REVIEW **Open access**

Data-Driven Materials Engineering: Inverse Design Strategies, Machine Learning Architectures, and Application Domains

Ahmed El-Kholy[1*], Nour Abdelrahman[1], Karim Hassan[2], Mona Saad[1]

Abstract

The advent of data-driven approaches has revolutionized materials engineering, enabling inverse design strategies that prioritize target properties to guide material synthesis and optimization. This review synthesizes recent advancements in machine learning architectures tailored for materials informatics, including graph neural networks and representation learning frameworks that capture atomic-scale interactions and multiscale phenomena. We examine the integration of high-throughput computations with experimental workflows, highlighting closed-loop systems that incorporate active learning and uncertainty quantification to accelerate discovery. Key application domains span energy materials, metamaterials, and catalytic systems, where multimodal datasets facilitate simulation-experiment synergies. By analyzing computational ecosystems, we underscore the shift from forward modeling to inverse paradigms, emphasizing autonomous laboratories that iteratively refine hypotheses through data feedback loops. Challenges in generalizability and data scarcity are contextualized within broader systems integration, offering a cohesive perspective on how these tools reshape materials design. This narrative integrates cross-study insights to propose unified frameworks for scalable, data-centric engineering, bridging theoretical models with practical implementations in computational materials science.

Keywords Autonomous discovery, Materials informatics, Inverse design, Graph neural networks, High-throughput computation, Machine learning architectures

*Correspondence:
Ahmed El-Kholy
ahmed.elkholy@gmail.com
[1] Department of Computational Materials Engineering, Faculty of Engineering, Alexandria University, Alexandria, Egypt
[2] Department of Materials Data Analytics, Faculty of Engineering, Ain Shams University, Cairo, Egypt

Introduction

Materials engineering has traditionally relied on empirical trial-and-error methods, often constrained by the vast chemical space and the complexity of structure-property relationships. The emergence of computational and data-driven strategies marks a paradigm shift, leveraging vast datasets and advanced algorithms to predict, optimize, and discover new materials with unprecedented efficiency [1, 2]. This transformation is rooted in the convergence of high-performance computing, machine learning (ML), and experimental automation, forming an ecosystem where data serves as the cornerstone for innovation [3, 4].

At the heart of this evolution is materials informatics, a discipline that applies information science principles to materials research, enabling the extraction of actionable insights from heterogeneous data sources [2, 5]. Early efforts focused on forward modeling, where known structures are simulated to predict properties using density functional theory (DFT) or molecular dynamics. However, inverse design strategies invert this process: starting from desired functionalities—such as high thermal conductivity

or catalytic activity—and computationally generating candidate materials that meet these criteria [6, 7]. This approach addresses the limitations of exhaustive screening by intelligently navigating design spaces, often incorporating constraints like stability and synthesizability [8, 9].

Machine learning architectures have been pivotal in realizing inverse design. Graph neural networks (GNNs), for instance, represent materials as graphs where nodes denote atoms and edges capture bonds, allowing for permutation-invariant predictions of properties like bandgaps or mechanical responses [7, 10-15]. These models excel in handling crystalline and molecular systems, outperforming traditional descriptors by learning hierarchical features directly from data [7, 12]. Representation learning further enhances this by embedding materials into latent spaces that preserve physicochemical similarities, facilitating transfer learning across datasets [10, 16]. In parallel, high-throughput computation generates large-scale databases, such as those from automated DFT calculations, which fuel ML models for rapid property screening [8, 17].

The integration of simulation and experiment is another critical facet, where data-driven methods bridge the gap between virtual predictions and real-world validation [18-20]. Multimodal datasets, combining computational outputs with experimental measurements, enable robust models that account for uncertainties inherent in both domains [21, 22]. Active learning systems iteratively select informative data points, optimizing resource allocation in discovery campaigns [18, 23]. Uncertainty quantification, often via Gaussian processes or ensemble methods, ensures reliable decision-making by quantifying prediction confidence [21].

Autonomous laboratories represent the pinnacle of this integration, employing robotic systems and closed-loop workflows to automate synthesis, characterization, and refinement [19, 20]. These platforms embody a feedback mechanism where ML models guide experiments, and experimental outcomes update models, accelerating cycles from hypothesis to validation [18, 20, 23]. Application domains illustrate the versatility: in energy materials, data-driven inverse design has expedited the development of batteries and photovoltaics [24]; in metamaterials, ML architectures optimize structures for tailored mechanical behaviors [3, 6]; and in catalysis, representation learning identifies active sites with minimal computational cost [25].

Despite these advances, the field grapples with challenges like data sparsity and model transferability, necessitating innovative strategies for scalable implementation [4, 22]. This review positions itself as a comprehensive synthesis of inverse design strategies, ML architectures, and their applications within computational and data-driven materials engineering. By providing an original integrative framework that structures the literature around workflow ecosystems—encompassing data generation, model architectures, and closed-loop integration—we aim to guide future developments toward more autonomous and efficient materials discovery paradigms. we aim to guide future developments toward more autonomous and efficient materials discovery paradigms **(Figure 1)**.

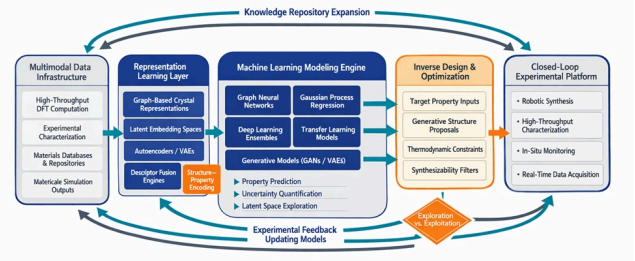

Figure 1. Closed-Loop Ecosystem of Data-Driven Materials Engineering Integrating Multimodal Data, Machine Learning Architectures, and Inverse Design Workflows.

Schematic representation of the integrated data-driven materials engineering ecosystem. Multimodal data infrastructures—comprising high-throughput computations, experimental datasets, and materials repositories—feed representation learning pipelines that encode structural and compositional features into machine-interpretable embeddings. These representations support diverse machine learning architectures, including graph neural networks, probabilistic models, and generative frameworks, which enable property prediction, uncertainty quantification, and latent space exploration. Inverse design engines translate target functionalities into candidate material structures, which are experimentally evaluated within autonomous laboratory platforms. Closed-loop feedback mechanisms iteratively update datasets and models through active learning and exploration–exploitation balancing, forming a self-optimizing discovery system.

Landscape of Computational & Data-Driven Materials Engineering The computational and data-driven materials engineering landscape encompasses a multifaceted ecosystem that integrates informatics, ML, and high-throughput methodologies to address complex design

challenges. This section synthesizes key components, organizing the discussion around data infrastructures, modeling paradigms, and integrative workflows, drawing on recent literature to highlight synergies and advancements [1, 2 , 4, 5].

Data Infrastructures and High-Throughput Computation

Foundational to contemporary data-driven materials engineering is the construction of robust, interoperable data infrastructures capable of sustaining high-throughput computational workflows and multimodal knowledge integration. These infrastructures function not merely as repositories but as epistemic backbones that structure how materials knowledge is generated, validated, and mobilized across discovery pipelines. High-throughput computational paradigms—particularly those anchored in density functional theory (DFT)—enable automated screening of expansive compositional and structural design spaces. Through algorithmically orchestrated simulation loops, thousands to millions of candidate materials can be evaluated for thermodynamic stability, electronic structure, and functional properties, generating statistically rich datasets that support large-scale trend analyses and surrogate modeling [8, 17].

Flagship initiatives such as the Materials Project exemplify this infrastructural paradigm, leveraging distributed computational resources and standardized workflows to curate open, queryable databases of computed materials properties [24]. These platforms have catalyzed a shift from hypothesis-limited experimentation toward data-saturated exploration, enabling machine learning (ML) systems to mine correlations across crystallographic, electronic, and thermodynamic descriptors. Beyond sheer scale, their importance lies in harmonization—standardized calculation protocols, metadata ontologies, and uncertainty annotations ensure interoperability across research groups and downstream ML pipelines.

The scope of data infrastructures further expands through multimodal dataset integration. Computational outputs are increasingly fused with experimental modalities—including spectroscopy, diffraction, and high-resolution microscopy—to produce composite representations that more faithfully capture real-world material complexity [4, 16]. Such multimodal convergence mitigates simulation-experiment gaps, embedding microstructural defects, synthesis conditions, and measurement noise into training corpora. The resulting datasets enable models to learn not only idealized quantum-mechanical behaviors but also processing-dependent variabilities, thereby enhancing translational reliability in applied discovery settings.

Representation Learning in Materials Data Ecosystems

Representation learning constitutes the critical translational layer that transforms heterogeneous raw data into structured, machine-interpretable formats. Rather than relying on manually engineered descriptors, modern approaches learn latent embeddings directly from structural and compositional inputs. Autoencoders, variational autoencoders (VAEs), and related generative compression architectures encode high-dimensional materials data into compact latent manifolds, preserving salient physicochemical features while enabling efficient similarity mapping, clustering, and anomaly detection [10, 12].

These latent spaces function as navigable cartographies of materials design domains, where geometric proximity reflects structural or functional relatedness. Such embeddings support accelerated screening by enabling nearest-neighbor searches, interpolation across compositional gaps, and density-aware sampling strategies.

In crystalline materials systems, graph-based representations have emerged as especially powerful. By encoding atoms as nodes and bonding environments as edges, graph formalisms naturally capture relational chemistry. Adjacency matrices and edge feature tensors represent coordination environments, bond lengths, and electronic interactions, allowing models to learn structure–property relationships invariant to translational, rotational, or permutational symmetries [7, 11, 12].

Graph representations have demonstrated broad applicability—from oxide perovskites to high-entropy alloys—where they consistently outperform traditional fingerprint descriptors in predictive tasks [9, 14]. Their advantage lies in hierarchical message passing: local bonding information propagates across graph layers to produce global material embeddings, enabling simultaneous learning of short-range chemistry and long-range structural order.

Machine Learning Architectures for Materials Informatics

Machine learning architectures tailored to materials science constitute the analytical core of data-driven discovery ecosystems. Among these, graph neural networks (GNNs) have emerged as a dominant modeling paradigm due to their capacity to operate on non-Euclidean data structures inherent to crystalline and molecular systems [10-15].

Through iterative message-passing operations, GNNs aggregate and transform node-level features across graph neighborhoods, constructing progressively abstract representations that culminate in global property predictions. Architectural variants such as crystal graph convolutional neural networks (CGCNNs) incorporate edge attributes—bond types, distances, coordination numbers—thereby enhancing resolution in predicting formation energies, elastic tensors, and electronic properties [7, 12, 14].

Transfer learning further amplifies the utility of these architectures. Models pretrained on large computational repositories can be fine-tuned for niche material classes or experimentally constrained datasets, enabling high predictive performance even under data scarcity [10]. This paradigm is particularly valuable for emerging materials domains where generating new ab initio data remains computationally prohibitive.

Complementary modeling strategies broaden the architectural landscape. Gaussian process regression (GPR) offers probabilistic prediction frameworks well suited to small datasets, embedding uncertainty directly into predictive outputs—an essential feature for risk-sensitive discovery planning [21]. Deep ensemble methods, which aggregate predictions across multiple neural networks, enhance robustness and mitigate overfitting in high-dimensional inference regimes [4, 22]. The diversity of machine learning architectures and their discovery functions are summarized in **Table 1**.

Table 1. Machine Learning Architectures in Data-Driven Materials Engineering: Functions, Data Inputs, and Discovery Roles.

Architecture Class	Core Function	Input Data Modalities
Graph Neural Networks (GNNs)	Structure–property prediction via graph message passing	Crystal structures, atomic graphs
Crystal Graph CNNs	Bond-aware graph convolution modeling	Bond distances, coordination environments
Gaussian Process Regression	Probabilistic regression with uncertainty estimates	Small experimental/computational datasets
Deep Learning Ensembles	Aggregated neural predictions	Multimodal materials datasets
Variational Autoencoders	Latent space generative modeling	Structural descriptors, compositions
Generative Adversarial Networks	Synthetic structure generation	Latent embeddings, structure datasets
Transfer Learning Models	Knowledge reuse across domains	Pretrained computational databases

In inverse design contexts, generative architectures—including VAEs and generative adversarial networks (GANs)—enable exploration of latent design spaces to propose novel material candidates [5, 6]. When constrained by physical priors or thermodynamic feasibility filters, these models transition from purely statistical generators to physics-aware design engines. Benchmark applications, such as bandgap optimization and optoelectronic material discovery, demonstrate how ML surrogates compress optimization timelines relative to brute-force simulation loops [8, 9, 24].

Inverse Design Strategies and Application Domains

Inverse design represents a paradigm inversion within materials engineering workflows. Rather than predicting properties from known structures, ML systems infer candidate structures that satisfy predefined functional targets [2, 6]. This property-to-structure mapping reframes materials discovery as a search problem within learned latent manifolds.

The potency of inverse design is particularly evident in architected and metamaterial systems. Here, data-driven topological optimization enables the creation of structures with unconventional mechanical responses, including negative Poisson's ratios and programmable anisotropy [3, 6]. By navigating high-dimensional design spaces inaccessible to analytical methods, ML systems uncover non-intuitive structural motifs that satisfy complex performance constraints.

Energy materials constitute another critical application frontier. In solid-state electrolyte discovery, inverse frameworks integrate ML screening with thermodynamic stability analyses to identify ion-conductive compounds across vast compositional chemistries [24]. Similarly, catalytic materials design benefits from surface-reaction learning models that optimize adsorption energetics and active-site geometries for processes such as CO oxidation [25].

Beyond bulk materials, inverse approaches extend to nanoscale and atomistic modeling domains. Machine-learned interatomic potentials—trained on ab initio datasets—enable large-scale molecular dynamics simulations with near-quantum accuracy [26, 27]. Neural network and kernel-based force fields capture anharmonic lattice dynamics, defect migration, and phase transformations at spatiotemporal scales inaccessible to first-principles calculations alone.

Scalability remains central across these domains. High-throughput infrastructures allow parallel candidate evaluation, while active learning frameworks iteratively refine search trajectories by prioritizing high-uncertainty regions in design space [4, 18]. This uncertainty-aware sampling maximizes informational gain per experiment or simulation cycle.

Simulation–Experiment Integration

A defining maturation marker of data-driven materials engineering is the progressive convergence of computational simulations and experimental validation within unified discovery architectures. Rather than operating as sequential silos, simulations and experiments now form bidirectional feedback systems that iteratively refine hypotheses and predictive models [18-20].

Autonomous laboratories exemplify this integration. Robotic synthesis and characterization platforms execute ML-guided experimental protocols, generating real-time validation data that feed back into model retraining loops [19, 20]. These cyber-physical systems operationalize closed-loop discovery, compressing the latency between prediction and validation.

Uncertainty quantification plays a pivotal coordinating role within these hybrid workflows. Bayesian inference frameworks propagate simulation uncertainties into experimental planning layers, enabling principled balancing between exploratory and exploitative experimentation [21, 23]. This ensures that experimental resources are allocated not merely to promising candidates but to epistemically informative ones.

Such simulation–experiment coupling has accelerated discovery in domains including thin-film photovoltaics, battery electrode materials, and catalytic interfaces, demonstrating measurable gains in optimization efficiency and discovery throughput [17-19].

Autonomous and Closed-Loop Discovery Systems

Autonomous and closed-loop discovery systems represent the operational apex of data-driven materials engineering infrastructures. These systems embody fully integrated platforms that automate iterative cycles of hypothesis generation, candidate evaluation, experimental execution, and model refinement [18-20, 23].

At their architectural core lie self-optimizing ML engines that dynamically update predictive models as new data streams emerge. Decision policies—often grounded in reinforcement learning or Bayesian optimization—steer

experimental selection toward regions of maximal performance potential or epistemic uncertainty.

Operationally, these platforms minimize human intervention while maximizing discovery throughput. Robotic synthesis modules, high-throughput characterization instruments, and adaptive ML controllers function as interoperable subsystems within unified cyber-physical ecosystems [1, 4]. Feedback latency is reduced from months to hours, fundamentally altering the tempo of materials innovation.

From a systems perspective, autonomous discovery infrastructures transform materials engineering from a linear workflow into a recursive learning organism—one capable of continuous self-improvement as it navigates vast compositional and structural landscapes.

Overall, the convergence of scalable data infrastructures, representation learning, advanced ML architectures, inverse design engines, and autonomous experimentation signals a paradigmatic transition in materials science. Data-driven systems no longer merely assist human discovery—they actively reconfigure its epistemic architecture, enabling accelerated, adaptive, and increasingly self-directed materials innovation [1, 3, 5].

Core architectures and machine learning integration

At the foundation of autonomous systems are ML architectures that drive decision-making in closed loops. GNNs and related models process incoming data to predict outcomes, while active learning algorithms select subsequent experiments based on information gain [10, 11, 18, 23]. Bayesian optimization, often coupled with Gaussian processes, formalizes this as an acquisition function that balances exploration (sampling uncertain areas) and exploitation (refining promising candidates) [21]. For instance, in materials synthesis, these architectures integrate real-time characterization data to update surrogate models, enabling adaptive parameter tuning [18, 20].

Closed-loop systems extend this by incorporating hardware automation, such as robotic synthesizers and in-situ analyzers, forming a feedback ecosystem [19, 20]. Representation learning ensures data from diverse sources—computational simulations, experimental spectra—is harmonized into unified embeddings, facilitating multimodal fusion [12, 16]. Uncertainty quantification is integral, with ensemble methods providing variance estimates that inform loop iterations, preventing premature convergence on suboptimal solutions [21, 22].

Operational workflows in discovery campaigns

Operational workflows in these systems follow a cyclical structure: data acquisition, model updating, hypothesis generation, and execution. High-throughput computation initializes the loop by generating baseline datasets, which ML models use to propose initial candidates [8, 17]. Autonomous laboratories then execute syntheses, with sensors providing immediate feedback to refine predictions [19, 20]. Active learning optimizes this by querying points that maximize model improvement, as demonstrated in accelerated thin-film discovery where loops reduced evaluation needs by orders of magnitude [18, 19].

In inverse design contexts, workflows invert property targets into actionable recipes, using generative models to explore design spaces [5, 6]. For metamaterials, this involves optimizing structural parameters via ML surrogates integrated with experimental validation loops [3, 6]. Energy materials discovery similarly benefits, with closed loops identifying stable electrolytes through iterative DFT-experiment cycles [24]. Nanocluster studies illustrate scalability, where ML force fields enable rapid simulations within loops for property optimization [26, 27].

Systems integration and application examples

Integration of simulation and experiment is paramount, with closed-loop systems bridging scales from atomic to macroscopic. Platforms like self-driving labs exemplify this, employing ML to orchestrate robotic workflows for materials like organic thin films, achieving discovery rates far exceeding manual methods [19, 20]. In catalytic design, loops combine computational screening with experimental testing, refining models for reactions like preferential CO oxidation [25]. Uncertainty-aware integration ensures robustness, using probabilistic frameworks to handle experimental noise [21, 22].

Application domains underscore the versatility: in battery materials, autonomous systems accelerate electrode optimization by closing loops between computation and electrochemical testing [17, 24]; in pharmaceuticals or

polymers, similar setups enable property-driven design [4, 5]. Cross-study analysis reveals common themes, such as the need for modular architectures that allow swapping ML components for task-specific adaptations [10, 16, 23].

Figure 1: Schematic of a closed-loop discovery system. The diagram illustrates a cyclical workflow beginning with a multimodal dataset repository (left), feeding into a central ML module comprising graph neural networks for representation learning and Gaussian processes for uncertainty quantification. Arrows depict data flow: from dataset to model training, then to active learning for experiment selection (top), interfacing with an autonomous laboratory (right) for synthesis and characterization. Feedback loops return experimental outcomes to update the dataset and model, with a decision node balancing exploration-exploitation via acquisition functions. Peripheral elements include high-throughput computation inputs and inverse design constraints, emphasizing systems integration.

These systems not only enhance efficiency but also foster innovation by enabling exploration of uncharted material spaces, positioning them as transformative tools in computational materials engineering [1-3].

Results and Discussion

Challenges and limitations

While data-driven materials engineering has advanced rapidly, several challenges persist that limit its full potential, particularly in inverse design, ML architectures, and integrated systems. This discussion synthesizes these hurdles through a systems-level lens, emphasizing computational workflows and cross-study insights to provide an original framework for understanding barriers in scalability, reliability, and integration [1, 2, 4, 22].

Data scarcity and quality remain foundational issues in materials informatics. High-throughput computations generate vast datasets, but these often suffer from biases toward stable, low-energy structures, underrepresenting metastable or novel materials critical for inverse design [8, 17, 24]. Multimodal datasets exacerbate this, as experimental data is sparse and noisy compared to simulations, leading to mismatches that hinder model training [4, 16]. For instance, in energy materials, discrepancies between DFT-predicted and measured properties arise from unaccounted environmental factors, compromising inverse strategies that rely on accurate mappings [24]. Representation learning attempts to mitigate this by creating unified embeddings, but incomplete graphs in GNNs—due to missing long-range interactions—can propagate errors across scales [10-12].

Model generalizability poses another significant challenge, especially for ML architectures applied across domains. GNNs excel in crystalline systems but struggle with amorphous or disordered materials where graph representations fail to capture entropy-driven behaviors [7, 10, 14]. Transfer learning frameworks aim to address this, yet domain shifts between datasets (e.g., from bulk to nanoscale) reduce efficacy, as seen in nanocluster force fields where trained potentials falter on unseen morphologies [10, 26, 27]. Uncertainty quantification helps by flagging out-of-distribution predictions, but methods like Gaussian processes scale poorly with dataset size, limiting their use in high-throughput contexts [21, 22]. Active learning systems offer partial solutions by focusing on informative samples, but in closed-loop setups, they can entrench biases if initial data is unrepresentative [18, 23].

Integration of simulation and experiment introduces workflow complexities. Autonomous laboratories streamline discovery, but hardware limitations—such as synthesis precision or characterization throughput—create bottlenecks that ML models cannot fully compensate [19, 20]. Closed-loop systems rely on seamless data flow, yet interoperability issues between computational tools (e.g., DFT software) and experimental platforms lead to delays or inaccuracies [18, 19]. In inverse design for metamaterials, this manifests as challenges in translating optimized virtual structures to fabricable prototypes, where manufacturing constraints are inadequately incorporated [3, 6]. Moreover, ethical and resource considerations arise: data-driven approaches democratize access but require substantial computational power, raising sustainability concerns in large-scale deployments [1, 5].

Scalability in application domains further highlights limitations. In catalysis, ML architectures predict active sites effectively for simple reactions but falter in multifaceted systems involving transient states [25]. Similarly, for energy materials, inverse strategies accelerate screening but overlook lifecycle factors like degradation, necessitating broader uncertainty frameworks [24]. Cross-study analysis reveals a common thread: over-reliance on supervised learning paradigms, which demand labeled data that is

costly to acquire, versus unsupervised or self-supervised alternatives that remain underexplored [4, 5, 16].

Addressing these challenges requires reframing them within an ecosystem perspective, where limitations are not isolated but interconnected—data quality affects model robustness, which in turn impacts closed-loop efficiency [2, 22]. This integrative view underscores the need for hybrid workflows that combine physics-informed constraints with data-driven flexibility, ensuring resilient systems amid real-world variabilities [9, 13].

Future research directions

Future trajectories in computational and data-driven materials engineering should prioritize enhancing inverse design robustness, evolving ML architectures, and deepening systems integration to overcome current limitations. This section outlines prospective avenues, synthesized from literature trends to propose an original roadmap centered on adaptive ecosystems and interdisciplinary convergence [1-5].

Advancing representation learning and ML architectures is paramount. Next-generation GNNs could incorporate dynamic graphs that evolve with simulation timesteps, better suiting time-dependent phenomena in catalysis or nanoclusters [11, 12, 25, 26]. Hybrid models blending GNNs with Gaussian processes might embed physical symmetries directly, improving generalizability and reducing data needs [7, 14, 21]. In inverse design, generative architectures could integrate multi-objective optimization, simultaneously targeting properties like stability and performance while enforcing synthesizability constraints [6, 8, 9]. Exploration of self-supervised learning on unlabeled multimodal datasets promises to alleviate data scarcity, enabling pre-training on vast computational repositories before fine-tuning on experimental subsets [4, 16, 24].

Closed-loop and autonomous systems offer fertile ground for innovation. Future developments could focus on meta-learning frameworks where systems learn to optimize their own workflows, adapting active learning strategies in real-time based on campaign progress [18, 22, 23]. Integration of edge computing in laboratories would enable on-device ML inference, reducing latency in feedback loops and supporting distributed discovery networks [19, 20]. Uncertainty quantification could evolve toward hierarchical models that propagate errors across simulation-experiment interfaces, informing risk-aware decision-making in high-stakes domains like energy materials [21, 22]. High-throughput computation might incorporate quantum computing hybrids for intractable problems, accelerating inverse searches in vast chemical spaces [5, 17].

Application domains will drive domain-specific advancements. In metamaterials, future inverse strategies could leverage reinforcement learning for iterative design refinement, incorporating real-time fabrication feedback [3, 6]. For catalytic systems, ML could predict ensemble behaviors under operando conditions, bridging static computations with dynamic experiments [25]. Nanoscale engineering might see ML force fields extended to reactive potentials, enabling closed-loop discovery of functional nanostructures [26, 27]. Broader ecosystems could foster open-source multimodal platforms, standardizing data formats to enhance collaboration and transferability across fields [2, 16].

A key direction is fostering human-AI symbiosis, where ML augments rather than replaces expertise, through interpretable architectures that explain design rationales [10, 13]. Sustainability integration—optimizing for low-carbon computations—and ethical guidelines for data usage will ensure responsible scaling [1, 5]. This roadmap envisions a unified framework where data-driven tools evolve into adaptive, intelligent systems, accelerating materials innovation through iterative, ecosystem-wide enhancements [2, 4].

Conclusion

In synthesizing the landscape of computational and data-driven materials engineering, this review highlights the transformative role of inverse design strategies, ML architectures like GNNs, and closed-loop systems in reshaping discovery paradigms. From materials informatics to autonomous laboratories, these tools integrate high-throughput computation with experimental workflows, leveraging active learning and uncertainty quantification to navigate complex design spaces. Application domains—encompassing metamaterials, energy systems, and catalysis—demonstrate practical impacts, where data-driven approaches accelerate innovation while addressing multifaceted challenges.

Despite limitations in data quality, model generalizability, and systems integration, future directions point toward hybrid, adaptive frameworks that promise greater efficiency

and scalability. By providing an original synthesis that structures the field around workflow ecosystems, this narrative underscores the shift toward autonomous, intelligent materials engineering, poised to unlock unprecedented advancements in science and technology.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 01 Jan 2023 Revised: 03 May 2023 Accepted: 03 Jul 2023
Published online: 18 September 2023

References

Himanen L, Geurts A, Foster AS, Rinke P. Data-Driven materials science: Status, challenges, and perspectives. Adv Mater. 2017;29(43).

Ramprasad R, Batra R, Pilania G, Mannodi-Kanakkithodi A, Kim C. Machine learning in materials informatics: Recent applications and prospects. npj Comput Mater. 2017;3(1).

Lee D, Choi J, Lee S, Mitchell B, Lee JH. Data-Driven Design for Metamaterials and Multiscale Systems: A Review. Advanced Materials. 2023;35(48). https://doi.org/10.1002/adma.202305254.

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).

Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A. Machine learning for molecular and materials science. Nature. 2018;559(7715).

Ha CS, Yao D, Xu Z, Liu C, Liu H, Elkins D, et al. Rapid inverse design of metamaterials based on prescribed mechanical behavior through machine learning. Nat Commun. 2023;14(1).

Xie T, Grossman JC. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. Phys Rev Lett. 2018;120(14).

Schmidt J, Shi J, Borlido P, Chen L, Botti S, Marques MAL. Predicting the thermodynamic stability of solids combining density functional theory and machine learning. Chem Mater. 2017;29(12).

Pilania G, Gubernatis JE, Lookman T. Multi-fidelity machine learning models for accurate bandgap predictions of solids. Comput Mater Sci. 2017;129.

Gupta V, Choudhary K, DeCost B, Tavazza F, Campbell C, Liao W-k, et al. Structure-aware graph neural network based deep transfer learning framework for enhanced predictive analytics on diverse materials datasets. npj Comput Mater. 2023;9(1).

Reiser P, Neubert M, Eberhard A, Torresi L, Zhou C, Shao C, et al. Graph neural networks for materials science and chemistry. Commun Mater. 2022;3(1).

Chen C, Ye W, Zuo Y, Zheng C, Ong SP. Graph networks as a universal machine learning framework for molecules and crystals. Chem Mater. 2019;31(9):3564-72.

Schütt KT, Sauceda HE, Kindermans PJ, Tkatchenko A, Müller KR. SchNet – A deep learning architecture for molecules and materials. J Chem Phys. 2018;148(24).

Park CW, Wolverton C. Developing an improved crystal graph convolutional neural network framework for accelerated materials discovery. Phys Rev Mater. 2020;4(6).

Louis SY, Zhao Y, Nasiri A, Wang X, Song Y, Liu F, et al. Graph convolutional neural networks with global attention for improved materials property prediction. Phys Chemistry Chem Phys. 2020;22(32).

Hatakeyama-Sato K, Tezuka T, Ujihira T, Oyaizu K. Integrating multiple materials science projects in a single neural network. Commun Mater. 2020;1(1).

Merchant A, Batzner S, Schoenholz SS, Aykol M, Cheon G, Cubuk ED. Scaling deep learning for materials discovery. Nature. 2023;624(7991).

Kusne AG, Yu H, Wu C, Zhang H, Hattrick-Simpers J, DeCost B, et al. On-the-fly closed-loop materials discovery via Bayesian active learning. Nat Commun. 2020;11(1).

MacLeod BP, Parlane FGL, Morrissey TD, Häse F, Roch LM, Dvorak KE, et al. Self-driving laboratory for accelerated discovery of thin-film materials. Sci Adv. 2020;6(20).

Szymanski NJ, Rendy B, Fei Y, Merckx R, Tan S, Zeng Y, et al. An autonomous laboratory for the accelerated synthesis of novel materials. Nature. 2023;624(7990).

Deringer VL, Bartók AP, Bernstein N, Wilkins DM, Ceriotti M, Csányi G. Gaussian process regression for materials and molecules. Chem Rev. 2021;121(16).

Li K, DeCost B, Choudhary K, Greenwood M, Hattrick-Simpers J. A critical examination of robustness and generalizability of machine learning prediction of materials properties. npj Comput Mater. 2023;9(1).

Kavalsky L, Hegde VI, Muckley E, Johnson MS, Meredig B, Viswanathan V. By how much can closed-loop frameworks accelerate computational materials discovery? Digit Discov. 2023;2(2).

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).

Karamad M, Magar R, Shi Y, Siahrostami S, Gates ID, Farhad S. Catalytic activity and stability over nanorod and octahedron-shaped CeO2 supported Pt catalysts for preferential oxidation of CO (PROX). ACS Appl Mater Interfaces. 2020;12(28).

Zeni C, Rossi K, Glielmo A, Fekete Á, Gaston N, Baletto F, et al. Building machine learning force fields for nanoclusters. J Chem Phys. 2018;148(23).

Glielmo A, Sollich P, De Vita A. Accurate interatomic force fields via machine learning with covariant kernels. Phys RevB. 2017;95(21).