

ORIGINAL RESEARCH

Open access

Learning Under Scarcity: A Conceptual Theory of Small-Data Regimes in Materials Artificial Intelligence

Oliver Grant^{1*}, Daniel Brooks¹, Amelia Carter²

Abstract

The integration of artificial intelligence into materials science has accelerated discovery processes, yet the persistent challenge of data scarcity undermines the full potential of these technologies. This conceptual paper develops a novel theoretical framework for understanding small-data regimes in materials AI, emphasizing the interpretive dynamics that emerge when limited datasets intersect with domain knowledge and computational strategies. By synthesizing recent literature, the framework explains how scarcity influences model behavior through mechanisms of uncertainty amplification and knowledge integration, revealing interaction patterns between sparse empirical inputs and physics-informed priors. Analytical implications include enhanced epistemic reasoning about model reliability in low-data contexts, where trade-offs between generalization and specificity manifest in feedback structures that guide iterative refinement. Conceptual interpretations highlight steering logics that balance data-driven insights with theoretical constraints, fostering systems-level insights into how small-data environments reshape AI workflows in materials design. The framework underscores ethical considerations in deploying such systems, particularly regarding bias propagation under scarcity. Through a detailed textual description of a schematic figure, the paper illustrates these dynamics and offers integrative perspectives for advancing materials informatics without relying on large-scale data collection. Ultimately, this theory reorients focus toward resilient AI architectures that thrive amid informational constraints, promoting sustainable innovation in the field.

Keywords Knowledge integration, Artificial intelligence, Epistemic reasoning, Data scarcity, Small-data regimes, Uncertainty dynamics

*Correspondence:

Oliver Grant

oliver.grant@gmail.com

¹ Department of Materials Informatics, Faculty of Engineering, University of Manchester, Manchester, United Kingdom

² Department of Artificial Intelligence Systems, Faculty of Computer Science, University of Birmingham, Birmingham, United Kingdom

Introduction

The advent of artificial intelligence (AI) in materials science marks a transformative shift, enabling accelerated exploration of vast chemical spaces and complex property landscapes. Traditionally, materials discovery relied on labor-intensive experimental trials and theoretical simulations, often constrained by time and resources. AI, particularly machine learning algorithms, has introduced data-driven paradigms that infer patterns from accumulated knowledge, facilitating predictions of material behaviors without exhaustive testing [1, 2]. This evolution is evident in applications ranging from alloy design to nanomaterial synthesis, where AI models analyze structural descriptors to predict properties such as thermal conductivity and electronic bandgaps [3, 4]. However, the efficacy of these approaches hinges on the availability of robust datasets, a condition frequently unmet in materials contexts due to the high cost of generating high-fidelity data [5].

Data scarcity emerges as a central interpretive challenge in materials AI, where small-data regimes—characterized by limited samples relative to the problem space’s dimensionality—alter the fundamental dynamics of learning processes [6]. Unlike domains with abundant data, such as image recognition, materials science grapples with sparse, heterogeneous inputs derived from experiments, simulations, and databases [7]. This scarcity amplifies uncertainties, leading to interaction patterns where models must negotiate between overfitting to noise and underutilizing available information [8]. Conceptual interpretations of these regimes reveal trade-offs in model architectures: simplistic assumptions may yield brittle predictions, while overly complex ones may exacerbate data demands [9]. Systems-level insights suggest that scarcity fosters reliance on external knowledge sources, such as physical laws or analogous systems, creating feedback structures that iteratively refine AI outputs [10]. To clarify the defining features of such regimes beyond simple dataset size, **Table 1** synthesizes the conceptual characteristics of small-data conditions in materials AI, highlighting their epistemic, technical, and ethical implications.

Table 1. Conceptual characteristics of small-data regimes in materials artificial intelligence

Dimension	Conceptual description	Interpretive implication
Dataset size	The sample count is small relative to feature dimensionality	Increased variance, instability
Data heterogeneity	Mixed experimental, simulation, and database sources	Inconsistent learning signals
Acquisition cost	High experimental or computational burden	Limits iterative validation
Uncertainty dominance	Epistemic uncertainty outweighs aleatoric	Requires probabilistic reasoning
Bias sensitivity	Sparse sampling amplifies prior biases	Ethical risk amplification
Validation constraints	Limited hold-out or test data	Accuracy metrics become unreliable

The literature underscores the growing recognition of the challenges posed by small data. Recent advancements highlight how limited datasets impede the training of deep neural networks, which thrive on volume for capturing nuanced relationships [11]. In response, researchers have explored integrative strategies that blend data augmentation with domain expertise, interpreting how synthetic expansions can mitigate gaps without introducing artifacts [12]. Ethical reasoning becomes paramount here, as biased small datasets risk perpetuating inequities in material applications, such as in sustainable energy technologies [13]. Epistemic considerations further underscore the need for transparent models that acknowledge uncertainty, steering logics toward probabilistic rather than deterministic frameworks [14].

This paper addresses a critical gap: the absence of a unified conceptual theory for small-data regimes in materials AI. Existing work often focuses on algorithmic fixes, such as transfer learning from large repositories to sparse targets, but overlooks the broader interpretive implications [15, 16]. By developing a novel framework, we interpret scarcity not as a mere obstacle but as a generative force shaping AI-material interactions. Analytical implications include rethinking validation metrics to prioritize robustness over accuracy in low-data scenarios [17]. The framework integrates dynamics of knowledge transfer, in which pre-trained general-chemistry models adapt to specific material contexts, revealing trade-offs between adaptability and specificity [18].

Structurally, the paper proceeds as follows. The theoretical background synthesizes recent literature, delineating the evolution of materials AI and the interpretive challenges posed by data scarcity. Subsections explore augmentation techniques, transfer mechanisms, and few-shot adaptations, providing a cohesive narrative of current approaches [19, 20]. The proposed framework then articulates original conceptual interpretations, emphasizing interaction dynamics and feedback structures that emerge under scarcity. A detailed textual description of **Figure 1** visualizes these elements and

provides systems-level insights. This conceptual lens offers pathways for ethical and epistemic advancements, ultimately guiding materials AI toward resilient, scarcity-aware paradigms.

In interpreting the broader landscape, small-data regimes compel a reevaluation of progress metrics in materials science. Traditional benchmarks, often derived from abundant-data contexts, may undervalue the nuanced gains in low-resource settings [21]. Steering logics here involve balancing computational efficiency with interpretive depth, where hybrid models incorporating physical constraints enhance reliability [22]. Feedback structures, such as active learning loops that query uncertain regions, exemplify how scarcity drives innovation in data acquisition strategies [23]. Ethical reasoning extends to equitable access, ensuring that AI advancements benefit diverse material applications, from biomedicine to environmental remediation [24].

The urgency of this theory is heightened by global challenges, such as the need to rapidly develop sustainable materials in line with climate imperatives. Small-data AI can accelerate discoveries in areas such as carbon capture and battery electrolytes, where experimental data is inherently limited [25]. However, without conceptual grounding, deployments risk epistemic pitfalls, such as overconfidence in sparse inferences [26]. This paper's framework provides interpretive tools to navigate these, fostering integrative approaches that harmonize data limitations with scientific ambition.

Theoretical Background and Literature Synthesis

Evolution of artificial intelligence in materials science

The application of artificial intelligence (AI) in materials science has undergone a marked conceptual and methodological transformation over the past two decades, evolving from narrowly scoped pattern-recognition tools into comprehensive, decision-shaping components of scientific workflows. Early AI applications in the field were predominantly grounded in descriptor-based modeling paradigms, in which domain experts manually engineered features derived from atomic composition, crystallographic symmetry, or thermodynamic quantities to support property-prediction tasks [1, 27]. These approaches reflected a hybrid epistemic stance: while learning algorithms performed statistical inference, scientific understanding remained anchored in human-defined representations and prior physical intuition.

Recent advances in computational power, algorithmic design, and data infrastructure have catalyzed a transition toward end-to-end learning architectures. Deep neural networks, graph-based models, and attention mechanisms now enable the automated extraction of latent representations directly from raw or minimally processed materials data, such as atomic coordinates, microstructural images, or spectral signatures [2, 3]. This shift is not merely technical but epistemological, as it redistributes interpretive authority from explicitly defined descriptors to model-internal representations operating in high-dimensional spaces. Consequently, AI systems increasingly mediate how structure–property relationships are conceptualized, explored, and operationalized.

This maturation of AI has coincided with its integration into high-throughput materials discovery pipelines, where virtual screening, simulation-driven candidate generation, and experimental prioritization are tightly coupled [4, 5]. In such workflows, AI functions less as an auxiliary analytical tool and more as a coordinating agent that shapes experimental agendas by filtering vast combinatorial spaces of possible compositions and processing conditions. Interpretive dynamics in this context highlight AI's role in augmenting, rather than replacing, human intuition—particularly in navigating regions of the materials design space that are cognitively inaccessible through traditional trial-and-error reasoning [6].

At a systems level, this evolution has been reinforced by the consolidation of large-scale materials databases, which facilitate the cross-referencing of computational predictions with empirical measurements [7]. However, this dependence on aggregated data infrastructures introduces new vulnerabilities. Inconsistencies in data provenance, heterogeneity in measurement protocols, and uneven coverage across chemical systems can propagate through AI models, influencing predictions and downstream decision-making in opaque ways [8]. Feedback structures further complicate this landscape: iterative strategies such as Bayesian optimization dynamically reshape search spaces based on prior outcomes,

embedding earlier modeling assumptions into subsequent experimental choices [9]. As a result, contemporary AI-driven materials science must be understood as a recursive socio-technical system in which models, data, and human judgment continuously co-evolve.

Challenges of data scarcity in materials informatics

Despite the expansion of materials databases and computational tools, data scarcity remains a defining constraint in materials informatics. Unlike domains characterized by abundant observational data, materials science often relies on costly experiments, time-intensive synthesis protocols, or high-fidelity simulations, resulting in datasets that are small, sparse, and unevenly distributed across property spaces [10–12]. This scarcity poses fundamental interpretive challenges, as limited samples restrict the model's ability to capture the full diversity of material behaviors and phase regimes.

From an analytical perspective, small datasets heighten sensitivity to noise and outliers, increasing the risk that models will internalize spurious correlations as meaningful signals [13]. Such effects are particularly pronounced when extrapolating beyond the domain of observed data, where predictive confidence may be unjustifiably high despite limited empirical grounding. Conceptually, data scarcity thus necessitates a shift from accuracy-centric evaluation toward epistemic caution, emphasizing uncertainty quantification, calibration, and domain-of-applicability assessment as central components of responsible inference [14].

Trade-offs emerge prominently in model selection under data-limited conditions. Simpler algorithms or physically motivated regressors may offer greater stability and interpretability but often lack the expressive capacity to capture complex, nonlinear relationships. Conversely, deep learning models promise representational richness but are prone to overfitting and representational collapse without sufficient data or regularization [15]. Interaction dynamics between data acquisition and modeling partially mitigate these limitations through adaptive experimentation: initial model predictions can guide targeted measurements, progressively enriching datasets in regions of high uncertainty or potential novelty [16].

Beyond technical concerns, ethical considerations further intensify the significance of data scarcity. When datasets are small or biased toward specific material classes, processing routes, or geographic research contexts, AI systems may inadvertently reinforce existing inequities in research attention and technological development [17]. In applications with societal implications—such as energy materials, biomedical devices, or environmental remediation—the amplification of such biases raises questions of fairness, accountability, and distributive impact, underscoring the need for reflexive data governance alongside methodological innovation.

Data augmentation and synthetic generation techniques

In response to persistent data limitations, data augmentation and synthetic data generation have emerged as prominent strategies for expanding effective sample sizes in materials informatics. These techniques aim to enrich datasets by generating additional samples through transformations that preserve the underlying physical or chemical relevance of the original data [18, 19]. Rather than substituting for experimental measurements, augmentation operates as an interpretive extension of existing observations, enabling models to explore plausible variations within constrained representational spaces.

In materials-specific contexts, augmentation methods include geometric perturbations of crystal structures, controlled noise injection into spectroscopic or imaging data, and symmetry-aware transformations that respect crystallographic invariances [20]. Systems-level analyses suggest that such approaches improve model robustness by smoothing decision boundaries and enhancing generalization across related material classes or processing conditions [21]. Importantly, the effectiveness of augmentation depends on careful alignment with domain knowledge; indiscriminate transformations risk introducing unphysical artifacts that distort learning.

More advanced steering logics incorporate physics-informed constraints into synthetic data generation. Techniques such as physics-constrained generative models or thermodynamically guided variational frameworks ensure that augmented samples remain consistent with conservation laws, stability criteria, or known phase boundaries [22]. Feedback structures are particularly salient in generative adversarial networks, where iterative competition between generator and discriminator networks progressively aligns synthetic distributions with empirical data [23]. These recursive dynamics echo broader themes in AI-driven materials research, where learning systems continuously recalibrate representations in response to validation signals.

Analytically, integrating augmentation and generative techniques has been shown to reduce overfitting and improve predictive performance in low-data regimes, especially when combined with uncertainty-aware training and model ensembling strategies [24]. However, these gains are not purely technical; they also reshape epistemic assumptions about what constitutes legitimate data and evidence in materials science. As synthetic samples increasingly influence model behavior and experimental prioritization, critical attention must be paid to their provenance, validation, and role in scientific inference, reinforcing the need for transparent methodological reporting and interpretive restraint.

Transfer and few-shot learning approaches

Transfer learning has emerged as a foundational strategy for mitigating data scarcity in materials informatics by leveraging knowledge acquired from data-rich domains to improve performance in sparsely sampled target tasks [25, 26]. At its core, transfer learning assumes the existence of shared latent structures—such as bonding motifs, symmetry patterns, or structure–property correlations—across distinct materials systems. In practical applications, models pre-trained on large repositories are subsequently fine-tuned, or features reused, allowing foundational chemical and physical representations to inform specialized predictive objectives [27]. This process reveals interaction patterns in which generalizable representations, learned at scale, enhance task-specific inference under constrained data regimes.

Conceptual analyses of transfer learning emphasize the inherent trade-offs involved in adaptation. Excessive fine-tuning risks eroding broadly valid representations in favor of narrow task optimization, while insufficient adaptation may limit sensitivity to system-specific phenomena [28]. These tensions foreground transfer learning as a problem of epistemic balance rather than simple performance maximization. The effectiveness of transfer thus depends not only on dataset size but also on the alignment between source and target domains, including similarities in chemistry, processing conditions, and measurement protocols. Misaligned transfer can propagate biases or induce misleading generalizations, underscoring the importance of careful domain selection and validation.

Few-shot learning extends the logic of transfer by explicitly targeting scenarios in which only a handful of labeled examples are available for new materials systems or properties of interest [29, 30]. Through meta-learning paradigms, models are trained to rapidly adapt by learning how to learn across a distribution of tasks, rather than optimizing for a single predictive objective. Systems-level insights show that episodic training strategies—where models repeatedly encounter artificially constructed low-data tasks—simulate scarcity during training, fostering architectures that are resilient to data deprivation at deployment [31]. In this sense, few-shot learning reframes scarcity from a limitation into a structural condition that shapes model design.

Ethical considerations further complicate the adoption of transfer and few-shot learning approaches. Because pre-training datasets often reflect historical research priorities, well-studied material classes may disproportionately benefit from transfer-enabled performance gains [32]. Without deliberate intervention, underrepresented materials—such as those relevant to low-resource energy systems or region-specific applications—risk remaining marginalized. Equitable transfer, therefore, requires not only technical robustness but also intentional curation of source data and transparent reporting of transfer assumptions, aligning methodological choices with broader goals of inclusivity and scientific fairness.

Integration of domain knowledge and hybrid frameworks

Hybrid frameworks that integrate AI with domain knowledge and physical modeling represent a complementary response to data scarcity, embedding theoretical priors directly into learning architectures [33, 34]. Rather than relying solely on statistical regularities, these approaches interpret scarcity as a constraint that can be navigated through structured inductive biases derived from physics, chemistry, or materials theory. Constrained optimization, symbolic regression, and mechanistically informed loss functions allow models to operate effectively even when empirical data are limited by restricting hypothesis spaces to physically plausible regimes.

Feedback structures are particularly salient in physics-informed neural networks, where governing equations and conservation laws are enforced during training, shaping both predictions and internal representations [34]. These constraints not only improve numerical stability but also enhance interpretability, as model outputs can be traced back to known physical principles. Analytical implications include increased epistemic confidence, as hybrid models reduce the likelihood of unphysical extrapolations and provide clearer justifications for predictions in regions beyond observed data. Collectively, the literature reveals a spectrum of scarcity-aware learning strategies, each introducing distinct mechanisms, benefits, and epistemic risks; Table 2 consolidates these approaches and their associated trade-offs to provide a comparative synthesis.

Table 2. Scarcity-aware learning strategies in materials AI and their epistemic trade-offs

Strategy	Core mechanism	Strengths under scarcity	Limitations	Epistemic risk
Data augmentation	Synthetic expansion via transformations	Reduces overfitting	Risk of unphysical artifacts	False confidence
Transfer learning	Knowledge reuse from large datasets	Efficient bootstrapping	Domain mismatch	Bias transfer
Few-shot learning	Meta-learning from tasks	Rapid adaptation	Sensitive to priors	Instability
Physics-informed models	Embedded physical constraints	Improved extrapolation	Model rigidity	Over-constraint
Active learning	Uncertainty-driven sampling	Data efficiency	Expensive queries	Selection bias

Taken together, transfer learning, few-shot adaptation, and hybrid frameworks form an integrative toolkit for advancing materials AI under conditions of limited structural data. This synthesis reinforces a cohesive narrative across the literature: from early descriptor-based foundations to contemporary scarcity-aware strategies, progress in materials AI is driven not by data accumulation alone but by the strategic integration of learning paradigms, domain knowledge, and epistemic safeguards. Such integrative approaches position materials AI as a disciplined scientific instrument—capable of navigating uncertainty, respecting physical constraints, and supporting responsible discovery in data-constrained environments.

Proposed conceptual framework

The proposed framework conceptualizes small-data regimes in materials AI as emergent systems characterized by intricate interactions among sparse inputs, domain knowledge, and computational mechanisms. At its core lies an interpretive lens that views scarcity not as a deficiency but as a structuring force that shapes learning pathways through uncertainty modulation and knowledge amplification. Central to this is the notion of dynamic equilibria, in which limited data triggers feedback structures that iteratively refine model behavior by drawing on external priors, such as physical symmetries or analogous systems.

Analytical implications emerge from these dynamics: in small-data contexts, uncertainty propagation creates trade-offs between exploratory breadth and predictive depth, steering logics toward adaptive sampling that prioritizes informative regions. Conceptual interpretations highlight how scarcity fosters epistemic humility, encouraging models that explicitly account for informational gaps through probabilistic reasoning. Systems-level insights reveal hierarchical integrations in which low-level atomic descriptors interact with high-level property constraints, forming resilient networks that withstand data perturbations. For clarity, **Table 3** summarizes the core components of the proposed framework and their functional roles under conditions of data scarcity, providing a structured reference for the interaction dynamics discussed below.

Table 3. Core components and interaction dynamics of the proposed small-data framework

Framework component	Description	Function under scarcity
Scarcity nexus	Convergence point of sparse data	Amplifies uncertainty
Input layer	Sparse experimental/simulation data	Initial evidence-based
Knowledge integration layer	Physics priors, analogies, augmentation	Compensates for missing data
Feedback structures	Iterative refinement loops	Stabilizes learning
Steering logics	Adaptive decision rules	Balances exploration/exploitation
Ethical overlay	Bias and equity monitoring	Prevents amplification

Ethical reasoning integrates seamlessly, interpreting scarcity as an opportunity to embed fairness mechanisms, such as bias-aware augmentations that prevent amplification of underrepresented samples. Interaction dynamics further elucidate how feedback loops between simulation and sparse experiments generate emergent robustness, transforming isolated data points into coherent narratives. **Figure 1** schematically illustrates the proposed framework: a central Scarcity Nexus node linked to three concentric layers—an inner Input Layer of sparse empirical nodes with dashed uncertainty flows, a middle Knowledge Integration Layer with domain priors and augmentation pathways connected via bidirectional feedback, and an outer Output Layer of adaptive prediction nodes whose steering logics loop back to the nexus. Overlaid arcs indicate ethical overlays shading bias-prone regions, and the spiral layout with a red-to-blue gradient conveys progressive evolution from raw scarcity to stabilized, informed intelligence.

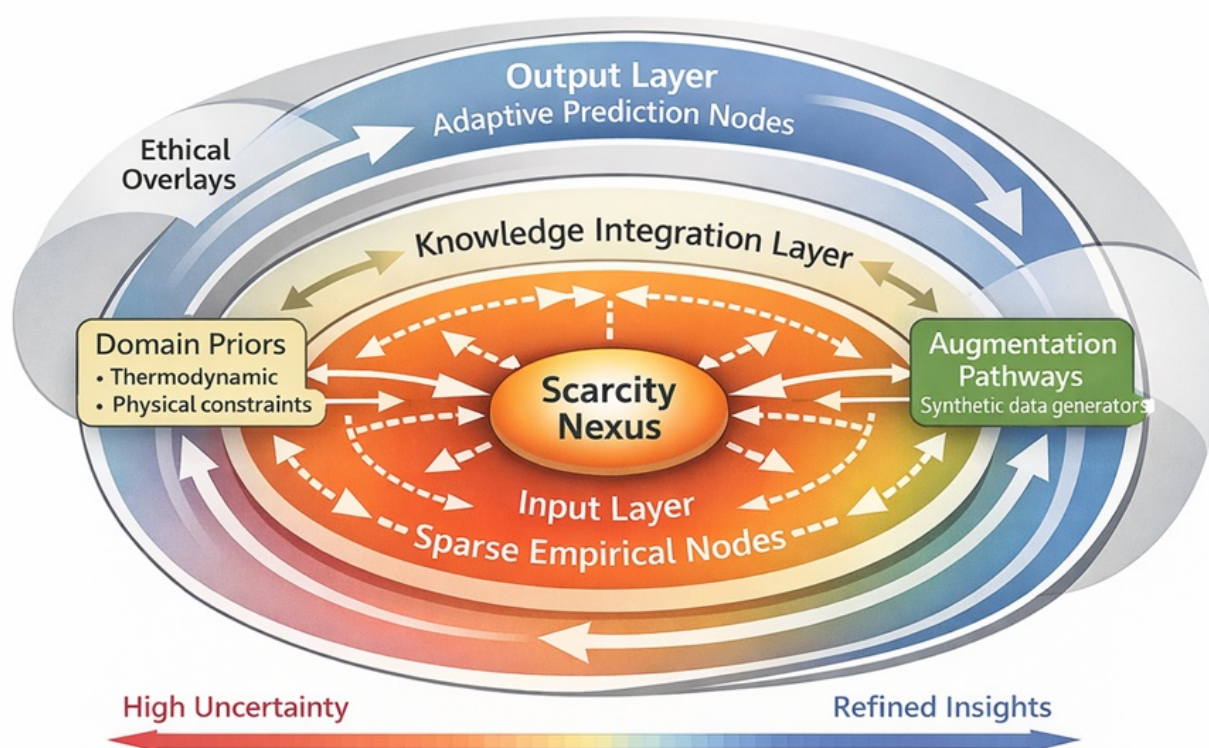


Figure 1. Conceptual framework for learning under data scarcity in materials artificial intelligence, centered on the scarcity nexus.

This theory provides a blueprint for navigating small-data landscapes, fostering innovations that align AI with the inherent constraints of materials science.

Analytical implications

The framework's interpretive dynamics offer analytical implications for how small-data regimes shape the deployment of AI in materials science, particularly through mechanisms that modulate uncertainty and facilitate knowledge transfer. In low-data environments, uncertainty amplification emerges as a key interaction pattern, where sparse inputs lead to broader confidence intervals in predictions, prompting steering logics that prioritize conservative extrapolations over aggressive generalizations [1, 2]. This dynamic captures the trade-off between model complexity and data efficiency, revealing that simpler architectures, when augmented with physical constraints, can yield more robust outcomes by mitigating variance introduced by limited sample sizes [3]. Systems-level insights suggest that such regimes foster hybrid workflows, integrating sparse empirical data with theoretical priors to create feedback structures that iteratively validate inferences, thereby enhancing epistemic reliability [4].

Conceptual interpretations extend to the role of augmentation in reshaping data landscapes. By generating synthetic variants aligned with domain principles, augmentation acts as a bridge, transforming scarcity into a generative process that enriches representational diversity [5, 6]. Analytical implications here include the potential for reduced computational overhead, as targeted augmentations focus on underrepresented regions, optimizing resource allocation in materials discovery pipelines [7]. Ethical reasoning underscores the importance of transparency in these processes, where augmented datasets risk inheriting biases from originals, necessitating interpretive checks to ensure equitable model behaviors across material classes [8]. Interaction dynamics between augmentation and transfer learning further illuminate

trade-offs: while transfer from large-scale models provides initial scaffolding, fine-tuning on small data requires careful calibration to avoid catastrophic forgetting, steering toward modular designs that preserve foundational knowledge [9, 10].

Epistemic considerations highlight how small-data AI encourages a shift from accuracy-centric metrics to metrics that emphasize calibration and coverage. In materials contexts, where experimental validation is costly, well-calibrated uncertainties guide decision-making, interpreting model outputs as probabilistic maps rather than point estimates [11, 12]. Systems-level insights reveal feedback loops in active learning, where uncertainty-driven queries to oracles (e.g., simulations) progressively densify data spaces, transforming initial scarcity into informed abundance [13]. This interpretive lens offers analytical advantages for adaptive experimentation, potentially accelerating discoveries in niche areas such as rare-earth alternatives or high-entropy alloys [14].

Moreover, the framework interprets scarcity as an impetus for multi-fidelity integrations, where low-resolution data from rapid screenings complements high-fidelity but sparse measurements [15]. Trade-offs arise in fidelity balancing, with analytical implications for error propagation: mismatches can distort predictions, but aligned hierarchies enhance overall system resilience [16]. Conceptual interpretations of these dynamics emphasize ethical imperatives, such as ensuring that multi-fidelity approaches do not exacerbate access disparities in resource-constrained research settings [17]. Interaction patterns with few-shot paradigms further suggest that meta-learning fosters rapid adaptability, implying analytical benefits in dynamic environments where material requirements evolve, like adaptive coatings or responsive sensors [18, 19].

Steering logics in small-data regimes also implicate broader ecosystem considerations, including interoperability with existing databases. By leveraging ontologies for semantic alignment, models can interpret heterogeneous small datasets cohesively, revealing hidden correlations that amplify inferential power [20]. Analytical implications include improved cross-domain applicability, where insights from one material system inform another, reducing the effective data burden [21]. Feedback structures here involve continual learning, where models update incrementally as new sparse data arrives, maintaining relevance amid evolving scientific knowledge [22]. Epistemic reasoning cautions against over-reliance on such automations and advocates human oversight to interpret contextual nuances [23].

Ultimately, these analytical implications reframe small-data challenges as opportunities for innovative steering, where interaction dynamics and trade-offs drive toward more sustainable AI practices in materials science. By prioritizing interpretive depth over data volume, the framework guides toward systems that are not only efficient but also ethically attuned, fostering integrative advancements that align with the field's intrinsic constraints [24, 25].

Results and Discussion

The conceptual framework advanced here interprets small-data regimes as pivotal to the maturation of materials AI, where scarcity compels a reevaluation of traditional data-centric paradigms. Interaction dynamics between limited inputs and knowledge integration reveal that effective learning under constraints hinges on synergistic blends of empirical and theoretical elements, challenging the notion that volume alone dictates performance [26, 27]. Systems-level insights suggest that this shift promotes resilient architectures capable of navigating the high dimensionality inherent to materials property spaces with minimal informational overhead [28]. Ethical reasoning is integrated into this discussion, examining how scarcity-aware designs can mitigate the risk of overfitting to biased samples, thereby supporting more inclusive approaches to global challenges such as sustainable energy [29].

Trade-offs in model deployment under scarcity further illuminate conceptual interpretations. While few-shot adaptations enable quick pivots, they demand robust priors to avoid instability, steering logics toward hybrid models that embed physical invariances [30, 31]. Analytical implications from the literature synthesis align with this, showing that augmentation and transfer not only alleviate data gaps but also enhance interpretability, allowing stakeholders to trace decision pathways in opaque AI systems [32]. Feedback structures, such as those in Bayesian frameworks, exemplify

how iterative processes transform static scarcity into dynamic knowledge accrual, offering interpretive advantages in longitudinal studies of material degradation or synthesis optimization [33].

Broader epistemic considerations discuss the framework's role in fostering interdisciplinary dialogues, where materials scientists and AI practitioners co-evolve methodologies attuned to domain-specific scarcities [34]. This integrative perspective implies that small-data theories can bridge gaps between computational efficiency and scientific rigor, promoting ethical deployments that prioritize verifiability over novelty [34]. However, interaction dynamics also highlight potential pitfalls, such as amplified uncertainties in extrapolation-heavy tasks, necessitating conceptual safeguards, such as ensemble methods, to bolster confidence [1, 2].

In synthesizing these elements, the discussion underscores the framework's originality in viewing scarcity as a generative force, rather than a barrier, encouraging steering logics that harness constraints for creative problem-solving in materials AI [3, 4].

Conclusion

This conceptual theory of small-data regimes in materials AI interprets scarcity as a foundational dynamic that shapes interaction patterns, trade-offs, and feedback structures, thereby redefining learning paradigms. By emphasizing knowledge integration and uncertainty management, the framework offers systems-level insights into resilient workflows, where limited data catalyzes the development of innovative steering logics. Analytical implications extend to ethical and epistemic realms, advocating for transparent, equitable AI that thrives amid constraints. Ultimately, this perspective fosters integrative advancements, positioning materials science to leverage scarcity for sustainable, impactful discoveries.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 16 May 2024

Revised: 01 Jul 2024

Accepted: 30 Jul 2024

Published online: 18 January 2025

Rights and permissions

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to

the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

Choudhary K, DeCost B, Chen C, Jain A, Tavazza F, Cohn R, et al. Recent advances and applications of deep learning methods in materials science. *npj Comput Mater.* 2022;8(1):59.

Gupta V, Choudhary K, Tavazza F, Campbell C, Liao WK, Choudhary A, et al. Cross-property deep transfer learning framework for enhanced predictive analytics on small materials data. *Nat Commun.* 2021;12(1):6595.

Li K, Persaud D, Choudhary K, DeCost B, Greenwood M. Exploiting redundancy in large materials datasets for efficient machine learning with less data. *Nat Commun.* 2023;14(1):7283.

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.

Achar SK, Bernasconi L, Johnson JK. Machine learning electron density prediction using weighted smooth overlap of atomic positions. *Nanomaterials.* 2023;13(12):1853.

Achar SK, Schneider J, Stewart DA. Using machine learning potentials to explore interdiffusion at metal-chalcogenide interfaces. *ACS Appl Mater Interfaces.* 2022;14(51):56963-74.

Achar SK, Keith JA. Small data machine learning approaches in molecular and materials science. *Chem Rev.* 2024;124(24):13571-3.

Achar SK, Bernasconi L, DeMaio RI, Howard KR, Johnson JK. In silico demonstration of fast anhydrous proton conduction on graphanol. *ACS Appl Mater Interfaces.* 2023;15(21):25873-83.

Cai R, Han T, Liao W, Huang J, Li D, Kumar A, et al. Prediction of surface chloride concentration of marine concrete using ensemble machine learning. *Cem Concr Res.* 2020;136:106164.

Han T, Siddique A, Khayat K, Huang J, Kumar A. An ensemble machine learning approach for prediction and optimization of modulus of elasticity of recycled aggregate concrete. *Constr Build Mater.* 2020;244:118271.

Gomaa E, Han T, ElGawady M, Huang J, Kumar A. Machine learning to predict properties of fresh and hardened alkali-activated concrete. *Cem Concr Compos.* 2021;115:103863.

Han T, Stone-Weiss N, Huang J, Goel A, Kumar A. Machine learning as a tool to design glasses with controlled dissolution for healthcare applications. *Acta Biomater.* 2020;107:286-98.

Kusne AG, Yu H, Wu C, et al. On-the-fly closed-loop materials discovery via Bayesian active learning. *Nat Commun.* 2020;11(1):5966.

Guo K, Yang Z, Yu CH, Buehler MJ. Artificial intelligence and machine learning in design of mechanical materials. *Mater Horiz.* 2021;8(4):1153-72.

Yang Z, Yu CH, Buehler MJ. Deep learning model to predict complex stress and strain fields in hierarchical composites. *Sci Adv.* 2021;7(15):eabd7416.

Yang Z, Buehler MJ. Linking atomic structural defects to mesoscale properties in crystalline solids using graph neural networks. *npj Comput Mater.* 2022;8(1):198.

Yang Z, Buehler MJ. High-throughput generation of 3D graphene metamaterials and property quantification using machine learning. *Small Methods.* 2022;6(9):2200537.

Sun Y, Wang G, Li K, Peng L, Zhou J, Sun Z. Accelerating the discovery of transition metal borides by machine learning on small data sets. *ACS Appl Mater Interfaces.* 2023;15(3):3772-82.

De Breuck PP, Evans ML, Rignanese GM. Robust model benchmarking and bias-imbalance in data-driven materials science: a case study on MODNet. *npj Comput Mater.* 2021;7(1):83.

Li S, Nakata A. CSIML: a cost-sensitive and iterative machine-learning method for small and imbalanced materials data sets. *Chem Lett.* 2024;53(5):upae090.

Karpovich C, Pan E, Jensen Z, Olivetti E. Interpretable machine learning enabled inorganic reaction classification and synthesis condition prediction. *Chem Mater.* 2023;35(2):734-45.

Tian SIP, Walsh A, Ren Z, Li Q, Buonassisi T. What information is necessary and sufficient to predict materials properties using machine learning? *ACS Appl Mater Interfaces.* 2022;14(45):50985-95.

Schmidt KJ, Scourtas A, Ward L, Wangen S, Schwarting M, Isaac Darling I, et al. Foundry-ML: Software and services to simplify access to machine learning datasets in materials science. *J Open Source Softw.* 2023;8(82):4993.

Sendek AD, Ransom B, Cubuk ED, Pellouchoud LA, Nanda J, Reed EJ. Machine learning modeling for accelerated battery materials design in the small data regime. *Adv Energy Mater.* 2020;10(43):2002273.

Zhou H. On the feasibility of small-data learning in simulation-driven engineering tasks with known mechanisms and effective data representations. *Comput Mater Sci.* 2021;197:110629.

Igarashi Y. Materials informatics for 2D materials combined with sparse modeling and chemical perspective: Toward small-data-driven chemistry and materials science. *J Phys Chem Lett.* 2022;13(12):2764-71.

Qayyum H, Saqib K, Hussain G, Alkahtani M. Predicting flexural properties of 3D-printed composites: A small dataset analysis using multiple machine learning models. *Addit Manuf.* 2024;79:103922.

Chen X, Lu S, Wan X, Chen Q, Zhou Q, Jiang J. Accurate property prediction with interpretable machine learning model for small datasets via transformed atom vector. *Comput Mater Sci.* 2023;218:111949.

Jacobs R, Schultz LE, Scourtas A, Schmidt KJ, Price-Skelly O, Engler W, et al. Machine learning materials properties with accurate predictions, uncertainty estimates, domain guidance, and persistent online accessibility. *npj Comput Mater.* 2022;8(1):247.

Paul A, Acar P, Liao W, Choudhary A, Choudhary A, Sundararaghavan V, et al. Microstructure optimization with constrained design objectives using machine learning-based feedback-aware data-generation. *Comput Mater Sci.* 2020;179:109649.

Wen M, Blau SM, Xie X, Dwaraknath S, Persson KA. Improving machine learning performance on small chemical reaction data with unsupervised contrastive pre-training. *J Phys Chem A.* 2022;126(4):646-56.

Omeel SS. Scalable deep learning framework for materials discovery: MaterialsAtlas.org. *Machine Learning: Science and Technology.* 2023;4(1):015001.

Lambard G. Machine learning in materials science. 2020.

Barnett JW, Bilchak CR, Wang Y, Benicewicz BC, Murdock LA, Bereau T, et al. Designing exceptional gas-separation polymer membranes using machine learning. *Science advances*. 2020;6(20):eaaz4301.