ORIGINAL RESEARCH **Open access**

Conceptual Foundations of Applied AI in Materials Science - Definitions, Assumptions, and Open Debates

Hiroshi Tanaka[1], Yuki Sato[1*], Kenji Mori[2], Rina Okabe[1]

Abstract

The rapid integration of artificial intelligence (AI) into materials science marks a profound shift in how materials are discovered, characterized, and optimized. Rather than functioning merely as a computational aid, AI increasingly operates as an epistemic instrument that reshapes scientific workflows, decision-making practices, and notions of explanation within the field. This narrative review examines the conceptual foundations underpinning applied AI in materials science, with a particular focus on core definitions, implicit and explicit assumptions, and unresolved debates that continue to shape the domain. Key AI paradigms—including supervised, unsupervised, and reinforcement learning—are situated within materials-specific contexts such as property prediction, structure–property mapping, and autonomous experimentation. The review critically interrogates foundational assumptions regarding data quality, representativeness, generalization, and model transferability, highlighting how these assumptions condition both the successes and failures of AI-driven materials research. Persistent debates surrounding interpretability, epistemic trust, ethical responsibility, and environmental sustainability are synthesized from recent literature published. By articulating both the transformative potential and the conceptual limitations of applied AI, this review underscores the necessity of rigorous validation, transparent reasoning, and interdisciplinary collaboration to ensure that AI contributes robustly and responsibly to materials innovation.

Keywords Artificial intelligence, Materials discovery, Machine learning, Explainable AI, Materials science, Ethics and sustainability

*Correspondence:
Yuki Sato
yuki.sato@gmail.com
[1] Department of Intelligent Materials Engineering, Faculty of Engineering, University of Tokyo, Tokyo, Japan
[2] Department of AI-Driven Materials Discovery, Faculty of Information Science, Kyoto University, Kyoto, Japan

Introduction

Materials science is an inherently multidisciplinary field concerned with elucidating the relationships between structure, composition, processing, and performance of materials, with far-reaching applications in energy storage, electronics, biomedicine, and environmental technologies. Historically, progress in materials development has depended on a combination of empirical trial-and-error experimentation, physics-based theoretical modeling, and computational simulations. While these approaches have yielded foundational advances, they are often constrained by high costs, long development cycles, and the combinatorial complexity of materials design spaces. As a result, translating a conceptual materials idea into a commercially viable technology has frequently required decades of iterative refinement and validation [1].

The rapid expansion of computational power, high-throughput experimentation, and digital data infrastructures has fundamentally altered this landscape. Within this context, artificial intelligence (AI) has emerged as a powerful enabling

technology capable of accelerating materials discovery and optimization by leveraging data-driven inference at unprecedented scales [2]. In particular, machine learning (ML) techniques enable models to extract complex, non-linear relationships from large, heterogeneous datasets, enabling the prediction of material properties, the identification of promising candidate compounds, and the optimization of synthesis and processing pathways. These capabilities offer the potential to reduce both time-to-discovery and resource consumption in materials research, positioning AI as a transformative tool rather than a mere computational adjunct [3].

Conceptually, the integration of AI into materials science is frequently framed within the paradigm of "data-intensive science," often described as the fourth paradigm of scientific discovery. This paradigm complements experimental observation, theoretical reasoning, and computational simulation by emphasizing knowledge generation through large-scale data analysis and algorithmic inference [4]. In materials science, this shift has enabled the systematic exploration of vast compositional and structural spaces that would be infeasible using traditional approaches alone. The paradigm is further exemplified by emerging research infrastructures such as self-driving laboratories (SDLs), in which AI systems autonomously coordinate experimental design, execution, data acquisition, and analysis in closed feedback loops, thereby compressing discovery timelines and reducing human intervention [5].

Despite these advances, the application of AI in materials science raises important conceptual and practical questions that extend beyond technical performance. AI-driven workflows implicitly rely on assumptions regarding data quality, representativeness, and completeness, as well as the robustness and generalizability of trained models across materials classes and application domains. When these assumptions are violated, AI systems may produce overconfident predictions, amplify hidden biases in datasets, or fail catastrophically when extrapolating beyond familiar regimes. Consequently, the increasing reliance on AI has intensified debates surrounding interpretability, epistemic trust, algorithmic bias, and the environmental sustainability of large-scale model training and deployment [6, 7]. These concerns are particularly salient in materials science, where mechanistic understanding, physical plausibility, and long-term reliability are central to scientific and technological progress.

Against this backdrop, a critical examination of the conceptual foundations of applied AI in materials science is both timely and necessary. This review aims to provide such an examination by synthesizing recent literature and clarifying the intellectual structures that underlie current practice. Specifically, the objectives are threefold: (1) to define core concepts and classifications of AI methodologies as they are applied within materials science; (2) to articulate the key assumptions —often implicit—that underpin AI-driven materials research; and (3) to examine the open debates that continue to shape the field's evolution, including questions of interpretability, ethical responsibility, and sustainability. By integrating insights from peer-reviewed studies, this review seeks to offer a balanced, analytically grounded perspective that informs researchers, policymakers, and practitioners in the judicious and responsible use of AI to advance materials innovation. **Figure 1** provides an organizing map of the review, showing how data and representations feed learning paradigms and outputs, and how foundational assumptions and open debates shape the transition from predictive utility to scientific legitimacy.

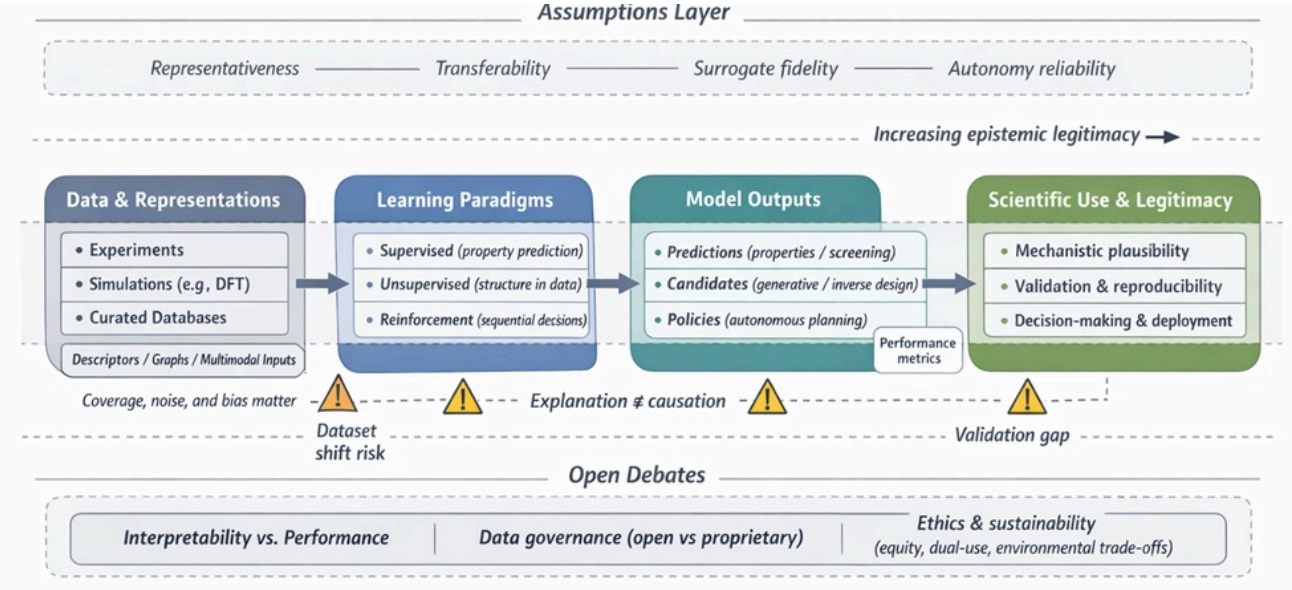

Figure 1. Conceptual map of the materials-AI epistemic pipeline: from data to scientific legitimacy through assumptions and debates

Definitions and classifications of AI in Materials science

Artificial intelligence broadly refers to computational systems designed to emulate aspects of human intelligence, including perception, reasoning, learning, and decision-making [2]. Within materials science, AI is not typically pursued as general intelligence, but rather as a collection of task-specific computational approaches aimed at extracting actionable knowledge from complex materials data. In practice, AI in this domain is predominantly operationalized through machine learning (ML), a subfield focused on algorithms that infer patterns and relationships directly from data without being explicitly programmed with physical rules or heuristics [3]. This data-driven orientation distinguishes AI-based approaches from traditional physics-based or rule-driven computational models that have historically dominated materials research.

Machine learning methodologies used in materials science are conventionally classified into three primary categories: supervised, unsupervised, and reinforcement learning [1]. Each category reflects a distinct mode of interaction between data, objectives, and learning signals, and each supports different stages of the materials research pipeline. For clarity, **Table 1** summarizes the principal AI paradigms used in materials science, their typical conceptual roles, and the epistemic limitations that motivate later debates on generalization and interpretability.

Table 1. Conceptual taxonomy of applied AI paradigms in materials science: roles, typical tasks, and epistemic limits

AI paradigm/family	Core learning signal	Typical materials-science uses (conceptual roles)	Representative model forms (as framed in this review)	Primary epistemic strength	Key limitation/ri (conceptual)
Supervised learning	Labeled targets (property, class)	Property prediction; screening; structure–property mapping; surrogate modeling	Regression/classification models; GNNs for crystals and atomic graphs	High predictive utility when labels are consistent, and coverage is adequate	Risk of spurious correlations; limit extrapolation beyond training distribution; opac at scale

Unsupervised learning	No explicit labels (structure in data)	Clustering of materials families; latent space organization; dimensionality reduction for exploration	Clustering; embedding/representation learning	Reveals hidden structure; supports hypothesis generation and dataset diagnostics	Can produce "apparent structu not aligned with physics; interpretability o clusters/embeddir may be ambiguo
Reinforcement learning	Reward from sequential interaction	Closed-loop optimization; experimental planning; autonomous decision policies in SDLs	RL agents integrated with automated platforms	Enables adaptive, sequential decision-making under resource constraints	Reward misspecification safety/uncertaint management; dependence or reliable automati pipeline
Deep learning (cross-cutting)	High-capacity representation learning	High-dimensional, non-linear mapping; multimodal learning (images/spectra/text)	Multi-layer neural networks; CNNs for images/spectra	Handles complexity and heterogeneous data; scalable feature learning	Data hunger; overfitting; reduc transparency without constrain or explanations
Generative models	Learning a data distribution / latent structure	Candidate generation; inverse design (property→material proposal); search-space expansion	VAEs, GANs	Shifts workflow from prediction to design; supports exploration of novel candidates	Validity/feasibilit gap; weak mechanistic grounding unles constrained; evaluation burde
Explainable AI (XAI)	Explanation is an objective layered on prediction	Feature attribution; interpretability support; trust calibration; scientific plausibility checks	Post-hoc or interpretable-by-design approaches	Improves transparency; supports scientific communication and validation	"Explanation" ma remain correlation it may not guaran causal/mechanis insight

Supervised learning is the most widely adopted paradigm in materials science applications. In this setting, models are trained on labeled datasets, where input representations—such as compositional descriptors, structural features, or microstructural images—are paired with known target outputs, including mechanical, electronic, or thermodynamic properties. The goal is to learn a mapping that enables the reliable prediction of properties for previously unseen materials. Recent advances in supervised learning architectures, particularly graph neural networks (GNNs), have been especially influential. By representing crystal structures as graphs composed of atoms as nodes and interatomic interactions as edges, GNNs naturally encode structural information and have demonstrated unprecedented accuracy in predicting properties such as formation energy and phase stability [8]. These developments illustrate how domain-appropriate representations can substantially enhance predictive performance.

Unsupervised learning, in contrast, operates on unlabeled data and seeks to uncover latent structure within datasets without predefined targets. In materials science, unsupervised methods are commonly employed for clustering chemically or structurally similar materials, identifying patterns in high-dimensional composition spaces, and performing

dimensionality reduction to enable visualization and exploratory analysis [9]. Such approaches are particularly valuable in early-stage discovery, where explicit property labels may be scarce or incomplete. By revealing intrinsic data organization, unsupervised learning can guide hypothesis generation and inform subsequent supervised modeling efforts.

Reinforcement learning represents a less mature but increasingly promising class of methods in materials research. In reinforcement learning, an agent learns to make sequential decisions by interacting with an environment and receiving feedback in the form of rewards or penalties. This paradigm is well-suited to optimization problems involving iterative decision-making, such as experimental planning, synthesis pathway optimization, or autonomous laboratory control. Emerging applications include automated synthesis and self-driving laboratory systems, where reinforcement learning agents iteratively refine strategies based on experimental outcomes [5]. Although still in its early stages of adoption, reinforcement learning introduces a dynamic, adaptive dimension to AI-assisted materials discovery.

Beyond these learning paradigms, deep learning (DL)—a specialized subset of machine learning characterized by multi-layered neural network architectures—has become increasingly prominent due to its ability to model complex, non-linear relationships in large, heterogeneous datasets [3]. Deep learning techniques are particularly effective for handling unstructured data, such as images of microstructures or spectra, and for learning high-capacity representations that capture subtle correlations across multiple length and time scales.

Within the deep learning landscape, generative models have attracted growing attention for their potential to move materials science beyond prediction toward creation. Approaches such as variational autoencoders (VAEs) and generative adversarial networks (GANs) learn compact latent representations of materials data. They can generate novel candidate structures or compositions that resemble, but are not identical to, known materials [11, 12]. These generative capabilities underpin the emerging paradigm of inverse materials design, in which desired target properties are specified first and AI models propose candidate materials likely to exhibit those properties [10]. This shift from descriptive and predictive modeling toward generative and design-oriented workflows represents a significant conceptual evolution in the field.

As AI systems increasingly influence scientific reasoning and decision-making, explainability has emerged as a central concern. Explainable artificial intelligence (XAI) refers to a collection of approaches aimed at rendering AI model predictions more transparent and interpretable, in contrast to opaque “black-box” models whose internal logic remains inaccessible [6, 13]. In materials science, XAI is particularly important because scientific credibility often depends not only on predictive accuracy but also on the ability to relate predictions to known physical principles and structure–property relationships. XAI methods seek to bridge this gap by highlighting influential features, revealing learned representations, or integrating domain knowledge directly into model architectures, thereby fostering greater trust and interpretability in AI-driven discoveries [13].

Together, these definitions and classifications clarify the conceptual landscape of applied AI in materials science. Rather than constituting a single methodological approach, AI encompasses a diverse set of learning paradigms and modeling strategies, each associated with distinct capabilities, assumptions, and limitations. Understanding these distinctions is essential for critically evaluating AI as a scientific tool and for situating subsequent debates over robustness, interpretability, and responsible deployment.

Key assumptions underpinning AI applications in materials science

The effectiveness of artificial intelligence in materials science rests on a set of foundational assumptions that are often implicit but critically shape model performance, reliability, and scientific value. Making these assumptions explicit is essential for understanding both the successes and the limitations of AI-driven materials research.

A first and central assumption concerns data representativeness. Data-driven AI models assume that historical datasets—whether derived from experimental measurements, computational simulations such as density functional theory (DFT), or

large-scale repositories including the Materials Project—adequately sample the relevant chemical and structural design space [15]. This assumption underpins the expectation that trained models can generalize beyond the specific materials included in their training sets. In practice, however, materials datasets are frequently sparse, unevenly distributed, and biased toward well-studied compositions or stable phases. The quality, diversity, and curation of input data therefore play a decisive role in determining model robustness, as noisy, incomplete, or skewed datasets can lead to systematically erroneous or overconfident predictions [4].

A second key assumption concerns the model's transferability across material domains. Many AI workflows rely on the premise that learned representations or parameters can be transferred from one materials class to another—for example, from organic to inorganic systems—through techniques such as transfer learning [11]. This assumption is critical for scalability, particularly in domains where labeled data is scarce. However, transferability implicitly assumes that underlying physical regularities are shared across domains and that essential principles, such as energy conservation or locality in interatomic interactions, are faithfully captured by the model architecture [16]. When these conditions are not met, transfer learning may introduce hidden inconsistencies or mask domain-specific physics.

A third assumption concerns computational efficiency and surrogate accuracy. AI-based approaches are often justified because they dramatically reduce computational cost relative to first-principles simulations. For example, machine-learned interatomic potentials aim to approximate quantum-level accuracy at a fraction of the computational expense. This strategy assumes that reference DFT calculations provide sufficiently reliable ground truth, despite well-known functional-dependent limitations, such as differences between generalized gradient approximations (e.g., PBE) and more advanced meta-GGA functionals (e.g., r²SCAN) [8]. Consequently, AI models may inherit systematic errors present in their training labels, raising questions about the ultimate fidelity of AI-accelerated predictions.

Closely related is the assumption of seamless autonomy in AI-driven experimental systems. In self-driving laboratories (SDLs), AI is expected to integrate effectively with robotics, instrumentation, and high-performance computing infrastructures to autonomously plan experiments, interpret results, and refine hypotheses [17, 18]. This presupposes not only technical interoperability but also that algorithmic decision-making can substitute for human judgment without compromising reproducibility or scientific rigor. While early demonstrations are promising, the reliability of fully autonomous discovery pipelines remains an open question.

Collectively, these assumptions have enabled notable breakthroughs, including the identification of stable topological insulators and the discovery of ion-selective membrane materials [19, 20]. However, they are not universally valid. A persistent risk is overinterpreting data-driven correlations as physically meaningful insights. When AI models are treated as epistemic authorities rather than heuristic tools, the danger is that causal mechanisms are obscured rather than illuminated, particularly in the absence of explicit physical constraints or interpretability mechanisms [6].

Historical evolution and current state of AI in materials science

The use of artificial intelligence in materials science has evolved over several decades, beginning with early statistical and rule-based models but accelerating markedly after 2010 due to advances in data availability, computational infrastructure, and deep learning methodologies [21]. Early AI applications primarily focused on property prediction and materials classification tasks, leveraging relatively simple descriptors and regression models. These efforts demonstrated the feasibility of data-driven materials modeling but were often limited in scope and generalizability.

By the late 2010s and early 2020s, the field experienced a conceptual shift toward generative and design-oriented applications of AI. Rather than merely predicting properties of known materials, AI models increasingly sought to propose novel compositions and structures with targeted functionalities [11]. This transition was enabled by advances in representation learning, scalable architectures, and access to large curated datasets.

Several landmark initiatives illustrate the current state of the field. The Open Catalyst Project, for example, aims to develop generalizable machine learning models for heterogeneous catalysis by combining large-scale datasets with open benchmarking frameworks [16]. Similarly, large-scale efforts such as GNoME have demonstrated AI's capacity to dramatically expand the known crystal structure space, generating millions of candidate materials for downstream validation [8]. These projects exemplify the growing ambition of AI-driven materials research, moving from incremental acceleration toward large-scale exploration.

At present, AI plays a central role in autonomous and semi-autonomous discovery workflows. Self-driving laboratories integrate AI-based hypothesis generation, experimental execution, and data analysis to enable closed-loop materials discovery [5, 10]. In parallel, high-throughput computations combined with machine learning have broadened access to materials design tools across domains such as photonics and thermoelectrics, lowering entry barriers and enabling rapid prototyping [22, 23]. Informatics platforms further support this ecosystem by curating and structuring data for inverse design and optimization tasks [10].

Despite these advances, the field remains in a transitional phase. AI systems predominantly function as augmentative tools that enhance human decision-making rather than fully autonomous scientific agents. Expert knowledge remains essential for framing research questions, validating predictions, and interpreting results within broader theoretical and experimental contexts [14].

Methodological approaches: Machine learning techniques in practice

In practice, machine learning techniques in materials science span a wide spectrum of modeling strategies, reflecting the diversity of materials problems and data modalities. Supervised learning models are commonly used for regression tasks, such as predicting electronic band gaps, elastic constants, or adsorption energies, while classification models support phase identification and materials screening [12]. These approaches form the backbone of many contemporary AI-driven materials workflows.

Graph neural networks (GNNs) have emerged as particularly powerful tools due to their ability to naturally represent atomic structures and interatomic interactions. By encoding relational information directly into model architectures, GNNs have achieved state-of-the-art performance in tasks ranging from stability prediction to the discovery of topological insulators [20]. Their success underscores the importance of physically meaningful representations in AI-based materials modeling.

Optimization-oriented methodologies further expand AI's practical impact. Bayesian optimization and active learning frameworks iteratively guide experimental or computational campaigns by selecting candidate materials that maximize information gain or minimize evaluation cost [15]. These approaches are especially valuable in scenarios where experiments are expensive or time-consuming, enabling more efficient exploration of complex design spaces.

Hybrid strategies that combine machine learning with physics-based simulations are increasingly adopted, operating under the assumption that such integration yields improved accuracy and robustness relative to purely data-driven or purely physics-based approaches [3]. In biologically relevant materials systems, AI has also been used to rationalize the design of smart materials by predicting interactions with biological environments, supporting applications in biomedical engineering and biosensing [24].

Nonetheless, methodological challenges persist. Materials datasets are often multimodal, encompassing numerical descriptors, images, spectra, and textual metadata. Addressing this heterogeneity requires sophisticated architectures, such as convolutional neural networks for image-based data and multimodal fusion strategies to integrate disparate information sources [7]. These challenges highlight that methodological sophistication alone does not guarantee scientific insight, reinforcing the need for careful alignment between modeling choices, data characteristics, and research

objectives. For clarity, **Table 1** summarizes the principal AI paradigms used in materials science, their typical conceptual roles, and the epistemic limitations that motivate later debates on generalization and interpretability.

Open debates: Interpretability, data quality, and ethical considerations

One of the most persistent and consequential debates in applied AI for materials science concerns interpretability. While many state-of-the-art models achieve remarkable predictive accuracy, they often function as opaque "black boxes," limiting insight into how predictions are generated. This opacity challenges the epistemic norms of materials science, which traditionally value mechanistic explanation alongside empirical performance. As a response, explainable artificial intelligence (XAI) has emerged as an active research direction, aiming to provide post hoc or intrinsically interpretable explanations for model outputs [6, 13]. However, the assumption that transparency alone resolves interpretability concerns remains contested. Materials systems are inherently complex and multiscale, and there is growing recognition that scientific explanation requires more than feature attribution or saliency maps—it demands causal, physically grounded insight that extends beyond statistical correlation [14].

Data quality and representativeness constitute a second major area of debate. Many AI-driven materials studies implicitly assume that available datasets sufficiently capture the relevant chemical and structural diversity needed for generalization. In practice, materials datasets are often biased toward well-characterized compounds, stable phases, or historically favored material classes. Such biases can undermine model robustness and lead to poor performance when extrapolating beyond the training distribution [4, 16]. These challenges are compounded by ongoing debates regarding data governance. Open data initiatives accelerate collective progress and reproducibility, yet concerns over intellectual property, competitive advantage, and data misuse continue to limit full data sharing [5, 10]. Striking a balance between openness and proprietary protection remains an unresolved tension shaping the pace and inclusivity of materials AI research.

Ethical and sustainability considerations are increasingly prominent in discussions of applied AI. Although not always explicitly foregrounded in early AI literature, the environmental costs of large-scale model training and high-throughput computation raise concerns about the carbon footprint of AI-enabled discovery. This introduces a paradox for a field that frequently positions itself as a driver of sustainable technologies. At the same time, the growing autonomy of AI systems introduces risks related to inequitable access, the concentration of technological power, and the potential misuse of AI in the design of harmful or dual-use materials. These issues highlight the need for governance frameworks that extend beyond technical optimization to encompass social responsibility and long-term impact [10]. At a deeper level, these debates converge on a fundamental question: whether AI contributes to genuine scientific understanding or merely delivers increasingly accurate predictions without explanatory substance [14].

Table 2 summarizes the foundational assumptions that implicitly structure materials AI workflows, highlighting how each assumption enables progress while introducing characteristic vulnerabilities that motivate current debates.

Table 2. Foundational assumptions in materials AI: enabling value, failure modes, and mitigation levers are discussed in this review

Assumption class	What is assumed (conceptual statement)	Why does it enable progress (claimed value)	Common failure mode/vulnerability (conceptual)	Practical mitigation lever (conceptual)	Key refs
Data representativeness	Historical datasets adequately cover relevant chemical/structural space	Enables generalization beyond seen materials; supports screening and	Dataset bias toward "popular" materials; sparse coverage; noisy measurements → overconfident errors	Data curation, diversity strategies, uncertainty-aware reporting,	[4, 15, 16]

		surrogate prediction		multimodal integration	
Label reliability	DFT/experimental labels are sufficiently "ground-truth" for learning	Makes high-throughput learning feasible; supports learned potentials and scalable screening	Systematic label bias from approximations (e.g., functional sensitivity) propagates into models	Benchmarking across label sources; multi-fidelity validation; explicit uncertainty	[8]
Transferability	Learned representations can transfer across domains/material classes	Reduces data burden; accelerates modeling for low-data regimes	Negative transfer when physics differs; hidden domain shift	Domain-aware transfer learning; physically motivated constraints; careful re-validation	[11, 16]
Efficiency superiority	AI surrogates deliver acceptable accuracy at lower computational cost	Replaces expensive simulation loops; enables exploration at scale	Surrogate drift; "fast wrong answers" dominate decision-making	Hybrid pipelines; selective high-accuracy recalibration; error auditing	[8]
Autonomy viability (SDLs)	AI + robotics + HPC can run closed-loop discovery reliably	Compresses timelines; improves reproducibility; reduces manual bottlenecks	Uncertainty mishandled; safety gaps; reproducibility breaks at interfaces	Human-in-the-loop oversight; safety constraints; robust automation standards	[17, 18]
Correlation-to-insight leap	Correlations can be treated as explanatory or mechanistically meaningful	Encourages interpretation and design decisions based on model outputs	Confusing predictive success with causal explanation → invalid scientific inference	XAI + physics alignment + mechanistic validation; explicit epistemic boundaries	[6, 14]
Open science accelerates progress	Broad data/model sharing improves reproducibility and discovery rate	Democratizes access; increases collective learning	IP constraints, uneven access, misuse risk; fragmented standards	Governance for data sharing; standardized reporting; controlled access where needed	[5, 10]
Responsible scaling is possible	Scaling models and computation can remain ethically and environmentally aligned	Supports large discovery campaigns and rapid innovation	Inequity via compute concentration; sustainability tensions; dual-use concerns	Governance frameworks; equitable infrastructure; resource-aware model design	[10]

Results and Discussion

The conceptual foundations of applied artificial intelligence in materials science, as examined in this review, reveal a field undergoing a profound transformation. AI is no longer confined to a supporting computational role but increasingly shapes how materials knowledge is generated, interpreted, and applied. The definitions, assumptions, and open debates discussed collectively illustrate both the promise and the fragility of this transformation. On one hand, machine learning techniques have demonstrated clear value in managing the combinatorial complexity of materials systems and in accelerating pathways from theoretical insight to practical application [5, 19]. On the other hand, the assumptions underpinning these techniques—particularly those concerning data quality, model generalization, and interpretability—expose structural vulnerabilities that demand critical scrutiny [9, 21].

A central point of discussion concerns the tension between predictive performance and scientific interpretability. Highly accurate black-box models have achieved notable success in tasks such as property prediction and materials screening, yet their opacity obscures the causal mechanisms that underpin scientific understanding [22, 25]. This limitation becomes particularly salient in inverse design workflows, where generative models propose novel structures or compositions without offering clear rationales for their suitability. In the absence of interpretable outputs, researchers may struggle to validate predictions, assess physical plausibility, or refine design strategies [3, 13]. Advocates of XAI argue that interpretability not only enhances trust but also enables knowledge transfer across material domains, thereby amplifying the long-term scientific value of AI systems [6, 17]. Evidence from catalysis research demonstrates that explainable models can uncover non-obvious descriptors—such as subtle electronic structure features—that elude traditional analysis, underscoring the potential of hybrid approaches that integrate domain knowledge with data-driven learning [2, 14].

Assumptions surrounding data representativeness and transferability further shape the reliability of AI-driven materials research. Materials datasets are inherently heterogeneous, with experimental data often affected by noise and inconsistency, and simulation-derived data constrained by methodological approximations [7, 26]. The widespread reliance on density functional theory as a reference standard assumes a level of accuracy that is sensitive to functional choice, introducing systematic biases that can propagate through AI models [19, 24]. Debates over data curation and sharing reflect broader tensions between accelerating discovery through openness and preserving competitive or proprietary interests [27, 28]. Techniques such as transfer learning have shown promise for mitigating data scarcity by leveraging large, general-purpose datasets for specialized tasks, including predicting rare-earth compound properties [9, 20]. Nevertheless, imbalanced datasets risk reinforcing existing research biases, privileging well-studied materials while marginalizing underexplored classes such as biodegradable polymers or low-resource materials systems [15, 23].

The increasing autonomy of AI systems, particularly in self-driving laboratories, introduces additional layers of complexity. These systems assume seamless integration between machine learning algorithms, robotic platforms, and experimental infrastructure. In practice, challenges remain in managing uncertainty, ensuring safety, and maintaining reproducibility when human oversight is reduced [1, 5]. Ethical concerns are amplified by the dual-use potential of accelerated materials design, where advances intended for societal benefit could also facilitate harmful applications [4, 13]. Sustainability considerations further complicate the picture, as the computational demands of large-scale AI models may conflict with the environmental objectives that materials science seeks to advance [14, 16]. In response, emerging proposals for “green AI,” including energy-efficient algorithms and resource-aware computing strategies, aim to align AI development with environmental stewardship [10, 25].

Methodological debates also persist regarding the scalability and efficiency of AI approaches. While machine learning offers clear advantages in high-throughput contexts, its effectiveness diminishes in low-data regimes common to specialized or emerging materials domains [11, 22]. Active learning and Bayesian optimization provide partial solutions by prioritizing informative experiments, as demonstrated in the efficient optimization of thermoelectric materials with minimal experimental input [3, 18]. Nonetheless, questions remain about AI’s capacity to navigate the vast chemical design space without succumbing to combinatorial explosion. Multi-fidelity strategies that combine inexpensive simulations with

targeted high-accuracy experiments have been proposed as a pragmatic compromise [6, 8]. In biologically relevant materials systems, additional challenges arise from the need to integrate multimodal data sources, including images, spectra, and textual annotations, a task that continues to test current modeling paradigms [7, 12].

The broader implications of these developments are substantial. AI has the potential to democratize materials research by lowering barriers to entry and enabling smaller laboratories to compete through access to open tools and shared resources [2, 29]. However, this democratization presupposes digital literacy and equitable access to computational infrastructure, raising concerns about a widening digital divide [13, 30]. Algorithmic bias may further entrench systemic inequalities by prioritizing materials aligned with existing data availability rather than societal need, underscoring the importance of diverse datasets and fairness-aware evaluation metrics [14, 31]. From a societal perspective, AI-enabled materials innovation holds promise for addressing global challenges such as energy storage and climate mitigation, but only if its underlying assumptions are rigorously examined and its ethical dimensions actively governed [21, 32].

In summary, the conceptual foundations of applied AI in materials science underscore its transformative potential while revealing critical tensions that must be addressed. Realizing AI's promise as a genuine scientific instrument—rather than a purely predictive engine—will require sustained interdisciplinary collaboration, standardized validation practices, and robust ethical frameworks. By confronting foundational assumptions and engaging openly with unresolved debates, the materials science community can ensure that AI catalyzes meaningful, responsible, and enduring scientific progress [5, 33].

Conclusion

This review has synthesized the conceptual underpinnings of applied AI in materials science, clarifying definitions, scrutinizing assumptions, and illuminating open debates. AI has revolutionized the field by enabling data-driven discovery, reducing development timelines from years to months. Yet, assumptions about data and models must be critically evaluated to avoid overreliance, while debates on interpretability, ethics, and sustainability guide responsible adoption.

Future directions should focus on developing robust, interpretable models using XAI and physics-informed learning to enhance generalization across scales. Expanding data infrastructure with global, open-access databases will address scarcity, incorporating multimodal and real-time experimental data. Autonomous systems like SDLs will evolve with advanced reinforcement learning, integrating safety protocols and human oversight. To tackle ethical concerns, frameworks for bias mitigation and green computing will be essential, promoting equitable access and environmental alignment. Ultimately, AI's integration promises breakthroughs in sustainable materials, but success hinges on collaborative, interdisciplinary efforts to resolve debates and validate assumptions, paving the way for a new era of innovation.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 16 Mar 2024 Revised: 29 Apr 2024 Accepted: 05 Jun 2024
Published online: 18 July 2024

References

Saidi P, Kaminsky B, Arruda DW. Data-centric learning framework for material-specific inverse design of architected materials. npj Comput Mater. 2023;9:44.

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

Goswami A, Deka B, Chattaraj D. A Review on Applications of Artificial Intelligence in Material Engineering. Adv Eng Mater. 2023;25(12):2300104.

Fung V, Ganesh P, Sumpter BG. Physically informed machine learning prediction of electronic density of states. Chem Mater. 2022;34:1302-13.

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

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.

Chan C H, Sun M, Huang B. Application of machine learning for advanced material prediction and design. EcoMat. 2022;4:e12194.

Batra R, Pilania G, Uberuaga BP, Ramprasad R. Multifidelity information fusion with machine learning: a case study of dopant formation energies in hafnia. ACS Appl Mater Interfaces. 2019;11:24906-18.

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

Liu X, Li Z, Zhang J, Tang J. Material machine learning for alloys: applications, challenges and perspectives. J Alloys Compd. 2022;921:165984.

Wang C, Fu H, Jiang L, Xue D, Xie J. A property-oriented design strategy for high performance copper alloys via machine learning. npj Comput Mater. 2019;5:87.

Lu S, Zhou Q, Guo Y, Zhang Y, Wu Y, Wang J. Graph neural networks accelerated molecular dynamics. J Phys Chem Lett. 2020;11:10568-73.

Reiser P, Neubert M, Eberhard A, Torresi Z, Zhou C, Shao C, et al. Graph neural networks for materials science and chemistry. Commun Mater. 2022;3:93.

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

Hey T, Butler K, Jackson S, Thiyagalingam J. Machine learning and big scientific data. Philos Trans A Math Phys Eng Sci. 2020;378(2166):20190054.

Wen C, Zhang Y, Wang C, Xue D, Bai Y, Antonov S, et al. Machine learning assisted design of high entropy alloys with desired property. Acta Mater. 2019;170:109-17.

Tao Q, Xu P, Li M, Lu W. Machine learning for perovskite materials design and discovery. npj Comput Mater. 2021;7:23.

Oliynyk AO, Antono E, Sparks TD. High-throughput machine-learning-driven synthesis of full-Heusler compounds. Chem Mater. 2016;28:7324-31.

Choudhary K, DeCost B, Cuban C, Behera RK, Saidi WA, Jonker BT, et al. Recent advances and applications of deep learning methods in materials science. npj Comput Mater. 2022;8(1):59.

Batra R, Song L, Ramprasad R. Emerging materials intelligence ecosystems propelled by machine learning. Nat Rev Mater. 2021;6:655-78.

Chen C, Zuo Y, Ye W, Li X, Deng Z, Ong SP. A critical review of machine learning of energy materials. Adv Energy Mater. 2020;10(8):1903242.

Pilania G. Machine learning in materials science: From explainable predictions to autonomous design. Comput Mater Sci. 2021;193:110360.

Ramprasad R, Behler J, Rossi M, Rupp M. Machine learning and energy minimization approaches for crystal structure predictions: A review and new horizons. Chem Mater. 2022;34:5695-709.

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

Hu Y, Sprague IS, Kalia RK, Nakano A, Vashishta P, Ramprasad R. Deep language models for interpretative and predictive materials science. APL Mach Learn. 2023;1(1):010901.

Dan Y, Zhao Y, Li X, Li S, Hu M, Hu J. Generative adversarial networks (GAN) based efficient sampling of chemical composition space for inverse materials design. npj Comput Mater. 2020;6:84.

Liu X, Zhang F, Hou Z, Mian L, Wang Z, Zhang J, et al. Self-supervised learning: Generative or contrastive. IEEE Trans Knowl Data Eng. 2023;35(1):857-76.

Lu S, Zhou Q, Ouyang Y, Guo Y, Li Q, Wang J. Accelerated discovery of superconducting crystals by machine learning. Nat Comput Sci. 2023;3:577-85.

Wen C, Wang C, Zhang Y, Antonov S, Xue D, Lookman T, et al. Modeling materials properties of lead-free solder alloys. J Mater Sci. 2020;55:13475-92.

Dunn A, Wang Q, Ganpule S, Sapkota D, Luo X, Cheng SS, et al. Benchmarking materials property prediction methods: the Matbench test set and Automatminer reference algorithm. npj Comput Mater. 2020;6:138.

Sun S, Ouyang R, Zhang B, Zhang TY. Data-driven discovery of thin-film photovoltaic materials via a Materials Genome approach: discovery of MoBi2(TeSe)4 and MoBiTe4 photocat hodes. J Mater Chem A. 2020;8:18913-23.

Takahashi K, Tanaka Y. Material synthesis and design principles for unconventional superconducting compounds. Phys Rev Mater. 2019;3:053403.

Meredig B. Five high-impact research areas in machine learning for materials science. MRS Commun. 2019;9:1035-9.