Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.079503

REVIEW

Data-Driven Materials Science Using Machine Learning and Computational Modeling

Manjodh Kaur[1], Princy Randhawa[2,*], Jitendra Jaiswal[2], Deepak Dubal[3], Ravindra N. Bulakhe[4,5], Deepanraj Balakrishnan[6] and Nithesh Naik[7,*]

[1]Department of Chemistry, School of Engineering, Dayananda Sagar University, Harohalli, Bengaluru, Karnataka, India
[2]Computer Science and Engineering (AI&ML), School of Engineering, Dayananda Sagar University, Harohalli, Bengaluru, Karnataka, India
[3]Faculty of Science, School of Chemistry & Physics, Queensland University of Technology, Brisbane, QLD, Australia
[4]Center for 2D Quantum Heterostructures, Institute for Basic Science (IBS), Sungkyunkwan University (SKKU), Suwon, Republic of Korea
[5]Symbiosis Centre for Nanoscience and Nanotechnology, Symbiosis International (Deemed University), Pune, India
[6]Department of Mechanical Engineering, College of Engineering, Prince Mohammad Bin Fahd University, Al-Khobar, Saudi Arabia
[7]Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, Karnataka, India

*Corresponding Authors: Princy Randhawa. Email: princy-aiml@dsu.edu.in; Nithesh Naik. Email: nithesh.naik@manipal.edu

Received: 22 January 2026; Accepted: 23 April 2026; Published: 15 June 2026

ABSTRACT: This review emphasizes the growing role of artificial intelligence (AI) in transforming the materials discovery process into a data-driven and autonomous approach. It systematically traces the evolution of scientific paradigms in materials science and examines how machine learning, generative models, and AI agents are revolutionizing the design, screening, and optimization of materials. A key contribution is a detailed, step-by-step machine learning framework that guides researchers through data collection, preprocessing, feature engineering, model development, and validation, utilizing publicly available materials databases and computational tools. Additionally, the review discusses the latest advances in generative AI and autonomous research systems, highlighting their potential to enable inverse design and closed-loop experiments. It includes a tutorial case study on sodium-ion battery materials to demonstrate practical application in formation energy prediction via machine learning, along with comparisons to high-throughput screening accuracy using density functional theory (DFT). The article also addresses current challenges such as data limitations, model interpretability, and physics-based approaches. Overall, this publication serves as both a conceptual and practical guide for integrating AI into materials research, aiming to accelerate the discovery process and improve efficiency.

KEYWORDS: Artificial intelligence; materials science; materials discovery; energy materials; computational modeling

1 Introduction

Over the past decade, materials science has undergone a paradigm shift driven by the rapid convergence of artificial intelligence (AI) and computational materials design. Recent research has proposed a fifth paradigm of scientific discovery, involving AI-driven, autonomous research systems that combine data, theory, and computation. However, this idea is not yet universally accepted and has not yet gained widespread adoption in the scientific community [1]. Fig. 1 illustrates the evolution of paradigms in materials science. All paradigms reflect changes in the study, comprehension, and design of materials [2]. Empirical science is

the first paradigm and is founded on observation and experimentation. Materials development is conducted through laboratory experiments and property measurements. It is through trial and error that knowledge is acquired. Early alloy development manifests this strategy through systematic composition and heat treatment testing. The second paradigm is theoretical science, which seeks to explain experimental outcomes using scientific laws and equations. Theoretical materials science, physics, and chemistry are applied to explain why materials behave as they do [3–5]. The principles of thermodynamics and solid-state physics can be used to explain phase stability, diffusion, and mechanical behavior in materials. Computational science is the third paradigm; it simulates material behavior on computers before experimentation. Density functional theory and molecular dynamics are two methods used to make approximations at the atomic level. The next level is data-driven science, which uses massive experimental and computational datasets to discover patterns and correlations. Machine learning (ML) models trained on materials databases can quickly filter through large collections of materials and efficiently provide target properties. The most recent paradigm, self-evolving scientific intelligence, is the main shift in materials research. Artificial intelligence systems are constantly self-educating through data, experiments, and feedback about operational processes. When new information arrives, these systems can automatically suggest new materials, make experimental decisions, and optimize models [4–7].

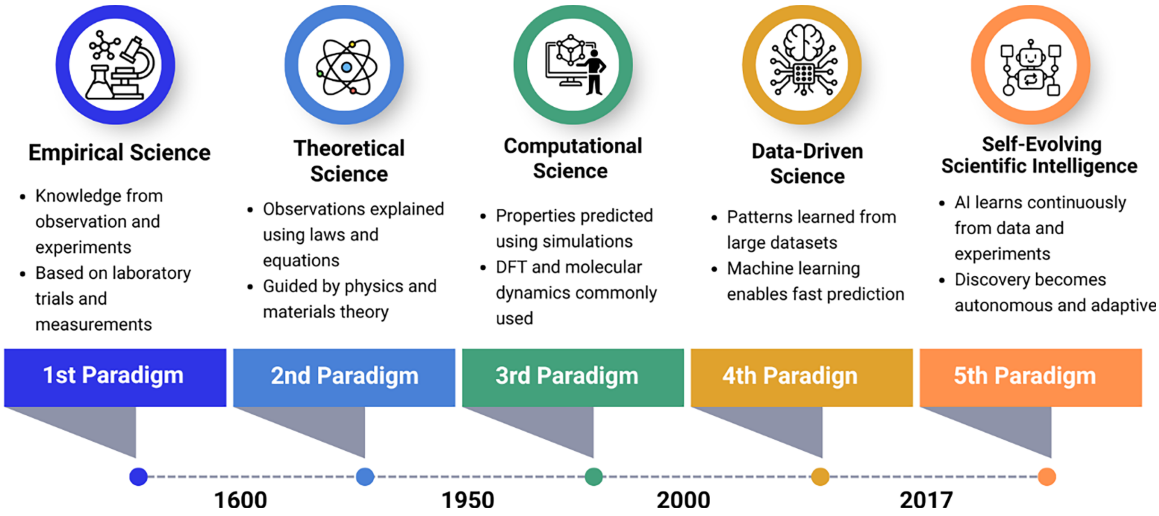

Figure 1: Evolution of scientific discovery paradigms.

The study of AI for material discovery seeks to answer a fundamental question: can machines autonomously learn the structure-property processing relationships that define material behaviour? The objective aims to create generative, reasoning systems capable of hypothesizing and designing novel materials *ab initio* [2]. In this context, researchers are exploring a spectrum of machine learning techniques from regression to generative modeling and reinforcement learning to map complex relationships among chemistry, structure, synthesis, and functionality. Such efforts build on the early vision articulated by Hill et al. [3] who proposed that materials informatics would transform discovery workflows by integrating multiscale modeling with experimental data.

It has become a branch of generative artificial intelligence (GenAI) in which models not only make predictions but also generate. Similarly, Park et al. [4] used text-guided diffusion models to explore crystal chemical space, while Kang and Kim [5] developed ChatMOF, a large language model (LLM) based system capable of generating and predicting metal-organic frameworks from textual prompts. All of these

developments indicate that the field of materials science is moving into a machine-creativity era. The research community is still considering the basic questions concerning the validation and ML model interpretability. Bhat et al. [6] highlighted that, though predictive, most of them were not successful. ML systems are not robust in the application and use of experimental data due to biases in the training data. Some form of approach is required to guarantee the physical consistency of AI models. In practice, the motive behind such efforts is to move AI from an exploratory to a decision-making role in the laboratory. Recent work further highlights that AI does not operate in isolation but in synergy with a global research ecosystem. Contemporary studies recognize the parallel evolution of automated experimentation ("self-driving laboratories") and LLM-powered scientific agents [7]. Piovarči et al. [8] argued that the fusion of reinforcement learning, high-throughput simulation, and automated synthesis forms a closed-loop system. The closed-loop system is capable of hypothesis generation, validation, and feedback, which together constitute the foundation of an autonomous materials research cycle. More recently, Pyzer-Knapp et al. [9] described the emergence of foundation models for materials discovery, which adapt the pretraining fine-tuning paradigm of natural language processing (NLP) to materials data, enabling generalizable embeddings across diverse materials systems. These studies collectively signal a convergence between materials informatics, computational chemistry, and AI research [10].

This review proposes a comprehensive pipeline for autonomous materials discovery that encompasses hypothesis generation, synthesis, and validation. Jain [11] evaluated the rapid increase in machine learning use in materials science, observing an annual growth factor of 1.67 in machine learning-related studies over the last decade. Nevertheless, despite this swift increase, practical implementation is hindered by limited data and restricted cross-domain transferability. The work aims to enhance computational screening and to transform the representation and reutilization of information across many fields. Researchers expect that integrating LLMs, graph models, and generative diffusion frameworks will elucidate the links among structure, properties, and functions [12]. As these models mature, their predictive scope continues to expand beyond inorganic crystals; AI frameworks are now being deployed in biomaterials [13] battery science [14], and sustainable composites [15] demonstrate that AI-guided hybrid models can design self-assembling peptides and other functional materials with unprecedented precision. Meanwhile, foundation models trained on multimillion compound datasets are being used to generate synthetic datasets, fill data gaps, and predict novel chemistries without explicit quantum calculations. Together, these results align with the researchers' expectation that AI will transform materials science into a predictive and autonomous discipline [16]. A major challenge remains the development of learning frameworks that can integrate numerical simulations, experimental datasets, and textual scientific knowledge, as Zeni et al. [17] illustrate, such integration bridges the gap between human reasoning and machine inference. The primary objective is to develop self-evolving materials intelligence systems that continually learn from global data sources and provide clear justifications. This study's subsequent sections will examine the architecture, datasets, and infrastructure that support AI in materials discovery.

Materials science majors find AI applications especially useful where the search space is too large for traditional experiments or computations. For example, alloy design can involve millions of element combinations, and unless a complete test experiment is feasible [18]. Likewise, predicting complex structure-property relationships often involves modeling nonlinear interactions among composition, crystal structure, and processing conditions [19]. While first-principles methods like density functional theory (DFT) offer accurate predictions, they are computationally intensive for large materials spaces [20]. Machine learning models mitigate these issues by learning patterns from existing data and quickly screening numerous candidate materials. Therefore, AI is particularly advantageous for problems with extensive compositional spaces, costly simulations, and intricate multi-parameter relationships [21].

Recent Scopus analytics show a sharp rise in ML for materials articles, increasing from a few hundred to thousands over the past 5 years. (Fig. 2a). Fig. 2b shows the pie chart depicting the percentage of Scopus publications categorized by scientific subject area. To improve the clarity of the pie chart, some of the field article numbers were merged or added to the multidisciplinary section. The Scopus search was performed in March 2026, and the keywords used to obtain the analytics were as follows: TITLE-ABS-KEY ("machine learning" AND "material discovery") AND PUBYEAR > 2019 AND PUBYEAR < 2027 AND (LIMIT-TO (LANGUAGE, "English")). These insights clearly demonstrate a strong interest in machine learning among researchers from multiple disciplines.

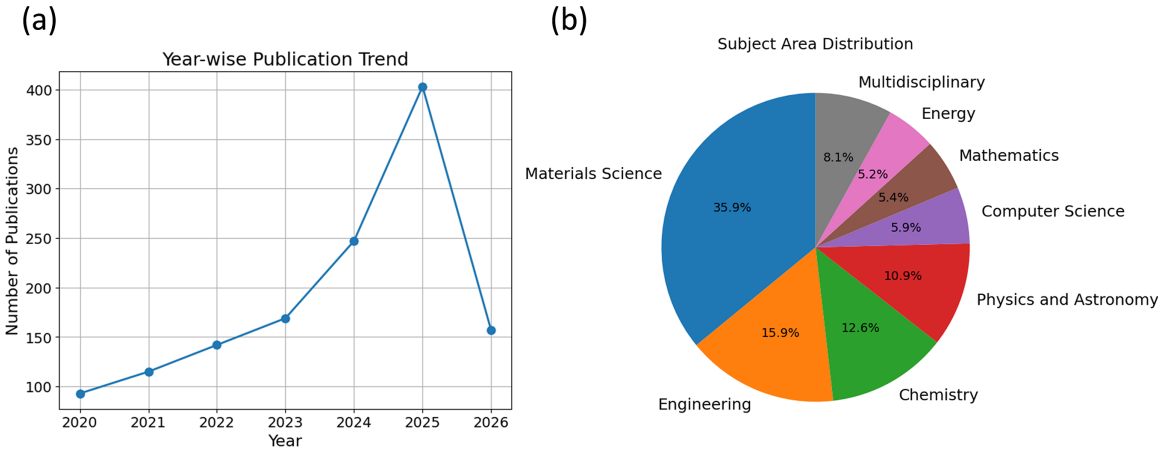

Figure 2: (**a**) Increase in the number of publications for the keyword search: machine learning for materials discovery from scopus analyses. (**b**) Pie chart showing the percentage of publications in major scientific fields related to machine learning.

2 Traditional vs. AI-Driven Materials Discovery

To better understand why traditional materials development is time-consuming and resource-intensive, it is helpful to examine the typical lifecycle of material discovery and deployment under conventional research paradigms, as shown in Fig. 3 [22]. Traditional materials research follows a sequential and time-intensive development pathway, beginning with material discovery and ending with large-scale deployment. This process typically involves seven stages: discovery, laboratory development, iterative optimization experiments, system-level design, certification, manufacturing, and final application. Each stage depends on the successful completion of the previous one, making the overall workflow slow and resource-intensive.

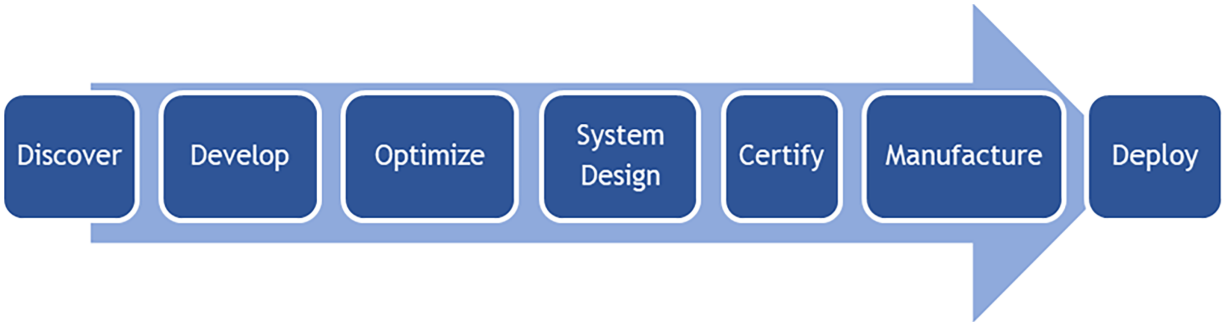

Figure 3: The process of finding new materials using traditional methods.

In many cases, the transition from initial discovery to first industrial use can take several decades, typically 10–20 years, particularly for safety-critical applications. The heavy reliance on trial-and-error testing, full testing, and regulatory approval further slows the pace of invention. The weaknesses of traditional methods suggest that better approaches could reduce development time without sacrificing reliability and performance. The rise of artificial intelligence has significantly transformed this linear workflow by integrating parallel, data-driven, and feedback-focused processes. AI-based materials research uses historical data, simulations, and experimental results to predict material behavior at the outset of the process, rather than building it step by step, as shown in Fig. 4. This shows that AI enables the entire materials research process to be connected and continuous. Discovery, exploration, and application flow seamlessly without separate stages. The process begins with AI-assisted discovery, where machine learning models analyze potential materials and predict their properties. This will help the researchers to quickly spot interesting material without conducting tiresome lab work. The selected materials are subjected to selective synthesis. At this point, AI helps select appropriate synthesis conditions and guides experiments.

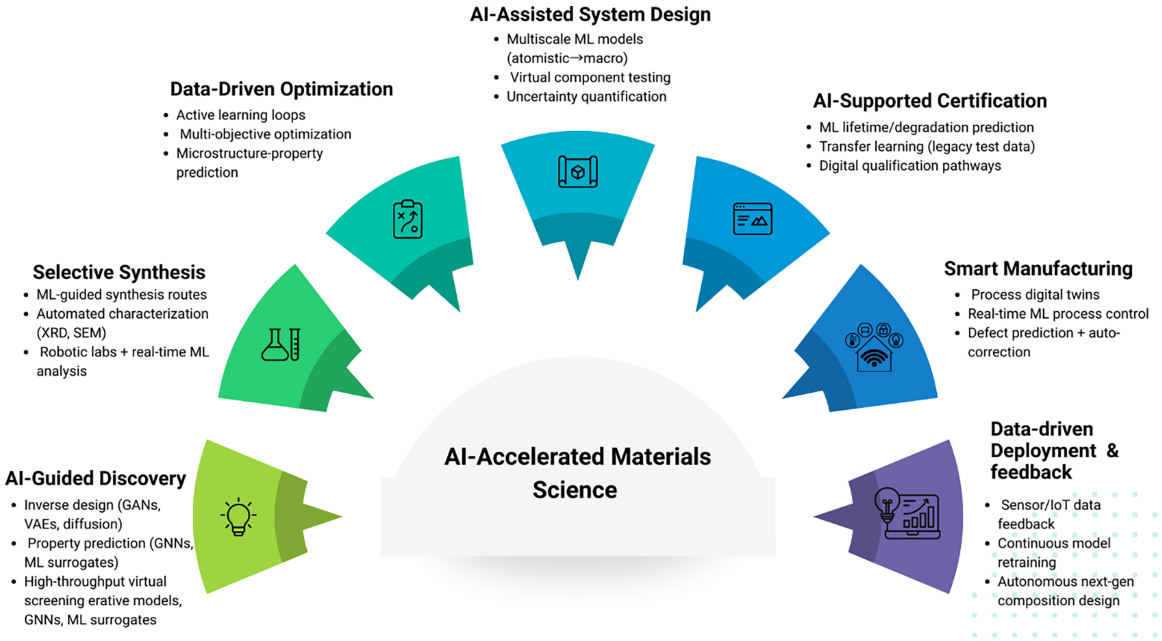

Figure 4: The process of finding new materials and their properties using AI-assisted methods.

Real-time data processing and automated characterization offer quick validation of expected materials. The experimental data are then used to optimize the data in a data-driven manner. In this respect, AI improves material composition and processing parameters based on new data. This measure reduces unnecessary experimentation and increases the efficacy of materials. After optimization, an AI-based system design is used to evaluate materials. Models are used to relate the properties of atoms at the atomic level to the behavior of components at the component level, and to predict the performance of components operating under different conditions. Virtual testing and uncertainty estimates can be used to assess reliability before actual use. This is followed by AI-assisted certification, in which machine learning algorithms determine material life and wear. The available experimental and historical data are reused to accelerate qualification and reduce the need for lengthy test cycles. In smart manufacturing, artificial intelligence monitors manufacturing activities in real time. Digital twins, sensors, and predictive models enable quality control, flaw detection, and the automatic adaptation of the production environment. Finally, the implementation and

feedback are data-based and finalize the process. The models encompass data collected during the use of materials, which support ongoing learning and improvement. Such feedback allows the creation of great materials in the next generation.

This shift was required to address the limitations of slow experimentation, high resource use, and low scalability inherent to traditional approaches, thereby enabling faster, more reliable, and cost-effective materials fabrication. Several reasons can support the choice of machine learning technologies over traditional discovery techniques. Such aspects include the ability of machine learning algorithms to process and analyze larger, more complex datasets, their capacity to identify patterns and relationships that traditional approaches might miss, and their ability to make more accurate predictions that are also more scalable. In addition, machine learning procedures enable automation and continuous improvement, significantly increasing the speed and precision of discoveries. These grounds are discussed in detail as explained below. The traditional approach in materials research has relied on intuition, assumptions, and trial-and-error experimentation, often influenced by human bias and restricting the exploration of possible compositions. Although it is a basic method, it lacks the scalability to explore the complexity of modern chemical space.

The AI approach replaces the empirical method with information-driven, closed-loop models that explicitly forecast relationships between structures, features, and functions. Recent developments have employed graph neural networks and attention-based transformers for catalytic activity prediction, achieving high accuracy that greatly exceeds that of traditional density functional theory (DFT). This transformation signifies a movement from descriptive to predictive science, consistent with the AI-driven discovery methodology. It employs inverse design frameworks, surrogate modelling, and generative AI technologies, including GANs, VAEs, and diffusion models, to enable rapid exploration of vast compositional spaces.

Conventional materials workflows rely on limited datasets, typically made of single experimental or computational records, and are not systematically integrated or standardized. This disaggregation limits generalization and reproducibility. On the other hand, AI-based materials science leverages large, multimodal inputs (text, images, spectroscopy, simulations, and so on) in consistent, structured formats. Examples of such initiatives include the Materials Graph Library (MatGL) [23] and the Open MatSciML Toolkit [24]. These systems enable data-driven optimization, where active learning loops, multi-objective optimization, and the prediction of microstructure-property relations form the basis for efficient model refinement. Old materials discovery, proposal to synthesis, characterization, and validation typically take several years, and active learning loops, multi-objective optimisation, and prediction of microstructure-property relations are often slow, feedback-driven, and manual processes. With the assistance of AI, more features can be explored at any given time by automating the design workflow. Nikolaev et al. [25] suggest that the autonomy in materials research combines active learning with robotic synthesis, cutting discovery time from months to hours. With reinforcement learning, the synthesis parameters are modified in real time, enabling closed-loop feedback and optimization. This solution supports the concept of intelligent manufacturing, where digital twins, process control, and defect prediction collaborate to continuously enhance production quality. In traditional materials science, much of the experimental design is done by humans, whether through design, measurement, or analysis of results. Such reliance on specialists may lead to delays and inconsistencies. Another AI substitute is algorithmic automation. The future alternative provided by AI is algorithmic automation, which is scalable. Recent technologies, including Chemistry Bayesian Optimization with LLM-Enhanced Multi-Agent System (ChemBOMAS [26]), applies Bayesian optimization and large language models (LLMs) to propose synthesis methods on its own. This is an AI-guided system design that bridges atomistic-level simulation and macroscale testing, and uncertainty quantification will be integrated into virtual component testing.

DFT and empirical models can provide valid predictions for well-understood systems, but face limitations in transferability and in complex chemical environments. In the meantime, quantum-chemical-level models are now accessible via deep learning at significantly reduced cost. Classical methods, such as density functional theory (DFT), require thousands of computationally intensive calculations to find a semiconductor with a 1.5 eV bandgap. On the other hand, more developed AI systems, such as Materials Transformer [27] have the ability to scan through millions of possible materials within a few minutes and predict band gaps to the accuracy of approximately 0.02 eV, comparable to the accuracy of DFT, but at much lower cost and time. These models facilitate AI-assisted certification by combining them with uncertainty quantification and digital qualification paths. They rely on predictive lifetime and degradation models to minimize reliance on large-scale physical validation, thereby increasing reliability under high-risk conditions [28].

Classical first-principles or molecular dynamics simulations are computationally intensive and are restricted to supercomputers, making them inelastic. Surrogate models and learning-based force fields enhance the computational efficiency of artificial intelligence techniques. The MACE-MP (Message Passing Atomic Cluster Expansion-Materials Project) potential reached density functional theory (DFT) level accuracy with an order of magnitude less computational effort. It enabled nanosecond-scale simulations previously unrealizable with *ab initio* methods. Moreover, sparse Gaussian processes and equivariant transformers are energy-efficient architectures that reduce energy consumption during training by 80%–90% compared to regular deep neural networks [29]. These models are data-driven, optimizing performance at any scale and incorporating IoT data to provide real-time updates. Conventional experimental methods require expensive reagents, are time-consuming, and entail high instrument operating costs.

Significant discoveries such as superconductors and perovskites resulted from chance experiments rather than planned research. AI turns this randomness into purposeful discovery by employing generative diffusion and reinforcement learning frameworks that strike a balance between novelty and feasibility, for example, Yao et al. [30] introduced "intentional serendipity," in which AI-driven searches balance innovation and functionality in the design of photonic materials. By measuring design uncertainties and novelty, AI reduces reliance on luck while still supporting creative exploration. Uncertainty arises from limitations in synthesis, unstable configurations, or ambiguous interpretations of data by conventional tools. Recent AI methods employ risk-aware models that quantify uncertainty and use probabilistic inference. For example, the BayesMat framework provides confidence intervals for the predicted thermal conductivities, which help determine a safe design space before synthesis [31]. The active-learning techniques aim to address areas of uncertainty, reducing experimental risks and waste, and promoting a risk-sensitive design methodology consistent with the reliability standards of the aerospace and biomedical industries. It is essential to have a full comprehension of complete systems. Traditional studies usually focus on the types of materials or the additions to properties, which is too narrow. AI extends this by examining a variety of properties, compositions, and environments at different scales and modes. For example, Materials Generative pre-trained transformer (MatGPT) is an incremental algorithm that combines text-mined scientific and numerical data with graph embeddings to analyze materials, including metals, ceramics, and polymers [32]. This aligns with the philosophy of designing AI-assisted systems, deploying them using numerical data, and providing feedback, with constant retraining using real-world sensor data expected to deliver progressive upgrades.

AI-based techniques are superior to classical techniques, particularly in high-dimensional material space or when solving multi-objective problems. For example, the development of new battery materials requires balancing thermodynamic stability, electronic conductivity, ionic diffusion, and electrochemical performance [33]. Conventional trial-and-error algorithms are often ineffective at searching such a complex design space. By contrast, machine learning models can process large amounts of data to reveal previously

unknown correlations between descriptors and conduct rapid screening of candidates [34]. Additionally, active learning strategies enable these models to iteratively identify the most informative experiments, thereby reducing the number of experimental iterations needed [35].

Table 1 indicates a shift in materials science, from people using their intuition to thinking to using machines to help them think, and summarizes the advantages of ML over traditional methods. The traditional methods, though still useful, are limited by their small size, manual testing, and fragmented data. Automated cycles of hypothesis generation, testing, and concluding AI-based approaches combine simulation, experimentation, and theory. It is not only a technological but also a conceptual change that discovery becomes more predictive, reproducible, and transparent. With the development of autonomous laboratories, generative networks, and foundation models, materials discovery is being brought closer to full automation, risk awareness, and collaboration across every corner of the globe. Machine learning has the potential to address the shortcomings of conventional trial-and-error materials development by providing a data-driven, closed-loop workflow that combines experimental data, simulations, and predictive modeling [36].

Table 1: Comparison of traditional vs. AI approaches.

Parameter	Traditional Approach	AI Approach
Methodology	Trial and Error	Data-Driven Predictive Modeling
Data Utilization	Limited data usage	Utilizes vast datasets
Speed of Discovery	Slow and time-consuming	Rapid discovery and design
Labor Intensive	Rely heavily on human expertise	Automation of tasks through algorithms
Prediction Accuracy	Variable and often inaccurate	High accuracy in predicting properties
Resource Consumption	High resource consumption	Efficient use of computational power
Costs	High experimentation costs	Reduced costs through simulation
Serendipity Dependent	Often relies on serendipitous findings	Systematic and controlled
Risk Management	Risky due to uncertainty	Risk reduction through data analysis
Innovation Pace	Slower innovation cycles	Accelerated innovation and development
Application Scope	Limited understanding of material behaviour	Comprehensive material insights

3 Machine Learning Workflow in Materials Science

The connections between data, models, and experiments include systematic workflows that are crucial to making good use of machine learning in materials science. Fig. 5 demonstrates a general machine learning workflow that encompasses data gathering, preprocessing, model generation, prediction, and feedback across many disciplines of materials science. Information is essential in ML processes [37]. In general, the quality and quantity of the data are decisive factors in the success of ML. The role of data preprocessing and feature engineering is emphasized, as it transforms data to make it easier to understand relationships among material physicochemical properties, predict material characteristics, and build predictive models. The studies apply natural language processing (NLP) to incorporate it in machine learning materials data preprocessing pipelines. Rather than rejecting failed or incomplete experiments, the study uses them as informative counterexamples, allowing the model to learn the limits of failure conditions (e.g., incorrect temperature ranges, stoichiometric ratios, or precursor sequences). Shaaban et al. [38] achieved a validation accuracy of almost 87 percent by training ML models on both successful and unsuccessful synthesis data. The failure data added to the system improved its ability to identify infeasible synthesis pathways before

experimental execution, thereby greatly increasing the efficiency of data cleaning and preprocessing and turning noise into a signal for predicting structure. Each of these workflow steps has been discussed in the following sections.

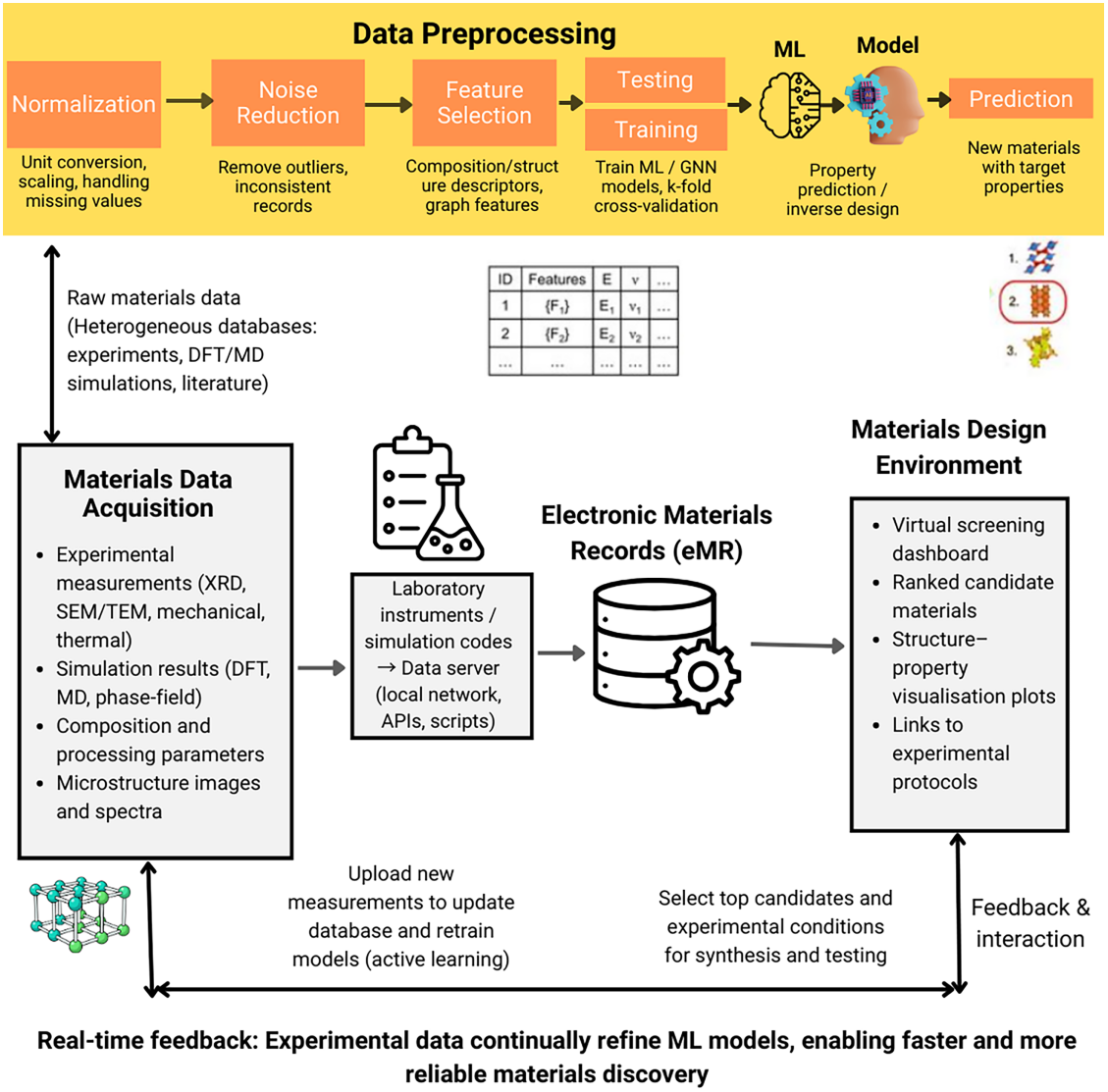

Figure 5: Comprehensive ML workflow for materials science with prediction and feedback-based refinement.

(i) Data Acquisition

High-quality, representative data is essential for dependable machine learning models. Collecting and validating data points across various material systems is crucial. The success of machine learning in materials discovery relies entirely on high-quality, representative datasets. Open-access materials databases have transformed data collection by offering structured, reliable data sources. Table 2 lists the major open-access databases, focusing on metals, polymers, ceramics, composites, and complex molecular structures. The datasets can provide unavailable information to predict and optimize the properties of new materials, thereby enabling ML models to be trained to create new materials.

Table 2: Central database sets for different types of materials, with relevant information.

Database	Material Types	Free/ Paid	Primary Research Objectives	Typical Properties	Recommended Research Use-Cases
MatWeb	Metals, polymers, ceramics, composites	Free + Paid	Property benchmarking, materials selection	Mechanical, thermal, and electrical properties	Baseline property prediction, ML regression training
MakeItFrom	Metals, polymers, ceramics	Free	Comparative materials analysis	Property comparisons, trade-off analysis	Feature screening, decision-tree models
Total Material	Metals, alloys, composites	Paid	Alloy design & selection	Mechanical, fatigue, corrosion	Process optimisation, alloy ML models
Matmatch	Metals, plastics, composites	Free	Supplier-linked materials selection	Property + supplier metadata	Industry-driven materials screening
CES Granta EduPack	Teaching datasets	Paid	Education & materials informatics	Property charts, Ashby plots	Descriptor understanding, ML pedagogy
ASM Materials Platform	Engineering alloys	Paid	Failure analysis & performance prediction	Fatigue, creep, corrosion	Lifetime prediction models
CINDAS	Aerospace alloys	Paid	Thermophysical modeling	Thermal expansion, conductivity	Multiphysics simulations
COD	Crystal structures	Free	Structure–property learning	CIF structures	Structure-based ML, GNNs
ICSD	Inorganic crystals	Paid	Crystal chemistry analysis	Lattice, symmetry	DFT validation, phase discovery
CSD	Organic & MOF structures	Paid	Molecular/crystal design	Packing, topology	MOF & organic ML models
AMCSD	Minerals	Free	Geomaterials research	Crystal structures	Earth materials modeling
PDB	Biomolecules	Free	Bio-materials modeling	Protein structures	Bio-inspired materials ML

(Continued)

Table 2 (continued)

Database	Material Types	Free/ Paid	Primary Research Objectives	Typical Properties	Recommended Research Use-Cases
Materials Project	DFT inorganic materials	Free	Property prediction & screening	Bandgap, voltage	Battery materials discovery
OQMD	DFT materials	Free	Phase stability prediction	Formation energies	Thermodynamic ML models
AFLOW	HT-DFT materials	Free	Automated materials discovery	Elastic, electronic props	Large-scale ML training
NOMAD	Computational materials	Free	Reproducible DFT workflows	Raw DFT outputs	AI-DFT benchmarking
NIST JARVIS	2D/3D materials	Free	ML-ready materials prediction	DFT + ML descriptors	GNN, surrogate modelling
C2DB	2D materials	Free	2D materials discovery	Electronic & magnetic props	Spintronics & battery anodes
SuperConductor (NIMS)	Superconductors	Free	Tc prediction	Transition temperature	Regression & physics-ML
MAPTIS	Aerospace materials	Free/Paid	Extreme-environment materials	High-T, radiation	Reliability modeling
CIRMS Data	Radiation materials	Mixed	Radiation damage prediction	Defect evolution	Nuclear materials AI
ICDD PDF	XRD patterns	Paid	Phase identification	Diffraction patterns	Image-based ML (XRD CNNs)
PoLyInfo (Polymer Database)	Polymers	Free	Polymer property prediction	Tg, modulus	Polymer ML
Polymer Property DB	Polymers	Free	Polymer selection	Mechanical & thermal	QSAR-type ML
CAMPUS Plastics	Commercial polymers	Free	Industry-grade selection	Processing data	Process optimization
NanoMine	Nanocomposites	Free	Structure–property learning	Filler dispersion	Image + tabular ML
Materials Commons	Materials data	Free	Data sharing & provenance	Experimental metadata	Reproducible research

(Continued)

Table 2 (continued)

Database	Material Types	Free/ Paid	Primary Research Objectives	Typical Properties	Recommended Research Use-Cases
GitHub Materials Lists	Meta-datasets	Free	Dataset discovery	Links & metadata	Rapid dataset access
OMDB	Electronic materials	Free	Electronic property prediction	DOS, Band structure	Condensed-matter ML
NREL	Photovoltaic materials	Free	Discovery of energy materials discovery	Efficiency data	Solar materials AI
Battery Materials Genome (DOE)	Battery materials	Free	Battery performance prediction	Voltage, diffusion	Na-ion/Li-ion ML models

By combining data from materials databases with training ML models, researchers can create more robust, predictive models. Bringing heterogeneous data sources together facilitates collaborative work and aligns with the principles of open science and domain-specific data curation to ensure transparent, reproducible data. As it is, a good example is the Computational 2D Materials Database (C2DB) by Haastrup et al. [39] which has the thermodynamic, elastic, magnetic, and structural properties of 1500 two-dimensional materials. The authors have maintained the database's source as open and allowed the application of ML models to it, and have presented a large number of potential new 2D materials.

(ii) Data Cleaning

Once the data has been gathered, raw datasets often contain duplications, blank spaces, and anomalies that may conceal significant trends. Thus, data cleaning is required to increase the accuracy and efficiency of models. This is typically done by sampling data, removing outliers, correcting them, discretizing, and normalizing. Data sampling will ensure smaller, more representative subsets are used to train the model, but at the cost of statistical integrity. Calibration eliminates noise and averts model bias, as demonstrated in the ML-directed defect prediction works by Lu et al. [40], who leveraged curated sets of first-principles simulations to better predict the bandgap and stability of hybrid perovskites. Discretization simplifies continuous attributes so that categorical trends can be learned. Normalization scales features to a fixed range required by algorithms such as neural networks and support vector machines, which rely on gradient-based optimization.

(iii) Feature Engineering

Feature engineering transforms raw or cleaned data into descriptive variables, or descriptors, that capture the essential physics or chemistry of a materials system. These descriptors serve as an intermediary between raw data and machine learning (ML) algorithms, helping models connect a material's composition or structure to its properties in a physically meaningful way [41]. New developments in automated feature engineering (AFE), such as deep learning and symbolic regression, have enabled the generation of descriptors directly from atomic structures or compositional data. To create proper descriptors, one still needs physical knowledge so that they are understandable and applicable across various instances. An effective descriptor must (1) be low-dimensional to avoid overfitting, (2) be a unique value of the relationship between

the material and its properties, (3) give similar values when comparable materials are used, and (4) be computationally inexpensive compared to the property of interest [42]. Descriptor design strategies fall into two categories: human-engineered descriptors, such as elemental electronegativity, ionic radius, and bond valence parameters, and algorithmically generated descriptors, such as convolutional voxel descriptors, which learn structure-property relationships directly from atomic configurations.

For example, in predicting the band gap of inorganic compounds, it was demonstrated that using only the chemical formula as input provides little insight into the underlying physics [43]. Nevertheless, by converting the composition into a set of features, including average electronegativity, mean atomic radius, and valence-electron concentration, the ML model can effectively determine the effects of differences in bonding and electronic structure on the band gap. This example demonstrates that features for which ML algorithms are designed can help learn the basic principles of materials without altering their original properties. Thus, feature engineering remains an essential process in materials informatics, connecting human-domain expertise with machine learning to identify governing principles within complex data.

(iv) Model Building

After feature engineering, the subsequent step is model building, during which machine learning algorithms learn the relationship between material descriptors and the target properties. The choice of algorithm is critical to prediction accuracy, interpretability, and generalization [44]. In materials science, algorithms are selected based on data size, feature type, and the problem's physical nature [45]. Linear models, e.g., linear regression, ridge regression, and lasso regression, are commonly used as baseline algorithms [46].

These models presuppose a linear correlation between attributes and target characteristics and are well interpretable. They especially come into play to explain the effects of individual material descriptors on properties; however, they fail to a large extent to handle complex, nonlinear relationships [47].

Among the most popular machine learning models are decision trees, random forests, and gradient boosting algorithms, all of which are based on trees. Decision trees are easy to visualize and simple, but they can overfit. Random forests enhance stability by combining multiple trees and work well with medium-sized data. Gradient boosting algorithms like eXtreme Gradient Boosting (XGBoost) and Light Gradient Boosting Machine (LightGBM) go further to correct errors sequentially and are therefore extremely useful for composition-based and process-related materials information [48].

Another successful type of algorithm, primarily used with small to medium data sets, is support vector machines (SVMs) [49]. SVMs can capture nonlinear relationships between features and attributes by using kernel functions. They have been successfully applied to phase classification, defect detection, and property prediction. Neural networks are typically used when the quantity and complexity of the material representation are large. Fully connected neural networks can model very nonlinear relationships, but require careful tuning and sufficient data. Graph neural networks (GNNs) like the Crystal Graph Convolutional Neural Network (CGCNN), Modality AGnostic NETwork (MEGNet), and ALIGNN operate directly on atomic structures as input and, preferably, predict more physically relevant results via a training, validation, and testing set split [50–52]. Before training, the dataset is divided into training, validation, and testing sets [53]. The model parameters are fitted using the training set. The validation set helps optimize hyperparameters and prevent overfitting. The test set will give a fair assessment of model performance with unknown data. Normally, the standard ratios are 70%–80% for training, 10%–15% for validation, and 10%–15% for testing [53].

In cases where datasets are small, k-fold cross-validation is the method of choice, as it allows evaluating the model multiple times on different data splits and obtaining more accurate performance estimates [54]. Python provides a rich environment for executing machine learning algorithms in materials science [55].

Scikit-learn is commonly used for linear models, decision trees, random forests, SVMs, and model evaluation. LightGBM and XGBoost are popular for high-performance gradient boosting. In deep learning, TensorFlow and PyTorch provide general-purpose frameworks for building neural networks. The most popular Python libraries for structure-based learning include CGCNN, MEGNet, PyTorch Geometric, and ALIGNN, which operate on material representations and are usually applied to the material domains, model architectures, and tasks described in Table 3 [52,56,57]. These libraries serve as the foundation for the practical implementation of AI-based workflows, including data preprocessing, model training, and prediction. The above-described workflow is a viable pipeline for the material discovery process when incorporating machine learning. In the first stage, scientists obtain information on materials databases or experimental outcomes. This is followed by cleaning and standardizing the dataset to address anomalies and missing data. After that, feature engineering transforms raw data into physical descriptors. Then, machine learning models are trained and tested on appropriate training and testing data. Lastly, the trained model would be used to identify potential candidate materials and to guide additional computational or experimental work. The workflow approach is a stepwise methodology to apply machine learning to the materials research process.

Table 3: Compilation of major Python libraries for the materials properties analyses.

Library	Typical Materials Domain(s)	Typical Model Architectures	ML Tasks
Pymatgen [1]	Crystalline solids, battery materials, diffusion, phase diagrams, etc.	Not an ML model	Data preprocessing for ML
M3Gnet [23]	Inorganic crystals; ML interatomic potentials	GNN with explicit 3-body interactions (Materials Graph Network)	Regression (E, F, stress), surrogate MD
Scikit-learn [46]	Tabular materials data; composition & hand-crafted features	Linear/Ridge/Lasso, SVM, kNN, RF, GB, PCA, k-means, GMM	Regression, classification, clustering, probability
LightGBM [48]	Large descriptor spaces, similar to XGBoost	Histogram-based boosted trees	Regression, classification, ranking
CGCNN [51]	Crystalline materials	Crystal Graph CNN	Regression, some classification
ALIGNN [52]	Crystals, 2D materials, MOFs, phonons, defect and vacancy energies (JARVIS, MP, QM9)	Line Graph Neural Network (bonds + angles)	Regression
Dscribe [55]	Solids, molecules, surfaces	Descriptors for SVM, GPR, kernels, GNNs	Regression, classification, clustering

(Continued)

Table 3 (continued)

Library	Typical Materials Domain(s)	Typical Model Architectures	ML Tasks
TensorFlow [56]	Crystals, molecules, process–property data; DFT/MD surrogate	DNN, CNN, RNN/LSTM, Transformers, GNN, VAE, GAN	Regression, classification, generative, probabilistic DL
MEGNet [58]	Molecules, crystals (MP, QM9)	Graph Neural Networks	Regression, multi-task
PyTorch [59]	Similar to TensorFlow, widely used for crystal/molecule GNNs, MLIPs	DNN, CNN, RNN, Transformers, GNN	Regression, classification, generative, probabilistic
Keras [60]	Rapid prototyping; small–medium datasets	DNN, CNN, LSTM, simple Transformers	Regression, classification, generative
XGBoost [61]	Feature-based property prediction (band gap, stability, mechanics)	Gradient-boosted trees	Regression, classification
Matminer [62]	Inorganic crystals, alloys (MP, OQMD, AFLOW)	Uses scikit-learn/XGBoost	Feature generation; regression/classification
ASE (Atomic Simulation Environment) [63]	Atomistic systems: surfaces, catalysis, bulk, clusters	Interfaces with ML potential & DFT	Simulation; training data; MLIP deployment
GPAW [64]	Electronic-structure datasets (DFT, TDDFT)	DFT reference generator	Produces labels for supervised ML
MODNet [65]	General properties: small datasets	Sparse feed-forward NN with feature selection	Regression, multi-target
SchNetPack [66]	Molecules, crystals, atomistic systems	Continuous-filter CNNs, symmetry functions	Regression, generative
PyTorch Geometric (PyG) [67]	Generic GNN framework for molecules/crystals	GCN, GAT, GIN, message passing	Regression, classification, link prediction
JAX-MD [68]	Differentiable molecular & materials MD	Classical + neural (GNN) potentials	Regression, differentiable simulation

4 Different Applications of Machine Learning in Material Science

Machine learning plays a role across various areas of materials science, including energy materials, structural alloys, electronic materials, polymers, and catalytic systems. Machine learning models are useful for predicting electrode stability and ionic diffusion rates in battery development [69]. AI is used in the design of alloys and microstructure in the field of metallurgy. Machine learning in Polymer informatics is used to predict glass transition temperatures and mechanical properties [70]. Several studies have also documented comparisons of ML model accuracy in predicting enthalpy and formation energy for chemical alloy systems. One such model for predicting the formation of binary alloys based on composition, lattice type, and lattice configurations is described below and summarized in Table 4. The paper by Nyshadham et al. [71] marks a turning point in the incorporation of machine learning into the field of computational materials discovery. It demonstrates how AI-enhanced surrogate modeling can be used within traditional DFT workflows to address DFT's inherent constraints without sacrificing predictive power. The researchers tested this hypothesis using the DFT-10B dataset, a highly curated collection of 15,950 binary crystal structures of 10 metals (AgCu, AlFe, AlMg, AlNi, AlTi, CoNi, CuFe, CuNi, FeV, and NbNi). The dataset comprised unrelaxed crystal structures of face-centered cubic (fcc), body-centered cubic (bcc), and hexagonal close-packed (hcp) crystal lattices, with eight atoms per cell. The DFT-computed formation enthalpies were also used to describe the structures and were determined using the Perdew-Burke-Ernzerhof (PBE) generalized gradient approximation. The dataset provided an ideal test bed for evaluating how results can be generalized across lattice types and alloy systems, given the dataset's great chemical and structural diversity. The authors have taken into account five representative methods of modeling, including traditional and state-of-the-art machine learning methods, the traditional method of the cluster expansion, two models of the Many-Body Tensor Representation (MBTR) with the kernel ridge regression (KRR) and the deep neural networks (DNN), a Smooth Overlap of Atomic Positions (SOAP) representation with the Gaussian Process (GP) regression, and a Moment Tensor Potential (MTP) with the polynomial regression. The authors did not aim to optimize hyperparameters or adapt the approaches to a specific dataset, but rather to underscore the overall applicability, strength, and reproducibility of the methods across systems. This technique predated subsequent developments in fundamental theories of materials science.

Table 4: Overview of AI/ML methods and applications in materials science.

Reference	Datasets	Extracted Features	Research Gaps	ML Models and Algorithms	Prediction
[71]	Binary alloys	Band gap, formation enthalpy, elastic constants	Dataset size	KRR, GPR	Formation enthalpy prediction
[72]	10,000 pairs of perovskite oxides	Atomic, structural, electronic, and molecular	Overfitting, poor generalization	GCNNs	Lattice constants, Atomic ordering
[73]	Not specified	Band gap	Limited practical validation; scalability unclear	RF (R^2 = 0.685, RMSE = 0.87 eV)	Band gap
[74]	Open datasets	Structural maps (tolerance, octahedral parameters)	Ignore electronegativity, covalency	SISSO	Phase diagrams, perovskite formability

(Continued)

Table 4 (continued)

Reference	Datasets	Extracted Features	Research Gaps	ML Models and Algorithms	Prediction
[75]	NIST, ASM, Knovel	Composition, processing, microstructure	Small datasets; unbalanced coverage	ANN, DNN, CNN, regression, Naïve Bayes	Phase, grain size, porosity, fatigue
[76]	MP, JARVIS	Degradation metrics	Small, incomplete datasets	XGBoost, RF, ALIGNN	Band gap, energy
[77]	66,981 polymers	SMILES (1024-bit vectors)	Feature enrichment, dataset augmentation	Lasso, Elastic Net, DT, XGB, SVR	T_g, T_d, T_m
[78]	DFTB energy data	HOMO, LUMO, band gap	Needs broader validation	Pure2DopeNet, ResNet, ViT	Electronic properties
[79]	AIMD optical datasets	Bond topology, atomic features	Not discussed	ChemGNN, PyG	Band gap in g-C_3N_4
[80]	SUNSET (30,000 multi-shell UCNP spectra)	Shell thickness, dopant, UV intensity	Representation and training data limits	RF, CNN, GNN	Inverse UCNP design
[81]	MoS_2 supercapacitor data	d-spacing, ion size, molarity, hydration energy	Limited 2D datasets	XGBoost, RF, SHAP	Capacitance
[82]	350 EA values	200 RDKit descriptors	Limited design strategies	kN regressor, GB, RF	Electron affinity
[83]	Stress-strain literature	Modulus, strain rate, grain size	Hidden structure-property links	KME model	Dislocation density evolution
[84]	1000 polymers	400–600 descriptors	Lack of integrated frameworks	RF, bagging, GB	Thermal conductivity
[85]	149,952 perovskites	O_p/M_d band centers	Avoid full DFT	CGCNN	OER catalyst screening
[86]	5741 magnetic materials	Elemental vectors (518 features)	Excludes AFM/non-collinear	LightGBM	New magnetic materials

The results showed highly homogeneous accuracy, with a mean error of less than 10 MeV/atom and a relative formation energy error of less than 2.5%. These values almost matched the DFT accuracy in all alloy systems and representation types. The analysis demonstrated that the machine learning algorithm (KRR, GP, or DNN) was less important than the quality and symmetry-conserving character of the atomic representation used to model the local environment. Rotational, translational, and permutational invariance were represented as MBTR, SOAP, and MTP, respectively, and led to their representation in structural variations that pose problems for discrete lattice-based methods, such as cluster expansion. A significant advance was the development of models trained on multiple alloys simultaneously. These multi-alloy models were as good as, or slightly better than, single-alloy models, even though they ought to be less accurate, and average errors were below 1 meV per atom. This indicates that a single surrogate model can reveal common trends across diverse chemical systems, and that cross-domain material predictors can be

transferred. These models can be used to predict formation enthalpies for a variety of lattice structures, as they are trained on unrelaxed structures and are useful for high-throughput pre-screening of many lattice configurations, where structural relaxation is expensive. The study demonstrates that surrogate models can achieve the same accuracy as DFT at a fraction of the cost, providing a scalable method for scaling up high-throughput materials design. These models can immediately predict the formation energy of thousands of candidate compounds, whereas traditional workflows require an independent DFT calculation for each new material. Such models can be integrated with inverse design, in which structure and composition are informed by desirable properties rather than the other way around, enabling rapid prediction. Since the proposed surrogate approach to modeling has been the subject of numerous studies in catalysis, alloy design, and semiconductor discovery, generalized interatomic potentials are important for mediating the performance gap between atomic-scale accuracy and device-scale performance. Along with these progresses came limitations outlined in the study, including the fact that the dataset included only unrelaxed structures (which might lead to formation energy predictions that do not agree with fully relaxed DFT or experimental measurements). Finally, the case study shows how materials science has been transformed by automating and data-driven reasoning rather than manual, hypothesis-driven exploration.

5 Use of Generative AI in Material Science

Generative Artificial Intelligence (Generative AI) introduces a new capability in materials science by enabling the generation of new material compositions, structures, and design candidates, rather than only predicting properties of existing materials [87]. Traditional machine learning models answer forward problems, such as predicting properties from known inputs. In contrast, generative AI addresses inverse problems, where the goal is to design materials that satisfy target properties. Different generative algorithms are used depending on the application; for example, Generative Adversarial Networks (GANs) consist of two neural networks, the generator and the discriminator, which are trained simultaneously in a competitive setting [88]. The generator generates artificial data samples that should be similar to data from real materials, and the discriminator differentiates between real and generated data. GANs are trained through this form of competition, where the outputs are naturalistic. GANs are especially good in image-related tasks in which visual realism matters. They are widely used to produce microstructure images, synthesize realistic material datasets, and aid image-based materials analysis. For example, Scanning Electron Microscopy (SEM) microstructure images generated by GANs have been used to supplement small experimental datasets, enhancing the strength and quality of machine learning models trained on small data [89]. Diffusion models are another type of data generator that produces meaningful, structured samples of new data by successively applying random noise to images in a series of learned denoising autoencoders. Diffusion models are more stable to train and less susceptible to model failure than GANs [90]. These models are particularly well suited to form complex, high-dimensional structures, making them appealing for materials science work that requires atomic-scale representations, as in Fig. 6.

Crystal structure generation, alloy design, and exploration of inorganic materials have been addressed using diffusion models. They can generate new atomic configurations that satisfy symmetry and stability constraints, making the materials they produce more realistic when trained on crystal structure databases [91]. Graph-based generative models model materials using graphs, where atoms are nodes and chemical bonds or interactions are edges. Structural and bonding information is automatically incorporated into this representation, which is essential for proper modeling of materials with complex atomic structures. These types of models are quite useful for the production of crystal and molecular structures, as well as for inorganic materials with complex bonding environments. By learning patterns from known materials, graph-based generative models can suggest new structures that adhere to realistic chemical and structural rules. This

structure-aware approach enhances physical plausibility compared to models that rely solely on vector-based methods [92,93]. Large language model-based generation is structure-aware and enhances physical plausibility to models that operate solely on numerical information, and not on text-based and symbolic information. These systems are trained on extensive amounts of scientific literature and can handle research papers, reports, and descriptions of experiments. In materials science, large language models are useful for generating material hypotheses, synthesizing existing information, proposing synthesis pathways, and relating information across various studies. They may be used to assist researchers in navigating complex literature and converting high-level objectives into research actions. Large language models, combined with predictive models and databases, can thus serve as smart assistants to enhance accessibility, efficiency, and knowledge integration in the materials research process [94].

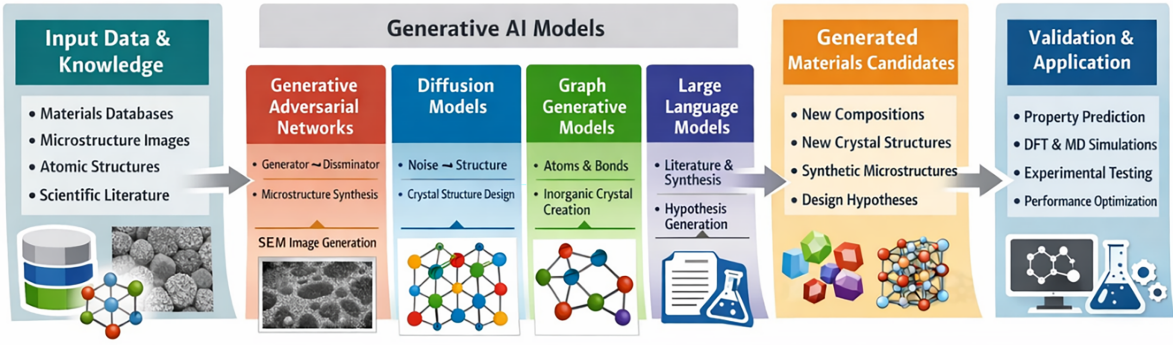

Figure 6: Generative paradigms in material science.

Problems of GenAI in Materials Science

The key issue is that the quality, diversity, and completeness of training data are highly relied upon by the generative models. If the available datasets are sparse, or biased toward certain material classes, the produced materials may be low in novelty or fail to yield physically meaningful solutions [95]. Moreover, not all generative models directly impose physical, chemical, or thermodynamic limitations, potentially leading to the creation of physically theoretically sound but, in reality, physically unstable materials [10]. The other weakness is model interpretability and trust. Generative AI models can be viewed as black boxes, and the process by which a specific material design was produced is difficult to comprehend, as is the set of factors that contributed to the final result [96]. This lack of transparency can undermine confidence in model predictions, particularly in safety-critical or high-cost materials applications. Additionally, computational cost is a concern, as training diffusion models, graph-based generators, and large language models demand significant resources and time, which can limit access for some research groups [97]. Experiments or high-fidelity simulations should validate the results of the generative AI systems. These materials must be considered hypotheses, not final solutions, because issues such as experiment ability, synthesis limitations, and practical implementation cannot be guaranteed. In the case of large language model-based systems, another challenge is the potential for partial or incorrect recommendations, as models rely on the literature and may disseminate outdated or context-specific information. To successfully use generative AI in materials science, domain knowledge must be combined with physical constraints and validation techniques to yield dependable, useful solutions [98]. In addition to data quality issues, one challenge in using generative AI in materials science is ensuring that generated candidates comply with basic physical and chemical constraints. Most models are trained solely on statistical patterns in the dataset, which can result in the generation of materials that appear mathematically possible but violate principles such as thermodynamic stability, crystal symmetry, or realistic bonding. To alleviate this, recent studies are increasingly using physics-informed

machine learning in generative systems. For example, thermodynamic quantities generated from structures can be filtered (e.g., formation energy or energy above the convex hull), and graph neural networks that are aware of symmetry can be used to ensure that the predicted crystal structures obey physically significant bonding patterns [87].

The seriousness of this issue is that it is difficult to generalize these models to materials that are not encountered during training. Computational database-trained models can perform well near chemical systems but fail when predicting stable compounds in unexplored regions of materials space. This highlights the relevance of quantifying uncertainty and of active learning methods, which continuously improve models by refining them with new data, whether generated or experimentally verified and tested [87].

Finally, the results of generative AI should be treated as hypotheses rather than discoveries. These materials usually require additional justification, such as high-fidelity calculations (e.g., density functional theory) or experimental synthesis, to confirm their thermodynamic stability, structural viability, and functionality [99].

6 AI Agents and Autonomous Materials Research

AI agents are systems that can make decisions, take actions, and learn from results with minimal human involvement. Unlike standard machine learning models that only predict or generate results, AI agents manage multiple steps of the materials research process in an organized way [100]. They observe data, decide on next steps, and improve their decisions over time. In materials science, AI agents are used to create autonomous research workflows. These agents aim to achieve a specific goal, such as improving material property or finding an optimal composition. To do this, they interact with predictive and generative models, as well as with experiments or simulations. Common decision-making algorithms used by AI agents include reinforcement learning, Bayesian optimization, and active learning. These algorithms help the agent [101] decide whether to explore new material options or focus on the most promising ones, as shown in Fig. 7.

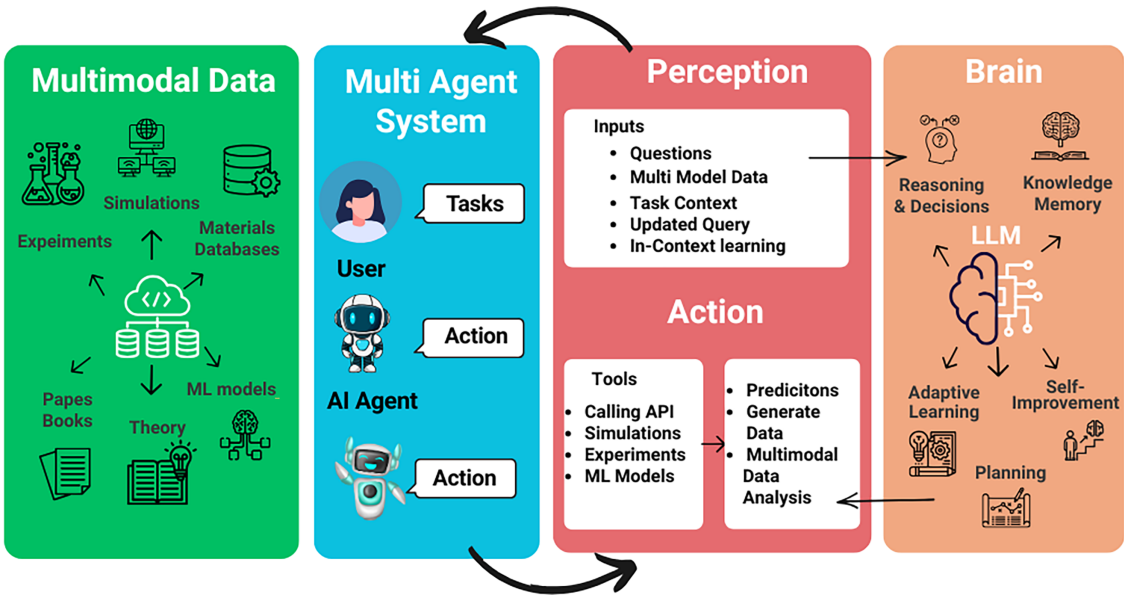

Figure 7: Multimodal multi-agent AI system for autonomous materials research.

Fig. 7 presents a multimodal, multi-agent AI framework designed to support intelligent and autonomous workflows in materials science. The framework incorporates knowledge of heterogeneous materials, including experimental and simulation data, materials databases, scientific literature, physical theory, and machine-learned models. A combination of these sources creates a comprehensive body of knowledge that integrates information-based and physics-based insights into materials systems. The interaction between human researchers and AI agents can be enabled through a multi-agent layer [102]. The researcher states the tasks in this layer, including property prediction, materials optimization, and inverse design, and notes that various AI agents perform specific actions. Individual agents can specialize in analysis, simulations, or experimental planning, enabling complex materials problems to be tackled in an integrated and effective manner. The perception component handles all incoming information, such as research questions, multimodal materials data, and task context. It helps the system identify the problem correctly, and it will update in accordance with new queries or data made available. This situational perception enables the framework to respond dynamically throughout the research process. A large-language-model-based intelligence core is capable of central reasoning and decision-making. This element incorporates reason and planning, knowledge storage and retrieval, adaptive learning, and personal enhancement. It links previous scientific understanding to new information generated, establishes the next steps, and makes them scientifically based rather than isolated predictions, as it allows the higher levels of reasoning in the sciences. The action component connects the AI system with computational and experimental devices. It enables physics simulations, experimental work, API calls, and the assessment of machine learning models. Such moves generate predictions, generate new data, and analyze multimodally. The feedback on the results is consistently entered into the system, allowing learning and refinement anew. Look at a new design for a battery electrode material.

A researcher states the objective, including the high energy density and stability. The AI agents process available experimental and simulation data, generate new candidate compositions using generative models, and simulate plans to assess their performance. The reasoning core then picks up the most promising candidates and proposes experimental validation measures. The outcome of the experiments is then fed back into the system, enhancing future predictions and design choices [103].

Physical modeling is uncommon in AI agents that do not involve working in real-world autonomous research systems. On the contrary, their machine learning results are checked with first-principles simulations or experiments in a closed-loop system. An example here is using AI-generated candidate material, which can first be evaluated using a surrogate model, followed by verification using density functional theory or automated synthesis. This hybrid approach will ensure that important physical limitations-thermodynamic stability, reaction kinetics, and synthesis accessibility are incorporated into the discovery process.

7 Case Study: Formation Energy-Driven Optimization of Sodium-Ion Battery Materials Using the Materials Project Database

This section describes the step-by-step procedure for extracting data from online databases and developing a machine learning workflow to predict materials and their desired properties. Instead of running quantum-mechanical simulations, the goal of this study is to develop a data-driven model that learns the patterns of material stability directly from existing computed materials data. A detailed workflow for initiating a machine-learning analysis of online databases is shown in Fig. 8.

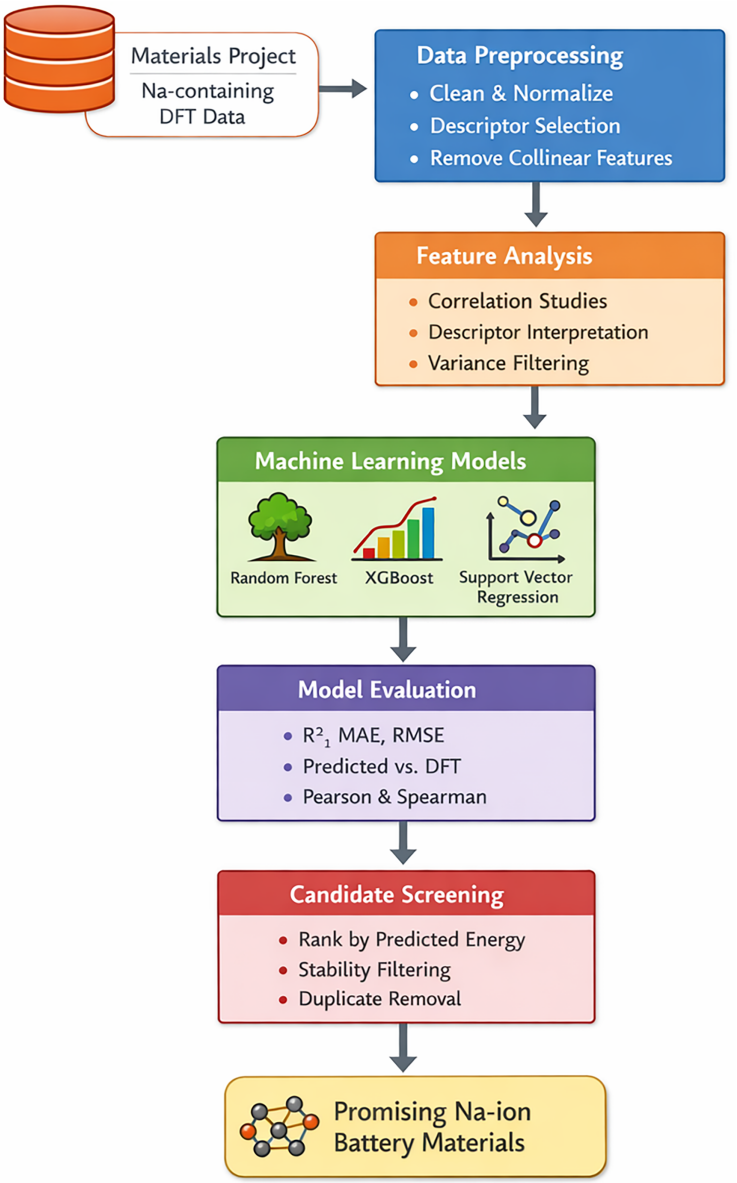

Figure 8: ML workflow for formation energy prediction and stability screening of sodium-containing materials using materials project data.

The steps are as follows: Jupyter Notebook was used to install essential Python libraries, including pymatgen, emmet-core, and mp-api. Compatibility patches were applied to enable seamless data extraction from online databases. A user account was created on the Materials Project website, and an API key was obtained for integration with the Python libraries. A query for sodium-based materials was performed, specifying that each compound must contain at least one sodium atom and no more than five total atoms. From about 1000 compounds retrieved from the database, 563 were selected based on the criterion that their energy above the hull was less than 0.05 eV. For these 563 inorganic sodium-containing compounds, 154 descriptors were extracted using density functional theory (DFT) and composition-based feature engineering. These descriptors are electronic, thermodynamic, and compositional properties applicable to sodium-ion battery materials. Fields containing identifiers were dropped to prevent data leakage, leaving only

physically useful numerical descriptors. This data was reduced by parsing chemical formulas for actinides, fluorides, and rare earths to eliminate those materials. This move narrowed the model to a sodium-ion battery chemistry of interest in electrochemical reactions. Noise, null, and missing entries were also removed from the dataset, leaving 415 valid samples. The primary regression variable was formation energy (eV/atom), which measures the thermodynamic and synthetic viability of sodium-ion cathode materials. The first set of features consisted of 151 numerical descriptors. Features with low variance (<0.01) were dropped, reducing the number of descriptors to 132. Fig. 9 presents the distribution of the DFT-calculated formation energies of the sodium-containing compounds in the Materials Project. Such a wide range of values indicates the presence of both stable and metastable materials, providing a strong basis for supervised machine learning.

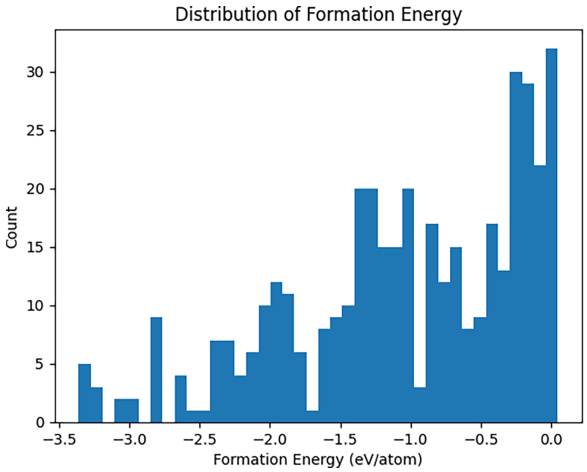

Figure 9: Distribution of formation energies for sodium compounds from the materials project, supporting supervised machine learning.

A Pearson correlation analysis was conducted between each remaining descriptor and formation energy to discover those that are highly correlated with thermodynamic stability. The strongest correlations were observed with descriptors based on electronegativity-, valence-, and space-group-related properties, which provide preliminary physics-based insights into what influences stability in sodium-based materials. To avoid redundancy, highly correlated descriptors were selected using an absolute correlation cutoff of 0.90. The features that exceeded this threshold were discarded, yielding a final number of 67 independent features. This set comprises compositional statistics, electronic structure descriptors, bonding features, and structural measures that provide complementary physical data for sodium-ion battery materials.

Pearson and Spearman correlation analyses were conducted between the chosen descriptors and formation energy. The maximum linear correlation was observed with MagpieData's maximum Electronegativity, explaining 78.08% of the variance in formation energy. Other descriptors showing strong correlations include deviations in electronegativity, p-valence features, and space-group metrics. The Spearman analysis indicated strong monotonic but nonlinear relationships, particularly for electronegativity- and valence-related descriptors, which justifies the use of nonlinear machine learning models. Pearson correlation for the top 10 descriptors is shown in Fig. 10.

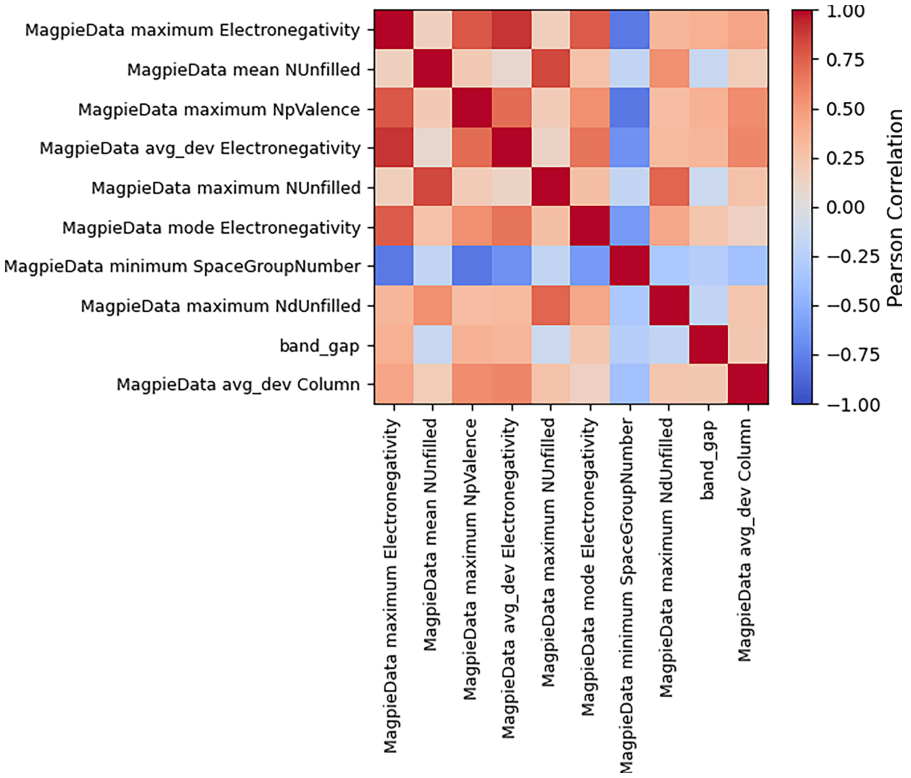

Figure 10: Correlation matrix of the top ten machine-learning descriptors.

The curated dataset was split into training and test sets at 80/20, yielding 332 training and 83 test samples. Feature scaling was applied where required. Three regression models were trained and evaluated: Random Forest regression, XGBoost regression, and Support Vector Regression (SVR) as a baseline. Tree-based ensemble models were selected for their ability to capture nonlinear interactions among descriptors common in materials datasets. Model performance for the prediction of formation energies was evaluated using the coefficient of determination (R^2), mean absolute error (MAE), and root-mean-square error (RMSE). A comparison of the different machine learning models' R^2, MAE, and RMSE values is shown in Table 5.

Table 5: Comparison of ML models for formation energy prediction.

Model	R^2	MAE (eV/atom)	RMSE (eV/atom)
Random Forest	0.958	0.107	0.154
XGBoost	0.965	0.092	0.140
SVR	0.961	0.089	0.149

Machine learning models accurately predict the formation energy of sodium-containing compounds, indicating that thermodynamic stability can be learned from composition-based descriptors. The three models have R^2 values greater than 0.95, indicating that the selected features are strong predictors of the key chemical factors that affect stability. Of all the listed methods, XGBoost provides the best overall results, with the largest R^2 = 0.965, the lowest mean absolute error of 0.092 eV per atom, and the lowest RMSE of 0.140 eV per atom, making it suitable for high-throughput screening. The random forest model is also accurate and captures nonlinear trends, but it shows higher error at the extremes. Support Vector Regression

(SVR) has the smallest average error, yet it is more susceptible to bigger errors. On the whole, these results indicate that ensemble machine learning algorithms can reproduce DFT formation energies with high accuracy, allowing them to rank stability and reliably discover battery-relevant sodium materials. Although machine learning models can predict formation energy with high accuracy, thermodynamic stability alone does not guarantee successful experimental synthesis or optimal experimental performance. Other critical factors, like ionic diffusion rates, structural stability over cycles, and electrode-electrolyte interactions, are also essential when assessing potential battery materials. Thus, the model introduced here should be viewed as a preliminary screening method that identifies promising candidates for further computational and experimental studies.

Fig. 11 further demonstrates that there is a close overlap between the DFT-computed formation energies and those obtained by machine learning. Table 6 displays the top ten descriptors that the machine-learning model could most use to predict the formation energy of sodium-containing materials. The strongest aspect is MagpieData's maximum electronegativity, with more than half of the total importance (0.5537) and negative Pearson (−0.7808) and Spearman (−0.8509) correlations. This shows that the more electronegative constituent elements are more likely to have lower formation energies and, hence, be more thermodynamically stable. Some descriptors related to the electronic structure, such as NpValence, NdUnfilled, and NUnfilled, show strong negative correlations, indicating that the degree of valence-electron configuration and the presence of unfilled electronic states significantly impact stability.

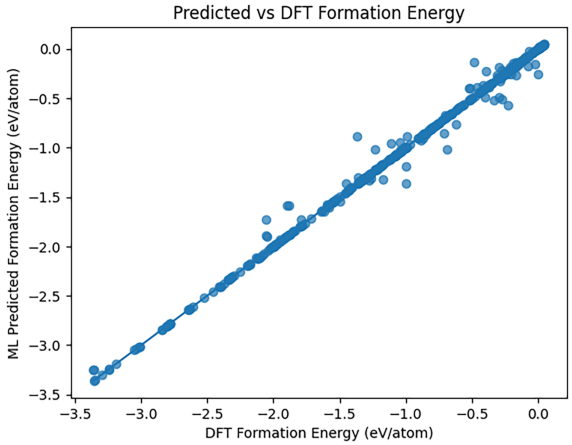

Figure 11: Comparison between DFT-calculated and machine-learning-predicted formation energies.

Table 6: Feature importance and correlation of the top ten descriptors for formation energy prediction.

Descriptor	Feature Importance	Pearson Corr.	Spearman Corr.
MagpieData maximum Electronegativity	0.5537	−0.7808	−0.8509
MagpieData mean NUnfilled	0.1295	−0.256	−0.2541
MagpieData maximum NpValence	0.1291	−0.6872	−0.8113
MagpieData avg_dev Electronegativity	0.0215	−0.772	−0.8205
MagpieData maximum NUnfilled	0.0215	−0.3691	−0.3248
MagpieData mode Electronegativity	0.0114	−0.6896	−0.7319
MagpieData minimum SpaceGroupNumber	0.0114	0.7242	0.7541

(Continued)

Table 6 (continued)

Descriptor	Feature Importance	Pearson Corr.	Spearman Corr.
MagpieData maximum NdUnfilled	0.0108	−0.5313	−0.5144
band_gap	0.0107	−0.3694	−0.4378
MagpieData avg_dev Column	0.0094	−0.3671	−0.3837

Descriptors related to electronegativity distribution (average deviation and mode) further confirm that chemical heterogeneity within a compound plays a key role in determining formation energy. The space-group number exhibits a positive correlation with formation energy, suggesting that structural symmetry contributes to stability trends and influences how atomic arrangements affect thermodynamic behaviour. The inclusion of band gap among the top descriptors indicates a meaningful relationship between electronic properties and material stability.

Machine-learning predictions were reattached to the original compounds to enable materials screening. Compounds were ranked by predicted formation energy, with more negative values indicating higher thermodynamic stability. Additional stability filtering, using energy above the hull <0.05 eV, was applied to retain only chemically plausible candidates. This process yielded a concise list of highly stable sodium-containing compounds suitable for further investigation. Table 7 shows the top 10 predicted materials suitable for Na-ion batteries based on redox and electronic properties. The literature also supports these findings: $NaTi_8O_{13}$ has been synthesised and structurally characterised previously, and hybrid $NaTi_8O_{13}$/$NaTiO_2$ nanoribbons have demonstrated promising Na-storage properties, supporting our identification of $NaTi_8O_{13}$ as a viable sodium intercalation host [104]. The Wadsley-Roth-derived $NaNb_{13}O_{33}$ has been shown to exhibit high conductivity and rapid insertion behavior in recent studies, indicating that $NaNb_{13}O_{33}$-type niobates are promising for high-rate electrodes [105]. Although $NaTaO_3$ is predominantly reported for photocatalysis, its perovskite-related structure and dopability are well documented and can guide future doping strategies to tune electronic properties for electrochemical applications [106].

Table 7: Top predicted low-formation-energy sodium compounds for further investigation.

	Material_ID	Formula	Predicted_Formation_Energy	Energy_Above_Hull
397	mp-760024	$NaTi_5O_{10}$	−3.358289	0
372	mp-860798	$NaAl_{11}O_{17}$	−3.351149	0.008215
288	mp-757433	$NaTi_4O_8$	−3.297521	0.023439
92	mp-28649	$NaTi_8O_{13}$	−3.245294	0.00493
268	mp-1221025	$NaTi_3O_6$	−3.24337	0.011656
245	mp-7914	$NaScO_2$	−3.190947	0
135	mp-4675	$NaTaO_3$	−3.050717	0.009339
181	mp-672212	$NaNb_{13}O_{33}$	−3.014563	0
119	mp-4514	$NaNbO_3$	−2.843733	0
280	mp-557406	NaB_3O_5	−2.784302	0

This case study serves as a tutorial, outlining the workflow for a materials researcher to follow and demonstrating how to use machine learning models with their collected experimental and theoretical

datasets. Apart from sodium-ion battery materials, the machine learning approach outlined here can be broadly applied to other materials discovery challenges. The same data-driven process, covering database extraction, descriptor creation, feature selection, model training, and stability screening, can be tailored to different materials systems. For instance, in polymer informatics, the prediction of thermal or mechanical properties, and in alloy design, identifying compositions with specific strength [107,108]. Models of electronic structure descriptors have also been used to identify candidate materials with desirable band gaps and charge-transport properties in semiconductor research [109]. As these examples indicate, this methodology is a flexible AI-based discovery pipeline rather than a system-based approach, underscoring its wide applicability across materials science.

Predicted and DFT-calculated formation energies across the entire dataset were evaluated using parity and statistical measures. The most robust compounds have strongly negative predicted formation energies and characteristic profiles of descriptors typical of a high electronegativity contrast, predominantly p-valence character, and an intermediate mismatch in atomic size. This method, together with the energy above the hull-based stability filtering, produces a short list of chemically allowable and thermodynamically stable candidates to be considered in the future as first-principles materials and experimentally validated for use in the sodium-ion battery. The parity analysis and statistical correlation measures were used to measure the agreement between machine-learning-predicted and DFT-calculated formation energies. There is a strong linear relationship, with a Pearson correlation coefficient of r = 0.9972, indicating almost perfect agreement between predicted and reference values. The Spearman rank correlation coefficient (ρ = 0.9948) indicates that the model correctly maintains the relative ordering of compounds containing sodium across the entire dataset, which is essential for stable materials screening. The large coefficient of determination (R^2 = 0.994) is also evidence that the model captures more than 99 percent of the variation in DFT formation energies. All of these findings demonstrate that the machine-learning model achieves nearly DFT accuracy and faithfully reproduces the underlying thermodynamic trends that govern the behavior of sodium-based materials, confirming its usefulness for high-throughput discovery and screening of sodium-ion battery candidates. This data confirms that the machine-learning model achieves almost-DFT accuracy and maintains the ranking of stability across materials containing sodium.

8 Conclusion

This review provides an in-depth analysis of the current materials discovery revolution, driven by artificial intelligence and machine learning, and how it is overcoming old trial-and-error methods with data-driven, autonomous research approaches. It summarizes three viewpoints: a conceptual summary of AI-driven materials discovery, a practical machine-learning workflow of materials informatics, and a tutorial example of sodium-ion battery materials. The paper initially traces the historical development of discovery paradigms in materials science and how machine learning, generative models, and autonomous systems are transforming the materials design process. In comparison to conventional processes, AI workflows can be used to explore complex material spaces more rapidly, make predictions more accurately, and combine experimental, computational, and literature data into unified pipelines. The comparative analysis of conventional and AI-aided discovery highlights how predictive modeling, active learning, and closed-loop experiments can accelerate the identification of promising materials. Second, the study's workflow is structured as follows: data collection, preprocessing, feature engineering, model building, and validation. It contains databases of popular materials, machine-learning software, and Python packages, which serve as a helpful guide for any materials researcher looking to use AI in their work. It also stresses that meaningful, physically significant descriptors and high-quality datasets are needed for reliable, interpretable models. Third, this workflow is used to predict the formation energies of sodium-based compounds of interest in sodium-ion battery

cathodes, using the Materials Project in a tutorial case study. Ensemble models, particularly XGBoost, were found to be highly accurate with an R^2 of over 0.96 and a mean absolute error of under 0.09 eV per atom. The descriptors based on composition were effective at describing important chemical variables that affect stability. An analysis of the importance of the features revealed that the distribution of electronegativity and valence electrons plays a major role in deciding the stability of sodium compounds. The fact that machine-learning results are similar to those of DFT calculations suggests that these models can be used to quickly screen thermodynamic results. Nevertheless, the adoption of AI in materials discovery remains a challenge despite the progress made. These are data quality concerns and data variety concerns, model interpretability, extrapolation to other materials systems, and improved integration between predictions and experimental validation. It is expected that future research will focus on physics-informed machine learning, multimodal data binders, probabilistic predictions, and autonomous laboratories capable of discovery in a closed loop. On the whole, this paper demonstrates that machine learning is a scalable and efficient approach for accelerating the discovery of materials without sacrificing scientific insight. With integrated databases, meaningful physical descriptors, and sophisticated algorithms, AI processes can reduce computational costs and shorten development times. With the development of generative and foundation models, as well as autonomous systems, AI-assisted discovery will be increasingly predictive, automated, and collaborative as it combines with experimental and computational materials science.

Acknowledgement: The authors have listed the support of their institutions that supported this research. The paper has used materials science database resources, such as the Materials Project, and open-source software collections, such as pymatgen, matminer, and scikit-learn, to extract data and perform machine learning analysis. They further acknowledge the wider materials informatics community by making datasets and computational tools freely available, enabling data-driven materials research.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: Manjodh Kaur: Conceptualization, Methodology, Investigation, Writing—Original Draft; Princy Randhawa: Supervision, Methodology, Validation, Writing—Review & Editing; Jitendra Jaiswal: Data Curation, Formal Analysis, Visualization, Writing—Original Draft; Deepak Dubal: Resources, Formal Analysis, Investigation, Writing—Original Draft; Ravindra N. Bulakhe: Software, Data Curation, Validation, Writing—Original Draft; Deepanraj Balakrishnan: Investigation, Resources, Visualization, Writing—Original Draft; Nithesh Naik: Conceptualization, Supervision, Project Administration, Writing—Review & Editing. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data associated with this study are provided in the manuscript and are available upon reasonable request from the corresponding authors.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

ML	Machine Learning
AI	Artificial Intelligence
AFLOW	Automatic FLOW for Materials Discovery
DFT	Density Functional Theory
TDDFT	Time dependent Density functional Theory
LLM	Large Language Model
ChatMOF	Chatbot for Metal-Organic Framework

NLP	Natural Language Processing
MatSciML	Materials Science Machine Learning
ChemBOMAS	Chemistry Bayesian Optimization with a Large Language Model (LLM)-Enhanced Multi-Agent SystemGenAI (Generative Artificial Intelligence)
ALIGNN-Mat	Atomistic Line Graph Neural Network for Materials
MACE-MP	Message passing Atomic Cluster Expansion-Materials Project
GNoME	Graph Networks for Materials Exploration
BayesMat	Bayesian Matting
MatGPT	Materials Generative Pre-trained Transformer
MatGL	Materials Graph Library
Matweb	Materials Web
MakeItFrom	Make it from (a material)
CES Granta Edupack	Cambridge Engineering Selector Grante Education Pack
ASM	American Society for Metals
CINDAS	Centre for Information and Numerical Data Analysis and Synthesis
COD	Crystallographic open database
ICSD	Inorganic Crystal Structure Database
CSD	Cambridge Structural Database
AMCSD	American Mineralogist Crystal Structure Database
PDB	Protein Bank Database
Materials Project	
OQMD	Open Quantum Materials Database
AFLOW	Automatic FLOW for Materials Discovery
NOMAD	Novel Materials Discovery
NIST JARVIS	National Institute of Standards and Technology-Joint Automated Repository for Various Integrated Simulations
C2DB	Computational 2D Matrials Database
SuperConductor (NIMS)	National Institute for Materials Science
MAPTIS	Materials and Processes Technical Information System
CIRMS Data	Council on Ionizing Radiation Measurements and Standards
ICDD PDF	International Centre for Diffraction Data's Powder Diffraction File
PoLyInfo	Polymer Database
CAMPUS Plastics	Computer Aided Material Preselection by Uniform Standards
OMDB	Organic Materials Database
NREL	National Renewable Energy Lab Solar Database
Battery Materials Genome	DOE-Department of Energy
AFE	Automated feature Engineering
XGBoost	eXtreme Gradient Boosting
LightGBM	Light Gradient Boosting Machine
SVM	Support Vector Machines
GNN	Graph Neural Networks
CGCNN	Crystal Graph Convolutional Neural Network
MEGNet	Multimodal Graph Neural Network
PyTorch	Python torch
ASE	Atomic Simulation Environment
DScribe	DescriptorScribe
MOD-Net	Model-operator-data network
M3Gnet	Materials 3-body Graph Network

JAX-MD	Just After eXecution-Molecular Dynamics
MBTR	Many-Body Tensor Representation
KRR	Kernel Ridge Regression
GPR	Gaussian Process Regression
MTP	Moment tensor Potential
MeV	Mega electron volt
DNN	Deep Neural Network
SOPA	Smooth Overlap of Atomic Positions
OC20	Open Catalyst 2020 dataset
OC22	Open Catalyst 2022 dataset
ACSF	Atom Centered Symmetry Functions
RF	Random Forest
PBE	Perdew–Burke–Ernzerhof functional
RMSE	Root Mean Square Error
SISSO	Sure Independence Screening and Sparsifying Operator
SMILES	Simplified Molecular Input Line Entry System
LASSO	Least Absolute Shrinkage and Selection Operator
Elastic Net	Elastic Net Regularization
DT	Decision Tree
T_g	Glass Transition Temperature
T_d	Decomposition Temperature
T_m	Melting Temperature
OLED	Organic Light-Emitting Diode
HCEP	Highest-Occupied Crystal Orbital Energy
DFTB	Density Functional Tight Binding
HOMO	Highest Occupied Molecular Orbital
LUMO	Lowest Unoccupied Molecular Orbital
Pure2DopeNet	Pure-to-Doped Network
ResNet	Residual Neural Network
ViT	Vision Transformer
NAS	Neural Architecture Search
SUNCAT	Stanford–SLAC Joint Center for Artificial Photosynthesis
UCNP spectra	Upconversion Nanoparticle Spectra
TEM	Transmission Electron Microscopy
GAN	Generative Adversarial Networks
HOPV database	Harvard Organic Photovoltaics Database
SHAP model	SHapley Additive exPlanations model
SSE	Solid-State Electrolyte
EA	Electron Affinity
KME	K-Means Ensemble
SEM	Scanning Electron Microscopy
MAE	Mean Absolute Error

References

1. Ioannidis Y. The 5th paradigm: AI-driven scientific discovery. Commun ACM. 2024;67(12):5. doi:10.1145/3702970.
2. Ong SP, Richards WD, Jain A, Hautier G, Kocher M, Cholia S, et al. Python materials genomics (pymatgen): a robust, open-source python library for materials analysis. Comput Mater Sci. 2013;68:314–9. doi:10.1016/j.commatsci.2012.10.028.

3. Hill J, Mulholland G, Persson K, Seshadri R, Wolverton C, Meredig B. Materials science with large-scale data and informatics: unlocking new opportunities. MRS Bull. 2016;41(5):399–409. doi:10.1557/mrs.2016.93.
4. Park H, Onwuli A, Walsh A. Exploration of crystal chemical space using text-guided generative artificial intelligence. Nat Commun. 2025;16(1):4379. doi:10.1038/s41467-025-59636-y.
5. Kang Y, Kim J. ChatMOF: an artificial intelligence system for predicting and generating metal-organic frameworks using large language models. Nat Commun. 2024;15(1):4705. doi:10.1038/s41467-024-48998-4.
6. Bhat N, Birbilis N, Barnard AS. Unsupervised learning and pattern recognition in alloy design. Digit Discov. 2024;3(12):2396–416. doi:10.1039/d4dd00282b.
7. Zhang H, Li R, Zhang Y, Xiao T, Chen J, Ding J, et al. The evolving role of large language models in scientific innovation: evaluator, collaborator, and scientist. arXiv:2507.11810. 2025. doi:10.48550/arXiv.2507.11810.
8. Piovarči M, Foshey M, Xu J, Erps T, Babaei V, Didyk P, et al. Closed-loop control of direct ink writing via reinforcement learning. ACM Trans Graph. 2022;41(4):1–10. doi:10.1145/3528223.3530144.
9. Pyzer-Knapp EO, Manica M, Staar P, Morin L, Ruch P, Laino T, et al. Foundation models for materials discovery—current state and future directions. npj Comput Mater. 2025;11(1):61. doi:10.1038/s41524-025-01538-0.
10. Pan X, Xie Y, Li C, He Y, Zhang Y, Wang Y, et al. Convergence of computational materials science and AI for next-generation energy storage materials. J Electron Mater. 2026;55(1):45–114. doi:10.1007/s11664-025-12511-4.
11. Jain A. Machine learning in materials research: developments over the last decade and challenges for the future. Curr Opin Solid State Mater Sci. 2024;33(1):101189. doi:10.1016/j.cossms.2024.101189.
12. Park YJ, Jerng SE, Yoon S, Li J. 1.5 million materials narratives generated by chatbots. Sci Data. 2024;11(1):1060. doi:10.1038/s41597-024-03886-w.
13. Pugliese R, Badini S, Frontoni E. Generative artificial intelligence for advancing discovery and design in biomateriomics. Intell Comput. 2025;4(1):117. doi:10.34133/icomputing.0117.
14. Meng K, Long R. A universal machine learning framework driven by artificial intelligence for ion battery cathode material design. JACS Au. 2025;5(8):3833–45. doi:10.1021/jacsau.5c00526.
15. Xue X, Dhumras H, Thakur G, Shukla V. Integrating artificial intelligence and sustainable materials for smart eco innovation in production. Sci Rep. 2025;15(1):36942. doi:10.1038/s41598-025-20803-2.
16. Van MH, Verma P, Zhao C, Wu X. A survey of AI for materials science: foundation models, LLM agents, datasets, and tools. arXiv:2506.20743. 2025. doi:10.48550/arXiv.2506.20743.
17. Zeni C, Pinsler R, Zügner D, Fowler A, Horton M, Fu X, et al. A generative model for inorganic materials design. Nature. 2025;639(8055):624–32. doi:10.1038/s41586-025-08628-5.
18. Torralba JM, Meza A, Kumaran SV, Mostafaei A, Mohammadzadeh A. From high-entropy alloys to alloys with high entropy: a new paradigm in materials science and engineering for advancing sustainable metallurgy. Curr Opin Solid State Mater Sci. 2025;36(5):101221. doi:10.1016/j.cossms.2025.101221.
19. Attari V, Arroyave R. Decoding non-linearity and complexity: deep tabular learning approaches for materials science. Digit Discov. 2025;4(10):2765–80. doi:10.1039/D5DD00166H.
20. Nyangiwe NN. Applications of density functional theory and machine learning in nanomaterials: a review. Next Mater. 2025;8(4):100683. doi:10.1016/j.nxmate.2025.100683.
21. Zhao L, Zong H. AI-driven decoding of material dynamics: from machine learning potentials and interpretability to generative prediction. Adv Mater. 2025;3:e14626. doi:10.1002/adma.202514626.
22. Liu Y, Zhao T, Ju W, Shi S. Materials discovery and design using machine learning. J Mater. 2017;3(3):159–77. doi:10.1016/j.jmat.2017.08.002.
23. Ko TW, Deng B, Nassar M, Barroso-Luque L, Liu R, Qi J, et al. Materials Graph Library (MatGL), an open-source graph deep learning library for materials science and chemistry. npj Comput Mater. 2025;11(1):253. doi:10.1038/s41524-025-01742-y.
24. Miret S, Lee KLK, Gonzales C, Nassar M, Spellings M. The open MatSci ML toolkit: a flexible framework for machine learning in materials science. arXiv:2210.17484. 2022. doi:10.48550/arXiv.2210.17484.
25. 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. 2016;2(1):16031. doi:10.1038/npjcompumats.2016.31.

26. Han D, Ai Z, Cai P, Lu S, Chen J, Ye Z, et al. ChemBOMAS: accelerated BO in chemistry with LLM-enhanced multi-agent system. arXiv:2509.08736. 2025. doi:10.48550/arXiv.2509.08736.
27. Fu N, Wei L, Song Y, Li Q, Xin R, Omee SS, et al. Materials transformers language models for generative materials design: a benchmark study. arXiv:2206.13578. 2022. doi:10.48550/arXiv.2206.13578.
28. Batatia I, Benner P, Yuan C, Elena AM, Kovács DP, Riebesell J, et al. A foundation model for atomistic materials chemistry. J Chem Phys. 2025;163(18):184110. doi:10.1063/5.0297006.
29. Wang G, Wang C, Zhang X, Li Z, Zhou J, Sun Z. Machine learning interatomic potential: bridge the gap between small-scale models and realistic device-scale simulations. iScience. 2024;27(5):109673. doi:10.1016/j.isci.2024.109673.
30. Yao L, Samantray S, Ghosh A, Roccapriore K, Kovarik L, Allec S, et al. Operationalizing serendipity: multi-agent AI workflows for enhanced materials characterization with theory-in-the-loop. arXiv:2508.06569. 2025. doi:10.48550/arXiv.2508.06569.
31. Zhu C, Bamidele EA, Shen X, Zhu G, Li B. Machine learning aided design and optimization of thermal metamaterials. Chem Rev. 2024;124(7):4258–331. doi:10.1021/acs.chemrev.3c00708.
32. Wang Z, Chen A, Tao K, Han Y, Li J. MatGPT: a vane of materials informatics from past, present, to future. Adv Mater. 2024;36(6):2306733. doi:10.1002/adma.202306733.
33. Palacin MR. Battery materials design essentials. Acc Mater Res. 2021;2(5):319–26. doi:10.1021/accountsmr.1c00026.
34. Zuccarini C, Ramachandran K, Jayaseelan DD. Material discovery and modeling acceleration via machine learning. APL Mater. 2024;12(9):090601. doi:10.1063/5.0230677.
35. Steed CA, Kim N. Deep active-learning based model-synchronization of digital manufacturing stations using human-in-the-loop simulation. J Manuf Syst. 2023;70(2):436–50. doi:10.1016/j.jmsy.2023.08.012.
36. Gao T, Huang H, Liu Y. Machine learning-driven nanoscale synthesis for electrocatalytic performance: from data-driven methodologies to closed-loop optimization. Adv Mater. 2025;2507:e08263. doi:10.1002/adma.202508263.
37. Abraham BM, Gogotsi Y. Machine learning toolkits and frameworks for materials design. WIREs Comput Mol Sci. 2026;16(2):e70067. doi:10.1002/wcms.70067.
38. Shaaban M, Al-Hamidi Y, El-Borgi S, Krishnamoorthy A. Machine learning-driven *in situ* defect monitoring and real-time process control in directed energy deposition: techniques, challenges, and future prospects. Mater Today Commun. 2026;51(2):114767. doi:10.1016/j.mtcomm.2026.114767.
39. Haastrup S, Strange M, Pandey M, Deilmann T, Schmidt PS, Hinsche NF, et al. The computational 2D materials database: high-throughput modeling and discovery of atomically thin crystals. 2D Mater. 2018;5(4):042002. doi:10.1088/2053-1583/aacfc1.
40. Lu T, Li H, Li M, Wang S, Lu W. Inverse design of hybrid organic-inorganic perovskites with suitable bandgaps via proactive searching progress. ACS Omega. 2022;7(25):21583–94. doi:10.1021/acsomega.2c01380.
41. Lee S, Chen C, Garcia G, Oliynyk A. Machine learning descriptors in materials chemistry used in multiple experimentally validated studies: oliynyk elemental property dataset. Data Brief. 2024;53(1871):110178. doi:10.1016/j.dib.2024.110178.
42. Collins CR, Gordon GJ, von Lilienfeld OA, Yaron DJ. Constant size descriptors for accurate machine learning models of molecular properties. J Chem Phys. 2018;148(24):241718. doi:10.1063/1.5020441.
43. Ward L, Agrawal A, Choudhary A, Wolverton C. A general-purpose machine learning framework for predicting properties of inorganic materials. npj Comput Mater. 2016;2(1):16028. doi:10.1038/npjcompumats.2016.28.
44. Owens CB, Mathew N, Olaveson TW, Tavenner JP, Kober EM, Tucker GJ, et al. Feature engineering descriptors, transforms, and machine learning for grain boundaries and variable-sized atom clusters. npj Comput Mater. 2025;11(1):21. doi:10.1038/s41524-024-01509-x.
45. 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. doi:10.1038/s41524-019-0221-0.
46. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in python. J Mach Learn Res. 2011;12:2825–30.

47. Yazdani Sarvestani H, Nadigotti S, Fatehi E, Aranguren van Egmond D, Ashrafi B. Beyond order: perspectives on leveraging machine learning for disordered materials. Adv Eng Mater. 2025;27(22):2402486. doi:10.1002/adem.202402486.
48. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: a highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst. 2017;30:1–9.
49. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273–97. doi:10.1007/BF00994018.
50. 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. doi:10.1021/acs.chemmater.9b01294.
51. Xie T, Grossman JC. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. Phys Rev Lett. 2018;120(14):145301. doi:10.1103/PhysRevLett.120.145301.
52. Choudhary K, DeCost B. Atomistic line graph neural network for improved materials property predictions. npj Comput Mater. 2021;7(1):185. doi:10.1038/s41524-021-00650-1.
53. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. New York, NY, USA: Springer; 2009. doi:10.1007/978-0-387-84858-7.
54. Fold cross validation—an overview|ScienceDirect topics [Internet]. [cited 2026 Mar 7]. Available from: https://www.sciencedirect.com/topics/computer-science/fold-cross-validation.
55. Himanen L, Jäger MOJ, Morooka EV, Federici Canova F, Ranawat YS, Gao DZ, et al. DScribe: library of descriptors for machine learning in materials science. Comput Phys Commun. 2020;247(26):106949. doi:10.1016/j.cpc.2019.106949.
56. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. TensorFlow: a system for large-scale machine learning. arXiv:1605.08695. 2016. doi:10.48550/arXiv.1605.08695.
57. 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):93. doi:10.1038/s43246-022-00315-6.
58. Almeida Gouvêa R, De Breuck PP, Pretto T, Rignanese GM, Santos MJL. Combining feature-based approaches with graph neural networks and symbolic regression for synergistic performance and interpretability. arXiv:2509.03547. 2025. doi:10.48550/arXiv.2509.03547.
59. PyTorch–HPC2N support and documentation [Internet]. [cited 2026 Jan 2]. Available from: https://docs.hpc2n.umu.se/software/libs/PyTorch/?utm_source=chatgpt.com.
60. Keras: deep learning for humans [Internet]. [cited 2026 Jan 2]. Available from: https://keras.io/.
61. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016 Aug 13–17; San Francisco, CA, USA. p. 785–94. doi:10.1145/2939672.2939785.
62. Ward L, Dunn A, Faghaninia A, Zimmermann NER, Bajaj S, Wang Q, et al. Matminer: an open source toolkit for materials data mining. Comput Mater Sci. 2018;152:60–9. doi:10.1016/j.commatsci.2018.05.018.
63. Hjorth Larsen A, Jørgen Mortensen J, Blomqvist J, Castelli IE, Christensen R, Dułak M, et al. The atomic simulation environment—a python library for working with atoms. J Phys Condens Matter. 2017;29(27):273002. doi:10.1088/1361-648X/aa680e.
64. Mortensen JJ, Larsen AH, Kuisma M, Ivanov AV, Taghizadeh A, Peterson A, et al. GPAW: an open Python package for electronic structure calculations. J Chem Phys. 2024;160(9):092503. doi:10.1063/5.0182685.
65. De Breuck PP, Hautier G, Rignanese GM. Materials property prediction for limited datasets enabled by feature selection and joint learning with MODNet. npj Comput Mater. 2021;7(1):83. doi:10.1038/s41524-021-00552-2.
66. Schütt KT, Kessel P, Gastegger M, Nicoli KA, Tkatchenko A, Müller KR. SchNetPack: a deep learning toolbox for atomistic systems. J Chem Theory Comput. 2019;15(1):448–55. doi:10.1021/acs.jctc.8b00908.
67. Fey M, Lenssen JE. Fast graph representation learning with PyTorch geometric. arXiv:1903.02428. 2019. doi:10.48550/arXiv.1903.02428.
68. Schoenholz SS, Cubuk ED. JAX, M.D. A framework for differentiable physics. J Stat Mech. 2021;2021(12):124016. doi:10.1088/1742-5468/ac3ae9.
69. Liu Y, Esan OC, Pan Z, An L. Machine learning for advanced energy materials. Energy AI. 2021;3:100049. doi:10.1016/j.egyai.2021.100049.

70. Brierley-Croft S, Olmsted PD, Hine PJ, Mandle RJ, Chaplin A, Grasmeder J, et al. Polymer informatics method for fast and accurate prediction of the glass transition temperature from chemical structure. Macromolecules. 2025;58(13):6407–17. doi:10.1021/acs.macromol.5c00178.
71. Nyshadham C, Rupp M, Bekker B, Shapeev AV, Mueller T, Rosenbrock CW, et al. Machine-learned multi-system surrogate models for materials prediction. npj Comput Mater. 2019;5(1):51. doi:10.1038/s41524-019-0189-9.
72. Peng J, Damewood J, Karaguesian J, Lunger JR, Gómez-Bombarelli R. Learning ordering in crystalline materials with symmetry-aware graph neural networks. arXiv:2409.13851. 2024. doi:10.48550/arXiv.2409.13851.
73. Prateek S, Garg R, Kumar Saxena K, Srivastav VK, Vasudev H, Kumar N. Data-driven materials science: application of ML for predicting band gap. Adv Mater Process Technol. 2024;10(2):708–17. doi:10.1080/2374068x.2023.2171666.
74. Guomundsson B, Lorna G. Automated design using machine learning in materials engineering—an explicit forecasts. J Comput Intell Mater Sci. 2023;1:56–66. doi:10.53759/832x/jcims202301006.
75. Hu M, Tan Q, Knibbe R, Xu M, Jiang B, Wang S, et al. Recent applications of machine learning in alloy design: a review. Mater Sci Eng R Rep. 2023;155(1):100746. doi:10.1016/j.mser.2023.100746.
76. Li K, Persaud D, Choudhary K, DeCost B, Greenwood M, Hattrick-Simpers J. Exploiting redundancy in large materials datasets for efficient machine learning with less data. Nat Commun. 2023;14(1):7283. doi:10.1038/s41467-023-42992-y.
77. Malashin IP, Tynchenko VS, Nelyub VA, Borodulin AS, Gantimurov AP. Estimation and prediction of the polymers' physical characteristics using the machine learning models. Polymers. 2023;16(1):115. doi:10.3390/polym16010115.
78. Polat C, Kurban M, Kurban H. Multimodal neural network-based predictive modeling of nanoparticle properties from pure compounds. Mach Learn Sci Technol. 2024;5(4):045062. doi:10.1088/2632-2153/ad9708.
79. Chen C, Xu E, Yang D, Yan C, Wei T, Chen H, et al. Chemical environment adaptive learning for optical band gap prediction of doped graphitic carbon nitride nanosheets. Neural Comput Appl. 2025;37(5):3287–301. doi:10.1007/s00521-024-10775-1.
80. Sivonxay E, Attia L, Spotte-Smith EWC, Sanchez-Lengeling B, Xia X, Barter D, et al. Inverse design of complex nanoparticle heterostructures via deep learning on heterogeneous graphs. ChemRxiv. 2025. doi:10.26434/chemrxiv-2024-1dw4q-v2.
81. Rahmanian E, Sajedi-Moghaddam A, Hoveizavi MT, Aboutalebi SH. Electrolyte hydration energy as a universal descriptor for ion-specific capacitance: insights from interpretable machine learning. Adv Powder Mater. 2026;5(1):100361. doi:10.1016/j.apmate.2025.100361.
82. Tahir MH, Naeem S, Moussa IM. Designing high electron affinity small molecule acceptors through comprehensive chemical library generation. Synth Met. 2026;316:117986. doi:10.1016/j.synthmet.2025.117986.
83. Luo J, Gu Y, Wang Y, Ma X, El-Awady JA. Uncertainty-aware machine learning framework for predicting dislocation plasticity and stress-strain response in metallic alloys, part I: FCC systems. Acta Mater. 2026;302(12):121610. doi:10.1016/j.actamat.2025.121610.
84. Wu S, Kondo Y, Kakimoto MA, Yang B, Yamada H, Kuwajima I, et al. Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm. npj Comput Mater. 2019;5(1):66. doi:10.1038/s41524-019-0203-2.
85. Kim C, Yoon M, Lee JH. Accelerated discovery of OER catalysts in Pnma perovskites via machine learning with minimal DFT structure relaxation. Comput Mater Sci. 2026;261(11):114320. doi:10.1016/j.commatsci.2025.114320.
86. Verma A, Jami J, Bhattacharya A. Accelerating magnetic materials discovery using interaction matrix-based machine learning descriptors. Comput Mater Sci. 2026;262(11):114395. doi:10.1016/j.commatsci.2025.114395.
87. Liu Y, Yang Z, Yu Z, Liu Z, Liu D, Lin H, et al. Generative artificial intelligence and its applications in materials science: current situation and future perspectives. J Mater. 2023;9(4):798–816. doi:10.1016/j.jmat.2023.05.001.
88. Alverson M, Baird SG, Murdock R, Ho S, Johnson J, Sparks TD. Generative adversarial networks and diffusion models in material discovery. Digit Discov. 2024;3(1):62–80. doi:10.1039/D3DD00137G.
89. Lambard G, Yamazaki K, Demura M. Generation of highly realistic microstructural images of alloys from limited data with a style-based generative adversarial network. Sci Rep. 2023;13(1):566. doi:10.1038/s41598-023-27574-8.

90. Yang K, Schwalbe-Koda D. A generative diffusion model for amorphous materials. npj Comput Mater. 2026;12(1):29. doi:10.1038/s41524-025-01901-1.
91. Hong T, Yang J, Cao G. Crystal structure prediction based on diffusion model and graph network optimization. J Phys Mater. 2025;8(3):035011. doi:10.1088/2515-7639/adeaed.
92. Rønne N, Aspuru-Guzik A, Hammer B. Generative diffusion model for surface structure discovery. Phys Rev B. 2024;110(23):235427. doi:10.1103/physrevb.110.235427.
93. Nordhagen E, Sveinsson HA, Malthe-Sørenssen A. Tailoring frictional properties of surfaces using diffusion models. J Phys Chem C. 2025;129(32):14559–64. doi:10.1021/acs.jpcc.5c02768.
94. Khastagir S, Das K, Goyal P, Lee SC, Bhattacharjee S, Ganguly N. LLM meets diffusion: a hybrid framework for crystal material generation. arXiv:2510.23040. 2025.
95. Uddin M, Arfeen SU, Alanazi F, Hussain S, Mazhar T, Arafatur Rahman M. A critical analysis of generative AI: challenges, opportunities, and future research directions. Arch Comput Meth Eng. 2026;33(2):1763–93. doi:10.1007/s11831-025-10355-z.
96. Chakraborty S, Björk J, Dahlqvist M, Rosen J, Heintz F. A survey of AI-supported materials informatics. Comput Sci Rev. 2026;59(4):100845. doi:10.1016/j.cosrev.2025.100845.
97. Malica C, Novoselov KS, Barnard AS, Kalinin SV, Spurgeon SR, Reuter K, et al. Artificial intelligence for advanced functional materials: exploring current and future directions. J Phys Mater. 2025;8(2):021001. doi:10.1088/2515-7639/adc29d.
98. Xie E, Wang X, Siepmann JI, Chen H, Snurr RQ. Generative AI for design of nanoporous materials: review and future prospects. Digit Discov. 2025;4(9):2336–63. doi:10.1039/d5dd00221d.
99. Fawzy SM, Ali MKM, Allam NK. Artificial intelligence-driven materials design for next-generation sustainable energy technologies. ACS Sustain Chem Eng. 2026;14(10):4745–61. doi:10.1021/acssuschemeng.6c01084.
100. Yao T, Huang J, Yan Y, Yang Y, Wang Z, Shao X, et al. From large language models to AI agents in energy materials research: enabling discovery, design, and automation. AI Agent. 2025;1(1):9. doi:10.20517/aiagent.2025.03.
101. Ansari M, Moosavi SM. Agent-based learning of materials datasets from the scientific literature. Digital Discov. 2024;3(12):2607–17. doi:10.1039/D4DD00252K.
102. Qu X, Damoah A, Sherwood J, Liu P, Jin CS, Chen L, et al. A comprehensive review of AI agents: transforming possibilities in technology and beyond. arXiv:2508.11957. 2025. doi:10.48550/arXiv.2508.11957.
103. He Y, Ruan S, Wang D, Lu H, Li Z, Liu Y, et al. Intelligent decision-making driven by large AI models: progress, challenges and prospects. CAAI Trans Intell Technol. 2025;10(6):1573–92. doi:10.1049/cit2.70084.
104. Akimoto J, Takei H. Synthesis and crystal structure of $NaTi_8O_{13}$. J Solid State Chem. 1991;90(1):147–54. doi:10.1016/0022-4596(91)90180-P.
105. Allen JL, Ren X, Nguyen CK, Horn DC, Sun HH, Tran DT. High conductivity and rate capability of $NaNb_{13}O_{33}$ Wadsley-Roth phase as a fast-charging Li-ion anode. ChemElectroChem. 2023;10(20):e202300267. doi:10.1002/celc.202300267.
106. Sudrajat H, Kitta M, Ito R, Nagai S, Yoshida T, Katoh R, et al. Water-splitting activity of La-doped $NaTaO_3$ photocatalysts sensitive to spatial distribution of dopants. J Phys Chem C. 2020;124(28):15285–94. doi:10.1021/acs.jpcc.0c03822.s001.
107. Ishikiriyama K. Machine learning prediction of heat capacity of polymers as a function of temperature. Polymer. 2025;339:129171. doi:10.1016/j.polymer.2025.129171.
108. Sha W, Li Y, Tang S, Tian J, Zhao Y, Guo Y, et al. Machine learning in polymer informatics. InfoMat. 2021;3(4):353–61. doi:10.1002/inf2.12167.
109. Ottomano F, Goulermas JY, Gusev V, Savani R, Gaultois MW, Manning TD, et al. Assessing data-driven predictions of band gap and electrical conductivity for transparent conducting materials. Digit Discov. 2025;4(7):1794–811. doi:10.1039/D5DD00010F.