REVIEW **Open access**

Small-Data and Sparse-Regime Learning in Materials AI — Methods, Assumptions, and Limits

Maria Hernandez[1*], Carlos Vega[1], Lucia Torres[2]

Abstract

The integration of artificial intelligence (AI) and machine learning (ML) into materials science, often referred to as materials informatics or materials AI, has accelerated the discovery, design, and optimization of advanced materials. However, materials science frequently operates in small-data and sparse-regime conditions, where datasets are limited in size (often tens to hundreds of samples), high-dimensional, imbalanced, or sparsely populated due to the high cost, time, and complexity of experimental measurements and high-fidelity simulations. This narrative review synthesizes recent advances in methods tailored to these constraints, categorizing approaches at the data-source level (e.g., literature extraction, database construction, high-throughput workflows), algorithmic level (e.g., support vector machines, Gaussian process regression, ensemble models, imbalanced learning techniques), and strategic level (e.g., active learning, transfer learning). Key assumptions underlying these methods are examined, including similarity between source and target domains for transfer learning, representativeness of initial samples and reliable uncertainty quantification in active learning, and the validity of physical priors or inductive biases in physics-informed approaches. The review also addresses inherent limits, such as risks of overfitting, poor generalization beyond the training distribution, sensitivity to data quality and noise, challenges in uncertainty calibration, and dependence on domain expertise. By highlighting successful applications in property prediction, alloy design, and perovskite optimization, this work elucidates the current capabilities and boundaries of small-data and sparse-regime learning in materials AI, guiding researchers navigating data-limited environments.

Keywords Materials informatics, Transfer learning, Machine learning, Active learning, Small-data learning, Sparse-regime learning

*Correspondence:
Maria Hernandez
maria.hernandez@outlook.com
[1] Department of Materials Science and Machine Learning, Faculty of Engineering, Polytechnic University of Valencia, Valencia, Spain
[2] Department of AI Materials Simulation, Faculty of Engineering, University of Seville, Seville, Spain

Introduction

Materials science has long been constrained by the slow, costly, and labor-intensive nature of traditional experimentation and physics-based modeling. The discovery of new materials with tailored properties—for energy storage, catalysis, electronics, or structural applications—typically relies on iterative trial-and-error workflows or computationally intensive simulations such as density functional theory (DFT). While these approaches have delivered foundational insights, their scalability remains limited when navigating the vast combinatorial design spaces characteristic of modern materials research. In response to these challenges, materials informatics has emerged as a transformative paradigm, leveraging machine learning (ML) to extract structure–property relationships from data, accelerate property prediction, enable large-scale screening, and guide experimental decision-making under resource constraints [1-3].

The rapid expansion of open materials databases—such as the Materials Project, OQMD, and JARVIS—has enabled the development of high-capacity ML models that achieve impressive predictive accuracy within well-populated chemical and structural domains. Recent advances in graph-based representations and deep learning architectures have further strengthened this trajectory by offering transferable, symmetry-aware descriptions of atomic environments across the periodic table [4, 5]. At scale, such models benefit from statistical averaging, dense sampling of configuration space, and extensive coverage of compositional diversity, enabling robust interpolation in regimes well represented by training data [6, 7].

However, a substantial fraction of scientifically and technologically important materials problems do not reside in this large-data regime. Instead, they are characterized by limited sample sizes, fragmented datasets, and sparse coverage of high-dimensional design spaces. In many experimental contexts, available data may number only in the tens to hundreds, while feature spaces—encompassing composition, structure, processing history, and environmental conditions—remain high-dimensional and sparsely sampled. Data scarcity arises from multiple, often compounding, sources: experimental measurements demand specialized instrumentation, strict environmental control, and long acquisition times; high-throughput synthesis and characterization pipelines remain restricted to specific material classes; first-principles calculations scale poorly with system size and chemical complexity; and the aggregation of heterogeneous data sources introduces inconsistencies in quality, resolution, and annotation [1, 2, 8-11].

Learning under such constraints presents challenges fundamentally distinct from those encountered in data-rich settings. Small-data and sparse-regime learning exacerbate the risks of overfitting, unstable generalization, and sensitivity to noise, while limiting models' ability to extrapolate beyond observed regions of material space. Sparse sampling further amplifies class imbalance—such as the underrepresentation of high-performance or metastable materials—and increases vulnerability to distribution shifts, where models trained on common compounds are deployed to predict properties of chemically or structurally novel systems [10, 12, 13]. These conditions undermine naive scaling strategies and call into question assumptions that performance improvements can be achieved simply by increasing model complexity or computational resources.

To address these limitations, the materials AI community has developed a diverse set of strategies tailored to data-constrained regimes. These include small-data-optimized learning workflows that emphasize careful feature engineering, noise control, and workflow design [1, 9, 14]; transfer and multi-task learning frameworks that leverage knowledge learned from related properties or datasets [5, 12, 15]; active learning and Bayesian optimization approaches that adaptively select the most informative experiments or simulations [11, 16-19]; multi-fidelity modeling techniques that combine data sources of varying accuracy and cost [20]; and physics-informed and interpretable models that embed physical constraints and domain knowledge to reduce hypothesis space and improve epistemic reliability [18, 21, 22]. Collectively, these approaches reflect a shift from brute-force data accumulation toward information-efficient learning paradigms.

Despite these advances, important questions remain regarding the assumptions, limits, and failure modes of small-data and sparse-regime materials AI. The effectiveness of transfer learning depends on the relevance and alignment of source and target domains; active learning strategies are sensitive to the quality of uncertainty quantification and to exploration bias; physics-informed models may encode incomplete or approximate physical priors; and explainability methods often trade predictive performance for interpretive clarity [18, 21-24]. Moreover, recent discussions highlight the need for explicit governance of data quantity and quality, challenging the implicit belief that "more data" is always beneficial in materials ML workflows [6, 10].

This narrative review synthesizes peer-reviewed literature to critically examine methods, assumptions, and limitations of small-data and sparse-regime learning in materials AI. The objectives are threefold:

(1) to categorize and describe key methodological strategies for operating under data scarcity;

(2) to analyze the theoretical and practical assumptions that underpin these approaches; and

(3) to delineate their performance boundaries, epistemic risks, and potential failure modes.

The review is organized thematically, covering challenges inherent to data scarcity, strategies to enhance information efficiency, algorithmic adaptations, the integration of domain knowledge, and overarching assumptions that shape model reliability. By doing so, it aims to inform best practices and identify directions for future methodological development in data-constrained materials research.

Main Text

The challenge of data scarcity and sparse regimes in materials science

Data scarcity in materials science manifests in ways structurally distinct from those in other machine learning–driven fields. Whereas domains such as computer vision or natural language processing routinely rely on datasets containing millions of labeled examples, materials datasets are typically artisanal—derived from targeted experiments, specialized characterization campaigns, or computationally intensive simulations. As a result, materials data frequently reside in what is best described as a small-data and sparse-regime setting, characterized by low sample-to-feature ratios, uneven coverage of chemical and compositional space, and non-negligible levels of noise and uncertainty [1, 2, 8].

In this context, small data commonly refers to datasets with fewer than approximately 10^2–10^3 samples, while sparse regimes arise when high-dimensional descriptors—encompassing composition, crystal structure, local atomic environments, electronic features, and processing metadata—are populated by only a few observations per local region of feature space [1, 9, 10]. Importantly, sparsity is not merely a function of dataset size, but of representation density, such that even moderately sized datasets may remain sparse relative to the dimensionality and heterogeneity of the underlying materials landscape.

These conditions impose fundamental limits on machine learning performance. Models trained on small datasets are especially prone to overfitting, learning spurious correlations or noise rather than generalizable structure–property relationships [1, 2]. Sparse sampling further compromises extrapolative reliability, particularly in out-of-distribution scenarios where models trained on common compounds are tasked with predicting properties of chemically or structurally novel materials [10, 12, 13]. Class imbalance—ubiquitous in materials discovery, where rare high-performance or metastable phases are underrepresented—further degrades predictive accuracy and robustness for scientifically valuable targets [1, 11]. Compounding these issues, data heterogeneity arising from the combination of experimental and computational sources introduces systematic biases, while label noise stemming from measurement uncertainty or simulation approximations amplifies epistemic uncertainty [2, 9]. The diverse origins and technical consequences of data scarcity and sparsity in materials AI are summarized in **Table 1**

Table 1. Sources, characteristics, and learning implications of data scarcity in materials science

Source of scarcity	Manifestation in materials data	Learning consequence	Epistemic risk
Limited experimental throughput	Tens–hundreds of samples per study	Overfitting, unstable models	Spurious correlations
High simulation cost (DFT, MD)	Sparse coverage of composition–structure space	Poor extrapolation	False confidence
High-dimensional descriptors	Low sample-to-feature ratio	Curse of dimensionality	Unreliable feature attribution

Class imbalance	Rare high-performance or metastable materials	Biased predictors	Systematic underprediction
Data heterogeneity	Mixed experimental and computational labels	Distribution shift	

These challenges are particularly acute in targeted materials design applications, such as high-entropy alloys, halide perovskites for photovoltaics, or heterogeneous catalysts, where exhaustive enumeration of composition–structure–property space is computationally or experimentally infeasible [4, 14, 15]. In such settings, naive data accumulation strategies are insufficient, and methods that maximize information extraction from limited observations or strategically acquire new data become essential.

Enhancing data availability at the source level

One class of responses to data scarcity focuses on increasing data availability and utility at the source level. Natural language processing (NLP) and text-mining techniques have been employed to extract structured materials data from the scientific literature, including composition–property pairs, synthesis conditions, and processing protocols [9]. These approaches have enabled the semi-automated population of materials databases and the recovery of otherwise inaccessible experimental knowledge. However, such strategies rely on assumptions of consistency, representativeness, and reporting quality that are often violated due to publication bias and incomplete metadata [1, 10].

Complementary efforts leverage high-throughput computation and automation to generate synthetic data at scale. First-principles databases have been widely used for model pre-training, data augmentation, and benchmarking, while robotic experimentation platforms enable closed-loop discovery pipelines that couple prediction, synthesis, and characterization [16, 17, 19]. Notably, Bayesian and active learning–driven experimental workflows have demonstrated the ability to concentrate experimental effort on informative regions of materials space, thereby improving data efficiency [11, 16]. Nevertheless, these approaches remain constrained by computational cost, domain specificity, and the fidelity gap between simulated and experimental data [20, 22].

Algorithmic approaches tailored to small data

Beyond data acquisition, a substantial body of work has focused on algorithmic strategies explicitly designed for small-data and sparse-regime learning. Classical machine learning methods such as support vector machines, random forests, gradient boosting models, and symbolic regression exhibit favorable bias–variance trade-offs under limited data due to built-in regularization and ensemble averaging [1, 7, 8]. Gaussian process regression (GPR), in particular, has gained prominence in materials science for its ability to provide calibrated uncertainty estimates alongside predictions, a feature that is critical for decision-making under sparsity [18].

Ensemble learning approaches further enhance robustness by combining multiple weak or complementary learners, thereby reducing sensitivity to noise and to the choice of representation. Such strategies have been successfully applied to property-prediction tasks in chemically complex systems, including bandgap estimation and alloy design, where single-model performance is unstable [4, 18, 20]. In parallel, techniques for handling imbalanced datasets—such as resampling strategies, cost-sensitive learning, and uncertainty-aware acquisition—have been employed to mitigate performance degradation on rare but scientifically important material classes [17, 19].

Despite their advantages, these methods rely on assumptions that limit their applicability. Many presume relatively low label noise, meaningful feature representations, and manageable computational cost for hyperparameter optimization or ensemble construction [1, 9]. When these assumptions are violated, gains in robustness may plateau or reverse, underscoring the need for principled workflow design and explicit governance of data quantity and quality [10, 14].

Strategic machine learning: Active learning and transfer learning

Strategic machine learning approaches aim to maximize information gain under data constraints by altering how data are selected, reused, or contextualized, rather than by increasing dataset size. Among these, active learning (AL) and transfer learning (TL) have emerged as central paradigms for research on small-data and sparse-regime materials.

Active learning operates through an iterative loop in which a model identifies the most informative samples to query—via simulation or experiment—based on uncertainty, expected improvement, or information-theoretic criteria. In materials science, AL is often coupled with Bayesian optimization frameworks, where acquisition functions balance exploration of under-sampled regions with exploitation of promising candidates [16-19]. Such closed-loop strategies have demonstrated substantial gains in data efficiency across property optimization and discovery tasks, enabling comparable or superior performance with orders of magnitude fewer evaluations than random or grid-based sampling [11, 16].

The effectiveness of AL in materials settings is closely tied to the quality of uncertainty quantification, as acquisition decisions rely on calibrated estimates of epistemic uncertainty in sparse regions of feature space [18, 19]. When uncertainty is misestimated—due to poor initial models, noisy labels, or representation mismatch—query efficiency degrades, and AL may repeatedly sample uninformative or misleading regions [17]. Moreover, AL workflows implicitly assume that newly acquired data are representative of the target distribution and that the underlying property landscape is sufficiently smooth for local sampling to yield global gains, assumptions that may fail in chemically complex or discontinuous systems [6, 10].

Transfer learning addresses data scarcity by leveraging knowledge learned from related tasks, properties, or materials domains. In materials informatics, TL is commonly implemented by pre-training models—often graph-based neural networks—on large, heterogeneous datasets and fine-tuning them on small, task-specific targets [5, 9, 15]. Cross-property and multi-task transfer learning frameworks further extend this idea by exploiting shared physical structure across correlated properties, enabling improved predictive performance in regimes where direct supervision is limited [5, 12].

Despite its promise, TL rests on the critical assumption of domain alignment between source and target tasks. When feature distributions, chemical spaces, or underlying physical mechanisms diverge substantially, transferred representations may introduce systematic bias or lead to negative transfer, degrading performance relative to training from scratch [1, 10, 13]. Recent work emphasizes the importance of explicitly diagnosing transferability and avoiding indiscriminate reuse of large pre-trained models, particularly in small-data regimes where spurious correlations can dominate learning dynamics [6, 10].

Integration of domain knowledge and physics-informed approaches

A complementary strategy for mitigating data scarcity involves embedding domain knowledge and physical constraints directly into learning workflows. Physics-informed machine learning (PIML) frameworks incorporate governing equations, conservation laws, or empirical relationships into model architectures or loss functions, thereby restricting the hypothesis space and reducing data requirements [22]. Such constraints are especially valuable in sparse regimes, where unconstrained models may otherwise converge to physically implausible solutions.

In parallel, descriptor engineering and symbolic regression approaches aim to construct compact, interpretable representations that encode known physical structure, reducing effective dimensionality and enhancing generalization [1, 9]. These methods emphasize transparency and scientific interpretability, enabling models to function as hypothesis generators rather than black-box predictors [21, 24]. Hybrid approaches that combine physically motivated descriptors with flexible learning architectures further reflect an emerging consensus that data efficiency and interpretability are often coupled rather than competing objectives.

However, physics-informed and knowledge-driven methods are not without limitations. Their effectiveness depends on the accuracy, completeness, and applicability of the imposed priors; oversimplified or context-inappropriate constraints can

bias models and suppress genuine discovery [22]. Moreover, many materials systems involve multi-scale, non-equilibrium, or poorly understood phenomena that resist concise physical formalization, limiting the universality of such approaches [3, 7].

Assumptions and limitations across methods

Across active learning, transfer learning, and physics-informed strategies, small-data materials AI rests on a set of shared assumptions: that available data are representative of the target domain; that uncertainty estimates meaningfully reflect epistemic ignorance; that source and target distributions are sufficiently aligned for knowledge transfer; and that embedded physical constraints remain valid in unexplored regimes [1, 10, 18, 22]. When these assumptions are violated, models may exhibit overconfidence, unstable extrapolation, or amplification of existing biases. **Figure 1** illustrates a diagnostic flow linking common failure modes in small-data and sparse-regime materials AI to their underlying assumptions and corresponding mitigation strategies.

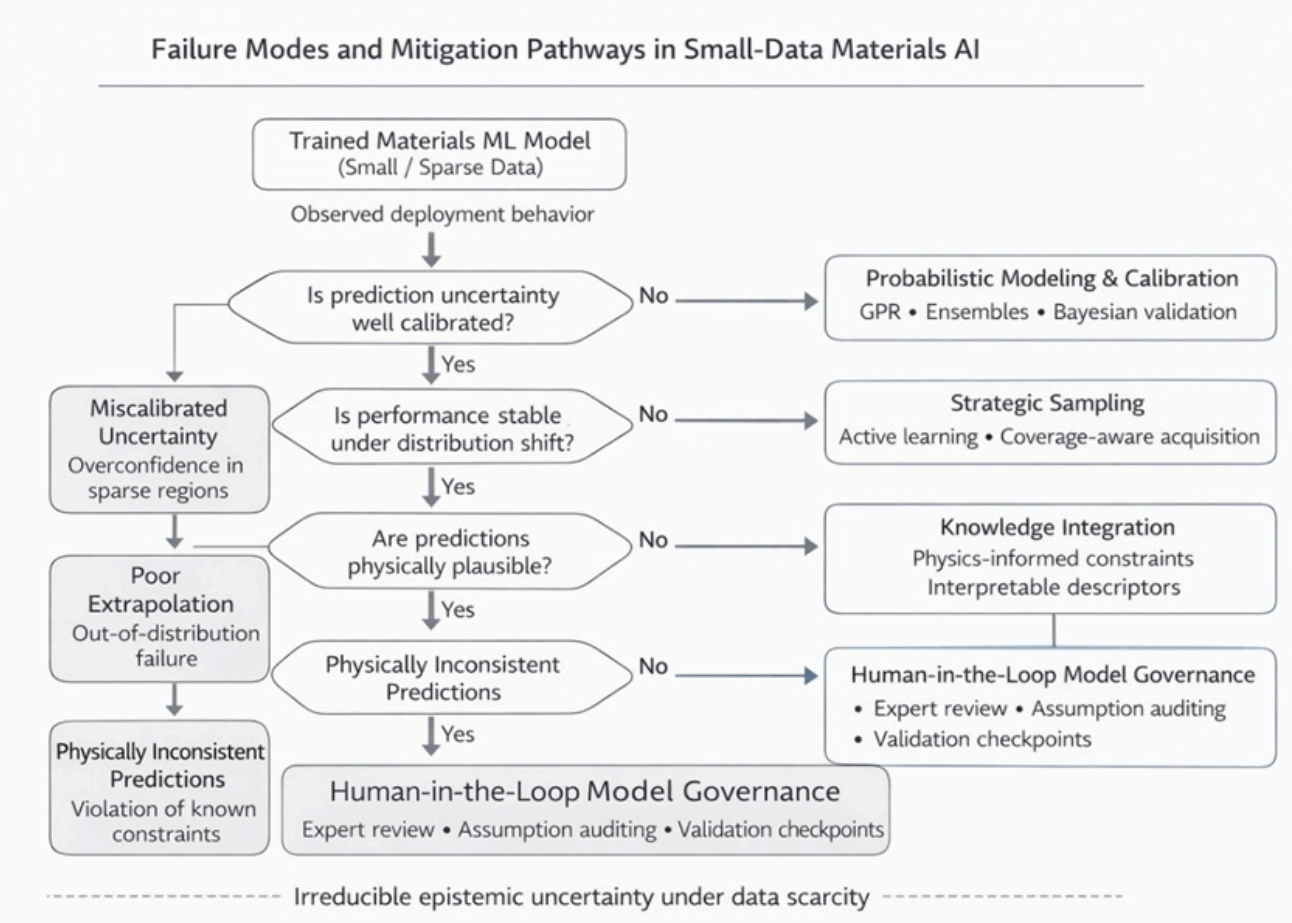

Figure 1. Diagnostic flow of small-data failure modes and mitigation paths in materials AI

Common failure modes include sensitivity to outliers, degradation in sparsely sampled regions of feature space, and heavy reliance on expert judgment for representation design, validation, and interpretation [2, 9, 21]. Generalization beyond the training manifold remains a persistent challenge, underscoring the limits of purely algorithmic solutions. These considerations motivate growing interest in hybrid human–AI workflows, where strategic learning methods are complemented by domain expertise, uncertainty-aware decision-making, and explicit governance of data and modeling choices [3, 10, 13]. The core assumptions, advantages, and characteristic failure modes of major small-data learning strategies are synthesized in **Table 2**.

Table 2. Comparative overview of small-data and sparse-regime learning strategies in materials AI

Strategy class	Representative methods	Core assumption	Strength in small data	Dominant limitation
Classical ML	SVM, RF, gradient boosting	Informative features, moderate noise	Strong regularization	Feature sensitivity
Probabilistic models	Gaussian process regression	Smooth property landscapes	Calibrated uncertainty	Poor scalability
Ensemble learning	Bagging, stacking	Error diversity among learners	Robustness to noise	Computational cost
Active learning	Bayesian optimization	Reliable uncertainty estimates	Data efficiency	Query bias
Transfer learning	Graph NN fine-tuning	Domain alignment	Knowledge reuse	Negative transfer
Physics-informed ML	PINNs, constrained loss	Valid physical priors	Improved extrapolation	Prior misspecification
Symbolic/interpretable ML	SISSO, symbolic regression	Low-dimensional physics	Interpretability	Limited expressiveness

Results and Discussion

The synthesis of recent literature indicates that small-data and sparse-regime learning has become a structural cornerstone of materials artificial intelligence, reflecting the persistent reality that most materials problems are characterized by limited, noisy, heterogeneous, and unevenly distributed data [1-3, 8, 14]. Across experimental and computational settings, data scarcity is not an exceptional case but the dominant operating condition, motivating a shift away from scale-centric paradigms toward information-efficient learning strategies. **Figure 2** presents a conceptual synthesis of small-data and sparse-regime learning in materials AI, highlighting how data constraints, methodological choices, and epistemic assumptions interact to shape model reliability.

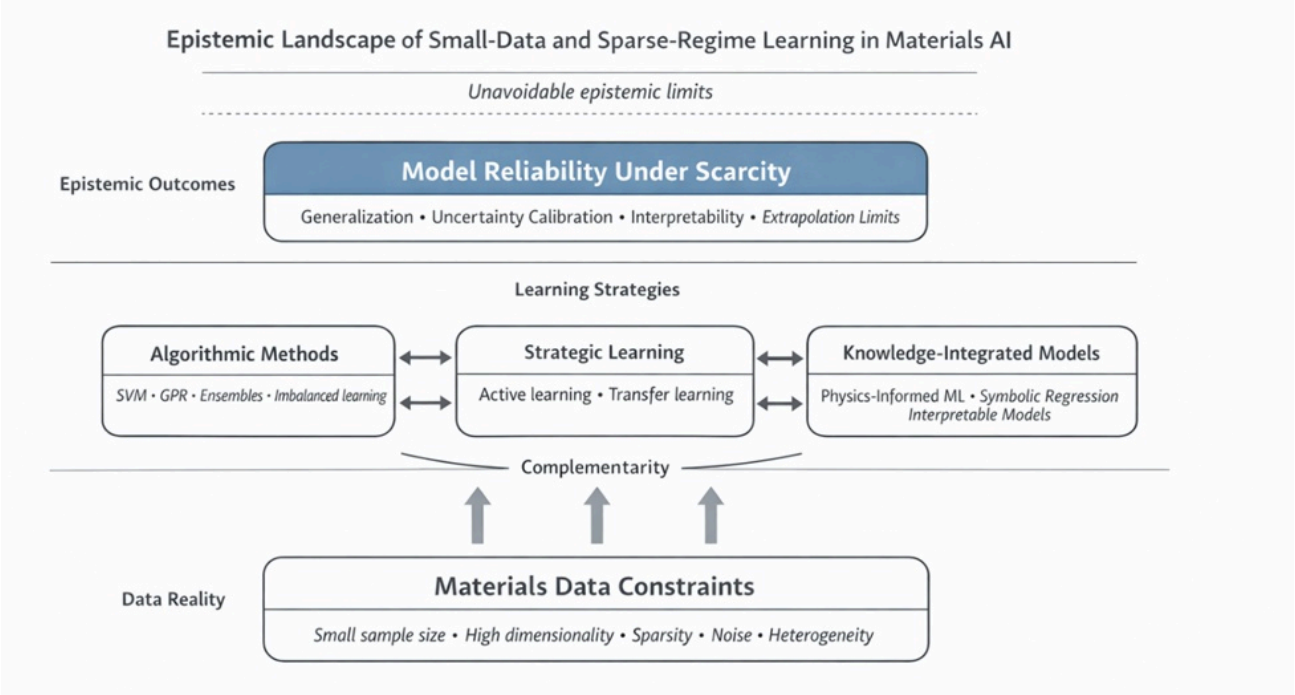

Figure 2. Epistemic landscape of small-data and sparse-regime learning in materials AI

At the data-source level, efforts to mitigate scarcity have focused on expanding and repurposing available information. Natural language processing and text-mining techniques have enabled the extraction of structured data from the literature, particularly for synthesis protocols and experimentally measured properties that are difficult to obtain through high-throughput computation alone [10]. In parallel, automated and semi-autonomous experimental platforms have demonstrated the feasibility of closed-loop discovery pipelines, where Bayesian active learning guides sequential experimentation and concentrates resources on informative regions of materials space [16, 17, 19]. While these approaches improve data efficiency, they implicitly assume consistency in reporting standards and representative sampling; in practice, publication bias, incomplete metadata, and fragmentation across databases continue to limit the reliability and coverage of aggregated datasets [1, 10].

From an algorithmic perspective, methods designed for small-data regimes emphasize regularization, uncertainty awareness, and robustness. Gaussian process regression (GPR) remains a prominent tool due to its principled treatment of uncertainty and favorable performance in sparse, high-dimensional settings [18]. Ensemble-based approaches, including random forests and gradient boosting models, further mitigate overfitting by averaging across multiple hypotheses, while support vector machines offer stability under limited sample sizes [1, 7, 8]. However, the effectiveness of these models depends critically on the quality of the representations and the noise levels; severe label uncertainty or poorly chosen descriptors can erode their advantages and lead to brittle predictions [2, 9].

Beyond static modeling choices, strategic learning paradigms such as active learning (AL) and transfer learning (TL) play a central role in coping with data scarcity. AL has proven particularly effective in guiding expensive simulations or experiments, enabling substantial reductions in the number of required evaluations by prioritizing samples that maximize expected information gain [11, 16, 17, 19]. Extensions to multi-objective and constrained optimization further reflect the realities of materials design, where trade-offs between competing properties must be navigated under limited budgets [18, 19]. TL, especially through pre-trained graph-based models and cross-property learning frameworks, enables knowledge reuse across related tasks and domains, achieving competitive accuracy with only tens to hundreds of labeled samples [5, 12, 15]. Hierarchical and modular transfer strategies, such as AtomSets, underscore that transferability is structured rather than universal, depending on alignment between representations, chemistry, and underlying physical mechanisms [5].

The integration of domain knowledge provides an additional axis of robustness in data-constrained regimes. Physics-informed machine learning frameworks constrain hypothesis spaces using governing equations, conservation laws, or empirical relationships, reducing data requirements and improving physical plausibility [22]. Descriptor engineering and symbolic regression approaches similarly aim to encode physical structure into compact, interpretable representations, enhancing generalization when purely data-driven models fail to extrapolate reliably [1, 9, 21, 24]. Nevertheless, these approaches rely on the availability and validity of physical priors; oversimplified or context-inappropriate constraints risk suppressing emergent behavior in complex materials systems [3, 7].

Despite substantial progress, fundamental limitations remain. Overfitting remains a dominant risk in small-data regimes, exacerbated by high dimensionality, class imbalance, and distribution shifts between training and deployment [1, 10, 14]. Extrapolation beyond the training manifold is particularly fragile, with uncertainty estimates—whether derived from GPR, ensembles, or AL acquisition functions—sometimes poorly calibrated in sparsely sampled regions [18, 20]. Data quality issues, including noise, inconsistencies, and hidden biases, propagate disproportionately when data are scarce. At the same time, reliance on expert-driven feature selection and workflow design introduces subjectivity that is difficult to standardize or audit [2, 21, 22]. In transfer learning settings, negative transfer remains a persistent concern when source and target domains diverge, undermining the assumption that larger proxy datasets necessarily confer epistemic advantage [6, 10, 13].

Collectively, these challenges highlight the need for rigorous validation practices, including domain-specific benchmarks, uncertainty-aware decision-making, and explicit human-in-the-loop oversight. Rather than replacing scientific judgment, small-data materials AI increasingly functions as a decision-support system, where hybrid workflows combine algorithmic efficiency with expert interpretation and governance [3, 10].

Conclusion

Small-data and sparse-regime learning has reshaped materials AI from a field implicitly dependent on large-scale datasets into one capable of delivering scientifically meaningful insights under realistic constraints. The methods reviewed here—spanning data-source enhancement, algorithmic adaptation, strategic learning, and physics-informed integration—have enabled accelerated exploration and design across diverse materials classes, often with dramatically reduced experimental or computational cost.

At the same time, the field remains methodologically immature in several respects. Progress to date has revealed not only what is possible under data scarcity, but also where current approaches falter. Future research directions include:

(i) the development of more robust and better-calibrated uncertainty quantification methods tailored to sparse, high-dimensional materials spaces;

(ii) advances in multi-fidelity and multi-task learning frameworks that more effectively exploit heterogeneous data sources without propagating bias;

(iii) exploration of foundation and multimodal models pre-trained on large materials corpora for zero- and few-shot adaptation, while critically assessing their transferability limits;

(iv) deeper integration of causal reasoning and physics-informed priors to improve extrapolation, interpretability, and scientific insight; and

(v) the establishment of standardized benchmarks and evaluation protocols for small-data materials tasks to enable reproducible and comparable progress across studies.

Finally, ethical and systemic considerations must not be overlooked. Bias amplification in underrepresented material classes, unequal access to high-throughput infrastructure, and the opacity of increasingly complex models pose risks to equitable and responsible innovation. Addressing these concerns will be essential if materials AI is to mature into a reliable, generalizable, and scientifically grounded discipline capable of advancing materials discovery even in the most data-constrained regimes.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 17 Jul 2025 Revised: 15 Aug 2025 Accepted: 19 Sep 2025
Published online: 18 January 2026

References

Xu P, Ji X, Li M, Lu W. Small data machine learning in materials science. npj Comput Mater. 2023;9:42.

Vanpoucke D. Small data materials design with machine learning: When the average model knows best. J Appl Phys. 2020;128:054901.

Jain A. Machine learning in materials research: Developments over the last decade and challenges for the future. Curr Opin Solid State Mater Sci. 2024;71:101118.

Chen C, Ong SP. A universal graph deep learning interatomic potential for the periodic table. Nat Comput Sci. 2022;2:718-28.

Chen H, Ong SP. AtomSets as a hierarchical transfer learning framework for small and large materials datasets. npj Comput Mater. 2021;7:184.

Merchant A, Batzner S, Schoenholz SS, Aykol M, Cheon G, Cubuk ED. Scaling deep learning for materials discovery. Nature. 2023;624(7990):80-5.

Unni R, Batra R, Ramprasad R. Advancing materials science through next-generation machine learning. Adv Mater. 2024;36:2401234.

Yuan L, Guo H, Li Q, Zhang H, Xu M, Zhang W, et al. Machine-learning-assisted material discovery of pyridine-based polymers for efficient removal of ReO4–. Environ Sci Technol. 2024;58(34):15298-310.

Herhausen D, Bernritter SF, Ngai EW, Kumar A, Delen D. Machine learning in marketing: Recent progress and future research directions. J Bus Res. 2024;170:114254.

Ji X, Xu P, Li M, Lu W. Data quantity governance for machine learning in materials science. Natl Sci Rev. 2023;10:nwad125.

Tran K, Neiswanger W, Yoon J, et al. Strategies for incorporating active learning into high-throughput density functional theory calculations. npj Comput Mater. 2023;9:54.

Jha D, Singh V, Ward L, et al. Cross-property deep transfer learning framework for enhanced predictive analytics on small materials data. Nat Commun. 2021;12:6593.

Yu Y, Xiong J, Wu X, Qian Q. From small data modeling to large language model screening: A dual-strategy framework for materials intelligent design. Adv Sci. 2024;11:2403548.
https://doi.org/10.1002/advs.202403548.

Ma Y, Xu P, Li M, Ji X, Zhao W, Lu W. The mastery of details in the workflow of materials machine learning. npj Comput Mater. 2024;10:141.

Lee S, Asahi R. Transfer learning for materials informatics using crystal graph convolutional neural network. Comput Mater Sci. 2021;190:110314.

Kusne AG, Yu H, Wu C, Zhang H, Hattrick-Simpers J, DeCost B, et al. On-the-fly closed-loop materials discovery via Bayesian active learning. Nat Commun. 2020;11(1):5966.
https://doi.org/10.1038/s41467-020-19597-w.

Frey N, Soklaski R, Rocca D, et al. Benchmarking active learning strategies for materials optimization and discovery. Oxf Open Mater Sci. 2022;2:itac006.

Butler KT, Oviedo F, Famprikis T. Gaussian process regression for materials and molecules. Chem Rev. 2021;121:10073-141.

Shahzad K, Mardare AI, Hassel AW. Accelerating materials discovery: Combinatorial synthesis, high-throughput characterization, and computational advances. Sci Technol Adv Mater Methods. 2024;4(1):2292486.

Pilania G, Ivady V, Zurek E, et al. Multi-fidelity machine learning models for accurate bandgap predictions of solids. Comput Mater Sci. 2021;188:110203.

Oviedo F, Ferres JL, Buonassisi T, Butler KT. Interpretable and explainable machine learning for materials science and chemistry. Acc Mater Res. 2022;3:1023-35.

Karniadakis GE, Kevrekidis IG, Lu L, Perdikaris P, Wang S, Yang L. Physics-informed machine learning. Nat Rev Phys. 2021;3:422-40.

Doncevic DT, Mitsos A, Guo Y, Li Q, Dietrich F, Dahmen M, et al. A recursively recurrent neural network (r2n2) architecture for learning iterative algorithms. SIAM J Sci Comput. 2024;46(2):A719-43.

Azari H, Nazari E, Mohit R, Asadnia A, Maftooh M, Nassiri M, et al. Machine learning algorithms reveal potential miRNAs biomarkers in gastric cancer. Sci Rep. 2023;13(1):6147.