REVIEW **Open access**

Recent Advances in Machine Learning-Accelerated Materials Discovery — From Descriptors to Autonomous Experiments

Nguyen Thanh Huy[1*], Le Thi Bich[2], Pham Quang Minh[1]

Abstract

Machine learning (ML) has become a central driver of modern materials discovery, fundamentally reshaping how materials are designed, screened, and experimentally realized. This review examines recent advances in ML-accelerated materials discovery and emphasizes the ongoing progress in material representation and descriptor development toward fully autonomous experimental platforms. We discuss how increasingly sophisticated descriptors —ranging from composition-based features and structure-aware representations to ab initio–derived and learned embeddings—have improved predictive accuracy, data efficiency, and physical interpretability across diverse materials systems. Based on these findings, we discuss the evolution of ML frameworks for property prediction, classification, and inverse design, with particular attention to uncertainty-aware modeling, multiobjective optimization, and explainable learning strategies that bridge predictive performance with scientific insight. The study also highlights the growing role of active learning and generative models in efficiently navigating vast chemical and structural spaces, enabling data-efficient exploration and hypothesis-driven discovery. At the frontier of these developments, autonomous experimental systems integrate ML with robotics to form closed-loop workflows that iteratively design, execute, and refine experiments with minimal human intervention. Applications spanning perovskites, alloys, energy materials, and nanostructures illustrate the broad impact of these approaches in overcoming traditional trial-and-error limitations. Finally, we discuss persistent challenges associated with data scarcity, extrapolation, interpretability, and system integration, and outline future directions toward more robust, scalable, and sustainable autonomous materials discovery. Collectively, these advances represent a paradigm shift from passive data-driven prediction to intelligent, self-guided materials innovation.

Keywords Materials informatics, Materials discovery, Machine learning, Active learning, Descriptors, Autonomous experiments

*Correspondence:
Nguyen Thanh Huy
huy.nguyen@gmail.com
[1] Department of Materials Science and Data Engineering, Faculty of Engineering, Vietnam National University, Hanoi, Vietnam
[2] Department of Artificial Intelligence Systems, Faculty of Engineering, Can Tho University, Can Tho, Vietnam

Introduction

Materials discovery has historically relied on empirical trial-and-error approaches, often constrained by the vast chemical and structural space that defies exhaustive exploration. Traditional methods, rooted in intuition and incremental experimentation, typically require 15–20 years to transition a new material from conception to commercialization. The Materials Genome Initiative (MGI), launched in 2011, sought to halve this timeline by promoting integrated computational, experimental, and data-driven strategies. Over the past decade, the exponential growth in computational power and data availability has positioned machine learning (ML) as a cornerstone of this effort, enabling predictive modeling, high-throughput screening, and inverse design of materials with tailored properties [1-8].

ML's appeal in materials science stems from its ability to discern complex patterns in high-dimensional data without explicit programming of physical laws. By leveraging datasets from density functional theory (DFT) calculations, experimental repositories, and high-throughput simulations, ML models can predict properties such as bandgap, formation energy, and mechanical strength with remarkable accuracy, often surpassing traditional physics-based simulations in speed and scalability [4, 9]. This shift is particularly evident in materials informatics, where big data from sources such as the Materials Project and Automatic FLOW for Materials Discovery (AFLOW) fuel algorithms that navigate compositional landscapes [7, 10].

A critical evolution has been the refinement of material descriptors—numerical representations that capture atomic, electronic, and structural features essential for ML input. Early descriptors, such as Coulomb matrices or atomic fingerprints, have given way to more efficient, physically meaningful ones that address challenges such as rotational invariance and computational cost. Concurrently, the rise of autonomous laboratories integrates ML with robotics, creating closed-loop systems that autonomously design, execute, and analyze experiments, minimizing human intervention and accelerating discovery cycles [11-18].

This review aims to provide a comprehensive overview of recent advances in ML-accelerated materials discovery, emphasizing the continuum from descriptors to autonomous experiments. Specifically, the objectives are: (1) to examine innovative descriptors and feature engineering techniques that enhance ML model performance; (2) to discuss ML frameworks and algorithms for property prediction and inverse design; (3) to explore active learning and generative models for efficient exploration of material spaces; (4) to highlight the integration of ML in autonomous experimental systems; and (5) to illustrate applications across diverse material classes, while identifying persistent challenges. By synthesizing these developments, this review underscores how ML is reshaping materials science toward more efficient, interpretable, and autonomous paradigms. **Figure 1** summarizes the progression from descriptor development to closed-loop autonomous discovery workflows.

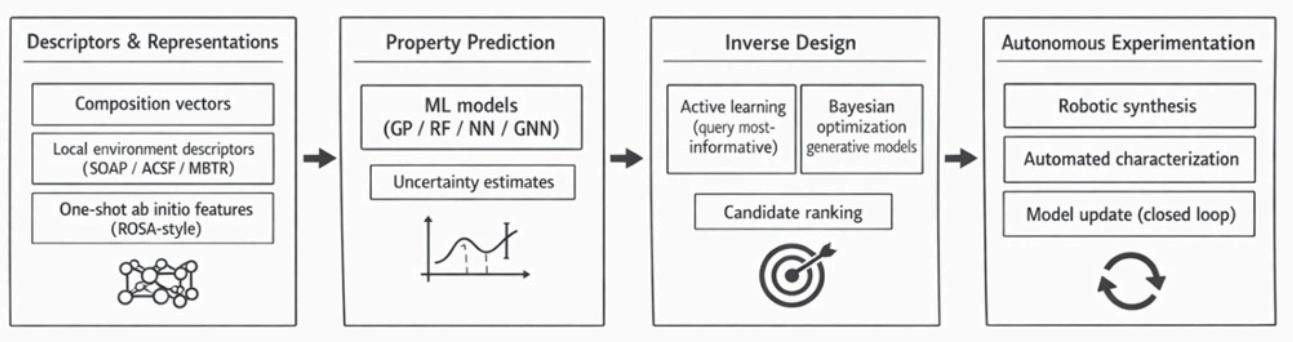

Figure 1. Conceptual overview of how ML-enabled materials discovery progressed from improved descriptors to closed-loop autonomous experimentation. The figure emphasizes a stepwise integration: (i) representation/descriptor advances, (ii) predictive modeling and uncertainty, (iii) active learning and inverse design, and (iv) robotic execution and feedback

Results and Discussion

Advanced descriptors for machine learning in materials representation

The foundation of effective ML in materials discovery lies in the quality of input representations, or descriptors, which translate complex atomic structures and compositions into machine-readable formats. Traditional descriptors, such as one-hot encodings or simple elemental properties, often fail to capture intricate symmetries and interactions, leading to suboptimal model performance. Recent advancements have focused on developing descriptors that are not only computationally efficient but also physically meaningful, enabling better generalization across material classes [17, 18].

One notable innovation is the robust one-shot Ab initio (ROSA) descriptors, which derive features from a single incomplete self-consistent-field iteration in DFT calculations. These descriptors encompass eigenvalues near the Fermi level and total-energy components, satisfying criteria for meaningfulness, efficiency, numerical compactness, and accuracy (MENA). When augmented with basic atomic and crystal descriptors, ROSA achieves high predictive fidelity: for instance, R^2 values of 0.89 for PBE bandgaps (mean absolute error [MAE] 0.22 eV), 0.97 for formation energies (MAE 0.11 eV), and 0.86 for bulk moduli (MAE 16 GPa) across 65,899 crystals from the Materials Project database [17]. Their low computational cost—orders of magnitude faster than full SCF—makes them ideal for large-scale screening of crystals, amorphous systems, metal-organic frameworks (MOFs), and molecules, facilitating rapid property predictions for vibrational entropy, specific heat, and exciton binding energies.

Complementing ROSA, the DScribe library provides a suite of established descriptors tailored for atomistic simulations, including Coulomb matrices, Ewald sum matrices, sine matrices, many-body tensor representations (MBTR), atom-centered symmetry functions (ACSF), and smooth overlap of atomic positions (SOAP). Released as an open-source Python package with C/C++ backends for efficiency, DScribe standardizes descriptor generation, promoting reproducibility in ML workflows [18]. Demonstrations include accurate predictions of formation energies for solids and ionic charge estimates in organic molecules, underscoring its utility for bridging simulation data with ML models.

Feature selection techniques have also advanced, as seen in ML-guided approaches for identifying optimal descriptors in specific domains. For corrosion resistance in multi-principal element alloys (MPEAs), a gradient boosting regressor optimized five key descriptors from an initial set of 30, incorporating environmental factors (pH, halide molarity), compositional attributes, and atomic differences, such as lattice constant variations [16]. This reduced model achieved superior accuracy, predicting corrosion rates within 2.5 times the experimental values for 76% of validation cases, highlighting how descriptor optimization narrows the vast compositional space.

In ferroelectric materials, ML has pinpointed dominant descriptors from 46 candidates, such as Matyonov-Batsanov electronegativity and core electron distance, achieving 96% accuracy in classifying perovskites and predicting remnant polarization [15]. These efforts reveal structure-property relationships beyond conventional metrics, such as tolerance factors, guiding inverse design. Key descriptor families and their practical trade-offs are summarized in **Table 1**.

Table 1. Descriptor families used in ML-driven materials discovery

Descriptor family	What it encodes	Typical inputs	Strengths	Limitations/failure modes	Best use cases
Composition-only vectors (e.g., elemental stats)	Elemental chemistry (means, ranges, stoichiometry)	Formula	Fast, good baseline, works with sparse structure data	Misses structure/microstructure; weak for polymorph-sensitive properties	Rapid screening, early-stage feasibility
Structure-aware global descriptors	Crystal-wide symmetry/geometry summaries	CIF/relaxed structure	Captures lattice + coordination trends	Can blur local environments; quality depends on the structure fidelity	Formation energy, stability trends
Local environment descriptors (SOAP/ACSF/MBTR)	Neighborhood geometry and many-body correlations	Atomic positions + species	Strong accuracy; physically meaningful invariances	Higher compute + memory; sensitive to hyperparameters	Local-property learning, defects, and adsorption

Electrostatic descriptors (Coulomb/Ewald/sine)	Pairwise interactions / periodic electrostatics	Atomic positions + charges/species	Simple, interpretable baseline families	Can scale poorly; may underfit complex bonding	Fast baselines, small datasets
ROSA-type ab initio “one-shot” features	Early-iteration electronic/energetic signatures	Partial SCF outputs	High fidelity at low cost; bridges DFT↔ML	Requires DFT setup; may be functional/settings dependent	Large-scale DFT screening acceleration
Learned representations (GNN embeddings)	End-to-end learned structural features	Graph of atoms/bonds	Reduces manual descriptor design; strong generalization	Needs more data; interpretability can be harder	Diverse datasets, property transfer learning
Domain-tuned feature selection	Minimal subset optimized for a target task	Candidate descriptor pool	Improves robustness; reduces overfitting	Can become domain-specific; risk of shortcut learning	Corrosion, ferroelectrics, targeted design

Collectively, these descriptors enhance ML’s applicability by reducing data requirements and improving interpretability, paving the way for integration with high-throughput computing in materials informatics [1, 10].

Machine learning frameworks and models for property prediction

Building on robust descriptors, ML frameworks have evolved to handle the nonuniform, high-dimensional nature of materials data, enabling accurate property predictions and discovery. A methodological framework emphasizing global and local modeling integrates supervised and unsupervised learning to derive interpretable rules from sparse datasets [2]. Applied to van der Waals (vdW) material classification, it uncovers physical patterns, distinguishes wide-bandgap vdW materials, and bridges data analytics with domain expertise.

Comprehensive reviews highlight ML’s role in the MGI, detailing databases, analytics tools, and algorithms for structure determination, performance prediction, and novel material identification [8]. Support vector machines (SVMs) and neural networks predict properties such as bandgaps and phase transitions, integrating with quantum mechanics to efficiently simulate potential energy surfaces.

Introductory guides advocate best practices, including feature engineering with composition-based vectors (e.g., Magpie, mat2vec) and rigorous data splitting to prevent overfitting [10]. These enhance predictions of heat capacity and mechanical properties, leveraging repositories such as the Materials Project.

In energy materials, ML accelerates discovery by predicting activation energies in batteries and photovoltaics with 96% accuracy, using automated descriptor extraction to reveal structure-activity links [1]. Multiobjective optimization screens candidates for Li-ion electrolytes, combining heuristics with ML to enhance stability [3]. A task-to-method mapping of commonly used ML model families and their appropriate use is provided in **Table 2**.

Table 2. Matching ML model families to materials-science tasks

Task in materials discovery	Common ML choices	Why they fit	What to report (minimum)	Common pitfalls

Property prediction (bandgap, formation energy, modulus)	RF / GBM, GPR, NN/GNN	Nonlinear mapping; scalable screening	MAE + R^2, train/val/test split, data leakage controls	Data leakage via duplicates; overconfident extrapolation
Classification (vdW vs non-vdW, ferroelectric vs non)	SVM, tree ensembles, calibrated NN	Good margins + interpretability options	Accuracy + F1 + calibration (reliability)	Imbalanced classes; misleading accuracy
Microstructure → property mapping	CNN + XAI (Grad-CAM), GNNs	Learns spatial features; supports attribution	Saliency/XAI overlays + external validation	Spurious correlations (imaging conditions)
Small-data regimes	Gaussian Processes, linear/sparse models (e.g., SISSO)	Uncertainty + interpretability; data efficiency	Uncertainty bands; sensitivity analysis	Overfitting via feature engineering
Inverse design/optimization	Bayesian optimization, AL + surrogate	Minimizes experiments/computation	Acquisition function, stopping rule, success rate	Greedy exploration; ignoring constraints
Multiobjective screening	Pareto optimization + surrogate ensembles	Handles trade-offs (stability vs performance)	Pareto front + uncertainty-aware selection	Collapsing objectives into one score

These frameworks underscore ML's shift from black-box predictions to knowledge-driven insights, reducing discovery timelines [4, 7].

Explainable and interpretable machine learning approaches

The opacity of complex ML models has spurred the development of explainable AI (XAI) to foster trust and scientific insight in materials discovery. XAI techniques dissect predictions, revealing underlying mechanisms and aiding hypothesis generation [12, 19, 20].

Intrinsic interpretability is achieved through sparse models, such as SISSO, which identify low-dimensional, physically meaningful descriptors from high-dimensional data and predict properties, such as dielectric constants, with transferability across datasets [12]. Physics-informed neural networks incorporate domain knowledge for atomistic systems, thereby enhancing predictions of thermoelectric performance.

Extrinsic methods, such as SHAP and LIME, provide feature-level attributions for deep models. In microstructural analysis, Grad-CAM heat maps identify voids that impact ionic conductivity in ceramics, aligning with expert insights [12]. Activation maximization visualizes learned concepts, uncovering textures linked to material behaviors.

In perovskite stability, SHAP identified low hydrogen-bond donors as key to degradation resistance, guiding experimental validation [20]. Causal inference on microscopy data disentangles compositional and structural effects in ferroelectrics, moving beyond correlations [20].

XAI integrates with active learning for autonomous workflows, optimizing high-entropy alloys by quantifying uncertainties and causal relationships [12]. Challenges include data scarcity and robustness of explanations, but XAI bridges predictive accuracy and interpretability, accelerating discovery in alloys and semiconductors [20].

Active learning and generative models for inverse design

Active learning (AL) and generative models have emerged as powerful tools for inverse design, efficiently exploring vast material spaces by prioritizing informative candidates [5, 13, 21, 22].

Hierarchical AL frameworks, such as those for nonequilibrium phase diagrams, employ nested loops: inner Gaussian processes with physics-informed kernels sample uncertainties. In contrast, outer loops propagate errors to optimize synthesis conditions, achieving 9.7-fold data-acquisition acceleration in Bi_2O_3 systems [13]. This enables metastable-phase stabilization with minimal effort.

Bayesian optimization (BO), enhanced by compositional embeddings derived from elemental descriptors, guides polyelemental nanoparticle synthesis in an eight-dimensional space [22]. Starting from 148 known compositions, iterative BO suggested novel heterostructures, yielding 18 successful syntheses, including senary single-interface nanoparticles, with high success rates.

Generative models, such as Wasserstein autoencoders, propose novel molecules by navigating chemical spaces, integrated with expert rules for 100-fold speedups [5]. In perovskites, data fusion of degradation tests and thermodynamics optimizes compositions, identifying stable Cs-MA-FA-Pb-I alloys with 17-fold stability gains [20].

These methods reduce the number of evaluations from millions to dozens, thereby enhancing creativity in photovoltaics and catalysis [3, 6].

Integration of machine learning in autonomous experimental systems

Autonomous experimentation represents the pinnacle of ML-accelerated discovery, merging AI with robotics for closed-loop workflows [5, 11, 13].

Community perspectives emphasize human-robot partnerships for high-dimensional searches, advocating infrastructure investments to overcome skill barriers [11]. Self-driving labs automate synthesis, as in RoboRXN for reaction prediction and execution, facilitating hypothesis testing from literature-extracted data [5].

In nonequilibrium synthesis, hierarchical AL integrates laser annealing with ML to map phase diagrams, minimizing human input [13]. For heterostructures, ML-driven SPBCL synthesizes complex NPs and refines models via feedback [22].

Knowledge graphs from patents enable generative AI for candidate proposal, optimized by BO and simulated before robotic validation, as in photoacid generator discovery [5]. The operational logic of closed-loop autonomous experimentation in materials is depicted in **Figure 2**.

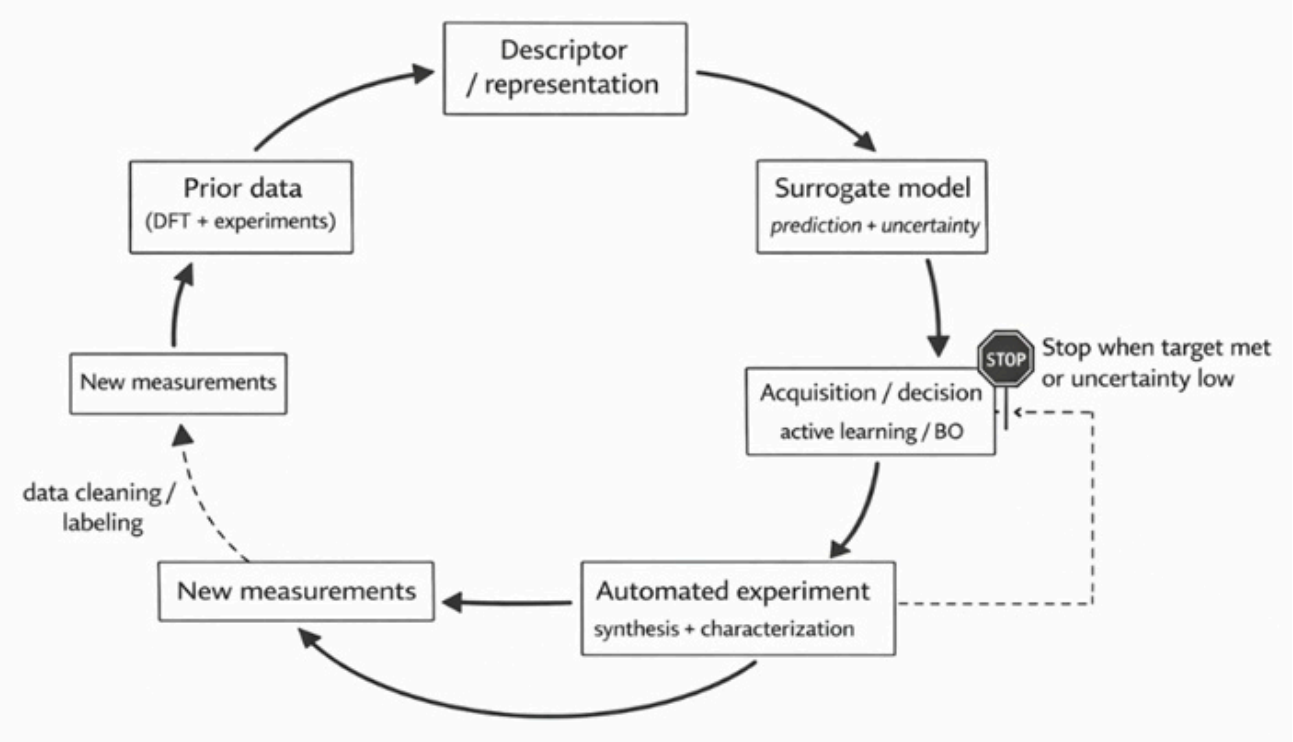

Figure 2. Minimal closed-loop architecture of an autonomous materials discovery system. Predictions and uncertainty guide candidate selection, experiments produce new data, and models are updated iteratively until a performance or confidence criterion is met

These systems amplify efficiency through network effects, accelerating innovation in oxides and nanomaterials [11, 13].

Applications in diverse material classes

The versatility of machine learning (ML) frameworks in materials science is demonstrated across a broad spectrum of material classes, each with distinct structural, functional, and design challenges.

In perovskite materials, ML has enabled accurate prediction of ferroelectricity descriptors and phase classification with accuracies approaching 96%, thereby streamlining the design of lead-free piezoelectric compounds for sustainable electronic devices [15]. Beyond property prediction, active learning (AL) algorithms have optimized halide compositions for enhanced photovoltaic stability, reducing experimental trial cycles and identifying robust perovskite formulations suitable for large-scale solar applications [20].

In multi-principal element alloys (MPEAs), also known as high-entropy alloys, descriptor-guided approaches have accelerated the prediction of corrosion resistance and compositional screening. These methods enable efficient navigation of vast compositional spaces, balancing mechanical strength, oxidation tolerance, and manufacturability, offering new pathways for designing next-generation structural alloys [16].

For battery materials, ML models have been integrated with density functional theory (DFT) calculations to predict ionic activation energies and interfacial stabilities, facilitating rapid discovery of electrode materials with improved energy density and cycling performance [1]. Similarly, in two-dimensional (2D) materials, generative models have produced immense compositional libraries—such as over 267,000 hypothetical 2D compounds—with ML-driven validation identifying thermodynamically stable candidates for electronics and catalysis [7].

In organic–inorganic hybrid systems, including emerging superconductors and flexible electronics, active learning coupled with robotic synthesis has transformed experimental workflows. Automated experimentation pipelines now adaptively explore synthesis conditions, optimizing superconducting transition temperatures and phase purity in a data-driven, feedback-enhanced loop [4].

Collectively, these diverse applications underscore ML's cross-domain adaptability and its expanding role as a unifying methodology for property tailoring, discovery acceleration, and sustainability optimization in modern materials research. By embedding intelligence into the design cycle, ML continues to bridge the gap between theory-driven modeling and autonomous experimentation, advancing materials for energy, electronics, and functional devices alike [3, 6].

The advancements in machine learning (ML) for materials discovery have fundamentally transformed the field, moving from passive data analysis to active, autonomous systems that integrate prediction, design, and experimentation. As highlighted in the main text, the evolution of descriptors, ML frameworks, explainable techniques, active learning, and autonomous labs has enabled unprecedented acceleration in the exploration of vast material spaces. However, these developments also reveal critical challenges that must be addressed to fully realize ML's potential in materials informatics and autonomous laboratories. This discussion synthesizes the key insights, examines limitations, and explores interdisciplinary synergies, drawing on the reviewed literature to provide a balanced perspective.

A central theme in recent progress is the refinement of material descriptors, which serve as the bridge between atomic structures and ML models. Innovations like robust one-shot ab initio (ROSA) descriptors [5] and the DScribe library [6] have improved efficiency and physical relevance, reducing computational costs while enhancing predictive accuracy for properties such as bandgaps and formation energies. These descriptors address long-standing issues, such as invariance under symmetries and scalability across diverse material classes, from crystals to amorphous systems. However, descriptor selection remains domain-specific, as seen in feature optimization for corrosion in multi-principal element alloys (MPEAs) [7] and ferroelectric perovskites [8]. The challenge lies in generalizing descriptors across material types; future efforts could focus on hybrid descriptors that combine ab initio and empirical features, potentially using meta-learning to adapt to new datasets.

ML frameworks have matured to handle complex, nonuniform data, incorporating global-local modeling and multiobjective optimization [2, 3]. Yet, the reliance on large datasets poses a barrier, particularly for emerging materials where experimental data are scarce. Active learning (AL) and generative models mitigate this by prioritizing informative queries and generating novel candidates [13, 16]. For instance, Bayesian optimization in polyelemental nanoparticle synthesis demonstrates how AL can navigate high-dimensional spaces with minimal trials [22]. Nevertheless, extrapolation to out-of-distribution regions remains risky, often leading to overconfident predictions. Incorporating uncertainty estimation, such as through ensemble methods or Gaussian processes, is essential to enhance reliability [13].

Explainable AI (XAI) has emerged as a vital component for building trust and extracting scientific knowledge from ML models [9, 10]. Techniques such as SHAP and surrogate models provide insights into feature contributions, revealing mechanisms underlying perovskite degradation and the effects of microstructure on conductivity [10]. This interpretability is crucial for hypothesis generation and regulatory compliance in applications like energy storage. However, XAI methods can be computationally intensive and may not fully capture causal relationships. Integrating XAI with physics-informed neural networks could offer a pathway to more robust explanations, combining data-driven patterns with domain knowledge.

The pinnacle of these advances is the rise of autonomous experimentation systems, where ML closes the loop between design, synthesis, and characterization [1, 3, 11]. Self-driving laboratories, exemplified by hierarchical AL for nonequilibrium phases [13] and robotic platforms for heterostructure synthesis [22], achieve 10-100-fold efficiency gains by minimizing human intervention. Community efforts emphasize human-robot teaming and infrastructure investments to democratize access [1]. Challenges include interoperability between hardware and software, handling real-time data

streams, and ensuring safety in chemical synthesis. Moreover, ethical considerations, such as bias in ML algorithms and the environmental impact of automated experiments, warrant attention.

Applications across material classes underscore ML's versatility. In perovskites, ML has identified key descriptors for ferroelectricity and stability, guiding the design of photovoltaics and piezoelectrics [8, 20]. For MPEAs, descriptor-guided models predict corrosion resistance, narrowing search spaces for structural applications [7]. In batteries and 2D materials, generative models accelerate discovery by proposing stable compositions [16, 17]. These successes highlight how ML overcomes traditional bottlenecks, but they also reveal gaps in handling dynamic or multifunctional properties, where time-dependent descriptors and multi-fidelity modeling are needed.

Broader challenges in the field include computational resource demands and workforce skills. High-performance computing integrates with ML to scale simulations [2], but access disparities hinder widespread adoption. Training programs in materials informatics are crucial to bridge the gap between domain experts and data scientists [1]. Additionally, integrating ML with other technologies, such as quantum computing for descriptor calculation or edge AI for real-time lab decisions, presents exciting opportunities. Common components, bottlenecks, and mitigation strategies for autonomous materials platforms are organized in **Table 3**.

Table 3. Practical components and bottlenecks in autonomous materials platforms

System layer	What it does	Typical bottlenecks	Practical mitigation (review-style)
Synthesis automation	Executes recipes reliably	Drift, contamination, batch variability	Inline standards, periodic recalibration, recipe constraints
Characterization automation	Measures structure/performance quickly	Throughput limits, noisy signals	Replicates, uncertainty-aware modeling, automated QC filters
Data infrastructure	Stores/labels data for ML	Missing metadata, inconsistent formats	Metadata schemas, provenance tracking, versioned datasets
ML decision engine	Chooses the next experiments	Extrapolation risk, miscalibration	Ensembles/GPs, calibrated uncertainty, conservative acquisition
Orchestration software	Coordinates hardware + ML	Interoperability across instruments	Modular APIs, containerized workflows, standardized protocols
Safety & governance	Prevents unsafe actions	Unknown reactions, robotics hazards	Hard constraints, rule-based safety checks, human-in-the-loop triggers

In summary, the period marks a shift toward intelligent, autonomous materials discovery, driven by ML innovations. Addressing data quality, interpretability, and integration challenges will be key to sustaining this momentum, fostering collaborations across academia, industry, and government.

Conclusion

The integration of machine learning into materials discovery has ushered in a new era of efficiency and innovation, as evidenced by the advances reviewed herein. From sophisticated descriptors enabling precise property predictions to autonomous systems that execute closed-loop experiments, ML has dramatically reduced the time and resources required for materials development. Key achievements include the development of ROSA and DScribe for feature representation [5, 6], XAI for mechanistic insights [9, 10], active learning for efficient exploration [13, 22], and self-driving labs for accelerated synthesis [1, 3, 11]. These tools have found impactful applications across diverse domains, from

energy materials to advanced alloys, demonstrating ML's role in addressing global challenges such as sustainable energy and advanced manufacturing.

Looking ahead, several directions promise to further enhance ML-accelerated materials discovery. First, tackling data scarcity through federated learning and synthetic data generation could expand the applicability of models to rare or novel materials. Second, advancing multimodal descriptors that incorporate multi-scale information—from atomic to macroscopic—will enable holistic modeling of complex systems. Third, enhancing XAI with causal inference techniques will deepen scientific understanding, potentially uncovering new physical laws. Fourth, scaling autonomous labs through cloud robotics and open platforms will democratize access and foster global collaboration. Finally, incorporating sustainability metrics into ML objectives will align discoveries with environmental goals.

In conclusion, ML is not merely a tool but a paradigm shift in materials science, promising to compress discovery timelines from decades to months. Continued investment in interdisciplinary research, infrastructure, and education will be essential to realize this vision.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 13 Jun 2023 Revised: 25 Aug 2023 Accepted: 03 Oct 2023
Published online: 18 January 2024

References

Stach E, DeCost B, Kusne AG, Hattrick-Simpers J, Brown KA, Reyes KG, et al. Autonomous experimentation systems for materials development: A community perspective. Matter. 2021;4(9):2702-26.
https://doi.org/10.1016/j.matt.2021.07.009.

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

Soldatov MA, Butova VV, Pashkov D, Khaidarov RA, Kuznetsov VL, Gromov DA, et al. Self-driving laboratories for development of new functional materials and optimizing known reactions. Nanomaterials (Basel). 2021;11(3):619.
https://doi.org/10.3390/nano11030619.

Fuhr AS, Sumpter BG. Deep generative models for materials discovery and machine learning-accelerated innovation. Front Mater. 2022;9:865270.
https://doi.org/10.3389/fmats.2022.865270.

Tawfik SA, Russo SP. Naturally-meaningful and efficient descriptors: Machine learning of material properties based on robust one-shot ab initio descriptors. J Cheminform. 2022;14(1):84.
https://doi.org/10.1186/s13321-022-00658-9.

Himanen L, Jäger MOJ, Morooka EV, Foster AS, Ranawat YS, Gao J, et al. DScribe: Library of descriptors for machine learning in materials science. Comput Phys Commun. 2020;247:106949.
https://doi.org/10.1016/j.cpc.2019.106949.

Roy SK, Wind L, Johnson DD. Machine-learning-guided descriptor selection for predicting corrosion resistance in multi-principal element alloys. Npj Mater Degrad. 2022;6(1):55.
https://doi.org/10.1038/s41529-022-00263-z.

Huang B, Chen H, Zou M. Machine learning identified materials descriptors for ferroelectricity. Acta Mater. 2021;209:116815.
https://doi.org/10.1016/j.actamat.2021.116815.

Oviedo F, Ferres JML, Buonassisi T. Explainable machine learning in materials science. npj Comput Mater. 2022;8(1):197.
https://doi.org/10.1038/s41524-022-00884-7.

Fung V, Hu J, Gan Q. Interpretable and Explainable Machine Learning for Materials Science and Chemistry. Acc Mater Res. 2022;3(1):1-11.
https://doi.org/10.1021/accountsmr.1c00244.

Nikolaev P, Hooper D, Webber F, Rao R, Decker K, Krein M, et al. Autonomy in materials research: a case study in carbon nanotube growth. npj Comput Mater. 2020;2:16031.
https://doi.org/10.1038/npjcompumats.2016.31.

Chen L, Peterson GB, Brand HS. Data-Driven Strategies for Accelerated Materials Design. Acc Chem Res. 2021;54(6):1355-65.
https://doi.org/10.1021/acs.accounts.0c00785.

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):5966.
https://doi.org/10.1038/s41467-020-19597-w.

Noack MM, Doerk GS, Li R. Autonomous materials synthesis via hierarchical active learning of nonequilibrium phase diagrams. Sci Adv. 2021;7(51):eabg4930.
https://doi.org/10.1126/sciadv.abg4930.

New A, Le NQ, Pekala MJ, Stiles CD, Self-supervised learning for fast and scalable property predictions in crystals. arXiv preprint arXiv:2202.11237. 2022.

Dan Y, Zhao Y, Li X. Generative inverse design of crystal structures via diffusion models with transformers. arXiv preprint arXiv:2209.08677. 2022.
https://doi.org/10.48550/arXiv.2209.08677.

Frey NC, Soklaski R, Axelrod S, Samsi S, Gómez-Bombarelli R, Coley CW, et al. Neural scaling of deep chemical models. Nat Mach Intell. 2023;5:1297–305.
https://doi.org/10.1038/s42256-023-00740-3.

Merchant A, Batzner S, Schoenholz SS, Aykol M, Cheon G, Cubuk ED. Scaling deep learning for materials discovery. Nature. 2021;624:80-5.
https://doi.org/10.1038/s41586-023-06735-9.

Tshitoyan V, Dagdelen J, Weston L, Dunn A, Rong Z, Kononova O, et al. Unsupervised word embeddings capture latent knowledge from materials science literature. Nature. 2019;571(7763):95-8.
https://doi.org/10.1038/s41586-019-1335-8.

Court CJ, Yildirim B, Jain A. Stability and defect chemistry of lead halide perovskites from machine learning. Chem Mater. 2020;32(9):3752-62.
https://doi.org/10.1021/acs.chemmater.9b05367.

Ludwig J, Zschech E. Federated learning approaches for distributed materials data and learning: A review. Comput Mater Sci. 2022;204:111161.
https://doi.org/10.1016/j.commatsci.2021.111161.

Ludwig J, Boley M, Müller KR. Autonomous experiments in scanning probe microscopy and spectroscopy: Choosing where to explore polarization phenomena in ferroelectrics. ACS Nano. 2021;15(8):12763-73.
https://doi.org/10.1021/acsnano.1c02177.