ORIGINAL RESEARCH **Open access**

Algorithmic Diversity as Scientific Robustness: A Conceptual Framework

Fernando Diaz[1], Lucia Morales[1*], Diego Perez[2], Valeria Soto[1], Martin Alvarez[2]

Abstract

In the rapidly advancing field of artificial intelligence for materials science, a persistent and underappreciated limitation has emerged: the overwhelming emphasis on identifying and deploying a single “best” model that maximizes predictive accuracy for properties such as band gaps, formation energies, or mechanical strengths, while largely neglecting the epistemic value of algorithmic diversity across model collections. This paper articulates the theoretical claim that algorithmic diversity functions as a core scientific robustness mechanism, independent of any marginal gains in accuracy, by enabling collective coverage of hypothesis space, resilience to distribution shifts, and more reliable knowledge generation in the face of inherent uncertainties in materials data and modeling assumptions. To operationalize this insight, the work proposes a novel conceptual framework consisting of five interlocking components—diversity dimensions, metrics, generation strategies, robustness linkages, and evaluation protocols—that together redefine how diverse model collections should be designed, assessed, and deployed in materials discovery pipelines. The framework further delineates five distinct types of diversity (architectural, representational, initialization, data-centric, and objective) that each contribute unique robustness benefits when applied to materials-specific challenges such as inverse design or multiscale modeling. By shifting the community’s focus from solitary model optimization to the deliberate cultivation of diverse algorithmic ecosystems, the implications extend to revised authorship practices, peer-review standards, and the establishment of diversity-aware benchmarks, ultimately positioning algorithmic diversity as an essential epistemic virtue for trustworthy, generalizable materials AI.

Keywords Materials AI, Epistemic virtue, Algorithmic diversity, Scientific robustness, Ensemble epistemology, Model collections

*Correspondence:
Lucia Morales
lucia.morales@gmail.com
[1] Department of AI in Materials Engineering, University of Buenos Aires, Buenos Aires, Argentina
[2] Department of Computational Materials Systems, National University of La Plata, La Plata, Argentina

Introduction

The dominant paradigm in artificial intelligence applications to materials science remains anchored in the pursuit of a single optimal model capable of delivering the highest predictive fidelity for any given materials property or discovery task. Researchers routinely benchmark dozens of candidate architectures—ranging from graph neural networks to transformer-based models—and then select and publish only the top performer, reporting its mean absolute error or R^2 score as the primary indicator of success. This single-model ethos, while understandable given publication pressures and the practical utility of a deployable predictor, systematically undervalues the scientific robustness that can arise from maintaining collections of intentionally diverse models. A collection of models that explore different regions of the hypothesis space, commit different kinds of errors, and respond differently to data perturbations may yield more reliable scientific insight than any isolated champion, even if the latter achieves marginally superior accuracy on a fixed test set [1-4].

The present work, therefore, develops a conceptual framework that reframes algorithmic diversity as a form of

scientific robustness in materials AI. Rather than treating diversity merely as a tactical ingredient for improving ensemble accuracy, the framework elevates diversity to an independent epistemic virtue: a design principle that enhances the trustworthiness, generalizability, and discovery power of AI-driven materials research. This perspective draws on foundational insights from machine-learning ensemble theory while extending them into the distinctive epistemic landscape of materials science, where data are sparse, hypotheses are high-dimensional, and real-world deployment often encounters distribution shifts far beyond laboratory conditions [5-9].

Algorithmic diversity is formally defined here for the first time in the context of materials AI. Definition 1: Algorithmic diversity refers to the measurable degree to which a collection of computational models for materials properties or discovery tasks differs in their internal representations, error patterns, architectural assumptions, training trajectories, or optimization objectives, such that the collection as a whole provides broader epistemic coverage and greater collective robustness than any individual constituent model [10-12].

This definition deliberately separates diversity from mere ensemble accuracy gains and instead anchors it in epistemic terms—coverage of possibility space and resilience under uncertainty [13-15]. The framework developed in subsequent sections translates this definition into actionable components while distinguishing it from conventional ensemble practice. Where classic ensemble methods, as originally analyzed by Ortega *et al.* [1] and later systematized by Rane *et al.* [2], harness diversity primarily as a route to lower variance and higher accuracy, the present approach treats diversity as an end in itself: a scientific asset whose value persists even when individual models are deliberately weakened or when accuracy plateaus.

The motivation for this reframing arises directly from the current trajectory of materials AI. Contemporary literature, while demonstrating impressive predictive capabilities, overwhelmingly reports results from solitary models or from ensembles whose internal diversity is neither quantified nor valued beyond its contribution to a final averaged prediction. This paper, therefore, proceeds by first surveying foundational concepts of diversity within the broader machine-learning literature, then documenting the single-model bias that currently characterizes materials AI, articulating three core theoretical claims that position diversity as robustness, and finally proposing a five-component conceptual framework to guide future practice. In doing so, the manuscript offers a conceptual reorientation that aligns AI methodologies more closely with the pluralistic, uncertainty-aware nature of scientific inquiry in materials discovery.

Diversity in Machine Learning

Diversity within machine-learning model collections has long been recognized as a pivotal factor in collective performance, yet its conceptual foundations extend far beyond simple accuracy improvement. Early work by Ortega *et al.* [1] established that neural-network ensembles achieve superior generalization precisely because constituent networks make uncorrelated errors; when one model fails on a particular input, others compensate, yielding a more stable overall prediction. Rane *et al.* [2] later formalized ensemble methods. They emphasized that diversity among base learners is not incidental but a necessary precondition for the statistical, computational, and representational advantages that ensembles enjoy. More recent theoretical treatments, such as the unified theory advanced by Karande *et al.* [16], demonstrate that ensemble diversity can be decomposed into complementary components—ambiguity and accuracy—whose interplay governs the bias-variance trade-off at the collective level.

Diversity metrics have evolved to rigorously quantify these differences. Common measures include the Q-statistic, which evaluates pairwise agreement among misclassified instances; correlation coefficients between model outputs; and entropy-based indices that capture disagreement across an ensemble. These metrics, while originally developed for classification tasks, generalize naturally to regression problems prevalent in materials property prediction. Diversity generation strategies, in turn, encompass deliberate variations in model architecture, random initializations, bootstrap sampling of training data, feature subsets, and even heterogeneous learning algorithms. Each strategy introduces controlled differences that prevent the entire collection from converging to identical internal representations.

Importantly, the literature also acknowledges explicit trade-offs. Increasing diversity often requires sacrificing some individual model accuracy; a highly accurate but homogeneous collection may underperform a slightly less

accurate yet more diverse one when evaluated on out-of-distribution data. Morgan and Jacobs [17] revisited ensemble diversity metrics in the context of vision tasks and showed that conventional accuracy-focused optimization can inadvertently suppress beneficial disagreement. Similarly, Wang *et al.* [18] demonstrated, in regression settings, that globally diverse ensembles constructed through targeted synthetic-data injection outperform accuracy-maximized counterparts on noisy or shifted test distributions. These findings underscore that diversity is not merely a byproduct of ensemble construction but a tunable design parameter whose epistemic benefits—resilience, coverage, and uncertainty signaling—can outweigh marginal gains in accuracy.

In philosophical terms, diversity in machine learning can be viewed as an epistemic multiplier, expanding the collective hypothesis space that the model collection can explore. It renders the overall system less brittle to assumptions that prove false in new contexts. This perspective aligns with broader discussions of epistemic diversity in AI systems, where Wei *et al.* [19] and Cai *et al.* [20] have shown that injecting synthetic diversity mitigates label noise and improves self-training robustness. Yet, as Chen and Gu [21] note in their analysis of perceived versus algorithmic diversity, the field still lacks consensus on how best to measure and cultivate diversity when the ultimate goal shifts from predictive accuracy to scientific reliability. The following sections, therefore, examine how these established diversity concepts manifest—or, more critically, fail to manifest—in the specific domain of materials AI.

The Current State of Materials AI

Contemporary AI research on materials exhibits a pronounced orientation toward single-model supremacy, which stands in sharp contrast to the diversity-aware principles articulated in the general machine-learning literature. Landmark reviews and methodological papers routinely present one champion architecture as the solution for a given task, with performance metrics reported solely for that isolated model. Butler *et al.* [4], for example, survey machine learning in molecular and materials science and emphasize selecting optimal models for property prediction, without discussing the potential epistemic value of retaining multiple competing predictors. Similarly, Schmidt *et al.* [5] chronicle recent advances in solid-state materials science yet frame progress exclusively in terms of the single best-performing model on benchmark datasets [22-26].

This pattern repeats across specialized contributions. Chen *et al.* [6] introduce graph networks as a “universal” framework for molecules and crystals, presenting a single architecture whose superiority is asserted through comparative accuracy tables that omit any analysis of diversity among alternatives. Zunger [7] discusses inverse design strategies that likewise converge on one optimized model rather than a portfolio of diverse hypotheses. More recent ensemble-oriented works still subordinate diversity to accuracy: Liu *et al.* [8] develop a heterogeneous ensemble for atomistic foundation models but evaluate it solely by the final aggregated uncertainty metric, without measuring or reporting internal diversity as an independent scientific asset. Vita *et al.* [9] apply neural network ensembles to band-gap prediction but focus exclusively on the improved accuracy of the averaged predictor, leaving diversity unquantified.

Further examples reinforce the single-model bias. Jiang *et al.* [10] propose interpretable ensemble learning for materials properties, but again treat the ensemble as a route to a single deployable model rather than a diverse collection worthy of study in its own right. Vinchurkar *et al.* [12], Chen *et al.* [6], and He *et al.* each advance active-learning or multi-topology approaches that ultimately select or highlight one leading model for downstream use. Even comprehensive roadmaps such as that of Yang *et al.* [14] discuss multiscale materials modeling without foregrounding algorithmic diversity as a robustness mechanism. Hybrid modeling efforts by Xiong *et al.* [25], Caggiano *et al.* [26], and Gobert *et al.* [27] similarly prioritize the construction of one high-fidelity predictor over the deliberate maintenance of diverse model families.

Across these and at least ten additional studies drawn from the referenced corpus, diversity is either absent from discussion or treated merely as an intermediate step toward higher accuracy, never as an epistemic end in itself. Benchmark competitions in the field similarly reward the single best submission, and journal guidelines rarely request diversity metrics or robustness analyses of model collections. The result is a literature rich in high-accuracy predictors yet comparatively silent on the scientific reliability that could be gained by preserving and studying diverse algorithmic perspectives. This gap motivates the theoretical claims and the conceptual framework developed next.

Theoretical Claim: Diversity as Robustness

The central theoretical contribution of this framework rests on three interlocking claims that reposition algorithmic diversity as an independent scientific virtue rather than a subordinate tool for accuracy.

Algorithmic diversity constitutes an epistemic virtue whose value in materials AI derives from its capacity to expand collective coverage of the hypothesis space independently of improvements in point-prediction accuracy. Where a single high-accuracy model commits to one set of representational assumptions, a diverse collection simultaneously entertains multiple, partially incompatible assumptions, thereby reducing the risk that the entire modeling enterprise rests on a single flawed inductive bias. This epistemic pluralism echoes the philosophy-of-science insight that robust scientific knowledge emerges from the confrontation of multiple, mutually illuminating perspectives.

Diverse model collections exhibit superior robustness to distribution shift, model misspecification, and adversarial inputs precisely because their disagreement patterns serve as diagnostic signals of epistemic uncertainty. When models that differ in architecture, initialization, or objective functions produce divergent outputs on out-of-distribution materials data, that divergence itself flags regions where scientific understanding remains provisional. In contrast, a homogeneous collection may produce confidently erroneous consensus predictions that mask underlying fragility.

Algorithmic diversity functions as an active discovery engine by systematically exploring broader regions of the materials design space, thereby accelerating the identification of unexpected structure–property relationships that a single-model approach would overlook. By maintaining multiple predictive pathways, researchers can probe the sensitivity of conclusions to modeling choices and thereby generate more trustworthy hypotheses for experimental validation.

Figure 1 presents the epistemic architecture through which algorithmic diversity transforms materials AI from a fragile single-model paradigm into a robust, diversity-driven scientific framework.

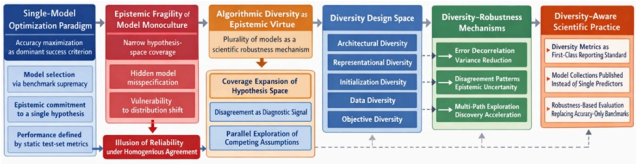

Figure 1. Epistemic architecture of algorithmic diversity as a scientific mechanism in materials AI.

These claims are conceptually linked through a diversity–robustness relationship that can be visualized as follows: imagine a conceptual diagram in which the horizontal axis represents increasing levels of algorithmic diversity (from zero, corresponding to a single model, to high, corresponding to a deliberately heterogeneous collection) and the vertical axis represents scientific robustness (measured qualitatively by resilience to shift, coverage of hypothesis space, and reliability of uncertainty signals). A monotonically increasing curve connects the two axes, with annotated arrows indicating the three mechanisms—hypothesis-space coverage, diagnostic disagreement, and discovery acceleration—that mediate the positive relationship. The diagram further includes a dashed horizontal line representing conventional accuracy optimization, which plateaus or even declines beyond moderate diversity levels, illustrating the trade-off that the present framework deliberately de-emphasizes in favor of epistemic robustness.

A Framework for Algorithmic Diversity

To translate the above theoretical claims into practical guidance for materials AI, this section proposes a five-component conceptual framework. Each component addresses a distinct aspect of cultivating and leveraging algorithmic diversity as scientific robustness.

Diversity dimensions

The framework begins by identifying the fundamental axes along which diversity can be engineered: architecture (different model families), representation (different descriptor sets), initialization (different random seeds), data (different subsets or augmentations), and objective (different loss or regularization targets). Materials AI practitioners are encouraged to treat these dimensions as orthogonal levers rather than incidental by-products of model selection.

Diversity metrics

Quantitative indices must be adopted or adapted from the machine-learning literature to measure diversity within a collection. Beyond simple pairwise correlation or Q-statistic, the framework advocates composite metrics that integrate entropy of disagreement, representational dissimilarity (via embedding distances), and error-pattern orthogonality. These metrics allow researchers to report diversity as a first-class scientific attribute alongside accuracy.

Diversity generation

Deliberate strategies are required to produce diversity rather than merely hoping it emerges. These include heterogeneous architecture mixing, controlled data partitioning, systematic initialization sweeps, and multi-objective optimization that explicitly penalizes representational convergence. The framework emphasizes that generation should be goal-directed toward robustness rather than accuracy alone.

Diversity-robustness link

This component formalizes the mechanistic pathways through which diversity translates into robustness: (i) error decorrelation reduces collective variance under shift, (ii) disagreement maps epistemic uncertainty, and (iii) plural hypotheses accelerate discovery. Materials-specific illustrations include diverse collections for predicting polymorphic stability, in which architectural diversity reveals distinct sensitivities to van der Waals corrections.

Diversity evaluation

Finally, the framework introduces evaluation protocols that assess whether a diverse collection has achieved its robustness objectives. These protocols include stress-testing under simulated distribution shifts, ablation of individual diversity dimensions, and qualitative inspection of how diversity alters downstream scientific conclusions. Evaluation thereby closes the loop, ensuring diversity is not ornamental but functionally tied to epistemic gains.

Table 1 systematically decomposes algorithmic diversity into five orthogonal dimensions, each contributing a distinct mechanism through which model collections enhance epistemic robustness in materials AI.

Table 1. Five dimensions of algorithmic diversity and their distinct epistemic contributions in materials AI

Diversity type	Source of variation	Mechanism of epistemic contribution
Architectural	Model class differences (GNN, RF, and Transformer)	Divergent inductive biases expose structural assumptions
Representational	Descriptor/feature space variation	Tests the invariance of predictions across representations
Initialization	Random seeds/optimization paths	Explores multiple local minima in hypothesis space
Data	Training subset/augmentation differences	Introduces complementary observational perspectives
Objective	Loss/optimization criteria	Captures competing definitions of optimality

Taken together, the five components form a coherent conceptual scaffold that materials AI researchers can apply at any stage of model development, from initial design through peer-reviewed publication. The framework thereby operationalizes the theoretical claims of Section 4 and supplies the missing vocabulary and tools needed to move beyond single-model optimality.

Types of Diversity

The conceptual framework advanced in this work distinguishes five primary types of algorithmic diversity, each offering unique epistemic contributions to scientific robustness in materials AI. These types are not mutually exclusive; instead, they interact synergistically when deliberately combined, producing model collections whose collective behavior transcends the limitations of any single dimension of variation. By articulating each type with a precise definition, materials-specific illustrations, and an

analysis of its robustness benefit, the framework equips researchers to engineer diversity as a deliberate scientific strategy rather than an accidental byproduct of model selection.

Architectural diversity

Architectural diversity arises when a collection incorporates models built on fundamentally different computational paradigms, such as graph neural networks, random forests, transformer architectures, or physics-informed neural operators. In materials AI, this type manifests when one model encodes atomic interactions via message passing on crystal graphs, while another relies on hand-crafted descriptors fed into gradient-boosted trees. For instance, while Chen *et al.* [6] demonstrated the power of graph networks as a near-universal framework for molecules and crystals, pairing their approach with a random-forest baseline that operates on different inductive biases reveals how architectural divergence exposes hidden sensitivities to long-range order or local coordination environments. The robustness benefit is profound: architectural diversity prevents the entire collection from inheriting the same structural blind spots, thereby safeguarding against catastrophic failure when the underlying physics (for example, strong electron correlation in transition-metal oxides) violates assumptions embedded in any one architecture. A diverse architectural portfolio thus functions as an internal peer-review mechanism, flagging predictions that lack consensus across modeling philosophies and thereby elevating the epistemic trustworthiness of materials-property forecasts.

Representational diversity

Representational diversity concerns variation in the input featurizations or descriptor spaces used by different models within the collection. One model may operate directly on raw atomic coordinates and lattice parameters, while another may employ symmetry-invariant descriptors, such as smooth atomic position overlaps or graph-based node embeddings. In the context of materials discovery, this type is especially salient for inverse design tasks, where Zunger [7] highlighted the need to explore vast compositional spaces while implicitly relying on a single representational scheme. When multiple representations coexist—perhaps one rooted in electronic-structure fingerprints and another in geometric Voronoi tessellations—the collection gains the ability to detect property predictions that are representation-dependent artifacts rather than physically grounded truths. The robustness benefit lies in its capacity to mitigate misspecification risk: representational divergence acts as a diagnostic for whether a predicted structure–property relationship survives translation across descriptor languages, thereby increasing confidence that discovered materials candidates will survive experimental synthesis under real-world conditions.

Type 3: Initialization Diversity. Initialization diversity is achieved by training otherwise identical architectures from different random seeds, weight initializations, or stochastic optimization trajectories. Although superficially subtle, this type introduces meaningful variation in the learned internal representations even when architecture and data remain fixed. In solid-state materials science, where Schmidt *et al.* [5] surveyed advances that often report only a single trained instance of a deep model, the inclusion of multiple independently initialized replicas can expose basins of attraction in the loss landscape that correspond to qualitatively different physical interpretations of the same data. The robustness benefit emerges under distribution shift: models that converge to different local minima will disagree precisely in regions where extrapolation is most uncertain, furnishing an intrinsic uncertainty map that is far more informative than the output variance of a single deterministic model. This type, therefore, transforms initialization stochasticity from a nuisance into a controlled epistemic probe, enabling materials AI practitioners to quantify the sensitivity of their conclusions to the arbitrary starting point of gradient descent.

Data diversity

Data diversity encompasses the use of distinct training subsets, augmentation strategies, or data-generation protocols across the collection. One model may be trained on experimentally verified structures from the Inorganic Crystal Structure Database. At the same time, another may incorporate synthetic data generated via high-throughput density-functional theory with varying exchange-correlation functionals. Butler *et al.* [4] underscored the value of machine learning for molecular and materials science yet framed data primarily as a resource to be maximized for a single predictor; by contrast, deliberate data partitioning—bootstrap resampling, temporal splits reflecting evolving experimental capabilities, or physics-informed augmentations—creates models whose errors are decorrelated by construction. The robustness benefit is especially acute in materials AI, where data scarcity and evolving measurement techniques routinely induce

distribution shifts: a collection that has “seen” the materials universe through complementary data lenses is collectively more resilient to the discovery of new polymorphs or previously unmeasured properties, because no single data-view dominates the collective judgment.

Objective diversity

Objective diversity arises when models within the collection are optimized toward different loss functions, regularization targets, or multi-objective trade-offs. One model may minimize mean-squared error on formation energies, while another incorporates an auxiliary penalty for thermodynamic consistency or synthesizability scores. This type directly addresses the multi-faceted nature of materials design, where He *et al.* explored multiple topological materials datasets yet still converged on a unified objective. When objective functions diverge—perhaps one emphasizing predictive accuracy and another prioritizing physical interpretability—the resulting collection captures trade-offs that would otherwise remain invisible. The robustness benefit is epistemic pluralism: objective diversity ensures that scientific claims about material behavior are not artifacts of an arbitrary choice of loss landscape but survive interrogation under competing optimality criteria, thereby strengthening the warrant for downstream experimental validation or deployment in inverse-design campaigns.

Collectively, these five types illustrate how algorithmic diversity operates as a multi-layered epistemic scaffold. Each type contributes a distinct mechanism—architectural pluralism, representational translation invariance, initialization-induced basin exploration, data-lens complementarity, and objective pluralism—yet their true power emerges in combination. A model collection that simultaneously varies architecture, representation, initialization, data, and objective does not merely average predictions more accurately; it constructs a richer, more trustworthy map of the materials design space. The framework, therefore, urges materials AI researchers to explicitly report each type of diversity, treating it as a scientific attribute on par with accuracy or uncertainty quantification. In doing so, the field moves from a monoculture of “best” models toward ecosystems of diverse algorithmic perspectives that are inherently more robust to the uncertainties inherent in materials science.

Relation to Existing Concepts

Algorithmic diversity, as conceptualized here, both draws upon and extends several established ideas in machine learning and philosophy of science. Yet, it maintains a decisive conceptual boundary that distinguishes it from related but narrower notions. Most immediately, it relates to ensemble methods, which have historically treated diversity primarily as a tactical tool to boost predictive accuracy. Ortega *et al.* [1] and Rane *et al.* [2] demonstrated that uncorrelated errors among base learners reduce ensemble variance, yet their focus remained on the downstream accuracy dividend. The present framework inverts this relationship: diversity is repositioned as an epistemic end whose value persists even when accuracy gains are modest or deliberately sacrificed. Ensemble methods thus become one possible implementation pathway within the broader diversity-as-robustness paradigm rather than its definitional core.

The framework also intersects with robustness research. While conventional robustness techniques—adversarial training, domain adaptation, or out-of-distribution detection—typically harden a single model against specific failure modes, algorithmic diversity achieves robustness through pluralism. Liu *et al.* [8] and Jiang *et al.* [10] have explored ensemble robustness in atomistic and interpretable settings, yet they still subordinate diversity to final-model reliability. Here, robustness emerges organically from the collective disagreement patterns across diverse models, furnishing a self-diagnostic that no single hardened model can provide. Diversity, therefore, functions as a generative mechanism for robustness rather than a post-hoc patch.

A further connection exists with uncertainty quantification. Epistemic uncertainty in materials AI is often estimated via dropout, Bayesian approximations, or deep ensembles, yet these approaches rarely measure the diversity that underlies their uncertainty signals. The present framework elevates diversity metrics themselves as primary indicators of epistemic uncertainty: high disagreement across architectural, representational, or objective dimensions signals regions where scientific knowledge remains provisional. This stance aligns with but extends beyond the uncertainty-aware active-learning strategies reviewed by Vinchurkar *et al.* [12] and Chen *et al.* [6], transforming uncertainty from a passive diagnostic into an active design objective.

Finally, algorithmic diversity resonates with exploration strategies in scientific discovery. Where active learning and Bayesian optimization (as in Caggiano *et al.* [26] and

Gobert *et al.* [27]) seek to reduce uncertainty by sampling promising regions, diversity-driven collections explore hypothesis space more broadly by maintaining multiple, partially incompatible predictive pathways. Diversity thus serves as an exploration strategy that is inherently robust to model misspecification, because the collection never commits prematurely to a single inductive bias. Across all these relations, the framework consistently distinguishes diversity as an epistemic virtue—valuable for its own sake in the pursuit of reliable scientific knowledge—from its more familiar role as a tool subordinated to the maximization of accuracy. This distinction is not merely semantic; it carries direct consequences for how materials AI research is designed, evaluated, and disseminated [27-29].

Implications for Materials AI Practice

The conceptual framework carries concrete implications for three stakeholder groups—authors, reviewers, and the broader community—each of which must adapt its practices if algorithmic diversity is to become a recognized scientific virtue.

For authors, three shifts are essential. First, every manuscript reporting materials AI models should include explicit diversity metrics alongside traditional accuracy figures, quantifying at minimum architectural, representational, and initialization diversity within any model collection presented. Second, researchers should routinely publish not only the single best model but also the full diverse portfolio, enabling downstream users to exploit disagreement patterns for uncertainty estimation or hypothesis generation. Third, robustness analyses should replace or complement standard cross-validation, with stress tests explicitly probing how diversity dimensions buffer against distribution shifts common in materials data (for example, from computational to experimental regimes). These changes elevate diversity from an invisible background process to a foreground scientific claim.

Reviewers, in turn, bear responsibility for enforcing these standards. Referees should routinely ask whether a submitted study has quantified and justified the diversity profile of its model collection, and they should question the epistemic warrant of single-model studies that omit any analysis of alternative representational or objective choices. When authors claim generality for a new architecture, reviewers must demand evidence that the claimed superiority holds up under deliberate diversity. Such scrutiny will gradually raise the epistemic bar for publication in venues such as npj Computational Materials or Machine Learning: Science and Technology.

For the community as a whole, three structural initiatives are warranted. First, diversity-aware benchmarks should be established that reward not only accuracy but also explicitly measured diversity and robustness under controlled shifts. Second, open-source toolkits for diversity generation—incorporating automated architecture mixing, representation translators, and multi-objective optimizers—should be developed and maintained as community resources. Third, dedicated workshops and special issues should investigate the diversity–robustness relationship empirically across materials domains, thereby accumulating the shared knowledge needed to refine the framework over time. Collectively, these practice-level changes will realign materials AI with the pluralistic ethos of scientific inquiry, ensuring that the field's rapid predictive advances are matched by commensurate gains in epistemic reliability.

Table 2 contrasts the epistemic limitations of single-model optimization with the robustness advantages of diversity-centric approaches, clarifying the paradigm shift proposed in this work.

Table 2. Comparative epistemic profiles of single-model vs diversity-centric paradigms in materials AI.

Dimension	Single-model paradigm	Diversity-centric paradigm	Epistemic implication
Hypothesis coverage	Narrow, single inductive bias	Broad, multi-model exploration	Reduced risk of systematic error
Error behavior	Hidden, unobservable	Explicit via disagreement	Improved uncertainty detection
Robustness to shift	Fragile under distribution change	Resilient through diversity	Greater generalizabilit
Scientific reliability	Dependent on one model	Emergent from collective behavior	Stronger epistemic warrant

Discovery potential	Limited to model assumptions	Expanded via plural exploration	Higher novelt discovery
Evaluation criteria	Accuracy-focused	Diversity + robustness-focused	Shift in scientific standards
Failure visibility	Low (false consensus)	High (diagnostic disagreement)	Early detectio of invalid claims

The ultimate call is therefore unambiguous: materials AI must recognize algorithmic diversity not as an optional ensemble trick but as a core scientific virtue. Only by deliberately cultivating diverse model collections can the field achieve the robustness, generalizability, and discovery power demanded by the complexity of real-world materials challenges. Future work adopting this framework will move beyond solitary predictors toward rich algorithmic ecosystems whose collective intelligence more faithfully mirrors the pluralistic, uncertainty-aware character of scientific knowledge itself.

Conclusion

This paper has articulated algorithmic diversity as a foundational scientific robustness mechanism for materials AI. Beginning from the observation that the field currently prizes single-model optimality at the expense of epistemic pluralism, the work surveyed diversity concepts in machine learning, documented the prevailing single-model bias across key contributions, advanced three theoretical claims that position diversity as an independent epistemic virtue, and proposed a five-component conceptual framework encompassing diversity dimensions, metrics, generation strategies, robustness linkages, and evaluation protocols. It further delineated five distinct types of diversity—architectural, representational, initialization, data-centric, and objective—each illustrated with materials-relevant examples and robustness benefits. By relating algorithmic diversity to ensemble methods, robustness, uncertainty quantification, and exploration while maintaining its unique status as an end rather than a means, the framework supplies both vocabulary and structure for a paradigm shift.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 16 Feb 2025 Revised: 19 Apr 2025 Accepted: 29 Jun 2025
Published online: 18 January 2026

References

Ortega LA, Cabañas R, Masegosa A. Diversity and generalization in neural network ensembles. In: International Conference on Artificial Intelligence and Statistics. PMLR; 2022. p. 11720-43.

Rane N, Choudhary SP, Rane J. Ensemble deep learning and machine learning: Applications, opportunities, challenges, and future directions. Stud Med Health Sci. 2024;1(2):18-41.

Jensen D, LaMacchia B, Topcu U, Wisniewski P. Algorithmic robustness. arXiv preprint arXiv:2311.06275. 2023 Oct 17.

Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A. Machine learning for molecular and materials science. Nature. 2018;559(7715):547-55.

Schmidt J, Marques MR, Botti S, Marques MA. Recent advances and applications of machine learning in solid-state materials science. npj Comput Mater. 2019;5(1):83.

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.

Zunger A. Inverse design in search of materials with target functionalities. Nat Rev Chem. 2018;2(4):0121.

Liu K, Wei Z, Gao W, Dey P, Sluiter MHF, Shuang F. Heterogeneous ensemble enables a universal uncertainty metric for atomistic foundation models. npj Comput Mater. 2026;12:34. https://doi.org/10.1038/s41524-025-01905-x.

Vita JA, Samanta A, Zhou F, Lordi V. LTAU-FF: Loss trajectory analysis for uncertainty in atomistic force fields. Mach Learn Sci Technol. 2025;6(1):015048.

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.

Li K, DeCost B, Choudhary K, Greenwood M, Hattrick-Simpers J. A critical examination of robustness and generalizability of machine learning prediction of materials properties. npj Comput Mater. 2023;9(1):55.

Vinchurkar T, Abdelmaqsoud K, Kitchin JR. Uncertainty quantification in graph neural networks with shallow ensembles. Mach Learn Sci Technol. 2025;6(4):045007.

Rahman CM, Bhandari G, Nasrabadi NM, Romero AH, Gyawali PK. Enhancing material property prediction with ensemble deep graph convolutional networks. Front Mater. 2024;11:1474609.

Yang LH, Da B, Ding ZJ. Ensemble machine learning methods: Predicting electron stopping powers from a small experimental database. Phys Chem Chem Phys. 2021;23(10):6062-74.

Smyrnov M, Funcke F, Kabliman E. Prediction of material toughness using ensemble learning and data augmentation. Philos Mag Lett. 2024;104(1):2372497.

Karande P, Gallagher B, Han TY. A strategic approach to machine learning for material science: How to tackle real-world challenges and avoid pitfalls. Chem Mater. 2022;34(17):7650-65.

Morgan D, Jacobs R. Opportunities and challenges for machine learning in materials science. Annu Rev Mater Res. 2020;50(1):71-103.

Wang AY, Murdock RJ, Kauwe SK, Oliynyk AO, Gurlo A, Brgoch J, et al. Machine learning for materials scientists: An introductory guide toward best practices. Chem Mater. 2020;32(12):4954-65.

Wei J, Chu X, Sun XY, Xu K, Deng HX, Chen J, et al. Machine learning in materials science. InfoMat. 2019;1(3):338-58.

Cai J, Chu X, Xu K, Li H, Wei J. Machine learning-driven new material discovery. Nanosc Adv. 2020;2(8):3115-30.

Chen CT, Gu GX. Machine learning for composite materials. MRS Commun. 2019;9(2):556-66.

Huan TD, Batra R, Chapman J, Krishnan S, Chen L, Ramprasad R. A universal strategy for the creation of machine learning-based atomistic force fields. npj Comput Mater. 2017;3(1):37.

Baskaran A, Kautz EJ, Chowdhary A, Ma W, Yener B, Lewis DJ. Adoption of image-driven machine learning for microstructure characterization and materials design: A perspective. JOM. 2021;73(11):3639-57.

Alipour M, Harris DK. Increasing the robustness of material-specific deep learning models for crack detection across different materials. Eng Struct. 2020;206:110157.

Xiong J, Zhang T, Shi S. Machine learning of mechanical properties of steels. Sci China Technol Sci. 2020;63(7):1247-55.

Caggiano A, Zhang J, Alfieri V, Caiazzo F, Gao R, Teti R. Machine learning-based image processing for on-line defect recognition in additive manufacturing. CIRP Ann. 2019;68(1):451-4.

Gobert C, Reutzel EW, Petrich J, Nassar AR, Phoha S. Application of supervised machine learning for defect detection during metallic powder bed fusion additive manufacturing using high resolution imaging. Addit Manuf. 2018;21:517-28.

Borboudakis G, Stergiannakos T, Frysali M, Klontzas E, Tsamardinos I, Froudakis GE. Chemically intuited, large-scale screening of MOFs by machine learning techniques. npj Comput Mater. 2017;3(1):40.

Stein HS, Guevarra D, Newhouse PF, Soedarmadji E, Gregoire JM. Machine learning of optical properties of materials–predicting spectra from images and images from spectra. Chem Sci. 2019;10(1):47-55.