ORIGINAL RESEARCH Open access

Synthetic Data as Scientific Intervention: A Conceptual Framework for Materials AI

Ravi Kumar[1*], Neha Sharma[1], Aniket Deshmukh[2], Arjun Nair[1]

Abstract

In the evolving landscape of materials artificial intelligence (AI), synthetic data emerges not merely as a technical augmentation but as a profound scientific intervention that reshapes the interpretive dynamics of knowledge generation. This manuscript develops a conceptual framework that interprets synthetic data as an intermediary layer facilitating interactions between empirical realities and algorithmic abstractions in materials science. This study synthesizes recent literature and examines how synthetic data influences epistemic trade-offs, such as those between data fidelity and model generalizability. It steers feedback structures within AI-driven discovery processes. The framework underscores systems-level insights into integrating generative models with domain-specific ontologies, highlighting ethical considerations in the curation of virtual datasets that mirror physical constraints without empirical grounding. Analytically, it explores the implications for accelerating materials innovation through enhanced representational capacities, while addressing potential distortions in scientific reasoning arising from over-reliance on simulated inputs. This interpretive approach reveals the transformative potential of synthetic data in reconfiguring the boundaries of human-AI collaboration, fostering a more reflexive understanding of material phenomena. Ultimately, the framework invites a reevaluation of data's role in scientific inquiry, emphasizing integrative logics that balance innovation with epistemological integrity in the pursuit of advanced materials.

Keywords Materials AI, Synthetic data, Scientific intervention, Conceptual framework, Epistemic trade-offs, Generative models

*Correspondence:
Ravi Kumar
ravi.kumar@outlook.com
[1] Department of Materials Engineering and Data Science, Faculty of Engineering, IIT Delhi, New Delhi, India
[2] Department of AI-Based Materials Design, Faculty of Engineering, IIT Bombay, Mumbai, India

Introduction

The integration of artificial intelligence into materials science represents a pivotal shift in how researchers conceptualize and navigate the complexities of material behaviors and properties. At its core, this integration relies on vast datasets that inform machine learning models, enabling patterns to emerge from the interplay of atomic structures, environmental conditions, and performance metrics. However, the scarcity of high-quality, experimentally derived data poses persistent challenges, prompting exploration of alternative strategies to enhance the robustness of AI applications. Among these, synthetic data stands out as a mechanism that extends beyond mere supplementation, functioning as a deliberate intervention in the scientific process. This intervention alters the interpretive lenses through which material phenomena are understood, bridging gaps between limited empirical observations and expansive computational possibilities.

In materials AI, the traditional reliance on experimental data often encounters bottlenecks in terms of cost, time, and accessibility. For instance, generating comprehensive datasets for rare-earth materials or high-entropy alloys requires extensive laboratory resources, limiting the scope of exploratory analyses [1-3]. Synthetic data, generated algorithmically through methods such as generative adversarial networks or diffusion models, introduces a layer of virtual representation that amplifies the available information landscape. This amplification is not neutral; it introduces dynamics where data authenticity intersects with predictive utility, compelling a reevaluation of how knowledge is constructed. The conceptual framing of synthetic data as an intervention highlights its role in modulating these dynamics, influencing not only model training but also the broader epistemic ecosystem of materials research.

Analytically, this intervention manifests in the trade-offs inherent to data-driven paradigms. On one hand, synthetic data enables the exploration of hypothetical scenarios, such as material responses under extreme conditions that are impractical to test empirically [4-7]. On the other hand, it raises questions about the fidelity of these representations to physical laws, potentially skewing interpretive outcomes if not carefully integrated. The framework proposed herein interprets these trade-offs through a systems perspective, where synthetic data serves as a feedback mechanism that iteratively refines AI models, enhancing their alignment with domain knowledge. This perspective draws from recent advancements in multimodal AI platforms that incorporate diverse data modalities, illustrating how synthetic inputs can enrich the semantic depth of materials databases [2, 4].

Furthermore, the ethical dimensions of this intervention warrant careful consideration. In an era where AI accelerates materials discovery for applications in energy storage, catalysis, and biomedicine, the generation of synthetic data must navigate issues of transparency and accountability [6, 8]. If synthetic datasets inadvertently perpetuate biases introduced by their generative algorithms, they could distort scientific interpretations, leading to misaligned research priorities. The framework addresses these by emphasizing integrative reasoning that connects data generation with ethical oversight, ensuring that interventions bolster rather than undermine the integrity of scientific inquiry.

From a broader perspective, synthetic data's role in materials AI reflects a paradigm in which human expertise and machine intelligence converge in novel ways. Traditional materials science emphasizes empirical validation, yet AI introduces abstraction layers that reinterpret data through probabilistic lenses [9, 10]. Synthetic data amplifies this abstraction, allowing for the simulation of vast parameter spaces that inform design strategies. This amplification fosters interaction dynamics between researchers and AI systems, in which synthetic inputs serve as probes to uncover latent relationships in material properties. For example, in electrocatalyst discovery, synthetic data can simulate multi-element compositions, guiding interpretive analyses toward optimal configurations without exhaustive experimentation [2, 11].

The literature synthesis that follows builds on these foundations, tracing the evolution of AI in materials science and the emergence of synthetic data as a key enabler. It integrates insights from generative AI applications, highlighting how they reshape knowledge flows [12, 13]. This sets the stage for the proposed framework, which interprets synthetic data as a multifaceted intervention that navigates the tensions between innovation and epistemological rigor.

In synthesizing these elements, the manuscript contributes to a deeper understanding of how synthetic data reconfigures the conceptual underpinnings of materials AI. By focusing on analytical implications rather than prescriptive guidelines, it illuminates the steering logics that govern effective integration, ultimately advocating for a reflexive approach to data-centric science. This introduction thus positions synthetic data not as a tool but as a transformative force in the interpretive fabric of materials research, inviting scholars to engage with its systemic ramifications.

Theoretical Background and Literature Synthesis

Evolution of AI in materials discovery

Since the early 2020s, the role of artificial intelligence in materials science has undergone a marked conceptual and functional transformation. Initial applications of machine learning were largely confined to retrospective pattern recognition —mining existing datasets to uncover correlations between composition, structure, and properties such as mechanical strength, thermal conductivity, or phase stability [1, 14, 15]. These early systems functioned primarily as analytical accelerators, enhancing efficiency without fundamentally reshaping scientific reasoning.

Over time, however, AI has evolved into a more integrative and interpretive infrastructure for materials discovery. Rather than merely identifying correlations, contemporary AI systems increasingly operate as epistemic bridges, linking heterogeneous data sources—computational simulations, experimental measurements, and domain heuristics—into coherent representations of material behavior. This shift reflects a broader reorientation in which AI is no longer treated as an auxiliary analytical tool but as an active participant in the construction of materials knowledge, capable of revealing systemic relationships that remain inaccessible through isolated human or computational reasoning alone.

Advances in deep learning architectures have played a pivotal role in enabling this transition. By accommodating the high dimensionality and hierarchical complexity intrinsic to materials data, modern neural models can encode interactions across compositional, structural, and property spaces with unprecedented flexibility [3, 16]. The scaling of these architectures to encompass millions of material descriptors, or “material narratives,” exemplifies how AI systems synthesize vast, fragmented knowledge repositories into unified latent representations [3, 9, 10, 12, 15, 17-19]. Yet this expansion foregrounds a critical trade-off: while larger models enhance generalizability and pattern coverage, they also increase the risk of interpretive dilution, in which learned representations drift away from physically meaningful constraints unless carefully curated and domain-aligned inductive biases guide them.

The integration of AI with robotics, automation, and high-performance computing has further extended its role from interpretation to orchestration. Closed-loop discovery systems now iteratively generate hypotheses, evaluate candidates, and update learning objectives through continuous data feedback [10]. These systems embody a new paradigm of scientific exploration in which AI not only processes information but actively steers inquiry toward unexplored or undersampled regions of materials space. Analytically, this introduces ethical and epistemic considerations: algorithmic prioritization influences which materials are studied, which hypotheses are pursued, and how resources are allocated [6]. As a result, human oversight becomes not merely a safeguard but a constitutive element of responsible discovery, ensuring that efficiency gains do not eclipse interpretive accountability.

Generative models and synthetic data generation

Generative artificial intelligence has emerged as a central mechanism through which materials AI extends beyond empirical limitations. Models based on diffusion processes, flow matching, or related generative paradigms infer latent distributions over structures and compositions, enabling the synthesis of plausible material configurations that may not yet exist in curated datasets [13, 16, 17]. Rather than functioning as simple data amplifiers, these models introduce explicit steering logics that mediate transitions between data scarcity and abundance, allowing systematic exploration of rare, unstable, or experimentally inaccessible material regimes [7, 8].

In domains such as metal–organic frameworks, crystalline solids, and complex inorganic compounds, generative approaches have been used to construct predictive datasets that inform design strategies and guide downstream screening [14, 15]. Conceptually, these methods illuminate epistemic trade-offs inherent in synthetic generation: the apparent realism of generated data depends critically on the adequacy of training priors and representational assumptions. When these priors are misaligned with domain ontologies or physical constraints, synthetic outputs risk introducing subtle but consequential distortions into discovery pipelines [5, 17].

At a systems level, the literature increasingly emphasizes the value of generative models as integrative instruments. By fusing multimodal inputs—structural graphs, compositional descriptors, thermodynamic signals—these systems enhance representational coherence and enable more expressive mappings between materials form and function [2, 9]. Synthetic data thus becomes part of a recursive feedback structure: generated candidates inform model refinement, which in turn reshapes the generative landscape. This recursive dynamic enhances robustness but simultaneously complicates the evaluation of validity, raising foundational questions about how virtual authenticity should be assessed relative to physical benchmarks.

Data preparation and integration in materials AI

The effectiveness of AI in materials science is ultimately contingent on the quality and structure of its data foundations. Recent scholarship highlights that performance gains increasingly arise not from algorithmic novelty alone, but from carefully designed data preparation pipelines that integrate empirical and synthetic elements into coherent learning ecosystems [9]. These pipelines require interpretive integration across diverse modalities, including spectroscopic signatures, microstructural images, simulation outputs, and curated metadata, transforming fragmented observations into unified knowledge representations [2, 4].

Small-data regimes—common in specialized or emerging materials domains—stand to benefit disproportionately from synthetic augmentation. By amplifying weak learning signals and smoothing sparse distributions, synthetic data can enhance generalizability without necessitating prohibitively costly experiments [15]. However, this benefit comes with analytical trade-offs. While increased data volume may stabilize learning dynamics, excessive reliance on synthetic patterns risks overfitting to artificial regularities that lack physical grounding [1, 3]. As such, the distinction between data enrichment and epistemic inflation becomes a central concern in materials AI.

Parallels with synthetic biology and related fields further illuminate these challenges. Across domains, synthetic data interventions reshape knowledge flows by mediating what counts as observable, learnable, and actionable [11, 19]. From a systems perspective, these intersections underscore shared epistemic vulnerabilities—particularly the potential for bias reinforcement and representational exclusion—while also highlighting opportunities for cross-disciplinary governance frameworks that promote inclusivity and reflexive oversight in AI-driven science. These roles of synthetic data across materials AI workflows are summarized in **Table 1**, which conceptualizes synthetic data as a structured intervention operating across generative, interpretive, and integrative layers.

Table 1. Modes of synthetic data intervention in materials AI

Intervention mode	Primary function	AI pipeline location	Epistemic benefit	Associated risk/trade-off
Data scarcity bridging	Augments sparse empirical datasets	Pre-training/data curation	Expands the learnable material space	Inflation of artificial regularities
Exploratory probing	Enables hypothetical or extreme-condition sampling	Generative modeling stage	Reveals latent interaction dynamics	Drift from physical plausibility
Representational smoothing	Regularizes noisy or imbalanced data	Feature learning/latent space construction	Improves generalizability	Loss of fine-grained physical detail
Feedback amplification	Reinforces iterative model refinement	Closed-loop discovery systems	Accelerates convergence	Bias reinforcement through recursion
Ontology alignment	Encodes domain constraints into generation	Constraint-aware generation	Maintains interpretive coherence	Over-constraining innovation
Access equalization	Reduces dependence on costly experiments	Cross-institutional deployment	Democratizes modeling capacity	Uneven epistemic authority

Ethical and epistemic considerations

The growing reliance on synthetic data in materials AI necessitates a sustained and reflexive examination of its epistemic implications. Generative processes inevitably encode assumptions embedded in training datasets, model architectures, and optimization objectives, shaping downstream interpretations in often opaque ways [6, 13, 17]. This underscores the need for integrative logics that explicitly align synthetic interventions with ethical principles, including transparency in data lineage, traceability of generative decisions, and clarity about the epistemic status of virtual observations [8, 18].

At the systems level, synthetic data participates in feedback structures that influence model evolution and scientific inference. These structures can either attenuate or amplify existing biases, depending on how synthetic outputs are validated, filtered, and reintegrated into learning loops [5, 7, 12]. Analytically, this dynamic reframes innovation as a balance between acceleration and epistemological caution, where the pursuit of efficiency must be weighed against the integrity of scientific reasoning [10, 14].

More broadly, ethical considerations extend to questions of accountability and access. While synthetic data holds promise for democratizing materials discovery by lowering experimental barriers, it also carries risks of misinformation, overconfidence, and misplaced authority if virtual results are uncritically equated with physical truth [11, 16, 19]. Collectively, these tensions signal a maturing field—one in which the responsible harnessing of synthetic interventions depends not only on technical sophistication, but on robust interpretive frameworks that anchor AI-enabled discovery within defensible epistemic and ethical boundaries.

Proposed conceptual framework

This framework conceptualizes synthetic data as a scientific intervention within materials artificial intelligence, rather than as a neutral technical augmentation. Synthetic data is interpreted as an active mediating layer that restructures interactions between empirically grounded knowledge and algorithmic abstraction. Through this intervention, materials AI systems are not merely supplied with additional inputs but are reconfigured in how they explore, integrate, and prioritize material possibilities under conditions of empirical limitation.

At the core of the framework is a set of dynamic feedback structures through which generative processes, domain priors, and interpretive constraints interact. Synthetic data operates within these structures as a transitional substrate—bridging empirical scarcity and computational abundance—while continuously negotiating epistemic trade-offs between physical fidelity, representational completeness, and exploratory reach. Rather than replacing observation, synthetic data reshapes how absence, uncertainty, and incompleteness are operationalized within AI-driven discovery.

Analytically, synthetic data serves as an intermediary representation layer that expands the effective search space for materials AI systems. By enabling controlled excursions beyond empirically sampled regions, this layer allows models to probe latent interaction dynamics that would otherwise remain inaccessible due to experimental or computational constraints [1, 3, 5, 7]. Crucially, these excursions are governed by steering logics that align generative outputs with physical ontologies, preventing unconstrained extrapolation while preserving creative latitude. Within this balance, synthetic data serves as a diagnostic instrument—revealing sensitivities, structural dependencies, and emergent regularities across complex material systems [9, 10, 12, 13].

From an ethical and epistemic standpoint, the framework foregrounds the reflexive consequences of synthetic intervention. The curation, validation, and reintegration of synthetic data directly influence the integrity of scientific reasoning, as virtual representations recursively shape subsequent model behavior and interpretive emphasis. While excessive reliance on synthetic datasets risks distorting feedback loops or amplifying latent biases, carefully governed augmentation can broaden access to advanced modeling capabilities and mitigate inequities arising from uneven experimental resources [6, 8, 11, 14]. The framework thus frames synthetic data as epistemically productive but conditionally legitimate, requiring safeguards to preserve rigor, traceability, and interpretive accountability [3, 15, 17, 18].

At a systems level, the framework positions synthetic data as a catalyst for structured human–AI collaboration. Generative models are interpreted not as autonomous discovery engines but as multimodal integrators that synthesize empirical priors, theoretical constraints, and computational inference into evolving conceptual representations [2, 4, 9, 12]. Through this integration, knowledge flows are reconfigured to prioritize coherence, reflexivity, and interpretive alignment over isolated predictive performance. Synthetic data thereby becomes a mechanism through which materials AI transitions from pattern amplification toward systems-level understanding. The proposed conceptual framework is visualized in **Figure 1**, which models synthetic data as a central intervention prism operating across generative, interpretive, and systems-integration layers.

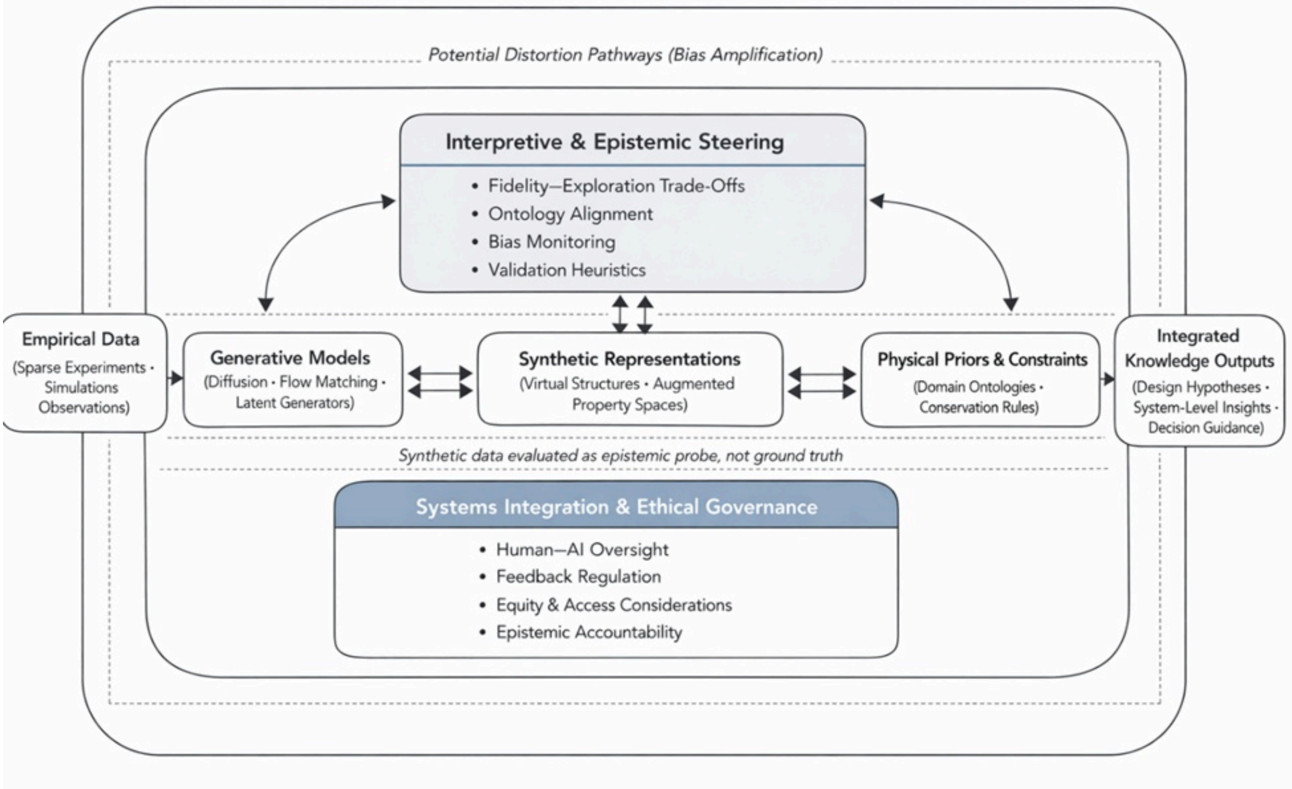

Figure 1. Schematic of a conceptual framework for synthetic data intervention in materials AI. The diagram illustrates the dynamic interplay between generative data creation, epistemic interpretation, and systems-level integration, highlighting iterative refinement and ethical feedback loops.

Analytical implications

Interpreting synthetic data as a scientific intervention reframes the analytical landscape of materials AI. Rather than serving as a passive supplement, synthetic data actively reshapes how epistemic trade-offs are navigated—particularly those involving representational fidelity, exploratory breadth, and interpretive confidence. By enabling controlled engagement with sparsely sampled regions of material, synthetic data allows AI systems to interrogate phenomena that remain empirically elusive, extending analytical reach without presupposing exhaustive observation [1, 3, 5, 15].

Within the framework, synthetic inputs act as catalytic elements inside recursive feedback structures. Generated representations modulate the alignment between probabilistic inference and domain-grounded understanding, enabling iterative recalibration of model behavior and interpretive emphasis [7-9, 13]. This dynamic foregrounds steering logics that privilege integrative coherence over isolated accuracy, reinforcing the view that analytical robustness in materials AI emerges from alignment across representations, constraints, and feedback—not from prediction alone. The central analytical consequences of treating synthetic data as a scientific intervention—along with their associated trade-offs and governance implications—are consolidated in **Table 2**.

Table 2. Analytical implications of synthetic data as a scientific intervention

Analytical dimension	Intervention effect	Steering logic introduced	Epistemic gain	Risk if ungoverned
Representational scope	Expansion beyond empirical coverage	Fidelity–exploration balancing	Access to latent material regimes	Epistemic dilution
Model–knowledge alignment	Iterative recalibration via feedback	Ontology-constrained generation	Improved interpretive coherence	Recursive bias locking
Knowledge boundaries	Blurring of empirical vs. virtual evidence	Reflexive provenance evaluation	Enhanced epistemic awareness	Misplaced epistemic authority
Scalability	Rapid exploration of compositional space	Volume–interpretability trade-off	Systems-level insight	Loss of traceability
Human–AI collaboration	Redistribution of interpretive agency	Oversight-integrated workflows	Democratized discovery	Unequal representational power
Ethical governance	Embedding values in data generation	Transparency and accountability	Responsible innovation	Invisible normative steering

Epistemologically, the intervention compels a reassessment of knowledge boundaries. By expanding the informational substrate available to AI systems, synthetic data introduces a reflexive analytical stance in which the provenance, scope, and limitations of virtual representations must be explicitly interrogated [10, 12]. This reflexivity exposes potential distortions arising from generative priors, including the amplification of latent structural regularities that may subtly privilege certain material classes or behaviors over others [6, 18].

At the systems level, the framework highlights trade-offs between scalability and interpretability. As synthetic volumes expand analytical coverage across compositional landscapes, the risk of epistemic dilution increases unless interpretive safeguards are maintained [2, 4, 16, 19]. The tension between generative diversity and physical plausibility thus emerges as a defining analytical dynamic—one that determines whether synthetic intervention enriches conceptual mappings or obscures material meaning.

Ethically, the analytical implications extend to stewardship of scientific knowledge production. Synthetic data holds potential to democratize access to advanced modeling, redistributing interpretive agency across institutions and communities [8, 11, 14, 18]. Simultaneously, biases embedded in generative architectures may perpetuate inequities if left unexamined, reinforcing dominant material paradigms at the expense of underexplored domains [6, 13, 17]. Addressing these tensions requires embedding ethical reflexivity directly within analytical practice rather than treating it as an external corrective.

Taken together, these implications position synthetic data as an active modulator of scientific reasoning in materials AI. By mediating between empirical scarcity and computational abundance, synthetic intervention reconfigures analytical processes toward deeper systems-level engagement, enabling more nuanced and responsible exploration of material phenomena through sustained epistemic reflection.

Results and Discussion

The interpretive lens afforded by the framework positions synthetic data as a transformative intervention that permeates multiple strata of materials AI inquiry. This positioning invites discussion of how such interventions recalibrate the interplay between human expertise and algorithmic abstraction, fostering collaborative dynamics that transcend traditional divisions of labor in scientific practice [9].

Central to this discussion is the recognition that synthetic data engenders feedback structures capable of self-reinforcement. As AI systems assimilate generated inputs, they evolve interpretive capacities that, in turn, inform subsequent data synthesis, creating loops that amplify both strengths and vulnerabilities [3]. These loops underscore the need for vigilant oversight, where analytical monitoring of divergence between synthetic and empirical distributions safeguards against cumulative epistemic drift.

The framework further illuminates ethical tensions inherent in scaling synthetic interventions. While generative approaches promise accelerated exploration of material design spaces, they simultaneously introduce accountability challenges related to the traceability of virtual influences on discovery outcomes [18]. Discussion thus centers on mechanisms for embedding transparency, such as provenance tracking within generative pipelines, to maintain trust in AI-mediated insights without stifling innovative potential [2, 12].

Systems-level insights reveal how synthetic data reshapes knowledge ecosystems in materials science. By augmenting sparse datasets, it enables more holistic interpretations of emergent properties in complex systems, from high-entropy alloys to functional interfaces [15]. Yet this reshaping necessitates consideration of interdisciplinary convergence, where insights from related domains—such as computational chemistry or data governance—inform adaptive strategies for synthetic integration [11, 19].

The trade-off between innovation velocity and epistemological caution is a recurring theme. The framework interprets the acceleration of discovery cycles as a double-edged dynamic: it compresses exploratory timelines. Still, it risks premature closure on interpretive possibilities if virtual representations dominate unchecked [5, 7, 9]. Discussion, therefore, advocates reflexive practices that periodically reassess the balance between synthetic enrichment and empirical anchoring, preserving the dialogic essence of scientific advancement.

Ultimately, the discussion frames synthetic data as an evolving element within the broader narrative of AI-augmented science. It encourages a shift toward integrative paradigms in which interventions are evaluated not solely by efficiency gains but by their contribution to deeper, more equitable understandings of material phenomena. This orientation positions materials AI as a reflexive enterprise, continually attuned to the epistemic and ethical ramifications of its data-centric foundations.

Conclusion

The conceptual framework developed herein interprets synthetic data as a scientific intervention that fundamentally reconfigures the interpretive dynamics of materials AI. By mediating between empirical constraints and algorithmic expansiveness, synthetic data fosters interaction structures that enhance systemic insights into material complexities while navigating inherent epistemic trade-offs.

This intervention, when approached through integrative reasoning, promotes a balanced evolution in knowledge-generation processes. It underscores the value of reflexive stewardship in curating generative practices, ensuring that virtual augmentations align with domain ontologies and ethical imperatives. The framework thus contributes to a maturing discourse on data-centric science, where synthetic elements serve as catalysts for collaborative, equitable advancement rather than substitutes for foundational inquiry.

In reflecting on these interpretive layers, the manuscript highlights the transformative potential of synthetic data to enrich the conceptual landscape of materials research. It invites ongoing engagement with the steering logics, feedback mechanisms, and ethical considerations that govern its integration, fostering a future where AI augments rather than supplants the nuanced pursuit of materials understanding.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 24 Aug 2021 Revised: 23 Sep 2021 Accepted: 22 Oct 2021
Published online: 18 January 2022

References

DeCost BL, Hattrick-Simpers JR, Trautt Z, Kusne AG, Campo E, Green ML. Scientific ai in materials science: A path to a sustainable and scalable paradigm. Mach Learn Sci Technol. 2020;1(3):035005.

Zhang Z, Ren Z, Hsu CW, Chen W, Hong ZW, Lee CF, et al. A multimodal robotic platform for multi-element electrocatalyst discovery. Nature. 2025;635:316-23.
https://doi.org/10.1038/s41586-025-09640-5.

Merchant A, Batzner S, Schoenholz SS, Aykol M, Cheon G, Cubuk ED. Scaling deep learning for materials discovery. Nature. 2023;624(7990):80-5.
https://doi.org/10.1038/s41586-023-06735-9.

Li H, Li L, Li Y, et al. This ai has chemical expertise — and helps synthesize 35 new compounds. Nature. 2026;637:240-5.
https://doi.org/10.1038/s41586-026-10131-4.

Liu Y, Yang Z, Yu Z, Liu Z, Liu D, Lin H, et al. Generative artificial intelligence and its applications in materials science: Current situation and future perspectives. J Materiomics. 2023;9(4):798-816.
https://doi.org/10.1016/j.jmat.2023.04.003.

Chavez-Angel E, Eriksen MB, Castro-Alvarez A, Garcia JH, Botifoll M, Avalos-Ovando O, et al. Applied artificial intelligence in materials science and material design. Adv Intell Syst. 2025;7(3):e2400986.
https://doi.org/10.1002/aisy.202400986.

Ye C, Wang Y, Xie X, Zhu T, Liu J, He Y, et al. Materials discovery acceleration by using conditional generative methodology. Npj Comput Mater. 2026;12(1):63.
https://doi.org/10.1038/s41524-025-01930-w.

Dao DA, Ha MQ, Vu TS, Takazawa S, Ishiguro N, Takahashi Y, et al. Material dynamics analysis with deep generative model. Digit Discov. 2025;4(12):3363-77.
https://doi.org/10.1039/D5DD00277J.

Lu Y, Wang H, Zhang L, Yu N, Shi S, Su H. Unleashing the power of ai in science-key considerations for materials data preparation. Sci Data. 2024;11(1):1039.
https://doi.org/10.1038/s41597-024-03821-z.

Pyzer-Knapp EO, Pitera JW, Staar PWJ, Takeda S, Laino T, Sanders DP, et al. Accelerating materials discovery using artificial intelligence, high performance computing and robotics. Npj Comput Mater. 2022;8(1):84.
https://doi.org/10.1038/s41524-022-00765-z.

Groff-Vindman CS, Trump BD, Cummings CL, Smith M, Titus AJ, Oye K, et al. The convergence of ai and synthetic biology: The looming deluge. Npj Biomed Innov. 2025;2(1):20.
https://doi.org/10.1038/s44385-025-00021-1.

Park YJ, Jerng SE, Yoon S, Li J. 1.5 million materials narratives generated by chatbots. Sci Data. 2024;11(1):1060.
https://doi.org/10.1038/s41597-024-03886-w.

Zeni C, Pinsler R, Zügner D, et al. A generative model for inorganic materials design. Nature. 2025;639(7930):624-32.
https://doi.org/10.1038/s41586-025-08628-5.

Kang Y, Kim J. Chatmof: An artificial intelligence system for predicting and generating metal-organic frameworks using large language models. Nat Commun. 2024;15(1):4705.
https://doi.org/10.1038/s41467-024-48998-4.

Xu P, Ji X, Li M, Lu W. Small data machine learning in materials science. Npj Comput Mater. 2023;9(1):42.
https://doi.org/10.1038/s41524-023-01000-z.

Joung JF, Fong MH, Casetti N, Liles JP, Dassanayake NS, Coley CW. Electron flow matching for generative reaction mechanism prediction. Nature. 2025;645(7930):115-23.
https://doi.org/10.1038/s41586-025-09426-9.

Pozzi M, Noei S, Robbi E, Cima L, Moroni M, Munari E, et al. Generating and evaluating synthetic data in digital pathology through diffusion models. Sci Rep. 2024;14(1):28435.
https://doi.org/10.1038/s41598-024-79602-w.

Wendland P, Birkenbihl C, Gomez-Freixa M, Sood M, Kschischo M, Fröhlich H. Generation of realistic synthetic data using multimodal neural ordinary differential equations. Npj Digit Med. 2022;5(1):122.
https://doi.org/10.1038/s41746-022-00666-x.

Habib M, Singh S, Jan S, Jan K, Bashir K. The future of the future foods: Understandings from the past towards sdg-2. Npj Sci Food. 2025;9(1):138.
https://doi.org/10.1038/s41538-025-00484-x.