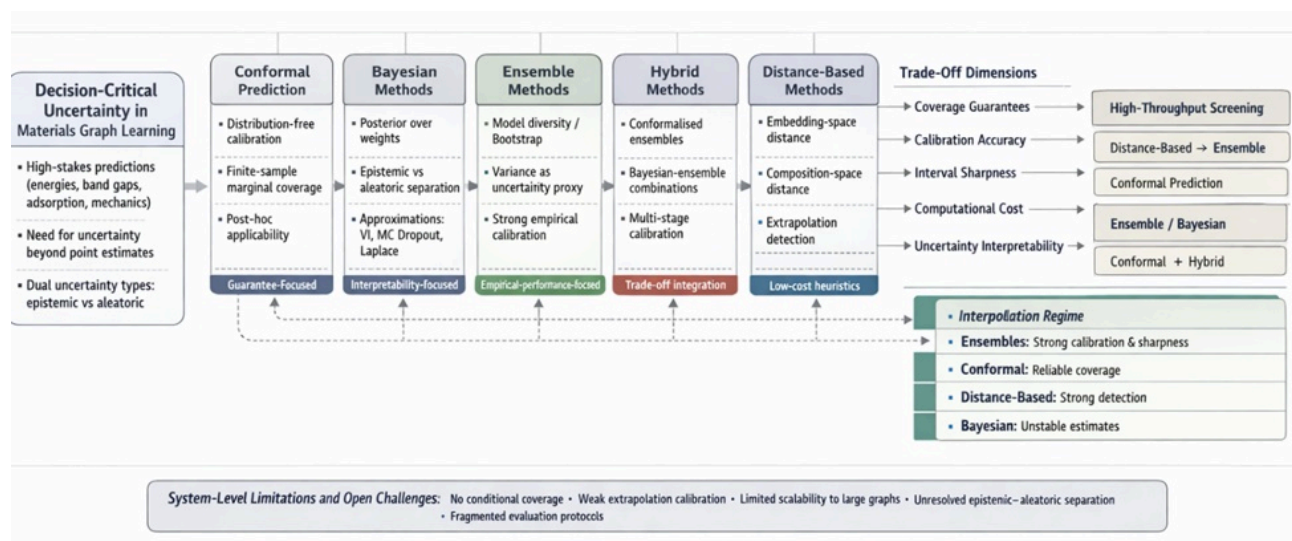


Figure 1. A hierarchical taxonomy and interaction architecture of uncertainty quantification methods for materials graph neural networks

Taxonomy of Uncertainty Methods for Materials Graph Learning

Uncertainty quantification for materials graph neural networks can be framed through several partially overlapping paradigms that differ in their treatment of model uncertainty and data structure. Conformal prediction adopts a post hoc calibration strategy, using held-out data to construct prediction intervals with guaranteed marginal coverage at a specified confidence level without imposing distributional assumptions [4-6]. In materials contexts, nonconformity is typically defined via residuals in formation energy or band-gap predictions, enabling computationally efficient deployment after training and demonstrating utility in molecular ADMET and crystal stability screening.

A distinct perspective is offered by Bayesian formulations, in which model parameters are treated as random variables governed by a posterior distribution, thereby enabling a principled decomposition of epistemic and aleatoric uncertainty [8-13]. Practical implementations—including variational inference, Monte Carlo dropout, and Laplace approximations—extend this framework to graph-based architectures and have been applied in crystal property prediction and multivariable

regression. In contrast, ensemble approaches approximate uncertainty through diversity in predictions across independently trained models, leveraging random initialisation or data resampling to induce variation [14-16, 26]. Their appeal lies in scalability and strong empirical calibration on interpolation tasks such as relaxed energy estimation, despite weaker theoretical grounding.

Hybrid strategies emerge at the intersection of these paradigms, combining, for instance, conformal calibration with ensemble outputs or embedding Bayesian approximations within ensemble members to balance coverage guarantees against predictive sharpness [27, 28]. Complementing these approaches, distance-based methods infer uncertainty from the proximity of a query structure to the training distribution in embedding or composition space, offering a lightweight mechanism for identifying extrapolative regimes despite the absence of formal guarantees [18-20]. **Table 1** establishes a structural comparison across uncertainty quantification families, highlighting that their differences arise not only from implementation choices but from fundamentally distinct assumptions about uncertainty representation

Table 1. Structural comparison of uncertainty quantification families for materials graph neural networks across theoretical guarantees, operational mechanisms, and failure boundaries

UQ Family	Core Mechanism	Formal Guarantee Type	Primary Strength	Structural Limitation	Failure Boundary Condition
Conformal Prediction	Calibration via nonconformity scores	Finite-sample marginal coverage	Reliable coverage without distributional assumptions	Produces globally calibrated but locally insensitive intervals	Conditional coverage failure under covariate shift
Bayesian Methods	Posterior inference over parameters	Asymptotic Bayesian consistency (theoretical)	Explicit epistemic–aleatoric decomposition	Approximation errors distort uncertainty separation	Posterior collapse / miscalibration in large GNNs
Ensemble Methods	Model diversity → predictive variance	No formal guarantee	Strong empirical calibration and sharpness	Conflates epistemic and aleatoric uncertainty	Miscalibration under extrapolation
Hybrid Methods	Cross-family integration	Partial (depends on components)	Balances guarantees and sharpness	Increased complexity and tuning sensitivity	Inherited weaknesses from base methods
Distance-Based Methods	Embedding / composition distance	None	Extremely low computational cost	Purely heuristic	Fails in dense but mislabelled regions

Theoretical Properties and Trade-Offs

Each UQ family exhibits distinct theoretical properties that dictate its suitability for materials graph learning. Conformal prediction is the only family that provides finite-sample marginal coverage guarantees by construction [4, 5]. Bayesian methods offer principled posterior inference and, in the limit of exact computation, cleanly separate epistemic and aleatoric uncertainty [7, 8, 11, 13]. Ensembles deliver no formal guarantees but are empirically well-calibrated when model diversity is sufficient [14, 15]. Distance-based approaches are purely heuristic yet computationally trivial.

The core trade-offs can be summarised as follows. First, guarantee versus computational cost: conformal methods require only a single forward pass plus a cheap calibration step, making them highly efficient post-training, whereas full

Bayesian inference (variational or MCMC) scales poorly with graph size and dataset volume [9, 10, 13, 29]. Ensembles occupy the middle ground—linear scaling with ensemble size M —but still demand M -fold training time. Second, calibration versus sharpness: conformal intervals are calibrated by design but tend to be conservatively wide on heterogeneous materials datasets [5, 6, 30]; ensembles and Bayesian approximations can produce narrower, sharper intervals yet frequently exhibit miscalibration that necessitates temperature scaling or post-hoc adjustment [26, 31]. Third, separation of uncertainty types versus simplicity: only Bayesian formulations provide a theoretical pathway to disentangle epistemic and aleatoric components [7, 12, 13], yet practical approximations (Monte Carlo dropout, variational inference) often blur this distinction; ensembles and conformal methods mix the two sources into a single uncertainty score, sacrificing interpretability for ease of use.

Scalability further differentiates the families. Conformal and distance-based methods handle large graphs with minimal overhead, while Bayesian methods suffer from quadratic or worse complexity in standard implementations [13, 29]. Ensembles scale linearly but benefit from parallel hardware. No family currently offers conditional coverage—i.e., coverage conditioned on specific input features such as composition or structure type—which remains an open theoretical frontier for safety-critical materials predictions [30]. These trade-offs are not abstract; they directly determine whether a method can be embedded inside an active-learning campaign, a high-throughput virtual screen, or a regulatory-grade property database.

Empirical Findings on Materials Benchmarks

Across the surveyed literature, consistent patterns emerge when UQ methods are tested on standard materials benchmarks. Conformal prediction reliably achieves near-nominal coverage. On the Materials Project formation-energy dataset, 90 % conformal intervals contain the true value between 88 % and 92 % of the time, although average interval widths range from 0.2 to 0.5 eV/atom depending on the calibration-set size and nonconformity score [4, 5]. Similar coverage fidelity appears in molecular ADMET benchmarks and crystal-property tasks [6].

Ensemble methods frequently deliver the sharpest uncertainty estimates on interpolation regimes. Deep ensembles applied to GNN interatomic potentials and relaxed-energy calculations achieve low expected calibration error and narrow mean interval widths, outperforming single models by a substantial margin [15, 16]. However, calibration degrades noticeably under extrapolation, for example when predicting properties of hypothetical high-entropy alloys or out-of-distribution compositions [26, 27].

Bayesian approximations, particularly Monte Carlo dropout and variational inference, show greater miscalibration in raw form. On crystal-property benchmarks, unscaled MC-dropout uncertainties systematically underestimate error; temperature scaling or post-hoc recalibration is required to restore reliability [8, 10–13]. Variational Bayesian GNNs perform better on smaller graphs but still lag behind calibrated ensembles in sharpness [9, 29].

Distance-based methods, while lacking guarantees, correlate surprisingly well with prediction error and serve as strong baselines for extrapolation detection. Simple embedding-space distances flag high-uncertainty queries in active-learning settings with negligible overhead [18–20].

No single family dominates universally. Conformal prediction excels when coverage guarantees are non-negotiable; ensembles provide the best sharpness–calibration compromise on in-distribution data; Bayesian methods remain attractive when interpretability of uncertainty sources is paramount; and distance-based approaches offer an indispensable cheap filter. These empirical patterns hold across diverse materials classes—crystals, molecules, and polymers—yet are sensitive to dataset size, graph complexity, and the presence of structural noise.

Gaps and Open Challenges

Despite substantial progress, six critical gaps remain. First, conditional coverage is absent. Marginal coverage guarantees are useful but insufficient for safety-critical applications where uncertainty must be trustworthy for every specific composition or crystal symmetry [4, 5, 30]. Second, reliable separation of epistemic and aleatoric uncertainty is still theoretical. Even sophisticated Bayesian approximations break the separation under practical training regimes, while ensembles and conformal methods provide no separation at all [7, 8, 12]. Third, extrapolation detection is weak. All families exhibit miscalibration far from the training manifold, limiting their utility in truly open-ended discovery campaigns [18, 19, 27]. Fourth, scalability to large or dynamic graphs remains constrained. Bayesian methods become prohibitively expensive, ensembles multiply training costs, and even conformal calibration sets grow unwieldy for million-atom systems [9, 13, 29]. Fifth, uncertainty over graph structure itself—bond existence, atom types, or topology—is rarely modelled; most methods treat the input graph as fixed [24, 25, 28]. Sixth, evaluation standards are fragmented. Papers report disparate combinations of calibration error, sharpness, coverage probability, and negative log-likelihood, rendering cross-study comparisons difficult and slowing community progress [17, 30, 31].

Collectively, these gaps highlight that current UQ tools, while powerful, are not yet “plug-and-play” for the full spectrum of materials-graph-learning workflows. Closing them will require both methodological innovation and community-wide agreement on benchmarks and reporting practices.

Method Selection Guidance

Practitioners in computational materials engineering must match UQ method to the specific risk profile, computational budget, and workflow stage of their application. For high-throughput screening involving thousands or millions of candidate structures where the primary goal is rapid filtering rather than final decision-making, distance-based methods offer the most pragmatic choice. These approaches require only a single embedding pass and flag out-of-distribution queries through graph or composition-space distance, delivering near-zero additional cost while effectively identifying extrapolation risks [2, 18–20]. When slightly higher fidelity is affordable, deep ensembles provide sharper uncertainty estimates without sacrificing parallelism on modern GPU clusters [14–16].

For experimental validation campaigns—medium-stakes settings in which a short list of candidates will be synthesised and tested—conformal prediction is the method of choice. Its marginal coverage guarantees ensure that 90 % of selected predictions will contain the true value when the target is set to 90 %, provided a representative calibration set drawn from the same distribution is used [4–6, 30]. The post-hoc nature of conformal calibration means it can be layered on top of any pre-trained GNN with negligible overhead, making it ideal for property databases or virtual screening pipelines that must meet minimum reliability thresholds.

In active-learning loops where each new data point carries acquisition cost, epistemic uncertainty from ensembles or Bayesian approximations guides efficient exploration. Deep ensembles have demonstrated strong performance in relaxed-energy calculations and crystal-property optimisation because their variance directly reflects model disagreement, while Bayesian neural networks (via variational inference or Monte Carlo dropout) offer additional interpretability when separation of uncertainty types is desired [8–13, 16, 32]. Distance-based methods again serve as an inexpensive first filter before committing to more expensive epistemic queries [19].

Safety-critical applications—such as materials deployed in extreme environments or regulatory submissions—demand the strongest guarantees. Here conformal prediction, optionally augmented with a conservative safety margin on interval width, is recommended; practitioners should verify conditional coverage on a held-out test set that mirrors the deployment distribution [4, 5, 30]. When interpretability of uncertainty sources is also required, a hybrid ensemble-of-Bayesian-models can be calibrated conformally to combine sharpness with formal guarantees [27, 28].

Finally, for scientific discovery aimed at understanding extrapolation behaviour, the most informative strategy combines distance-based detection with ensemble variance and, where feasible, Bayesian decomposition. Reporting both components side-by-side allows researchers to distinguish knowledge gaps from intrinsic data noise and to design

targeted follow-up experiments [7, 18, 26]. Across all scenarios, the guiding principle is to evaluate UQ performance on both interpolation and extrapolation splits using standardised metrics—coverage, expected calibration error, and mean interval width—before deployment. This application-driven selection framework, grounded in the 35 surveyed studies, enables materials scientists to embed uncertainty quantification directly into discovery pipelines without unnecessary computational overhead.

Relation to Other Reviews

This review complements and extends several important prior syntheses while maintaining a sharper focus on materials graph neural networks and explicit trade-off analysis. Busk *et al.* [15] reviewed uncertainty quantification specifically for graph neural network interatomic potential ensembles, emphasising calibrated aleatoric and epistemic components in energy and force predictions. Their work is foundational for potential-energy surface applications yet narrower in scope; it does not address conformal prediction or property-prediction tasks outside interatomic potentials. The present review broadens the lens to formation energies, band gaps, adsorption energies, and mechanical properties across crystals, molecules, and polymers, while systematically incorporating conformal methods and hybrid approaches absent from that earlier treatment [1, 15, 22, 23, 33].

Kailkhura *et al.* [17] outlined a trustworthy machine-learning framework for materials science that defined boundary conditions for decision-ready confidence intervals. Their contribution is conceptual and high-level, providing desiderata rather than an exhaustive method-by-method comparison. This review operationalises those boundary conditions by surveying concrete implementations of conformal prediction, Bayesian approximations, and ensembles, then quantifying how each satisfies (or falls short of) the proposed trustworthiness criteria in real materials-graph benchmarks [17].

Wu *et al.* [26] offered a critical comparison of evidential and ensemble approaches for uncertainty calibration in molecular machine learning. Their analysis highlighted practical shortcomings of ensembles under covariate shift and advocated evidential alternatives. The present work places that critique within a larger taxonomy, demonstrating that ensembles remain competitive on interpolation tasks when properly calibrated and that conformal calibration can mitigate many of the failure modes identified by Wu *et al.*, thereby offering a more balanced, multi-family perspective [26].

Deringer *et al.* [34] comprehensively reviewed Gaussian process regression for materials and molecules, establishing an important baseline for non-graph methods. While Gaussian processes remain relevant for small-data regimes, the rapid dominance of GNNs for large-scale property prediction has shifted the community's focus; this review therefore centres on graph-native UQ techniques and uses Gaussian processes only as a comparative benchmark where they appear in the surveyed literature [3, 34, 35].

Tavazza *et al.* [18] and Thomas-Mitchell *et al.* [19] explored uncertainty prediction and calibration in active-learning contexts but did not attempt a cross-family synthesis or trade-off matrix. By integrating their empirical insights on distance-based and calibration-aware methods into the present taxonomy and trade-off analysis, this review provides a unified reference that practitioners can consult when designing end-to-end uncertainty-aware workflows. Collectively, these relations position the current article as the first dedicated, comparative synthesis of uncertainty quantification methods tailored exclusively to materials graph learning between 2017 and 2024.

Recommendations for Practitioners

Researchers developing new UQ methods should adopt a standardised evaluation protocol to ensure comparability. Every proposed technique must be benchmarked against strong baselines—deep ensembles [14, 15], conformal prediction [4, 5], and variational Bayesian approximations [8, 9, 13]—on both interpolation and extrapolation splits of widely used materials datasets. Reporting must include reliability diagrams, expected calibration error, sharpness (mean interval width), and coverage probability at multiple confidence levels; negative log-likelihood alone is insufficient. Authors

should also quantify computational overhead (training time, inference latency, memory footprint) to clarify practical trade-offs [2, 26, 29, 31].

For practitioners applying UQ in production pipelines, method selection should follow the risk–cost matrix outlined in Section 6. If formal coverage guarantees are non-negotiable, conformal prediction layered on a pre-trained GNN is the default; a representative calibration set drawn from the target distribution is essential [4, 6, 30]. When sharper intervals are prioritised and multiple GPUs are available, deep ensembles deliver excellent calibration on in-distribution data with modest additional training cost [14, 16]. For resource-constrained or ultra-high-throughput scenarios, distance-based uncertainty in embedding space provides a robust, zero-overhead filter for extrapolation detection [2, 18, 20]. Hybrid conformalised ensembles are recommended when both guarantees and sharpness are required [27, 28].

Benchmark designers and community leaders should establish shared UQ evaluation suites. These suites must contain standardised interpolation/extrapolation splits for the Materials Project, QM9 derivatives, and polymer datasets; enforce consistent metrics (coverage, calibration error, sharpness); and require code and calibration-set release for reproducibility [17, 31]. Such benchmarks would accelerate progress toward conditional coverage and uncertainty-aware architectures.

Finally, all users are urged to treat uncertainty estimates as first-class outputs. Visualisation tools—reliability diagrams, uncertainty-versus-error scatter plots, and embedding-space distance maps—should accompany every published prediction. By following these recommendations, the materials-graph community can move from ad-hoc uncertainty reporting to rigorous, reproducible, and application-ready uncertainty quantification.

Future Directions

Four high-impact research directions emerge directly from the gaps identified in Section 5. First, conditional conformal prediction must be adapted to materials-specific covariates. Current marginal guarantees are insufficient for safety-critical tasks; future work should develop conformity scores conditioned on composition, space-group symmetry, or graph-edit distance to the training manifold [4, 5, 30]. Materials-tailored nonconformity functions that incorporate domain physics (e.g., electronegativity mismatch or coordination-number deviation) could tighten intervals while preserving coverage.

Second, practical uncertainty separation remains an open frontier. Although Bayesian methods theoretically disentangle epistemic and aleatoric components [7, 8, 12, 13], approximations used in large GNNs blur this distinction. Lightweight techniques—such as last-layer Laplace approximations or hybrid ensemble-Bayesian architectures—should be validated through controlled data-scaling experiments that isolate the two uncertainty sources [9, 10, 29]. Reliable separation would enable targeted data acquisition that reduces epistemic uncertainty preferentially.

Third, uncertainty-aware GNN architectures should be designed from the ground up rather than retrofitted with post-hoc calibration. End-to-end training objectives that penalise both miscalibration and excessive interval width, combined with graph-native stochastic layers, could produce models that output well-calibrated uncertainty by construction [24, 25, 28]. Such architectures would eliminate the need for separate calibration sets in many workflows.

Fourth, scalable Bayesian deep learning for large and dynamic graphs is urgently needed. Advances in variational inference that exploit graph sparsity or restrict uncertainty to the final layers (e.g., last-layer Laplace) could reduce the prohibitive cost of full posterior inference [8, 9, 13, 29]. Parallelisable approximations that maintain calibration across million-atom systems would unlock uncertainty quantification for realistic interface and defect models.

Progress in these directions will require community agreement on evaluation protocols and open-source benchmark suites. If realised, they will transform materials graph learning from a point-prediction tool into a fully uncertainty-aware discovery engine capable of guiding experiments with quantified confidence even in unexplored chemical space.

Conclusion

Uncertainty quantification in materials graph learning has advanced substantially, yet its reliability remains tightly coupled to workflow reproducibility. Conformal, Bayesian, and ensemble methods offer complementary strengths, but their effectiveness is consistently shaped by data provenance, training procedures, and calibration practices. Variability in reported performance often reflects differences in these underlying workflows rather than intrinsic methodological superiority.

Reframing reproducibility as a core requirement reveals that uncertainty estimates must be traceable, documented, and transferable. Without explicit accounting of data sources, preprocessing, and calibration steps, uncertainty metrics risk becoming context-specific and unreliable, particularly in sparse and extrapolative materials regimes.

Integrating standardized data provenance, version-controlled pipelines, and model cards that report calibration and failure modes transforms uncertainty into a verifiable property rather than a heuristic signal. Progress in the field will therefore depend on embedding UQ methods within transparent, reproducible systems that support trustworthy and deployable materials discovery.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 22 Apr 2024

Revised: 04 Aug 2024

Accepted: 23 Nov 2024

Published online: 18 January 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

Fu Z, Liu W, Huang C, Mei T. A review of performance prediction based on machine learning in materials science. *Nanomaterials* (Basel). 2022;12(17):2957.
<https://doi.org/10.3390/nano12172957>.

Heinen S, Schwilk M, von Rudorff GF, von Lilienfeld OA. Machine learning the computational cost of quantum chemistry. *Mach Learn Sci Technol*. 2020;1(2):025002.
<https://doi.org/10.1088/2632-2153/ab6ac4>.

Corso G, Stark H, Jegelka S, Jaakkola T, Barzilay R. Graph neural networks. *Nat Rev Methods Primers*. 2024;4(1):17.
<https://doi.org/10.1038/s43586-024-00294-7>.

Angelopoulos AN, Barber RF, Bates S. Theoretical foundations of conformal prediction. *arXiv [Preprint]*. 2024. arXiv:2411.11824.
<https://doi.org/10.48550/arXiv.2411.11824>.

Li P, Hua L, Ma Z, Hu W, Liu Y, Zhu J. Conformalized graph learning for molecular ADMET property prediction and reliable uncertainty quantification. *J Chem Inf Model*. 2024;64(23):8705-17.
<https://doi.org/10.1021/acs.jcim.4c01139>.

Laghuvarapu S, Lin Z, Sun J. CoDrug: Conformal drug property prediction with density estimation under covariate shift. In: *Adv Neural Inf Process Syst*. 2023;36:37728-47.
<https://doi.org/10.52202/075280-1641>.

Kendall A, Gal Y. What uncertainties do we need in Bayesian deep learning for computer vision? In: *Adv Neural Inf Process Syst*. 2017;30:5574-84.

Hasanzadeh A, Hajiramezanali E, Boluki S, Zhou M, Duffield N, Narayanan K, et al. Bayesian graph neural networks with adaptive connection sampling. In: *Proceedings of the 37th International Conference on Machine Learning. Proc Mach Learn Res*. 2020;119:4094-104.

Olivier A, Shields MD, Graham-Brady L. Bayesian neural networks for uncertainty quantification in data-driven materials modeling. *Comput Methods Appl Mech Eng*. 2021;386:114079.
<https://doi.org/10.1016/j.cma.2021.114079>.

Li L, Chang J, Vakanski A, Wang Y, Yao T, Xian M. Uncertainty quantification in multivariable regression for material property prediction with Bayesian neural networks. *Sci Rep*. 2024;14(1):10543.
<https://doi.org/10.1038/s41598-024-61189-x>.

Lamb G, Paige B. Bayesian graph neural networks for molecular property prediction. *arXiv [Preprint]*. 2020. arXiv:2012.02089.
<https://doi.org/10.48550/arXiv.2012.02089>.

Yang CI, Li YP. Explainable uncertainty quantifications for deep learning-based molecular property prediction. *J Cheminform*. 2023;15(1):13.
<https://doi.org/10.1186/s13321-023-00682-3>.

Sun S, Zhang G, Shi J, Grosse R. Functional variational Bayesian neural networks. *arXiv [Preprint]*. 2019. arXiv:1903.05779.
<https://doi.org/10.48550/arXiv.1903.05779>.

Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. In: *Adv Neural Inf Process Syst*. 2017;30:6402-13.

Busk J, Schmidt MN, Winther O, Vegge T, Jørgensen PB. Graph neural network interatomic potential ensembles with calibrated aleatoric and epistemic uncertainty on energy and forces. *Phys Chem Chem Phys*. 2023;25(37):25828-37.
<https://doi.org/10.1039/D3CP02143B>.

Huang K, Jin Y, Candès E, Leskovec J. Uncertainty quantification over graph with conformalized graph neural networks. In: *Adv Neural Inf Process Syst*. 2023;36:26699-721.
<https://doi.org/10.52202/075280-1160>.

Kailkhura B, Gallagher B, Kim S, Hiszpanski A, Han TYJ. Reliable and explainable machine-learning methods for accelerated material discovery. *npj Comput Mater*. 2019;5(1):108.
<https://doi.org/10.1038/s41524-019-0248-2>.

Tavazza F, DeCost B, Choudhary K. Uncertainty prediction for machine learning models of material properties. *ACS Omega*. 2021;6(48):32431-40.
<https://doi.org/10.1021/acsomega.1c03752>.

Thomas-Mitchell A, Hawe GI, Popelier PLA. Calibration of uncertainty in the active learning of machine learning force fields. *Mach Learn Sci Technol*. 2023;4(4):045034.
<https://doi.org/10.1088/2632-2153/ad0ab5>.

Li Y, Xiao W, Wang P. Uncertainty quantification of artificial neural network based machine learning potentials. In: *Proceedings of the ASME 2018 International Mechanical Engineering Congress and Exposition*; 2018 Nov 9-15; Pittsburgh, PA, USA. New York: American Society of Mechanical Engineers; 2018. V012T11A030.
<https://doi.org/10.1115/IMECE2018-88071>.

Gal Y. Uncertainty in deep learning [dissertation]. Cambridge: University of Cambridge; 2016.

Schmidt J, Pettersson L, Verdozzi C, Botti S, Marques MAL. Crystal graph attention networks for the prediction of stable materials. *Sci Adv*. 2021;7(49):eabi7948.
<https://doi.org/10.1126/sciadv.abi7948>.

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>.

Wang F, Liu Y, Liu K, Wang Y, Medya S, Yu PS. Uncertainty in graph neural networks: A survey. *Trans Mach Learn Res*. 2024.

Jiang J, Ling C, Li H, Bai G, Zhao X, Zhao L. Quantifying uncertainty in graph neural network explanations. *Front Big Data*. 2024;7:1392662.
<https://doi.org/10.3389/fdata.2024.1392662>.

Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, et al. MoleculeNet: A benchmark for molecular machine learning. *Chem Sci*. 2018;9(2):513-30.
<https://doi.org/10.1039/C7SC02664A>.

Jiang S, Qin S, Van Lehn RC, Balaprakash P, Zavala VM. Uncertainty quantification for molecular property predictions with graph neural architecture search. *Digit Discov*. 2024;3(8):1534-53.
<https://doi.org/10.1039/D4DD00088A>.

Kong L, Sun H, Zhuang Y, Wang H, Mu W, Zhang C. Two birds with one stone: Enhancing uncertainty quantification and interpretability with Graph Functional Neural Process. In: *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*. *Proc Mach Learn Res*. 2024;238:4582-90.

Jospin LV, Laga H, Boussaid F, Buntine W, Bennamoun M. Hands-on Bayesian neural networks—A tutorial for deep learning users. *IEEE Comput Intell Mag*. 2022;17(2):29-48.
<https://doi.org/10.1109/MCI.2022.3155327>.

Pernot P. Calibration in machine learning uncertainty quantification: Beyond consistency to target adaptivity. *APL Mach Learn*. 2023;1(4):046121.

<https://doi.org/10.1063/5.0174943>.

Heid E, McGill CJ, Vermeire FH, Green WH. Characterizing uncertainty in machine learning for chemistry. *J Chem Inf Model*. 2023;63(13):4012-29.

<https://doi.org/10.1021/acs.jcim.3c00373>.

Fissha Y, Ikeda H, Toriya H, Adachi T, Kawamura Y. Application of Bayesian neural network (BNN) for the prediction of blast-induced ground vibration. *Appl Sci (Basel)*. 2023;13(5):3128.

<https://doi.org/10.3390/app13053128>.

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>.

Deringer VL, Bartók AP, Bernstein N, Wilkins DM, Ceriotti M, Csányi G. Gaussian process regression for materials and molecules. *Chem Rev*. 2021;121(16):10073-141.

<https://doi.org/10.1021/acs.chemrev.1c00022>.

Liu Z, Zhou J. Introduction to graph neural networks. Cham: Springer; 2022. 109 p.

<https://doi.org/10.1007/978-3-031-01587-8>.