REVIEW Open access

From High-Throughput Computation to Autonomous Discovery: A Review of Closed-Loop Data Infrastructures in Materials Engineering

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

Abstract

The field of materials engineering has undergone a profound transformation through the integration of high-throughput computation and data-driven methodologies, evolving from traditional trial-and-error approaches to sophisticated closed-loop systems that accelerate discovery. This review synthesizes recent advancements in computational and data-driven materials ecosystems, focusing on the infrastructure enabling autonomous discovery. Key elements include materials informatics platforms that leverage machine learning for property prediction and inverse design, graph neural networks for representation learning, and high-throughput computational workflows that generate multimodal datasets. We examine the progression from static high-throughput screening to dynamic, closed-loop paradigms incorporating active learning, uncertainty quantification, and simulation-experiment integration. Autonomous laboratories represent a pinnacle of this evolution, where AI orchestrates iterative cycles of hypothesis generation, experimentation, and refinement. The synthesis highlights how these infrastructures bridge computational predictions with experimental validation, fostering inverse materials design and optimizing resource allocation in complex chemical spaces. Challenges in data interoperability and model generalizability are noted, alongside prospects for scalable, self-optimizing systems. Overall, this review positions closed-loop data infrastructures as foundational to next-generation materials engineering, promising accelerated innovation in areas like energy storage, catalysis, and structural materials. By integrating diverse literature, we provide a systems-level perspective on how these tools are reshaping the discovery landscape.

Keywords Autonomous discovery, Materials informatics, Graph neural networks, Machine learning in materials science, High-throughput computation, Closed-loop systems

*Correspondence:
Ravi Kumar
ravi.kumar@outlook.com
[1] Department of Materials Data Engineering, Faculty of Engineering, IIT Delhi, New Delhi, India
[2] Department of Computational Materials Systems, Faculty of Engineering, IIT Bombay, Mumbai, India

Introduction

Materials engineering has historically relied on empirical methods and intuition-driven experimentation, often resulting in protracted development cycles for new materials with targeted properties. The advent of computational techniques in the late 20th century marked a significant shift, enabling predictive simulations based on quantum mechanics and continuum models. However, these early computational approaches were limited by computational expense and the vastness of materials space, which encompasses an estimated 10^100 possible compositions and structures [1]. The integration of high-throughput computation in the early 2000s, facilitated by advances in density functional theory (DFT) and automated workflows, allowed for the rapid screening of thousands of candidate materials [2, 3]. This paradigm, often termed the "materials genome initiative," emphasized the creation of

large-scale databases through systematic calculations, such as those in the Automatic FLOW for Materials Discovery (AFLOW) or the Materials Project [4, 5].

The rise of data-driven approaches further accelerated this evolution, incorporating machine learning (ML) to extract patterns from computational and experimental data. Materials informatics emerged as a subdiscipline, applying informatics principles to materials science challenges [1, 2, 6]. Early applications focused on regression models for property prediction, using descriptors like atomic fingerprints or composition-based features to surrogate expensive simulations [3, 7]. For instance, kernel ridge regression and random forests were employed to predict band gaps, elastic moduli, and formation energies with accuracies rivaling DFT but at fractions of the computational cost [1-3]. This data-centric shift was propelled by the exponential growth in materials data, from high-throughput experiments and simulations, creating repositories that now exceed petabytes in scale [5, 6].

A pivotal development has been the incorporation of artificial intelligence (AI) infrastructures that not only predict but also guide discovery. Graph neural networks (GNNs) have become instrumental in representation learning, capturing the topological and chemical intricacies of materials structures [8, 9]. Unlike traditional feature engineering, GNNs learn hierarchical representations directly from atomic graphs, enabling transferable models across diverse materials classes [8, 9]. This has facilitated inverse design, where desired properties are input to generate candidate structures, inverting the conventional forward mapping [10, 11]. Generative models, such as variational autoencoders and generative adversarial networks (GANs), have expanded this capability by sampling chemical spaces efficiently, proposing novel compounds beyond existing databases [12].

The motivation for closed-loop data infrastructures stems from the need to address the inefficiencies in isolated computational or experimental silos. Traditional high-throughput computation generates vast data but lacks feedback mechanisms to refine searches adaptively [2, 3]. In contrast, closed-loop systems integrate prediction, experimentation, and learning in iterative loops, mimicking biological evolution or reinforcement learning paradigms [13]. This infrastructure enables autonomous discovery, where AI agents prioritize experiments based on uncertainty or novelty, optimizing exploration-exploitation trade-offs [6, 9, 12]. Such systems are particularly vital in high-stakes applications, like battery materials or catalysts, where rapid iteration can yield breakthroughs [5].

The scope of this review encompasses computational and data-driven ecosystems in materials engineering, with an emphasis on the transition from high-throughput tools to autonomous, closed-loop frameworks. We synthesize literature on key components: data ecosystems, representation learning, property prediction, inverse design, and multimodal integration. Priority is given to works demonstrating infrastructure-level advancements, such as AI-orchestrated workflows and simulation-experiment couplings [1, 4-15]. The major infrastructural layers enabling this transition from high-throughput computation to autonomous discovery are synthesized in **Table 1**.

Table 1. Core Infrastructural Components of Closed-Loop Data Ecosystems in Materials Engineering

Infrastructural Layer	Core Functions	Enabling Technologies	Discove Role
High-Throughput Computation Platforms	Automated property calculations; large-scale materials screening	DFT workflows, AFLOW, Materials Project pipelines	Genera foundatic datasets model trai
Materials Data Ecosystems	Aggregation, curation, dissemination of multimodal datasets	FAIR data schemas, NOMAD, ontology frameworks	Enabl interopera and larg scale benchmar
Representation Learning Architectures	Encode structural and chemical features into machine-readable formats	Graph neural networks, SOAP, Coulomb matrices	Captur structur proper relationsl
AI Property Prediction Systems	Rapid surrogate modeling of materials properties	Deep learning, Gaussian processes, ensemble ML	Replac expensi simulatic guide screeni
Inverse Design Frameworks	Generate materials	GANs, VAEs, latent	Accelera targete

	from target properties	optimization models	materia discove
Multimodal Integration Pipelines	Fuse simulation + experimental + literature data	Transfer learning, graph fusion models	Improv predicti robustne
Active Learning Engines	Select high-value experiments iteratively	Bayesian optimization, acquisition functions	Optimiz explorati exploitat
Autonomous Laboratory Systems	Execute AI-guided synthesis and testing	Robotics, automated reactors, sensing platforms	Enable s driver discove cycles

This review positions closed-loop data infrastructures as the nexus of computational materials engineering, bridging theoretical predictions with practical realization. By providing an original synthesis of interconnected workflows, it highlights how these systems are democratizing materials discovery, reducing reliance on expert intuition, and paving the way for AI-augmented laboratories [5, 6].

Landscape of Computational & Data-Driven Materials Engineering

Materials data ecosystems

The foundation of data-driven materials engineering lies in robust data ecosystems that aggregate, curate, and disseminate multimodal datasets from computational simulations and experiments. High-throughput computation has been instrumental in populating these ecosystems, generating extensive libraries of material properties through automated DFT calculations [2, 3, 5, 15]. For example, frameworks like the Materials Project and NOMAD repository compile thermodynamic, electronic, and mechanical data for millions of compounds, enabling community-wide access and benchmarking [1-3]. These ecosystems extend beyond static repositories to include dynamic pipelines that integrate real-time data ingestion from diverse sources, such as X-ray diffraction, spectroscopy, and microscopy [5, 6, 13]. Multimodal datasets, combining structural, compositional, and functional attributes, are critical for holistic modeling, as they capture interdependencies often overlooked in unimodal approaches [4, 6, 7].

Data quality and interoperability pose key considerations in these ecosystems. Standardization efforts, such as ontology-based schemas and FAIR (Findable, Accessible, Interoperable, Reusable) principles, facilitate seamless data exchange across platforms [1, 6, 13]. Machine learning pipelines benefit from such structured data, allowing for meta-analysis and transfer learning across materials classes [2, 3, 5]. Recent advancements emphasize uncertainty quantification in data generation, where Bayesian methods assess confidence in computational predictions, informing downstream model reliability [6, 9, 12]. This infrastructure supports the scalability of data-driven discovery, transforming disparate datasets into cohesive knowledge bases that fuel AI algorithms [8, 10, 14, 15].

Representation learning architectures

Effective representation learning is essential for translating complex materials structures into machine-readable formats that capture physicochemical essence. Traditional descriptors, like Coulomb matrices or SOAP (Smooth Overlap of Atomic Positions), have evolved into learned representations via deep neural networks [4-7]. Graph neural networks stand out for their ability to model atoms as nodes and bonds as edges, propagating features through message-passing mechanisms [4, 7-9]. For instance, atomistic line graph neural networks enhance this by incorporating bond-centric features, improving predictions for properties like adsorption energies and vibrational frequencies [8]. Global attention mechanisms further refine these architectures, allowing models to weigh distant interactions in extended structures [7].

Representation learning also addresses the challenge of transferability across chemical spaces. Physics-inspired representations, incorporating symmetries like rotational invariance and periodicity, ensure models generalize beyond training data [6]. This is particularly relevant for 2D materials and nanostructures, where multiscale hierarchies require adaptive embeddings [15]. Hybrid approaches blending DFT-derived features with learned ones offer a balance between interpretability and performance, enabling inverse mapping from properties to structures [10-12]. Overall, these architectures form the backbone of data-driven infrastructures, enabling efficient navigation of high-dimensional materials spaces [1-3, 6, 9].

AI-driven property prediction

AI has revolutionized property prediction in materials engineering, surrogating computationally intensive simulations with fast, accurate models. Machine learning techniques, from Gaussian processes to deep learning, predict a wide array of properties, including dielectric constants, thermal conductivities, and mechanical strengths [1-3, 11]. In solid-state materials, ML integrates with DFT to accelerate screenings, as seen in models for perovskite stability and superconductor critical temperatures [3, 5]. Benchmarking studies highlight GNNs' superiority for crystal property prediction, outperforming traditional ML on datasets like MatBench [9].

Uncertainty quantification enhances prediction reliability, using ensemble methods or Bayesian neural networks to estimate epistemic and aleatoric uncertainties [6, 9, 13]. This informs active learning strategies, where models query data points that maximize information gain [12]. For microstructure-dependent properties, convolutional networks process image-based inputs, linking processing conditions to performance [11, 15]. These predictive frameworks are integral to data infrastructures, providing rapid feedback loops that guide experimental prioritization [4, 7, 8, 10, 14].

Inverse design frameworks

Inverse design inverts the property-structure relationship, generating materials tailored to specific applications through optimization in latent spaces. Generative models like GANs efficiently sample composition spaces, proposing inorganic compounds with desired band gaps or elastic moduli [10, 12]. High-entropy alloys and ceramics exemplify this, where ML discovers stable phases in multicomponent systems previously intractable [5]. Representation learning underpins these frameworks, with autoencoders compressing structures into continuous manifolds for gradient-based optimization [4, 6].

Integration with high-throughput computation validates inverse proposals, coupling generative sampling with DFT refinement [2, 3, 10, 15]. Active learning refines design spaces iteratively, balancing exploration of novel regions with exploitation of promising candidates [9, 12, 13]. This infrastructure accelerates discovery in energy materials, such as electrolytes or photocatalysts, by focusing resources on high-potential candidates [1, 5, 11]. Challenges in scalability are addressed through hierarchical designs, starting from coarse-grained models and refining atomistically [7, 8].

Multimodal integration

Multimodal integration fuses data from simulations, experiments, and literature to create comprehensive models. This involves aligning disparate modalities, such as combining DFT energies with experimental spectra via transfer learning [5, 6, 13, 15]. Graph-based frameworks facilitate this by embedding multimodal features into unified graphs, enabling joint predictions [4, 6-8]. For instance, in 2D materials growth, multiscale models integrate quantum calculations with macroscopic simulations [15].

Simulation-experiment integration closes the feedback loop, using ML to calibrate models against real-world data and mitigate systematic errors [1, 3, 13]. Uncertainty-aware fusion ensures robust inferences, prioritizing modalities based on confidence [6, 9, 12]. These integrated infrastructures support end-to-end discovery pipelines, from virtual screening to lab validation [2, 10, 11, 14].

Autonomous & closed-loop discovery systems

Autonomous discovery systems represent the culmination of computational and data-driven infrastructures, enabling self-sustaining cycles of materials innovation without constant human intervention. These systems build on high-throughput computation by incorporating feedback mechanisms that adaptively refine searches [1-3]. Self-driving laboratories exemplify this, where robotic platforms execute experiments guided by AI algorithms, iterating on design hypotheses in real time [12, 13]. Key to their operation is the closed-loop architecture, which links prediction models with experimental actuators, allowing for continuous optimization [9, 14, 15].

Active learning lies at the core of these systems, selecting informative data points to minimize uncertainty and maximize discovery efficiency [6, 9, 12]. In a typical workflow, an initial model trained on existing data proposes candidates; experiments validate them, and results update the model via Bayesian optimization or reinforcement learning [10, 11, 13]. This approach is particularly effective in sparse data regimes, common in materials exploration, where random sampling is inefficient [3, 5]. Robotic experimentation automates synthesis and characterization, using platforms like automated flow reactors or high-

throughput screening robots to handle tasks from alloy mixing to property measurement [14, 15].

Simulation-experiment coupling strengthens autonomy by using computational surrogates to pre-screen candidates, reducing physical experiments to validation only [1, 2, 4, 8]. For example, DFT simulations inform ML models, which in turn guide robotic synthesis, creating a hybrid loop that leverages the strengths of both domains [3, 5, 15]. Uncertainty quantification ensures safe operation, flagging high-risk predictions for human oversight or further computation [6, 9, 13]. In inverse design contexts, closed-loop systems evolve designs iteratively, refining structures based on experimental feedback to converge on optimal solutions [10-12].

A conceptual formalization of closed-loop discovery can be expressed as an iterative process: Let M_t denote the machine learning model at iteration t trained on dataset D_t. The system selects the next experiment x $x_{\{t+1\}}$ via an acquisition function $a\left(x\right)$, such as expected improvement:

$$a\left(x\right) = E\left[max\left(0, f\left(x\right) - f^{*}\right)\right], \quad (1)$$

where $f\left(x\right)$ is the predicted objective and f^{*} is the current best. Post-experiment, $D_{\{t+1\}} = D_t \cup \left\{\left(x_{\{t+1\}},\ y_{\{t+1\}}\right)\right\}$, and $M_{\{t+1\}}$ is retrained [6, 9]. This balances exploration (high uncertainty regions) and exploitation (promising areas), accelerating convergence [12]. The integrated architecture linking data ecosystems, predictive models, experimental robotics, and adaptive learning cycles is illustrated in **Figure 1**.

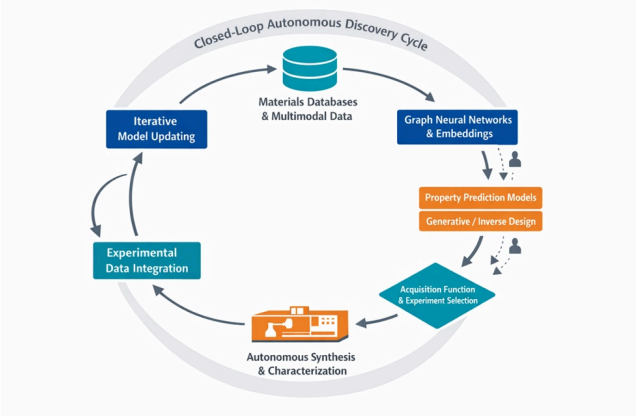

Figure 1. Closed-Loop Autonomous Discovery Architecture in AI-Driven Materials Engineering

Figure 1 Closed-loop autonomous discovery architecture in AI-driven materials engineering. The schematic depicts a cyclical infrastructure linking materials databases, machine learning prediction engines, inverse design modules, active learning acquisition systems, and robotic experimentation platforms. Iterative feedback updates data repositories and retrains predictive models, enabling adaptive exploration of materials space. Dashed pathways denote optional human oversight and simulation refinement layers.

Table 2. Operational Dynamics and Feedback Mechanisms in Autonomous Closed-Loop Discovery Systems

Closed-Loop Stage	Input Data	Computational Process	Phy
Dataset Initialization	Legacy simulations, experimental archives	Model pre-training	
Representation Encoding	Structural & compositional data	Feature learning via GNNs/embeddings	
Candidate Generation	Encoded materials space	Inverse design / generative sampling	
Acquisition Selection	Prediction + uncertainty metrics	Bayesian optimization / active learning	
Robotic Experimentation	Selected candidates	Automated synthesis & characterization	P
Data Assimilation	Experimental outputs	Database updating & cleaning	
Model Retraining	Updated dataset	Iterative ML optimization	
Convergence Assessment	Performance metrics	Exploration–exploitation balancing	Add

Applications span diverse materials classes, from high-entropy ceramics discovered via ML-guided

experimentation to 2D materials optimized through multiscale simulations [14, 15]. Challenges in hardware-software integration are mitigated by modular designs, allowing plug-and-play components [13]. Overall, these systems transform materials engineering into an autonomous discipline, promising order-of-magnitude speedups in discovery timelines [13].

Results and Discussion

The progression from high-throughput computation to autonomous discovery reflects a fundamental reconfiguration of materials engineering workflows, where data infrastructures transition from passive repositories to active, adaptive engines of innovation. High-throughput methods initially addressed the scale of chemical space by generating voluminous datasets [2, 3, 5], yet their static nature limited efficiency in targeting high-value candidates. The incorporation of representation learning—particularly graph neural networks—enabled more nuanced structural encodings that preserve atomic connectivity and symmetry, facilitating superior property predictions across diverse materials classes [4, 7-9]. These learned representations serve as critical bridges, allowing models to generalize beyond interpolation toward meaningful extrapolation in underrepresented regions [6, 12].

Inverse design frameworks further invert this paradigm, shifting focus from forward prediction to targeted generation. Generative approaches, including GANs, sample latent spaces efficiently to propose novel compositions while respecting physicochemical constraints [10, 12]. When coupled with active learning, these frameworks prioritize validation of uncertain or promising candidates, creating resource-efficient pathways that minimize experimental overhead [9, 12, 13]. Multimodal integration enhances robustness by fusing computational descriptors with experimental observables, mitigating discrepancies between simulation assumptions (e.g., 0 K DFT) and real-world conditions [5, 6, 13, 15]. Such fusion is especially valuable in bridging length and time scales, as seen in multiscale modeling of 2D materials growth or microstructure evolution [11, 15].

Closed-loop systems represent the integrative apex, where prediction, selection, execution, and feedback form self-reinforcing cycles. Active learning mechanisms, driven by uncertainty metrics, dynamically balance exploration of novel regions with exploitation of known optima [6, 9, 12]. This is formalized in iterative loops where acquisition functions guide experiment selection, progressively refining surrogate models [9]. Robotic platforms execute these decisions autonomously, enabling high-frequency iteration that compresses discovery timelines [13, 14]. The resulting infrastructures not only accelerate screening but foster emergent knowledge discovery, as iterative refinement uncovers non-intuitive structure-property linkages [4, 10, 11].

These advancements collectively redefine materials engineering as a data-centric, feedback-driven discipline. By synthesizing high-throughput origins with AI-orchestrated autonomy, the field moves toward scalable platforms capable of addressing grand challenges in sustainability, energy, and advanced manufacturing [1, 5]. The infrastructure emphasis—on interoperable data flows, uncertainty-aware decisioning, and simulation-experiment coupling—provides a blueprint for sustained acceleration beyond isolated tools.

Challenges

Despite substantial progress, several persistent challenges hinder the full realization of closed-loop data infrastructures in materials engineering.

Data scarcity and quality remain foundational limitations. Many materials classes, particularly those involving rare elements, defects, or extreme conditions, suffer from sparse datasets that impede robust model training [2, 3, 6]. High-throughput computation generates volume but often lacks diversity in underrepresented subspaces, leading to biased predictions and poor generalization [1, 5]. Experimental data, while valuable for grounding models, introduce inconsistencies due to varying protocols, instrumentation, or environmental factors [6, 13]. Multimodal integration exacerbates these issues, as aligning disparate sources (e.g., DFT energies with spectroscopic measurements) requires sophisticated preprocessing and uncertainty propagation [13, 15].

Uncertainty quantification poses another critical hurdle. While Bayesian approaches and ensemble methods estimate predictive confidence, translating these into actionable acquisition decisions in high-dimensional spaces remains nontrivial [6, 9, 12]. Overconfident models in extrapolation regimes can mislead closed-loop selection, wasting resources on suboptimal candidates [9]. In autonomous systems, uncalibrated uncertainties risk

compounding errors across iterations, potentially stalling convergence or favoring local optima [12].

Hardware-software integration in self-driving laboratories introduces practical barriers. Robotic platforms must handle heterogeneous operations—synthesis, characterization, handling volatile precursors—while maintaining precision and reproducibility [13-15]. Closed-loop operation demands real-time interfacing between prediction engines, actuators, and sensors, yet interoperability standards lag, complicating modular deployment [13]. High capital and maintenance costs further restrict accessibility, particularly for complex multimodal setups [14].

Algorithmic challenges include balancing exploration and exploitation in sparse regimes. Active learning acquisition functions perform variably depending on initial data quality and model architecture; poor starting conditions can lead to inefficient sampling [9, 12]. Inverse design faces additional difficulties in enforcing physical realism within generative models, where proposed candidates may violate stability or synthesizability constraints [10, 12].

Interpretability and generalizability also constrain trust and transferability. Deep architectures, while powerful, often function as black boxes, obscuring mechanistic insights essential for domain acceptance [6, 7]. Models trained on specific datasets struggle to extrapolate to new chemistries or processing conditions without domain adaptation [4, 8].

Collectively, these challenges underscore the need for resilient, uncertainty-aware, and interoperable infrastructures to realize the full potential of autonomous discovery [1-3, 5, 6, 12].

Future research directions

Addressing the identified challenges opens several promising avenues for advancing closed-loop data infrastructures.

Enhancing data efficiency through advanced active learning variants holds priority. Incorporating prior physical knowledge—via physics-informed architectures or constrained generative models—can mitigate scarcity by guiding exploration toward plausible regions [6, 12]. Hybrid strategies combining pool-based and generative sampling may further optimize under budget constraints, particularly for multimodal or high-cost experiments [9, 13].

Improving uncertainty quantification and calibration is essential for reliable autonomy. Developing scalable Bayesian deep learning methods or evidential frameworks could provide better-calibrated uncertainties, enabling safer decision-making in closed loops [6, 9]. Benchmarking extrapolation performance on out-of-distribution materials will standardize assessment, informing robust model selection [4, 8].

Advancing simulation-experiment integration requires tighter coupling. Real-time feedback mechanisms, where experimental discrepancies trigger model recalibration or additional high-fidelity computations, can reduce systematic biases [3, 5, 15]. Federated or transfer learning approaches may leverage distributed datasets while preserving proprietary information, accelerating community-wide progress [1, 13].

Hardware-software ecosystems demand modular, standardized designs. Open-source orchestration platforms and plug-and-play interfaces would lower barriers to entry, enabling customizable autonomous systems [13]. Investments in robust, flexible robotics capable of handling diverse materials workflows remain critical [14, 15].

For inverse design, enforcing synthesizability and stability constraints within generative frameworks—through reinforcement learning with domain rewards or multi-objective optimization—will yield more viable candidates [10-12]. Scaling to multi-property or multi-fidelity optimization will broaden applicability to complex engineering targets [4, 7].

Longer-term directions include human-AI collaboration paradigms, where infrastructures augment rather than replace expert intuition, and ethical frameworks ensuring equitable access and responsible deployment [1, 6]. As these elements mature, closed-loop systems could evolve into fully adaptive platforms, fundamentally transforming materials discovery into a rapid, intelligent process [2, 3, 5, 9, 12].

Conclusion

This review synthesizes the evolution from high-throughput computational screening to autonomous, closed-loop discovery infrastructures in materials engineering. Foundational high-throughput workflows generated essential datasets, while advances in representation learning—particularly graph neural networks—enabled

accurate, transferable property predictions. Inverse design and generative models inverted the discovery paradigm, proposing tailored candidates efficiently. Active learning and uncertainty-driven selection bridged prediction with experimentation, forming iterative cycles that optimize resource use. Multimodal integration and simulation-experiment coupling further strengthened these loops, addressing discrepancies across domains.

Autonomous systems exemplify this integration, orchestrating hypothesis generation, robotic execution, and model refinement in self-sustaining workflows. Despite challenges in data scarcity, uncertainty calibration, hardware integration, and interpretability, these infrastructures promise dramatic acceleration of materials innovation.

By framing the field through a systems lens—emphasizing interoperable data flows, adaptive decisioning, and feedback mechanisms—this synthesis highlights closed-loop platforms as transformative tools. Their maturation will democratize discovery, reduce reliance on serendipity, and address urgent needs in energy, sustainability, and advanced technologies. Ultimately, these developments position computational and data-driven materials engineering at the threshold of a new era of autonomous, intelligent innovation.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 19 May 2021 Revised: 03 Oct 2021 Accepted: 14 Dec 2021
Published online: 18 March 2022

References

Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A. Machine learning for molecular and materials science. Nature. 2018;559(7715):547-55. https://doi.org/10.1038/s41586-018-0337-2.

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):54. https://doi.org/10.1038/s41524-017-0056-5.

Schmidt J, Marques MRG, Botti S, Marques MAL. Recent advances and applications of machine learning in solid-state materials science. npj Comput Mater. 2019;5(1):83. https://doi.org/10.1038/s41524-019-0221-0.

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. https://doi.org/10.1021/acs.chemmater.9b01294.

Schleder GR, Padilha ACM, Acosta CM, Costa M, Fazzio A. From DFT to machine learning: Recent approaches to materials science–a review. J Phys Mater. 2019;2(3):032001.

Musil F, Grisafi A, Bartók AP, Ortner C, Csányi G, Ceriotti M. Physics-Inspired structural representations for molecules and

materials. Chem Rev. 2021;121(16):9759-9815. https://doi.org/10.1021/acs.chemrev.1c00021.

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 Chem Chem Phys. 2020;22(32):18141-8. https://doi.org/10.1039/D0CP01474E.

Chen C, Ong SP. Atomistic line graph neural network for improved materials property predictions. npj Comput Mater. 2021;7(1):141. https://doi.org/10.1038/s41524-021-00617-0.

Fung V, Hu G, Ganesh P, Sumpter BG. Benchmarking graph neural networks for materials chemistry. npj Comput Mater. 2021;7(1):84. https://doi.org/10.1038/s41524-021-00554-0.

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 design of inorganic materials. npj Comput Mater. 2020;6(1):84. https://doi.org/10.1038/s41524-020-00352-y.

Wu D, Fang F, Wang Y, Guo X, Zhao Y, Duan Y, et al. Machine-Learning microstructure for inverse material design. Adv Mater. 2021;33(29):2101207. https://doi.org/10.1002/adma.202101207.

Chen L, Zhang W, Nie Z, Li S, Pan F. Generative models for inverse design of inorganic solid materials. J Mater Inform. 2021;1(1):7. https://doi.org/10.20517/jmi.2021.07.

Oviedo F, Ferres JL, Agarwal TK, Cai J, Wang E, Lipkowitz G, et al. Interpretable and explainable machine learning models for materials discovery. npj Comput Mater. 2022;8(1):115. https://doi.org/10.1038/s41524-022-00809-2.

Kaufmann K, Maryanovsky D, Mellor WM, Zhu C, Rosengarten AS, Vecchio KS. Discovery of high-entropy ceramics via machine learning. npj Comput Mater. 2020;6(1):42. https://doi.org/10.1038/s41524-020-0317-6.

Momeni K, Ji HM, Ji Y, Hood RW, Hao J, Hu K, et al. Multiscale computational understanding and growth of 2D materials: A review. npj Comput Mater. 2020;6(1):97. https://doi.org/10.1038/s41524-020-00367-5.