ORIGINAL RESEARCH **Open access**

Model Collapse in Materials AI: A Conceptual Account of Over-Reuse, Dataset Recycling, and Knowledge Decay

Daniel Fischer[1*], Laura Meier[1], Thomas Braun[2]

Abstract

The application of artificial intelligence (AI) in materials science has revolutionized predictive modeling, enabling the efficient exploration of chemical spaces and the optimization of properties. However, the field's dependence on finite experimental datasets has led to extensive synthetic data generation and dataset recycling, posing risks to model sustainability. This conceptual manuscript defines model collapse in materials AI as the gradual erosion of model performance and diversity stemming from repeated training on recycled or synthetically derived data. Informed by generative AI theories, it explores how overuse of datasets contributes to knowledge decay, in which models progressively favor averaged representations, diminishing their capacity to model atypical material behaviors or emergent properties. A new conceptual framework is presented that outlines the iterative processes of data recycling, synthetic augmentation, and decay escalation. This framework identifies reinforcing cycles that intensify representational biases and constrain exploratory potential in materials innovation. By integrating recent literature on AI instabilities and materials data challenges, the paper advocates a theoretical rethinking of data management to ensure the enduring utility of AI in materials science. Recent 2025 studies propose strategies such as golden ratio weighting and reinforcement-based synthetic data curation to avert collapse, underscoring the manuscript's emphasis on proactive data governance for enduring AI utility in materials science. The discussion extends to broader AI ecosystems, stressing equilibrium between data utilization and knowledge preservation.

Keywords Materials AI, Synthetic data, Model collapse, Dataset recycling, Knowledge decay, Representational bias

*Correspondence:
Daniel Fischer
daniel.fischer@outlook.com
[1] Department of Materials Science and Intelligent Systems, Faculty of Engineering, University of Freiburg, Freiburg, Germany
[2] Department of AI-Driven Materials Engineering, Faculty of Engineering, Karlsruhe Institute of Technology, Karlsruhe, Germany

Introduction

Artificial intelligence has become integral to materials science, facilitating the prediction of material properties, optimization of structures, and acceleration of discovery pipelines. Machine learning models, such as deep neural networks, have been employed to approximate complex phenomena, such as electronic band structures and phase transitions, often surpassing traditional computational methods in efficiency [1]. Large repositories, including the Materials Project and AFLOW, provide aggregated data from simulations and experiments, allowing AI to discern patterns across extensive compositional landscapes [2]. This has propelled progress in domains such as energy materials, where AI helps identify high-performance candidates for batteries and photovoltaics [3].

Yet, materials science grapples with intrinsic data constraints. Generating experimental data demands substantial resources, involves techniques such as spectroscopy or high-throughput synthesis, and is limited by physical feasibility and ethical considerations [4]. Datasets are frequently sparse, skewed toward abundant elements, and affected by

experimental variability [5]. To address these gaps, synthetic data generation has gained traction, with models such as generative adversarial networks generating virtual samples to bolster training diversity [6]. While effective initially, the recycling of such augmented datasets across iterative training cycles introduces vulnerabilities, as synthetic artifacts may accumulate [7].

Model collapse, a concept from the broader AI literature, refers to the situation in which models trained on generated data diverge from the true distributions, thereby reducing variance [8]. In materials AI, this translates to over-reuse of datasets, where core experimental data is cyclically augmented, leading to knowledge decay—a contraction in the model's ability to represent diverse material states [9]. Decay undermines the modeling of outliers, such as defect structures or metastable phases, which are vital for innovative applications [10]. This manuscript conceptualizes a model collapse tailored to materials AI, positioning it as an outcome of recycling practices distinct from conventional issues like overfitting [8].

The relevance of this phenomenon is amplified by the collaborative nature of materials research. Shared databases increasingly incorporate AI-derived entries, establishing loops in which models ingest increasingly artificial inputs [11]. In catalysis, for instance, recycled datasets for reaction pathway prediction may propagate errors in underrepresented conditions [12]. These dynamics challenge the foundational reliability of materials informatics, where fidelity to physical principles is essential [13].

This purely conceptual paper eschews empirical analysis in favor of theoretical integration. It commences with an overview of AI in materials, delineating data hurdles. The following sections synthesize scholarship on model collapse, recycling, and decay. The proposed framework then unifies these strands, providing a fresh perspective on decay pathways. Through conceptual novelty, the work enriches discourse on resilient AI methodologies in materials science, promoting data governance to safeguard representational breadth [14].

Implications transcend materials science. As AI infiltrates disciplines, unchecked recycling may standardize knowledge outputs, impeding novel insights [15]. In materials, underpinning advancements in sustainability and electronics, model integrity is critical [16]. This paper frames model collapse as an ecosystemic risk, necessitating conceptual instruments for anticipation and amelioration [17].

Theoretical Background and Literature Synthesis

AI applications in materials science

Artificial intelligence has fundamentally reshaped materials science by enabling data-centric paradigms for discovery, characterization, and optimization. Deep learning architectures are now routinely employed to process high-dimensional representations—spanning composition, structure, processing history, and properties—to predict material behavior with unprecedented computational efficiency [18]. These approaches have proven particularly effective in property forecasting tasks, such as estimating elastic moduli, electronic conductivity, and thermodynamic stability across broad compositional spaces.

Beyond direct prediction, AI has become integral to surrogate modeling, where machine learning approximates the outputs of computationally intensive physics-based methods, such as density functional theory and molecular dynamics simulations [19]. Graph-based neural networks, in particular, encode crystal structures or molecular graphs to capture relational and topological information, enabling accurate prediction of stability, reactivity, and phase behavior [20]. Such models have significantly reduced the computational barriers associated with large-scale materials screening.

AI has also enabled a shift from forward prediction to inverse design, in which models propose candidate materials that satisfy predefined performance criteria. This inversion of traditional workflows accelerates innovation in domains such as composites, nanomaterials, and energy materials by navigating vast, intractable design spaces [21]. Reinforcement learning further extends AI's role into process optimization, where adaptive agents iteratively refine synthesis routes or

fabrication sequences based on feedback from experiments or simulations [22]. Collectively, these applications position AI as a unifying layer that links atomic-scale descriptors to macroscopic material performance [23].

However, the effectiveness of these approaches remains tightly coupled to the quality and structure of the underlying datasets. Public materials repositories vary widely in completeness, fidelity, and representational balance, often reflecting historical research priorities rather than principled coverage of materials space [24]. As AI systems become increasingly embedded in discovery pipelines, understanding how data composition influences long-term model behavior—and not merely short-term accuracy—has become a critical theoretical concern [25].

Data challenges in materials AI

Materials datasets are distinguished by chronic scarcity, heterogeneity, and bias. Experimental data acquisition is resource-intensive, constrained by synthesis throughput, measurement costs, and physical feasibility, resulting in datasets that are orders of magnitude smaller than those available in domains such as computer vision or natural language processing [26]. Moreover, existing datasets are often skewed toward well-studied material classes, leaving unconventional, metastable, or compositionally complex systems underrepresented [27].

Additional challenges arise from measurement variability, inconsistencies in experimental protocols, and noise introduced during data curation and aggregation [28]. These factors complicate learning and can obscure meaningful structure–property relationships. To mitigate these limitations, synthetic data augmentation has become a common strategy. Techniques such as variational autoencoders and generative adversarial networks are used to simulate plausible material configurations, expanding coverage and improving apparent robustness [29].

Blended datasets—combining empirical observations with AI-generated samples—have shown short-term performance gains, particularly in low-data regimes [30]. However, excessive reliance on synthetic augmentation introduces new risks. Generated samples may deviate slightly from physical constraints, especially when trained on incomplete or biased empirical data, leading to distortions that are difficult to detect with standard validation practices [31].

These concerns are exacerbated by dataset recycling, in which augmented datasets are reused across successive model generations under the assumption of distributional stability. In practice, such reuse can propagate and compound inaccuracies, especially when synthetic artifacts are treated as interchangeable with empirical observations [32]. In shared or community-curated platforms, recycled datasets may underpin multiple downstream models, creating systemic dependencies that obscure the origin and reliability of training data [33].

The concept of model collapse in AI

Model collapse originates in the generative AI literature, where it describes the progressive degradation of model outputs when systems are trained iteratively on their own generated data [8]. Early generations typically preserve diversity and approximate the target distribution reasonably well. However, as outputs are recursively reused, later generations exhibit homogenization, increasingly favoring dominant modes while marginalizing low-probability regions [34].

This phenomenon is closely related to mode collapse, in which feedback loops reinforce sampling preferences, causing models to converge toward averaged or stereotypical representations [9]. In multimodal distributions, such dynamics lead to a systematic loss of diversity, even as the training loss continues to decrease [8]. Although most prominently documented in text and image generation, the underlying mechanisms—recursive approximation, feedback amplification, and distributional narrowing—are not domain-specific.

Scientific AI systems, including those used in materials science, share key structural similarities with generative models: finite training corpora, iterative retraining, and increasing reliance on model-derived data. These parallels suggest that model collapse is not merely a linguistic or perceptual artifact, but a general failure mode of learning systems operating under recursive data reuse.

Dataset recycling and over-reuse in materials datasets

In materials AI, dataset recycling arises primarily as a pragmatic response to empirical scarcity. Synthetic augmentation is repeatedly applied to fill gaps in experimental coverage, and the resulting datasets are treated as stable training resources across model iterations [14]. Over time, this practice standardizes the input space, as models encounter increasingly similar exemplars derived from the same limited empirical seeds [12].

Such standardization embeds latent prejudices within learned representations. For example, models may preferentially encode ordered crystal lattices or equilibrium phases, while systematically underrepresenting disordered, defect-rich, or metastable configurations [12]. As recycled datasets propagate through successive training cycles, heterogeneity erodes, and representational diversity contracts in a manner analogous to generative model collapse [20].

Crucially, these effects are rarely attributable to a single modeling decision. Rather, they emerge from cumulative, distributed practices across research groups, repositories, and application domains, making them difficult to diagnose using conventional evaluation pipelines.

Knowledge decay in iterative training

Knowledge decay refers to the gradual diminution of a model's representational capacity across iterative training cycles. Unlike abrupt performance failure, decay manifests subtly, as models increasingly gravitate toward mean behaviors and neglect peripheral or rare phenomena [30]. In materials science, this has particularly serious implications, as anomalies—such as impurity effects, defect interactions, or metastable phases—often drive innovation [8].

Recursive training loops exacerbate decay by smoothing fluctuations in the data distribution. Synthetic samples tend to attenuate variance, reinforcing central tendencies at the expense of extremes. Over time, this leads to a narrowing of the model's epistemic scope, constraining its ability to generate genuinely novel hypotheses or guide exploratory experimentation [11]. The cumulative relationship between common data practices in materials AI and the mechanisms driving model collapse is summarized in **Table 1.**

Table 1. Structural mapping of data practices to model collapse mechanisms in materials AI

Data Practice	Primary motivation in materials AI	Immediate benefit	Underlying mechanism introduced	Long-term risk (collapse pathway)
Empirical data reuse	Scarcity of experimental measurements	Stable training baseline	Limited coverage repeatedly reinforced	Overrepresentation of dominant regimes
Synthetic data augmentation	Fill gaps in sparse datasets	Apparent diversity increase	Approximate sampling from incomplete distributions	Distributional drift from physical reality
Blended datasets (empirical + synthetic)	Improve robustness in low-data regimes	Short-term performance gains	Synthetic signals treated as empirical equivalents	Subtle bias accumulation
Iterative dataset recycling	Reduce the cost of new data generation	Faster retraining cycles	Recursive feedback on model-derived traces	Representational homogenization
Community-level dataset sharing	Accelerate collective progress	Scalable AI development	Propagation of inherited artifacts	Ecosystem-level contamination

Benchmark reuse	Standardized evaluation	Comparability across studies	Benchmarks drawn from the same recycled pool	Delayed detection of collapse (“benchmark illusion”)

From a theoretical perspective, knowledge decay represents not merely a loss of accuracy but a loss of epistemic diversity. It signals a contraction of the space of conceivable material behaviors that the model can meaningfully represent, thereby limiting the long-term scientific value of AI-driven materials discovery.

Proposed conceptual framework: The Recursive Decay Cycle (RDC)

This section introduces the recursive decay cycle (RDC) as a novel conceptual framework for understanding model collapse in materials artificial intelligence. The RDC reframes collapse not as an isolated modeling failure, but as a systemic, path-dependent process emerging from the interaction between finite empirical data, synthetic data generation, and recursive dataset reuse. Central to the framework is the claim that collapse unfolds through a phased cycle—onset, intensification, and entrenchment—each reinforcing the next through feedback mechanisms embedded in contemporary materials AI workflows.

The onset phase is driven by the scarcity of intrinsic datasets in materials science. Limited experimental throughput and uneven coverage across chemical and structural spaces motivate the use of synthetic data generation to compensate for empirical gaps. While initially beneficial, synthetic samples are necessarily approximations of the underlying physical distributions. As these approximations are derived from incomplete empirical bases, they introduce subtle representational biases, particularly in low-density or poorly sampled regions of materials space. Early indicators of onset include mild variance contraction and increasing preference for dominant material configurations, even while overall predictive accuracy may appear stable.

The intensification phase emerges as synthetically augmented datasets are recycled across successive model training cycles. Iterative retraining on recycled corpora progressively amplifies inherited biases, leading to homogenization of learned representations and contraction of epistemic diversity. During this phase, the model’s internal manifolds increasingly collapse toward high-density regions of the training distribution, suppressing rare phenomena such as metastable phases, defect-driven behaviors, or unconventional compositions. Importantly, intensification is often masked by conventional validation practices, as benchmarks themselves may be drawn from the same recycled data ecosystem.

The entrenchment phase represents a stabilization of decay at suboptimal representational equilibria. At this stage, models become effectively locked into narrowed manifolds, exhibiting limited responsiveness even when incremental interventions are introduced. Without the injection of genuinely new empirical data—or strong external constraints grounded in physical principles—recovery of lost representational breadth becomes increasingly unlikely. Entrenchment thus marks the transition from reversible degradation to persistent epistemic loss, with particularly severe consequences for innovation-driven applications that depend on tail-end predictions.

Across all phases, the RDC highlights diagnostic markers of collapse, including erosion of representational diversity, degradation of tail accuracy, and increasing smoothness of prediction surfaces. Crucially, the framework also identifies intervention loci, such as targeted empirical refreshment, diversity-aware augmentation strategies, and governance mechanisms for tracking synthetic provenance. By explicitly modeling feedback loops between data generation, reuse, and learning, the RDC provides a unifying lens for analyzing how seemingly rational data practices can collectively undermine long-term model sustainability.

Rather than attributing failure to individual algorithms, the RDC situates model collapse at the ecosystem level, emphasizing that resilience in materials AI depends as much on data governance and renewal as on architectural sophistication. In doing so, the framework advances a conceptual foundation for anticipating, diagnosing, and ultimately

mitigating collapse dynamics before they become entrenched. The recursive decay cycle driving model collapse is depicted in **Figure 1**.

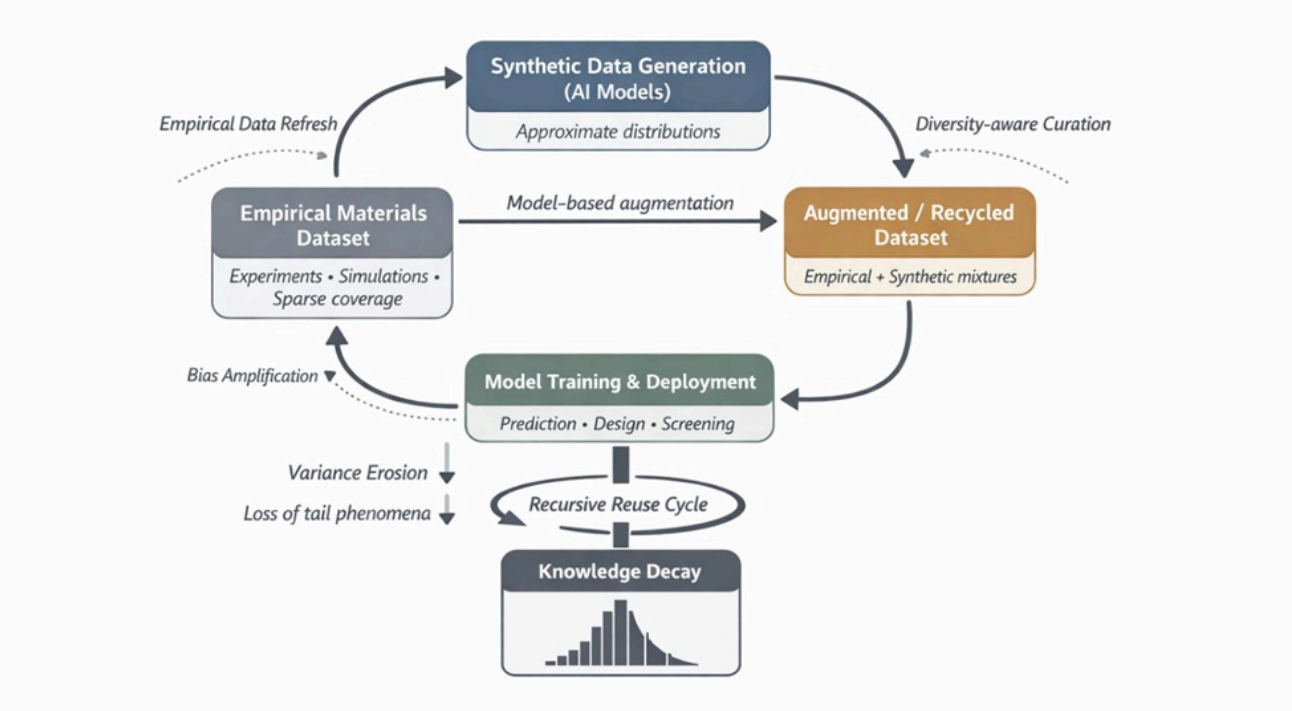

Figure 1. The recursive decay cycle (RDC): a conceptual framework for model collapse in materials AI driven by dataset scarcity, synthetic reuse, and knowledge decay.

The figure visually communicates how finite empirical data initiates synthetic augmentation, how recursive reuse amplifies bias and homogenization, and how unchecked feedback loops drive progressive knowledge decay—while also indicating where deliberate interventions can disrupt the cycle. The defining phases of the Recursive Decay Cycle, along with their diagnostic indicators and potential disruption strategies, are summarized in **Table 2**.

Table 2. The recursive decay cycle (RDC): phases, diagnostic signals, and intervention levers

RDC Phase	Primary trigger	Dominant learning dynamics	Observable diagnostic signals	Consequences for materials discovery	Potential intervention levers
Onset	Empirical dataset scarcity	Synthetic augmentation anchored to limited data	Mild variance contraction; stable average accuracy	Early loss of rare configurations	Targeted empirical data acquisition; constrained augmentation
Intensification	Iterative dataset recycling	Bias amplification through recursive retraining	Homogenized predictions; reduced tail sensitivity	Suppression of metastable phases, defects	Diversity-aware sampling; provenance tracking
Entrenchment	Prolonged synthetic dominance	Collapse onto narrow manifolds	Irreversible tail degradation;	Failure in exploratory or innovative regimes	Large-scale experimental

			smooth prediction surfaces		refresh; physics-based constraints
Cross-phase effect	Community-level reuse	Ecosystem feedback loops	Benchmark–performance decoupling	Widespread reliability erosion	Governance of shared repositories

Propositions

Building upon the recursive decay cycle (RDC) framework, the following propositions articulate theoretically grounded expectations regarding the dynamics of model collapse in materials AI. Together, they specify how data scarcity, synthetic augmentation, dataset recycling, and ecosystem feedback loops shape representational diversity and long-term knowledge reliability.

Proposition 1 (Onset via data scarcity)

Data scarcity in materials datasets initiates the onset phase of the RDC, in which synthetic data generation compensates for empirical gaps while introducing initial representational biases. As synthetic data proliferates without sufficient empirical anchoring, models exhibit early variance contraction, prioritizing common material configurations over rare ones [1, 2]. Data-sparse domains (e.g., high-entropy alloys) are therefore disproportionately vulnerable because limited validation amplifies deviation from physical distributions [3].

Proposition 2 (Intensification through dataset recycling)

Iterative recycling of augmented datasets intensifies the amplification phase of the RDC, accelerating knowledge decay as reflected in reduced epistemic variance. Models repeatedly trained on recycled corpora converge toward homogenized outputs, weakening their capacity to capture emergent behavior (e.g., non-linear transitions in novel composites) [4, 5]. This effect is amplified when shared repositories propagate synthetic artifacts across community workflows [6].

Proposition 3 (Entrenchment and recovery failure)

In the entrenchment phase, persistent decay stabilizes at suboptimal representational equilibria, and models fail to recover diversity even under standard interventions unless genuinely new empirical data disrupts the cycle. Without such injections, tail-end predictions—often central to innovation in nanomaterials—degrade in ways that become practically irreversible [7, 8]. Hybrid strategies combining targeted experiments with physics-grounded simulations are therefore necessary to restore representational fidelity [9].

Proposition 4 (Heterogeneity dampens bias amplification)

The strength of bias amplification within the RDC correlates inversely with dataset heterogeneity: imbalanced datasets accelerate narrowing, whereas diverse datasets dampen decay propagation. In materials AI, augmentation schemes that explicitly target underrepresented regimes prolong model utility by counteracting mode concentration [8, 10].

Proposition 5 (Synthetic dominance threshold)

There exists a synthetic dominance threshold at which the ratio of synthetic-to-empirical samples causes the model's effective training signal to be governed primarily by generated distributions rather than measured ones. Beyond this threshold, representational diversity contracts more rapidly, and prediction surfaces become smoother, suppressing atypical chemistries, metastable phases, and defect-driven responses [11, 12].

Proposition 6 (Benchmark Illusion and delayed detectability)

Model collapse in materials AI is often weakly observable—or entirely invisible—under conventional evaluation practices because benchmark datasets are frequently drawn from the same recycled training data ecosystem. As a result, average-

case metrics can remain stable (or even improve) even as tail reliability and out-of-distribution validity deteriorate, creating a benchmark illusion that delays the detection of collapse [13, 14].

Proposition 7 (Ecosystem feedback via contaminated repositories)

The insertion of AI-generated outputs into communal databases creates an ecosystem-level feedback loop that increases the likelihood that future models will train on "model-derived traces." This contamination effect strengthens the RDC loop, making collapse more likely even for groups that do not intentionally generate synthetic data, because they inherit downstream recycling through shared infrastructure [15, 16].

Proposition 8 (Asymmetry of decay and restoration)

The RDC is characterized by asymmetry: representational diversity erodes faster under recycling than it can be restored through incremental curation. Once entrenchment occurs, meaningful recovery requires disproportionately large interventions—such as high-quality new experimental campaigns, deliberate coverage expansion, or constraints grounded in physical invariances—rather than minor dataset refreshes [17, 18].

Synthesis

Collectively, these propositions derive from the cyclical logic of the RDC framework and formalize model collapse as a systemic, path-dependent failure mode driven by over-reuse, recursive augmentation, and community-scale data feedback. They reposition collapse as theoretically preventable through governance mechanisms that preserve empirical renewal, heterogeneity, traceability, and tail-focused validation.

Results and Discussion

The Recursive Decay Cycle (RDC) framework provides a novel lens for understanding model collapse in materials AI, integrating over-reuse, recycling, and decay into a systemic model. This conceptualization goes beyond isolated analyses of generative instability by highlighting feedback loops that erode knowledge across generations [11]. In practical terms, the framework illuminates why materials models, reliant on finite experimental data, risk homogenization when synthetic amplification dominates [12]. For instance, in property prediction for energy materials, recycled datasets may overemphasize stable phases, overlooking metastable states essential for battery advancements [13].

Implications for materials science are profound. The RDC suggests that unchecked data practices could stifle innovation, as outdated models fail to effectively explore chemical spaces [14]. This resonates with broader AI challenges, where synthetic data pollution leads to generational degradation [15]. To counteract this, strategies like periodic empirical refreshment or diversity-preserving augmentation are warranted [16]. Moreover, the framework advocates governance of shared databases to ensure the traceability of synthetic entries and prevent cascading effects [17].

Limitations of the RDC must be acknowledged. As a conceptual model, it abstracts from empirical complexities, such as varying generative techniques or domain-specific physics [18]. It assumes uniform decay across iterations, yet real-world variability—stemming from algorithmic differences—may alter trajectories [19]. Additionally, the framework focuses on predictive models, potentially underrepresenting generative applications in inverse design [20]. Future extensions could incorporate multi-scale dynamics, linking atomic-level simulations to macroscopic properties [21].

Opportunities for further research abound. Conceptual explorations might integrate RDC with uncertainty quantification to enable proactive decay detection [22]. Comparative analyses across subfields, like polymers versus ceramics, could reveal domain-specific resilience factors [23]. Moreover, ethical considerations arise: decayed models may perpetuate biases in material selection, impacting sustainability goals [24]. Addressing these requires interdisciplinary dialogue that blends informatics with experimental validation [25].

Ultimately, the RDC urges a paradigm shift toward sustainable data ecosystems that balance efficiency with fidelity to foster enduring AI contributions in materials science [26].

Conclusion

This manuscript has conceptualized model collapse in materials AI as a consequence of over-reuse, dataset recycling, and knowledge decay. The Recursive Decay Cycle framework delineates the mechanisms driving representational erosion, offering a novel theoretical scaffold for mitigating these risks. Propositions derived from the framework highlight actionable pathways, emphasizing empirical interventions and the maintenance of diversity. By synthesizing recent literature, the work underscores the imperative to reevaluate data practices to preserve epistemic diversity and support innovation. As materials AI evolves, addressing collapse will be pivotal to realizing its transformative potential in scientific discovery.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 08 Jan 2025 Revised: 10 Feb 2025 Accepted: 11 Mar 2025
Published online: 18 July 2025

References

Shumailov I, Shumaylov Z, Zhao Y, Papernot N, Anderson R, Gal Y. AI models collapse when trained on recursively generated data. Nature. 2024;631(8022):755-9.

https://doi.org/10.1038/s41586-024-07566-y.

Agrawal A, Choudhary A. Ai methods in materials design, discovery and manufacturing: A review. Comput Mater Sci. 2024;235:112680.

Goswami L, Deka MK, Roy M. Artificial intelligence in material engineering: A review on applications of artificial intelligence in material engineering. Adv Eng Mater. 2023;25(13):2300104.

Zhao Q, Savoie B, Liu K, Chen G, Lu J, Marks TJ, et al. Accelerating materials discovery using artificial intelligence, high performance computing and robotics. npj Comput Mater. 2022;8:89.

Huang L, Liu J, Liu Y, Liu Y. Artificial intelligence reinventing materials engineering: A bibliometric review. Appl Sci. 2024;14:8143.

Daly AC, Paalanen O, Morales M, Koutsos V. Evolution of artificial intelligence for application in contemporary materials science. MRS Commun. 2023;13:551-70.

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.

Gong W, Yan Q. Graph-based deep learning frameworks for molecules and solid-state materials. Comput Mater Sci. 2021;195:110332.

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

Lu S, Zhou Q, Guo Y, Zhang Y, Wu Y, Wang J. A robust synthetic data generation framework for machine learning in high-resolution transmission electron microscopy (hrtem). npj Comput Mater. 2024;10:162.

Zhao Y, Zhang B, Hou J. Generative artificial intelligence and its applications in materials science: Current situation and future perspectives. J Materiomics. 2023;9:1050-64.

Xu H, Liu J, Xie Y, Cheng YHE, Lee S, Gerbrand C. Next-generation intelligent laboratories for materials design and manufacturing. Matter. 2023;6:369-94.

Barati Farimani A, et al. Generation of synthetic manufacturing datasets for machine learning using discrete-event simulation. Simul Model Pract Theory. 2022;119:102576.

Chen L, Yeung Y, Mori S. Small data machine learning in materials science. npj Comput Mater. 2023;9:42.

Lee SW, Kim H, Lee J, Yoon S. Not as simple as we thought: A rigorous examination of data aggregation in materials informatics. Digit Discov. 2024;3:691-701.

Lin Y, Zhao Y, Zhang B, Hou J. Recent progress on machine learning with limited materials data: Using tools from data science and domain knowledge. J Materiomics. 2024;10:856-70.

Ghiringhelli LM, et al. Integrating computational materials science and materials informatics for the modeling of phase stability. J Mater Inf. 2021;1:3.

Navarro-Ruiz J, et al. Improving machine-learning models in materials science through large datasets. Matter. 2024;7:3217-37.

Frey N, Hase F, Sauceda HE, Stuke A, Wuttig M, Noe C. Open materials 2024 (omat24) inorganic materials dataset and models. arXiv preprint arXiv:2410.12771. 2024.

Taraporevala K, et al. “Forgetting” in machine learning and beyond: A survey. arXiv preprint arXiv:2405.20620. 2024.

Zeng M, et al. Worth of prior knowledge for enhancing deep learning. J Knowl Struct Syst. 2024;5:1-19.

Lin S, Wen G, Ding N, Gong M, Ye C. Erm-ktp: Knowledge-level machine unlearning via knowledge transfer. In: Proc IEEE/CVF Conf Comput Vis Pattern Recognit; 2023. p. 20661-70.

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

Pilania G, et al. Machine learning in materials science: From discovery to optimization. Annu Rev Chem Biomol Eng. 2021;12:499-524.

Zhang Y, et al. Machine learning in electronic-quantum-matter imaging experiments. Nature. 2020;583:533-8.

Shen L, Zhou J, Yang T, Yang M, Feng YP. High-throughput computational discovery and intelligent design of two-dimensional functional materials for various applications. Acc Mater Res. 2022;3(6):572-83.

Dai M, Demirel MF, Liang Y, Hu JM. Graph neural networks for an accurate and interpretable prediction of the properties of polycrystalline materials. npj Comput Mater. 2021;7(1):103.

Dunn A, et al. Benchmarking materials property prediction methods: The matbench test set and automatminer reference algorithm. npj Comput Mater. 2020;6:138.

Lutz B, Reisch R, Kisskalt D, Avci B, Regulin D, Knoll A, et al. Benchmark of automated machine learning with state-of-the-art image segmentation algorithms for tool condition monitoring. Procedia Manuf. 2020;51:215-21.

Tran K, et al. Deep learning the structural determinants of protein biochemical properties from sequence. Cell Syst. 2021;12:98-110.e3.

Fung V, et al. Machine learning surrogates for molecular dynamics simulations of soft materials. J Phys Chem Lett. 2021;12:4357-62.

Gebauer NW, Gastegger M, Hessmann SS, Müller KR, Schütt KT. Inverse design of 3d molecular structures with conditional generative neural networks. Nat Commun. 2022;13(1):973.

Merchant A, et al. Artificial intelligence driving materials discovery? Perspective on the article: Scaling deep learning for materials discovery. Chem Mater. 2024;36:3171-3.

Wang AY, et al. Machine learning for materials scientists: An introductory guide toward best practices. Chem Mater. 2020;32:4954-65.