ENGINEERING Energy 2026, 20(1): 10535
https://doi.org/10.1007/s11708-026-1053-5

REVIEW ARTICLE

Artificial intelligence for energy materials research: From classical machine learning to large models

Mingxi Jiang[#], Jie Zhou[#], Yanggang An, Zhengran Lin, Menghao Yang (✉)

Shanghai Key Laboratory for R & D and Application of Metallic Functional Materials, Institute of New Energy for Vehicles, School of Materials Science and Engineering, Tongji University, Shanghai 201804, China

HIGHLIGHTS

- Tracks the evolution of AI from classical machine learning to advanced large models.
- Analyzes AI applications in battery and electrocatalytic materials.
- Highlights the primary role of generative models in inverse design and material synthesis.
- Outlines future directions for next-generation intelligent development of energy materials.

Keywords:
Energy materials
Artificial intelligence (AI)
Machine learning (ML)
Generative models
Large language models (LLMs)

ABSTRACT

With the global energy system transitioning to renewable energy, high-efficiency energy storage and conversion technologies have become crucial. However, traditional research paradigms for the research and development (R&D) of energy materials such as batteries and electrocatalysts present the limitations in efficiency. This review systematically summarizes the progress of artificial intelligent (AI) in this field, ranging from classical machine learning (ML) to advanced representation methods such as graph neural networks (GNNs) and transformers that enable precise property prediction and structure generation. It also covers generative models for inverse design and large language models (LLMs) for knowledge extraction, along with key domain databases. Current challenges include limited interpretability and the underutilization of emerging AI technologies. Finally, this review discusses future directions such as the applications of multimodal language models, aiming to provide insights for accelerating high-performance energy materials innovation and advancing the global renewable energy transition.

1 Introduction

In the process of the global energy system's transition to renewable energy, high-efficiency energy storage and conversion technologies serve as the core support for overcoming development bottlenecks. Among these, batteries—acting as the primary carriers for electrical energy storage—have been widely applied in fields such as electric vehicles and energy storage power stations, covering multiple directions including lithium-ion batteries [1–3] and solid-state batteries [4–6]. Meanwhile, electrocatalytic materials (e.g., catalysts for hydrogen evolution and oxygen reduction reactions, ORR) are key to the industrial implementation of energy conversion technologies such as water electrolysis for hydrogen production and fuel cells, directly determining reaction efficiency and cost control [7,8].

Currently, optimizing the performance of batteries and electrocatalytic materials has become a research hotspot in the field of energy materials; however, their research and development (R&D) processes are still significantly constrained by traditional research paradigms. Reliance on the "experimental trial-and-error" approach requires cycles of several months to years and incurs high costs, while also making it difficult to reveal correlations between microstructures (e.g., ion transport pathways in battery electrolytes and active sites of electrocatalytic materials) and macroscopic performance [9]. Although theoretical calculations such as density functional theory (DFT) can assist in prediction [10–12], they are limited by structural dependence and computational efficiency, and struggle to handle complex scenarios such as battery

[#] These two authors contributed equally to this work.

✉ Corresponding author. E-mail: menghaoyoung@tongji.edu.cn (M. Yang)

Article history: Received 8 November 2025, Revised 2 January 2026, Accepted 15 January 2026, Available online 15 February 2026

Special Issue: Emerging Investigators in Energy Frontiers

electrolyte doping combinations and multicomponent design of electrocatalytic materials [13].

Over the past decade, the rapid development of artificial intelligence (AI) for science has driven a profound transformation in the materials science research paradigm, shifting from "experiment/theory-driven" to "data/intelligence-driven" approaches. This transformation has achieved remarkable results in the field of batteries and catalytic materials [14,15]. Benefiting from extensive experimental efforts and effective theoretical calculation methods, data-driven approaches based on a large volume of literature [16], combined with first principles calculations to construct multifunctional descriptors for machine learning (ML) frameworks [17,18], have become widely adopted research paradigm. Leveraging powerful capabilities in data mining, complex relationship modeling, and high-throughput prediction, AI technologies have achieved a leap from basic correlation analysis to precise structural regulation.

In the early stage, classical ML methods (e.g., support vector machines (SVMs) [19] and gradient boosting [20–22] were able to quickly establish quantitative correlations between battery electrolyte components and ionic conductivity, as well as between active sites of electrocatalytic materials and overpotential. With breakthroughs in deep learning (DL), models represented by deep neural networks (DNNs) have been more deeply applied in the field of energy materials. Subsequently, innovations in data representation and model expressiveness have emerged. Representation methods such as transformers and graph neural networks (GNNs) have played unique roles by enabling the automatic extraction of high-dimensional, physically interpretable features from diverse material data (e.g., crystal structures, microstructural images, and molecular sequences) without manual intervention. In particular, self-attention mechanisms embodied by transformers have endowed AI models with strong capabilities for understanding textual information [23]. In AI applications for energy materials, discriminative tasks serve as a vital link that bridges microscopic material features to macroscopic performance via quantitative feature-label mapping and application-oriented ranking. In recent years, the rise of generative models has further expanded the research paradigm for AI-assisted energy material discovery, shifting material design toward inverse design and structure generation.

In addition to the aforementioned methods, advances in language model technology have further broken the modality limitations of traditional AI. Small-scale pretrained language models based on transformers have demonstrated enhanced information integration and processing capabilities in specific application scenarios by integrating massive amounts of domain-specific knowledge, such as SMILES strings, literature texts, experimental data, and theoretical calculation results [24,25]. Furthermore, large language models (LLMs), represented by ChatGPT, have developed rapidly in recent years and are trained on massive volumes of general knowledge. In parallel with advances in AI technologies, high-quality domain-specific databases, one of the cornerstones of AI, also merit systematic discussion and organization. Public datasets in materials science, especially those related to energy materials, have accumulated to a certain extent and achieved increasing standardization.

However, current research still faces fragmentation issues: the interpretability of AI applications in material systems is generally limited; the potential of emerging AI technologies (e.g., generative AI and multimodal large models) has not been fully explored; and pretrained LLMs s lack sufficient capability for deep understanding of materials science knowledge. Against this backdrop, systematically reviewing the application logic, research progress, and challenges of AI technologies in batteries and electrocatalytic materials is of great significance for clarifying future research directions and accelerating innovation in energy materials.

In view of this, this paper takes the development of AI technologies as the main thread and conducts a comprehensive review focusing on key directions in energy materials research. As shown in Fig. 1, it illustrates core methods and representative models the AI for energy materials research. The scope centers on

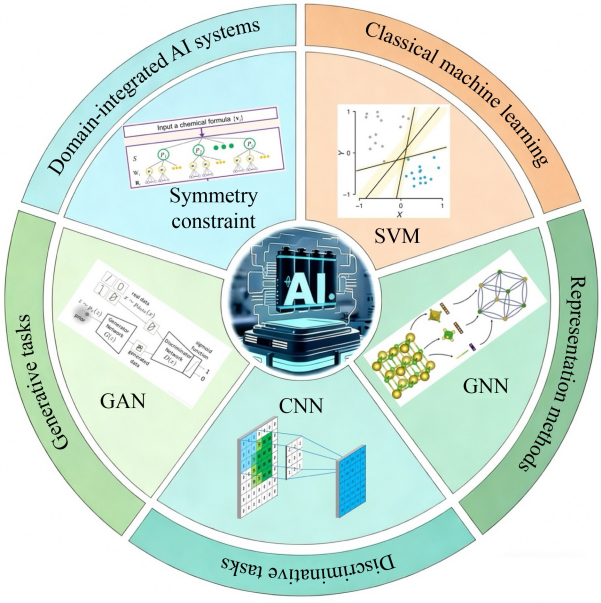


Fig. 1 Key methods and representative models of "AI for energy materials research".

battery materials, supplemented by representative electrocatalytic materials. In terms of content structure, it first outlines the basic applications of classical ML methods, followed by an in-depth discussion of advanced representation methods, including transformers and GNNs. It, then reviews the applications of generative models in inverse design and structure generation, as well as knowledge-driven AI systems based on chemistry and physics. Finally, it looks ahead to emerging application scenarios of cutting-edge language models and the construction of fundamental databases.

2 Classical machine learning

Before the boom of DL, traditional ML and statistical approaches had already exerted considerable influence in materials science. These methods, encompassing supervised learning, unsupervised learning, and statistical analysis, typically require relatively modest datasets, making them well suited to addressing the prevalent "small data" challenge commonly encountered in energy materials research [26–29].

2.1 Supervised learning

Supervised learning has long been applied in the field of AI for science. In 2018, Bzdok et al. [30] elaborated on the working principles of linear SVM, a classic supervised learning method, and its applications to pattern recognition problems in biology and medicine. They also discussed key practical issues, including hyperparameter tuning and the impact of dimensionality on model performance. As illustrated in Fig. 2, SVM classifies data points by maximizing the width of the margin that separates different classes. Supervised learning remains a fundamental tool for predicting the properties of energy materials, such as solid-state electrolytes (SSEs) and electrocatalysts. Traditional approaches, including linear and multivariate regression, paved the way for more advanced algorithms such as SVM, random forests (RF), and gradient boosting decision trees (GBDT) [31,32]. For instance, Mishra et al. [31] showed that ensemble learning techniques, especially RF, can markedly enhance the prediction of ionic conductivity in various inorganic solid electrolytes, underscoring the adaptability of supervised learning to diverse small datasets and its value in identifying promising candidate materials through predictive modeling.

DL models have demonstrated excellent performance in the inverse design of high-efficiency ORR/hydrogen oxidation reaction (HOR) catalysts, as well as electrode structures. Klein Moberg et al. [33] developed a DL-enabled online mass spectrometry approach to analyze the reaction products of individual catalyst nanoparticles. By leveraging a constrained denoising autoencoder to extract weak signals from noise, this method reduced the required catalyst surface area by approximately three orders of magnitude, enabling real-time reaction monitoring of single-particle catalysis in fuel cells.

In the field of catalysis, Li et al. [34] developed a DL framework combined with molecular dynamics

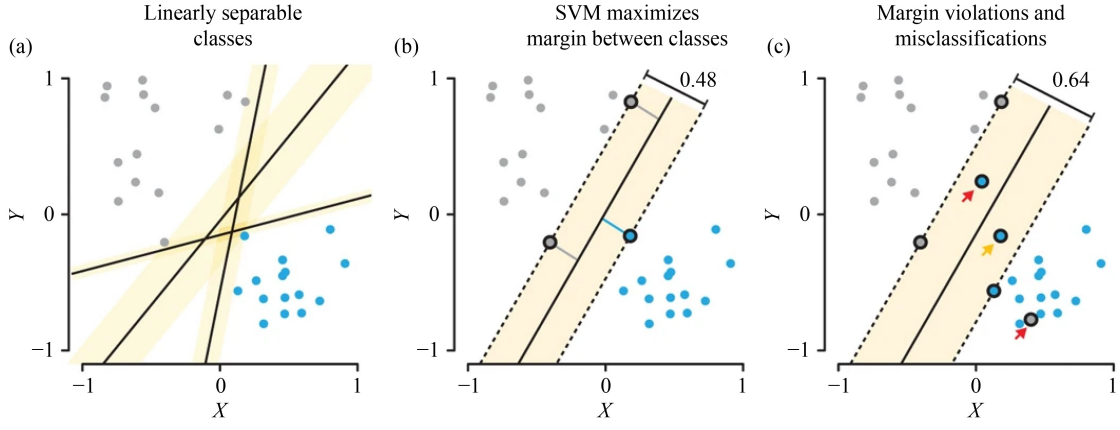


Fig. 2 SVM classification by maximization of the separating margin between classes.

(a) Points from two classes (gray, blue) that are perfectly separable by multiple separating lines (black), demonstrating the concept of a margin (orange highlight), defined as the rectangular region extending from the separating line to the perpendicularly closest data point; (b) selection by an SVM of the separating line (black) with the widest margin (0.48) (Points lying on the margin boundaries (black outlines) are the support vectors—the margin is unaffected by moving or adding other points outside this region); (c) margin violations and misclassification errors resulting from the imposition of a separating line on linearly nonseparable classes (Same data as in (b), with two additional misclassified points. The resulting margin is 0.64 with six support vectors, adapted from Bzdok et al. [30] under the terms of CC BY license).

simulations to model the density distribution and diffusion coefficient of water molecules in catalyst layers (CLs). This framework provides a molecular-scale perspective on intermolecular interactions, achieving low prediction error and high computational efficiency, thereby supporting CL structure optimization and the interpretation of water transport mechanisms.

Bi et al. [35] proposed a DL-based systematic framework for structural characterization and mass transport simulation of CO_2 reduction CLs. They used semantic segmentation models to extract parameters such as porosity and pore size distribution, achieving segmentation accuracies over 91%. Experimental validation further confirmed that an ionomer-to-carbon (I/C) ratio of 0.2 enhances gas diffusion and increases CO yield, offering valuable guidance for multiscale analysis.

2.2 Unsupervised learning

Unsupervised learning plays a vital role in exploring the complexity of material properties and classifications. Techniques such as clustering (e.g., *k*-means and hierarchical clustering) are commonly employed to categorize materials based on structural characteristics or phase diagrams, helping researchers identify inherent trends and data groupings [36]. Chen et al. [37] proposed a multiscale topological learning (MTL) framework that integrates algebraic topological modeling with AI-based unsupervised learning.

Figure 3 illustrates the workflow of this multiscale topological approach for discovering lithium superionic conductors (LSICs). In the initial step (Fig. 3(a)), the data collection stage filters lithium-containing materials from the ICSD-2019 database, yielding 2590 unique materials

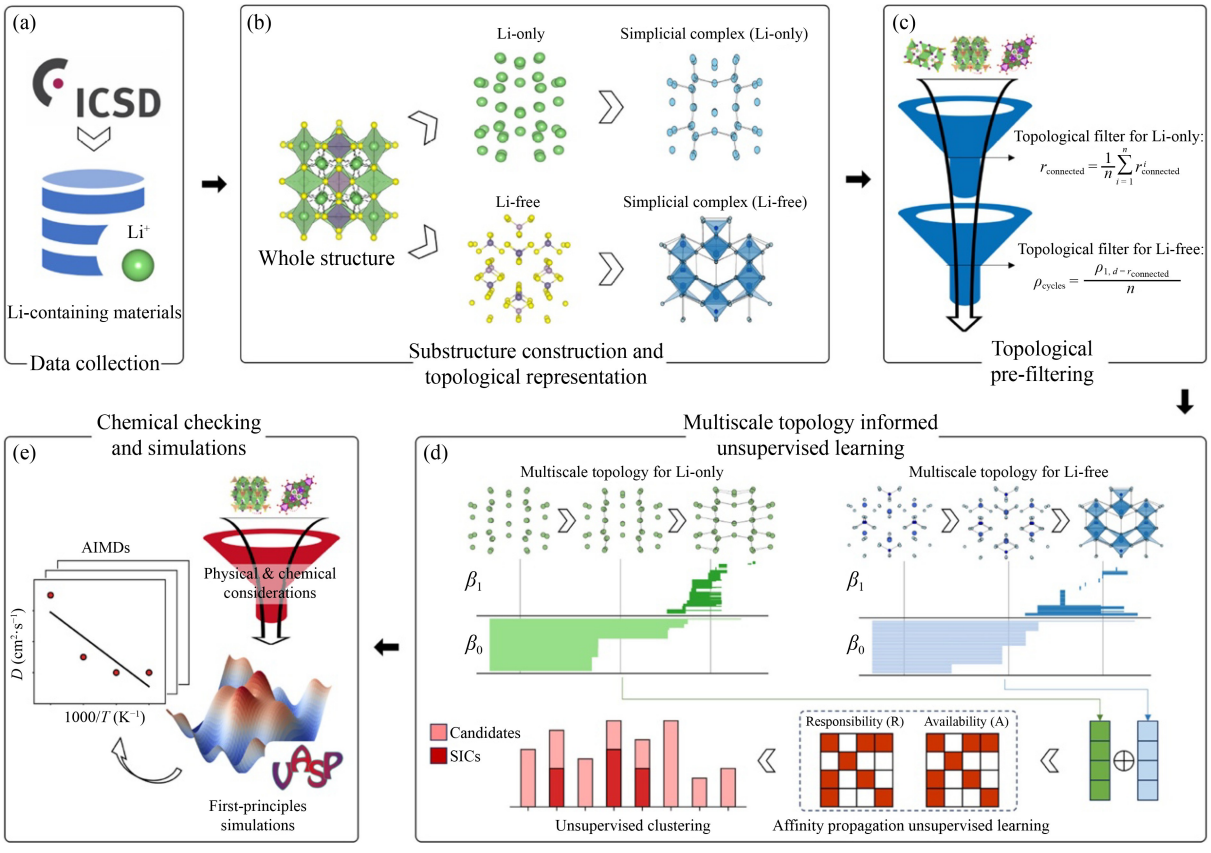


Fig. 3 Workflow for a MTL approach to discovering LSICs.

(a) Data collection and filtering of lithium-ion–containing materials from the ICSD database to identify potential candidates; (b) preliminary analysis of known LSIC structures, with lithium-only substructures (Li-only) and lithium-free frameworks (Li-free) modeled as simplicial complexes and analyzed independently; (c) topological representation of Li-only and Li-free substructures using simplicial complexes, capturing high-order interactions and extracting features such as connectedness ($r_{\text{connected}}$) and cycle density (ρ_{cycles}) to narrow the search space; (d) multiscale topological features extraction via persistent homology and affinity propagation clustering, grouping materials by topological similarity to highlight clusters enriched in LSIC candidates; (e) final physical and chemical validation, including first-principles analysis, for identification of the most promising LSIC candidates (adapted from Chen et al. [37] under the terms of CC BY license).

for analysis. Figure 3(b) shows the second stage, where a preliminary study of well-known LSIC structures is conducted. The third step involves a topological approach, where Li-free and Li-only substructures are represented as simplicial complexes to capture high-order interactions (0-simplices for atoms, 1-simplices for pairwise interactions, and 2-simplices for triplet interactions). This process yields key descriptors, including connectedness ($r_{\text{connected}}$) for Li substructure conductivity and cycle density (ρ_{cycles}) for the stability of the Li-free framework, which are used to filter candidates.

Subsequently, multiscale topological features are incorporated into unsupervised learning algorithms to cluster the remaining candidates according to topological similarity, thereby revealing underlying structural patterns. Finally, physical and chemical criteria are applied for validation, leading to the identification of viable LSIC candidates and demonstrating the effectiveness of this approach.

As a novel application of unsupervised learning, Zhang et al. [38] proposed an approach for screening and discovering new high-performance materials using limited material data. Unlike the common practice of using ML to establish quantitative correlations between material features and properties, they employed unsupervised learning to classify materials. By analyzing these classifications, they identify categories exhibiting desirable material characteristics and then use high-throughput computing to predict the properties of materials in those categories.

Classical ML represents the first wave of AI adoption in energy materials. Its advantages include effectiveness under data scarcity, the ability to provide interpretable insights, and seamless integration with experimental workflows. However, in materials research, raw data often extend beyond simple numerical values to include complex information such as images, spectroscopic data, textual content, and graph structures. In such cases, representation learning methods, represented by GNNs and transformers, can map complex data to low-dimensional, high-semantic vector spaces (i.e., embeddings), achieving excellent performance in addressing these challenging tasks.

3 Representation methods

In energy materials research, a core objective is to elucidate the relationship between material properties and structures. Both the molecular/crystalline structural data and the spectroscopic information or images obtained via characterization techniques often exhibit non-Euclidean characteristics or long-range dependencies [39,40]. While certain approaches can handle grid-based data such as images (e.g., convolutional neural networks, CNNs) or sequential data such as spectra (e.g., long short-term memory networks, LSTMs), they are generally limited in capturing the global structural dependencies inherent in materials.

Representation learning methods, exemplified by GNNs and transformers, offer revolutionary solutions for representing graph-structured and sequence/grid data respectively, and have become two main pillars of current "AI for energy materials" research. GNNs represent crystalline materials as graphs, with atoms as nodes and chemical bonds as edges. Through message-passing mechanisms, they capture both local and global topological relationships, enabling high-precision prediction of properties such as voltage, capacity, and ionic diffusion barriers. Transformer-based models, on the other hand, excel at processing sequence or grid-type materials data, with self-attention mechanisms that extend beyond local regions to capture long-range correlations.

3.1 GNNs

GNNs have demonstrated strong performance in accurately predicting key thermodynamic and electrochemical properties, such as formation energy, battery voltage, and ionic migration barriers. The crystal graph convolutional neural network (CGCNN), proposed by Xie and Grossman [41], was the first model to systematically apply graph convolution to crystal structures. On the Materials Project dataset, CGCNN achieved a formation energy prediction error of approximately 0.03 eV/atom, significantly outperforming traditional ML descriptors. Later, SchNet introduced continuous-filter convolution to further improve modeling of 3D geometric interactions, making it well-suited for predicting quantum chemical properties of molecules and crystals [42].

The atomistic line GNN (ALIGNN) model, developed by Choudhary et al. [43], significantly improved the prediction accuracy of various material properties by constructing both atomic graphs and bond-angle line graphs. On the materials project dataset, ALIGNN achieved a mean absolute error (MAE) of 0.022 eV/atom for formation energy and 0.218 eV for band gap, representing 43.6% and 43.8% improvements over CGCNN, respectively. This work highlights the potential of multi-scale graph representations.

The CGCNN framework proposed by Xie and Grossman [41] directly learns from the atomic connections within crystals and, after training on approximately 10^4 data points, can predict eight properties of crystals across diverse structures and

compositions with accuracy approaching that of DFT. Moreover, it can extract the contributions of local chemical environments to global properties, providing interpretable chemical insights for materials design. Figure 4 illustrates the CGCNN architecture, where a crystal graph is represented as an undirected multigraph, with nodes corresponding to atoms and edges representing the connections between them.

In battery electrode voltage prediction, the introduction of attention mechanisms has further enhanced the GNNs' sensitivity to long-range interactions. Louis et al. [44] proposed an attention-based GNN framework for voltage prediction. By integrating compositional and spatial information, this framework reduced the MAE of average voltages for lithium-ion battery electrodes to 0.3 V—significantly outperforming traditional models that rely solely on chemical composition. He et al. [45] combined transformer and GNN architecture to successfully screen sodium-ion cathode materials with voltages exceeding 5 V, such as Na $(NiO_2)_2$.

The high-throughput prediction capability of GNNs makes them a key accelerator for quickly screening high-performance energy materials from large candidate libraries. Law et al. [46] used an atomic line graph GNN to identify K_2MnS_2 from 7385 topological quantum materials as a promising cathode for potassium-ion batteries. This material has a theoretical capacity of 203.8 mAh/g, a volume change of only 6.4%, and an energy density of 564.5 Wh/kg, demonstrating the potential of GNNs in emerging energy storage systems [47]. Law et al. [46] proposed a GNN-based strategy leveraging upper-bound energy minimization. By quickly eliminating unstable structures, this approach increased the efficiency of stability evaluation for 285 functional materials severalfold, providing a new paradigm for high-throughput virtual screening.

3.2 Transformer models

While GNNs excel at capturing local geometric environments, transformer models have introduced powerful capabilities for modeling long-range dependencies in materials science by treating chemical structures as a type of language. These models leverage self-attention mechanisms to process sequential data, enabling both rapid property prediction and the generative design of novel structures.

In polymer informatics, polyBERT treats the chemical structure of polymers as a chemical language. Developed by Kuenneth and Ramprasad [48], this model utilizes a DeBERTa-based transformer encoder to generate dense numerical fingerprints. Figure 5 illustrates polymer informatics with polyBERT. The model is trained on a massive dataset of 100 million hypothetical polymer structures represented as polymer SMILES strings. Through this self-supervised training, polyBERT learns

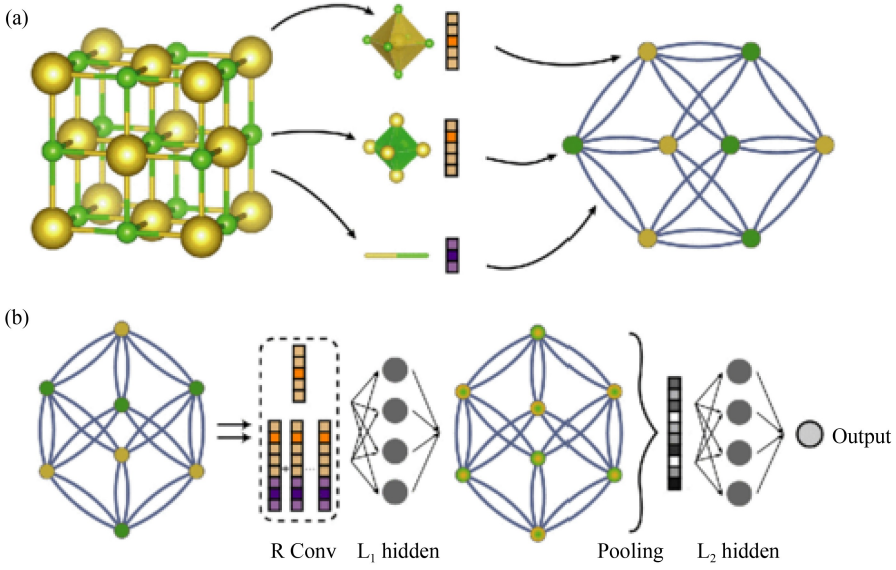
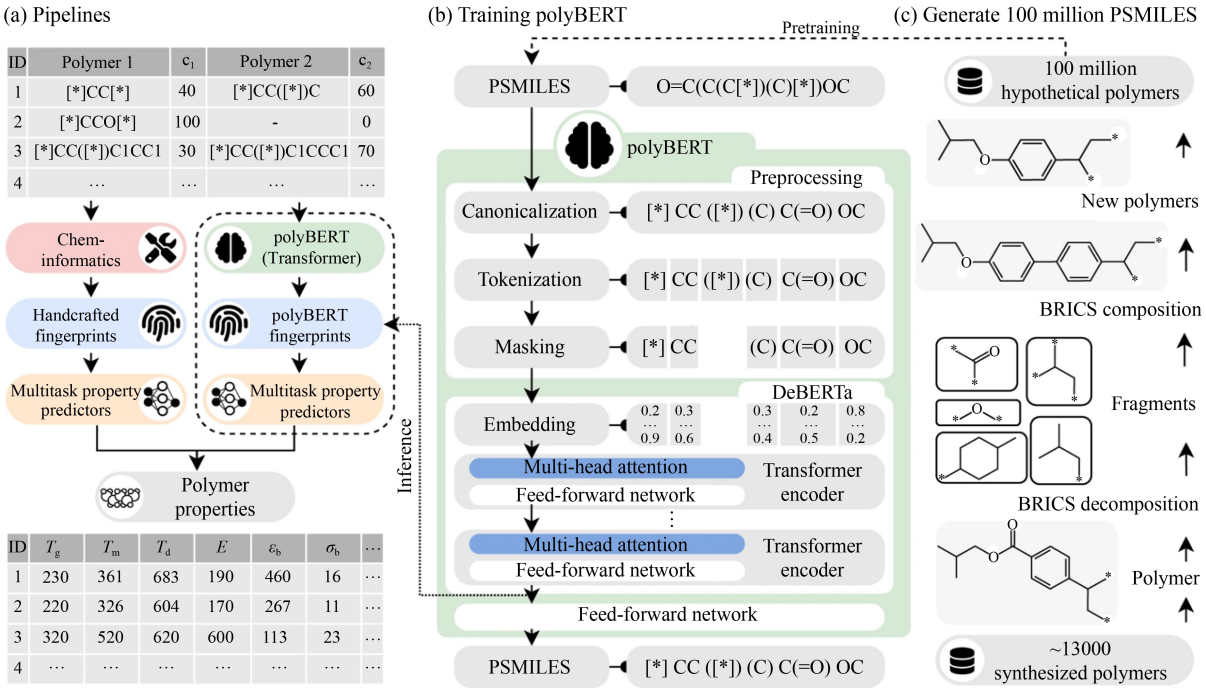


Fig. 4 Illustration of CGCNNs.

(a) Construction of the crystal graph, with crystals converted into graphs in which nodes represent atoms in the unit cell and edges represent interatomic connections (Nodes and edges are characterized by vectors corresponding to atomic and bond features, respectively); (b) structure of the convolutional neural network built on the crystal graph (R convolutional layers and L_1 hidden layers applied each node produce a new graph in which nodes represent local environments (After pooling, a crystal-level vector is passed to L_2 hidden layers, followed by the output layer for prediction) (adapted with permission from Xie and Grossman [41], copyright 2018, under the terms of CC BY license).



ID	Polymer 1	c_1	Polymer 2	c_2
1	[*]CC[*]	40	[*]CC([*])C	60
2	[*]CCO[*]	100	-	0
3	[*]CC([*])C1CC1	30	[*]CC([*])C1CCC1	70
4	…	…	…	…

ID	T_g	T_m	T_d	E	ε_b	σ_b	…
1	230	361	683	190	460	16	…
2	220	326	604	170	267	11	…
3	320	520	620	600	113	23	…
4	…	…	…	…	…	…	…

Fig. 5 Polymer informatics with polyBERT.

(a) Pipelines comparison between a traditional cheminformatics workflow with handcrafted fingerprints and a fully machine-driven polyBERT workflow, in which polymer SMILES strings are tokenized and processed by a transformer encoder to generate fingerprints; (b) polyBERT training procedure, including canonicalization, tokenization, and masking of PSMILES strings prior to processing by a DeBERTa-based transformer encoder; (c) hypothetical polymer library used for pre-training, consisting of 100 million hypothetical polymers generated by recombination of chemical fragments from synthesized polymers (adapted from Kuenneth and Ramprasad [48] under the terms of CC BY license).

the syntax and grammar of chemical strings, capturing chemically relevant information.

A key advantage of this approach is speed: the polyBERT pipeline generates fingerprints over two orders of magnitude faster than traditional handcrafted methods, such as Polymer Genome while maintaining comparable accuracy. These machine-learned fingerprints enable multitask DNNs to predict 29 distinct properties, including thermal characteristics like glass transition temperature T_g and melting temperature T_m, as well as electronic and optical properties such as dielectric constants and refractive index.

In the domain of inorganic crystals, CrystaLLM applies LLM techniques to the generation of crystal structures. Antunes et al. [49] introduced this generative decoder-only transformer, trained on millions of Crystallographic Information Files (CIFs). The model treats the standard CIF format, which encodes unit cell parameters and atomic coordinates, as a text sequence. By predicting the next token in a sequence, CrystaLLM learns to generate syntactically correct and physically plausible crystal structures.

CrystaLLM is distinguished by its ability to generate novel and stable materials not present in the training set. Studies have shown that the model can including ionic, semi-ionic and metallic compounds, with several lying on or near the convex hull of thermodynamic stability [49]. Furthermore, the framework allows integration with Monte Carlo tree search to actively explore low-energy structures. In this process, an auxiliary GNN named ALIGNN guides generation toward stable phases by estimating formation energies, combining generative modeling with property-aware optimization.

4 Discriminative tasks

As a core paradigm of ML in energy materials research, discriminative tasks serve as a critical bridge linking microscopic material features to macroscopic performance and application-oriented classification. By establishing quantitative mappings between structural descriptors and target labels, discriminative tasks are typically categorized into property prediction and classification, each playing distinct roles in different scenarios of energy material research and development.

Property prediction is a regression-based discriminative tasks aimed at predicting continuous energy-related

performance parameters from material features. For example, in screening electrolyte additives for aqueous Zn-ion batteries (AZIBs), Li et al. [50] employed the SISSO (sparse identification of nonlinear dynamics by sparse regression) method, based on symbolic regression to forecast the surface free energy of additive molecules. SISSO outputs explicit feature-combination formulas, directly revealing correlations between molecular properties and battery stability [50]. Using only 38 electrolyte additive molecules with different functional groups (e.g., alcohols, ethers, acids, amines) and elemental compositions, they successfully identified that the number of heavy atoms and liquid surface tension of additive molecules are key factors affecting battery stability. Ultimately, the selected 1,2,3-butanetriol additive increased battery Coulombic efficiency to 99.3%, significantly higher than the 96.4% achieved with pure zinc sulfate electrolyte.

Classification, in contrast, is a discrete discriminative task aimed at categorizing materials into predefined classes based on inherent characteristics. Certain material properties such as microstructure or operating conditions of energy devices, exhibit typical discrete features, making classification more suitable than regression. In the Daisy framework developed by Nandishwara et al. [51], a binary classification model built upon a 16-layer visual geometry group (VGG16) network automatically labels training data via Harris corner detection. This approach accurately differentiates defective (pinhole-containing) from defect-free samples in scanning electron microscope (SEM) images of Ag-Bi-I perovskite-based thin films, with a test accuracy of 97.1%—a 120-folds increase in efficiency compared to manual screening. This method eliminates the need for manual annotation, adapts to images at different magnifications, and provides essential technical support for rapid quality assessment and optimization of synthesis parameters in energy devices such as photovoltaic materials.

Beyond image analysis, classification is also well-suited for operational condition monitoring of energy materials. For instance, Tao et al. [52] proposed a federated learning framework using random forest as local models. By extracting 30 features from charge-discharge and $\mathrm{d}Q/\mathrm{d}V$ curves of retired batteries and integrating an innovative Wasserstein distance voting strategy, this framework efficiently sorted five types of cathode materials and batteries from seven manufacturers without historical operational data. Sorting errors were only 1% and 3% under homogeneous and heterogeneous data scenarios, respectively. Figure 6 shows this federated ML framework for retired battery sorting and recycling.

Discriminative tasks focus on learning the mapping between data features and categories, identifying decision boundaries across different numerical values or categories. With advances in AI, generative models have emerged that can learn the underlying distribution of material data to create novel structural designs for energy materials. Meanwhile, multimodal language models are capable of integrating cross-modal information, such as text, images, and atomic structures, to perform complex reasoning and analysis. These two types of models complement purely discriminative approaches in terms of core objectives, modeling logic, and application scenarios, which will be further explored in subsequent sections.

5 Generative tasks

The discovery of high-performance energy materials, from SSEs to efficient catalysts, which has traditionally been hindered by the vastness of the chemical space, spanning approximately 10^{60} to 10^{100} potential possibilities [53]. To overcome these bottlenecks, the field is shifting from discriminative ML which focuses on property prediction—to generative intelligence. The application of generative AI in energy materials discovery falls into three distinct paradigms: inverse design, generative-aided high-throughput screening, and structure generation.

5.1 Inverse design

Inverse design leverages generative models to directly map desired functional properties to corresponding material structures, transforming the traditional "trial-and-error" approach into a "demand-oriented" development workflow. Diffusion models and flow-matching frameworks have become powerful tools for the inverse design of inorganic materials, enabling efficient discovery of stable, high-performance candidates.

Zeni et al. [54] proposed MatterGen, a generative framework based on diffusion models, which further enhances the controllability of materials inverse design. As illustrated in Fig. 7, the core workflow of MatterGen consists of three key steps:

① Generating stable materials by reversing a physics-informed corruption process through iterative denoising of random initialized structures.

② Denoise atom types, coordinates, and lattices, using an equivariant score network pretrained on a large dataset of stable material structures, followed by fine-tuning with labeled property data via an adapter module that encodes target properties.

③ Generating materials constrained by desired

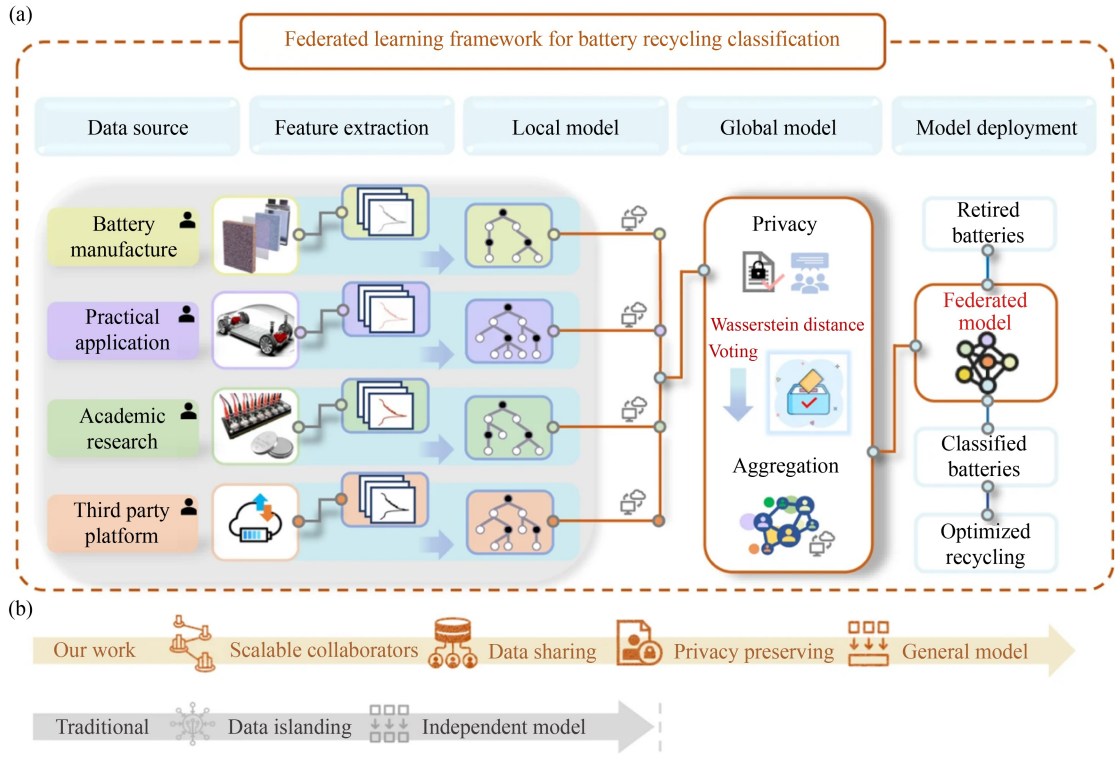


Fig. 6 Federated ML framework for retired battery sorting and recycling.

(a) Federated learning framework for battery recycling, with multiple data contributors, such as battery manufacturers (image courtesy of Addionics), practical application operators (battery pack in the floor pan of a Tesla; image courtesy of Tesla), academic research institutions, and third-party platforms; (b) comparison between the federated ML paradigm, enabling collaborative model training with preserved data privacy, and the traditional data-islanding paradigm (adapted from Tao et al. [52] under the terms of CC BY license).

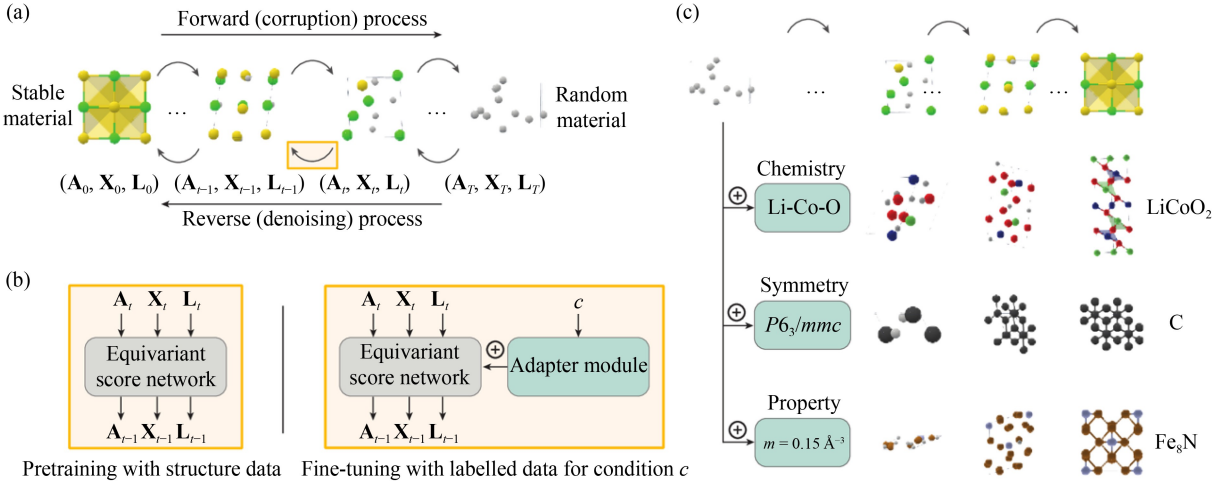


Fig. 7 Inorganic materials design with MatterGen.

(a) MatterGen framework for stable materials generation via reversal of a corruption process, using iteratively denoising of an initially random structure (Forward diffusion process independently corrupts atom types (A), coordinates (X), and the lattice (L) toward a physically motivated distribution of random materials); (b) pretraining of an equivariant score network on a large dataset of stable material structures for jointly denoising of atom types, coordinates, and lattice parameters, followed by fine-tuning with a labeled dataset through an adapter module that incorporates encoded material properties; (c) generation of materials satisfying target chemistry, symmetry, or scalar property constraints using the fine-tuned model (adapted form Zeni et al. [54] under the terms of CC BY license).

features, such as chemical composition, crystal symmetry, or scalar properties (e.g., magnetic density).

Notably, MatterGen supports conditional fine-tuning for multi-target properties (e.g., magnetic density and electronic properties), and the energy storage materials generated by this framework exhibit both structural novelty and thermodynamic stability.

In the context of battery cathode material discovery, researchers applied the pretrained MatterGen model to generate a diverse library of lithium-containing crystals [55]. By combining unsupervised clustering with GNN-based stability ranking, only a small number of DFT calculations were needed to validate two new types of high-voltage cathode materials. This approach demonstrates the potential of a "first generate, then predict".

Generative models have also been successfully applied to the inverse design of electrolyte solvents for advanced batteries. Gao et al.[56] integrated a custom generative algorithm with a virtual molecular database, producing 823 candidate solvent molecules for magnesium metal batteries. Using ML models and multi-dimensional screening criteria, such as lowest unoccupied molecular orbital energy (LUMO) and electrostatic potential (ESP), the candidate pool was narrowed to 18 solvents, improving screening efficiency by nearly 50-fold. When paired with the $Mg[B(hfip)_4]_2$ salt, two novel solvents (DOX and DMP) achieved an average Coulombic efficiency of 99.54% over 5200 cycles and a cumulative capacity exceeding 2000 mAh/cm^2. This work represents a milestone in the inverse design of functional organic molecules for energy storage applications.

Beyond crystal structure generation, researchers have also explored combine generative models with advanced crystal representation methods. MatterGPT, developed by Chen et al. [57], is a generative model based on GPT-2 that adopts the simplified line-input crystal-encoding system (SLICES). Unlike traditional CIF-based representation, SLICES reversibly and invariantly encodes crystal structures as character sequences. MatterGPT performs inverse design by learning the grammar of SLICES and the intrinsic correlations between crystal structures and properties.

For crystalline materials, Wang and You [58] summarized the current core methods for crystal structure generation and proposed approaches for constructing crystal representations that satisfy Periodicity, Invertibility, and Invariance, aiming to advance the development of more versatile and robust AI-driven design tools.

5.2 Generative-aided high-throughput screening

While inverse design focuses on directly targeting specific material properties, generative models are also widely applied to expand the search space for high-throughput screening of battery components. In this "generate-then-screen" workflow, generative models first create diverse libraries of chemically valid candidates, which are then filtered through predictive models or physical simulations.

In battery cathode discovery, researchers employed generative frameworks to construct a diverse library of lithium-containing crystals [55]. Although the candidates were AI-generated, the key discovery process relied on unsupervised clustering, ML force-field (MatterSim) screening, and DFT validation to identify thermodynamically stable, high-voltage materials. This approach demonstrates the concept of "collaborative screening," where generative models act as intelligent samplers to efficiently populate the screening funnel.

Similarly, Gao et al. [56] combined a custom generative algorithm with a virtual molecular database to produce 823 candidate solvent molecules for magnesium metal batteries. These candidates were subjected to a multi-dimensional screening funnel based on ML predictions of LUMO and ESP, reducing the pool to 18 promising solvents and improving screening efficiency by nearly 50-fold compared to random selection. When paired with the $Mg[B(hfip)_4]_2$ salt, two novel solvents, DOX and DMP, achieved an average Coulombic efficiency of 99.54% and a cumulative capacity exceeding 2000 mAh/cm^2. This work represents a significant milestone in the accelerated screening of functional organic molecules for post-lithium energy storage systems.

5.3 Structure generation

Beyond atomic-scale composition, generative models play a critical role in constructing realistic material structures at the mesoscale (microstructures) and in enhancing characterization data quality.

5.3.1 Microstructure reconstruction

Electrode microstructural features, such as porosity and tortuosity, directly affect battery rate capability and energy density. Generative models, including GANs and VAEs provide efficient solutions for reconstructing these complex 3D structures. For example, Gayon-Lombardo et al. [59] employed a GAN-based approach to stochastically reconstruct 3D multiphase electrode microstructures with periodic boundaries. The generated structures closely matched experimental statistical metrics, enabling rapid parameter scanning. Additionally, Kench et al. [60] developed a generative AI-based workflow for microstructure optimization. By integrating

a generative model into a Bayesian optimization loop, the framework successfully identified optimal manufacturing parameters to improve battery energy density.

5.3.2 AI-enhanced characterization

With the accumulation of characterization data, generative models are emerging as auxiliary tools for image enhancement in characterization workflows. Müller et al. [61] used CycleGAN to generate synthetic electrode structures and corresponding XCT images. Supplementing the training set with these AI-generated images increased the Dice coefficient for segmentation of the carbon black–binder domain from 0.38 to 0.58, significantly improving accuracy.

Furthermore, with the proliferation of *in situ* characterization technologies and the accumulation of extensive material datasets [62], generative models are poised to become an even more powerful tool for data augmentation and enhanced analysis, enabling more effective and efficient characterization of energy materials.

6 Domain-integrated AI systems

Current data-driven models often struggle with generalizability and interpretability because they are largely disconnected from fundamental scientific principles. A rapidly evolving frontier seeks to integrate domain knowledge ranging from basic physical laws (e.g., energy minimization, interatomic potentials) to accumulated chemical insights (e.g., synthesis protocols from literature) directly into AI architectures. This section reviews emerging AI systems that incorporate such constraints to achieve high-fidelity simulation and autonomous materials discovery.

To overcome the limitations of purely data-driven fitting, researchers are increasingly embedding physical constraints into ML frameworks. Notably, Yin and colleagues [63] at Soochow University have made significant strides in this direction by combining GNNs with global optimization algorithms governed by physical laws. In their seminal work on crystal structure prediction, Cheng, Yin and collaborators demonstrated an AI framework that enforces the principle of minimum energy, effectively solving complex structural search problems constrained by thermodynamic stability [63], as summarized in Fig. 8.

Building on this foundation, they further proposed a "full-space" inverse design approach and a multiobjective optimization framework for superhard C–N compounds [64,65]. These systems utilize universal machine learning potentials as the physical engine, enabling the AI to explore vast chemical spaces while strictly adhering to interatomic force rules. This approach represents a paradigm shift from screening existing databases to de novo design grounded in physics.

Similarly, embedding force-field constraints is essential for surface science. Wang's group at Tsinghua University introduced SurFF, a foundation force-field-based model tailored for intermetallic crystals [65]. Extending physics-grounded exploration from bulk crystal structures to explicit surfaces and facets, SurFF integrates a universal force field with an active-learning loop to efficiently map surface energetics, as summarized in Fig. 9.

By learning from high-throughput DFT calculations and covering over 12000 unique surfaces, SurFF achieves DFT-level precision (error ≈3 meV $Å^{-2}$) with providing a 10^5-fold acceleration. Importantly, SurFF integrates domain knowledge on surface energy and Wulff shapes, enabling large-scale prediction of explicit surface exposure for heterogeneous catalysts.

To tackle the challenge of linking synthesis, structure, and performance, Chen et al. [66] developed the corpus-aware automated text-to-graph catalyst discovery agent (CATDA) . Unlike traditional text-mining tools, CATDA employs a long-context LLM to process full documents, converting them into structured knowledge graphs with near-human fidelity (F1 = 0.983). By transforming unstructured literature into a computable resource, CATDA enables extraction of multistep synthesis protocols and reaction conditions that were previously inaccessible.

This work exemplifies the shift toward multi-modal large models, where AI agents not only predict material properties but also comprehend and reason through complex chemical knowledge, guiding the rational design of catalysts.

7 Discussion

While the previous sections have summarized applications of mainstream ML techniques, AI remains a rapidly evolving frontier, with numerous cutting-edge directions and fundamental challenges in its intersection with energy materials.

This section will focus on recently developed AI technologies, particularly language models, along with their applications and future prospects in energy materials research. It will then revisit the foundational role of materials science databases, which underpin the practical implementation of these AI technologies. Finally, it will highlight the supplementary contributions of this review, including the clarification of technical logics and mapping of cutting-edge research directions in the field.

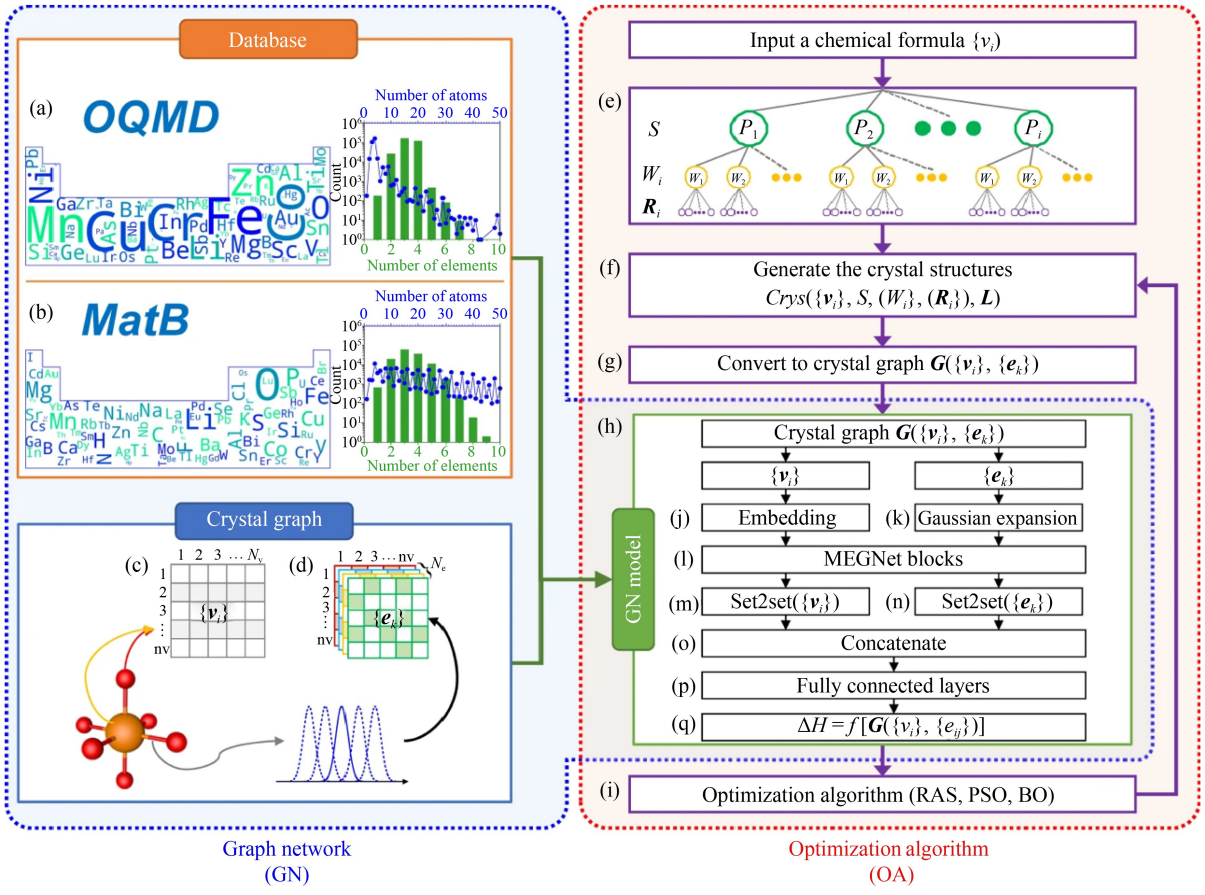


Fig. 8 Flowchart of GN-OA approach.

(a, b) Elemental composition and atomic count distributions in the two databases used for training the GN model: (a) OQMD and (b) MatB, with quantitative counts; (c, d) crystal graph input features: (c) embedded atomic numbers (1 to nv) for each compositional atom (1 to nv), and (d) Gaussian-expanded pairwise distance (1 to N_e) for each pair (i, j; 1 to nv); (e–g) structural generation workflow, including (e) symmetry constraints, (f) generation of crystal structures, and (g) conversion to crystal graphs; (h–q) GN model architecture: (h) combination of features, including (j) embedded atomic numbers, (k) Gaussian-expanded pair distances, (l) MEGNet blocks, (m–n) set2set layers for atomic and pair features, (o) concatenate layer, and (p) fully connected layer, yielding (q) correlation model between crystals and formation enthalpy; (i) optimization algorithm module (adapted from Cheng et al. [63] under the terms of CC BY license).

7.1 Applications of language models

Large-scale data accumulation provides the foundation for AI-driven research and development of new energy materials. However, the vast majority of material knowledge is published in the form of scientific literature. As noted in previous two sections, studies leveraging traditional ML or DL for materials research often rely on manually curated data from publications or laboratory-scale experiments. However, data acquisition in materials research is inherently challenging: whether through experiments, theoretical calculations, or manual literature mining, the process is time-consuming, limiting the efficiency of large-scale data collection. This has necessitated automatic acquisition of materials information, establishing a core requirement for the materials informatics. Natural language processing (NLP) provides a solution by enabling computers to understand and generate text, with natural language understanding (NLU) and natural language generation (NLG) being its two primary tasks [67].

7.1.1 Traditional NLP methods

NLP was applied early in energy materials research. For example, Huo et al. [68] combined unsupervised learning (LDA) with supervised learning (RF) to process over 2.28 million experimental paragraphs from materials science literature. Using Markov chain analysis to capture sequential relationships among LDA topics, they built a form of “machine intuition” for materials synthesis procedures. Kim et al. [69] proposed an unsupervised conditional variational autoencoder (CVAE) linking scientific literature with context-aware insights for

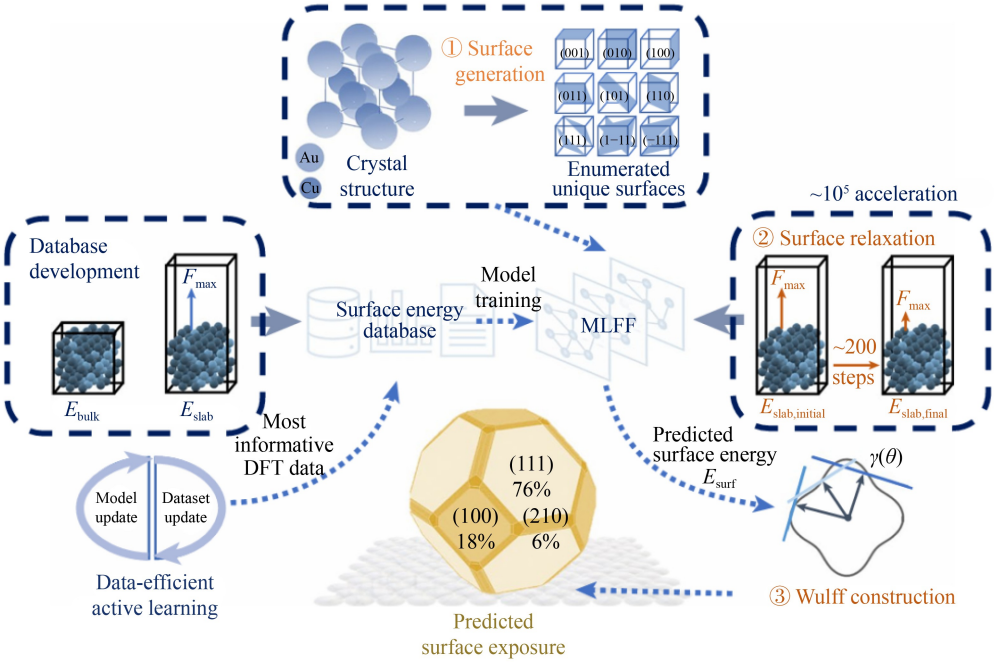


Fig. 9 Overall framework of the SurFF model for predicting surface exposure (adapted form Yin et al. [65] under the terms of CC BY license).

inorganic materials synthesis planning. While these studies demonstrated the potential of NLP in addressing specific, narrowly scoped problems [68–70], traditional NLP methods remain limited in flexibility and struggle with more complex tasks.

7.1.2 Pretrained language models

The advent of the transformer architecture in 2017 revolutionized NLP. Its self-attention mechanism overcomes the limitations of traditional models in capturing long-range textual dependencies and modeling cross-sentence semantic correlations, enabling large-scale pre-training. It was precisely this architectural innovation that directly catalyzed the rapid development of pre-trained language models. The application of pre-trained language models has endowed AI models with more powerful data extraction capabilities in energy materials research and development. For instance, Huang and Cole [24] developed BatteryBERT, a domain-specific model for battery materials. Fine-tuned on literature, it achieves semantic understanding of battery-related terminology and enables token-level classification of "material entity–property" pairs, successfully extracting key parameters of cathode, anode, and electrolyte materials from massive unstructured literature, producing structured datasets directly applicable for materials screening, thus addressing the critical bottleneck of transforming fragmented literature into usable databases.

7.1.3 True LLMs

LLMs expand on pretrained language models by scaling parameters and innovating training paradigms. They inherit advantages in text mining and information extraction while adding emergent capabilities, cross-modal reasoning, and powerful generative capabilities, providing comprehensive technical support for energy materials research. Wang et al. [71] developed a framework integrating LLMs, metadynamics (MetaD), and ML to facilitate the development of divalent hydride SSEs and uncover ion migration mechanisms [71]. In their study, the LLM identified that hydride SSEs containing neutral molecules (e.g., NH_3, CH_3NH_2) exhibit low activation energies, guided MetaD simulation parameters, and extracted eight key descriptors to construct an ML regression model, thereby forming a closed-loop workflow linking data mining, ML-based prediction, and MetaD simulation. .

In recent years, multimodal LLMs have been applied for full-process energy materials research and development. Zhang et al. [72] proposed CRESt (copilot for real-world experimental scientists), a multimodal robotic platform for electrocatalyst discovery. Figure 10 illustrates CRESt's workflow. It integrates a user interface, an LMM-powered back end, and robotic actuators, fusing chemical compositions, text embeddings, and microstructural images. CRESt autonomously explores material research and development strategies, performs automated

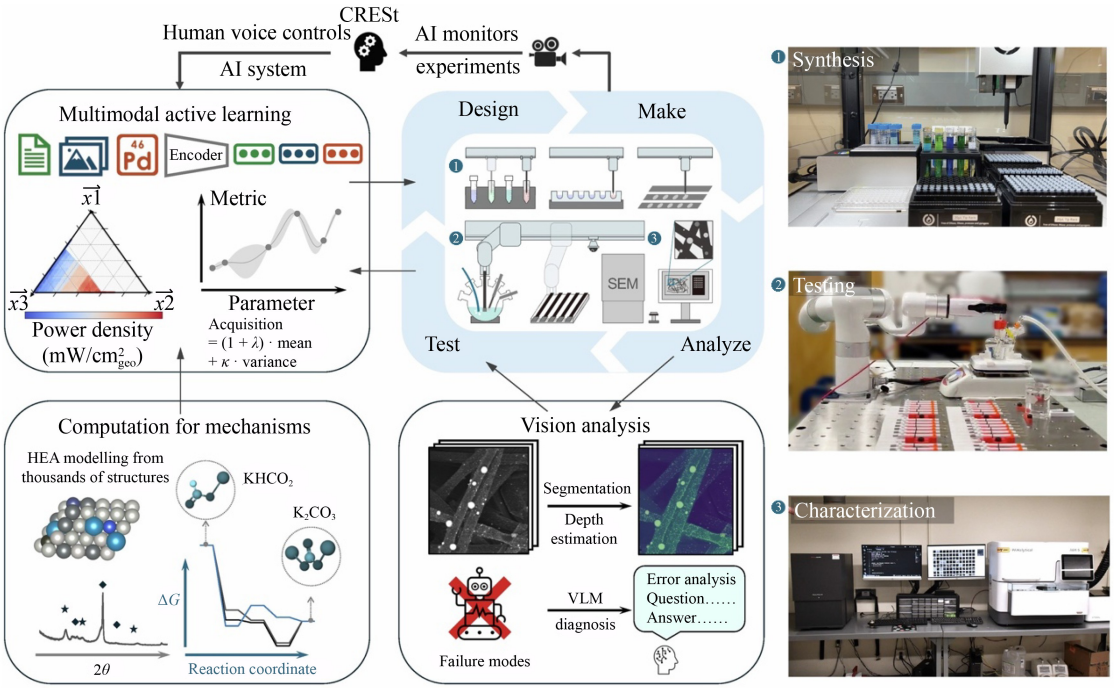


Fig. 10 Workflow of electrocatalyst discovery guided by CRESt. (adapted from Zhang et al. [72] under the terms of CC BY license).

material design, high-throughput synthesis, characterization, and electrochemical performance optimization. It also monitors experiments via cameras and diagnoses experimental anomalies using vision-language models (VLMs). Within three months, CRESt independently explored over 900 catalyst chemical compositions and conducted 3500 electrochemical tests.

In summary, LLM technology, built upon advances in NLP, has been widely applied to data extraction and analysis in energy materials research. With the continued development of multimodal fusion techniques, these models are expected to further enhance the integration and interpretation of diverse material information. Some researchers have already proposed systematic evaluation methods for multimodal data [73], and explored the processing of spectroscopic information [74]. However, approaches that incorporate more complex characterization data, such as microstructural information from scanning electron microscopy, remain lacking. Overall, the adoption of LLMs promises to significantly improve the efficiency of energy materials research and development in the future.

7.2 Energy material database

As materials science increasingly shifts toward a data-driven paradigm, a variety of representative datasets with different functions have been created and widely used, providing a crucial data foundation for both academic research and industrial applications. First, several large-scale general databases are noteworthy. Materials Project, led by the Lawrence Berkeley National Laboratory in the United States, is a large-scale open-source database built on DFT, covering crystal structures, electronic structures, and thermodynamic properties for over one million inorganic substances. AFLOW, developed under the auspices of Duke University, is an open-source materials database and automated computational framework storing structural data for more than 3.56 million materials and 700 million first-principles calculation datasets, encompassing systems such as inorganic compounds and multi-component alloys. OQMD (Open Quantum Materials Database) provides DFT calculation data for over one million materials, focusing on thermodynamic stability and electronic band structure information of inorganic crystals and alloys. NOMAD is an emerging materials data platform notable not only for aggregating more than 12 million computational simulation datasets but also for supporting long-term archiving, sharing, and mining of materials data, while attaching traceable metadata to data to meet the requirements of scientific research reproducibility.

In addition, although some specialized databases are relatively smaller in scale, they may offer greater applicability in the field of energy materials. The Universal Battery Database (BatteryDB), developed by Dahn et al., is an open-source lithium-ion battery data management platform that stores experimental data such as cycling

tests and electrochemical impedance spectroscopy, supports hybrid modeling combining MLwith physical mechanisms, and enables real-time data updates. In the field of SSEs, Yang et al. [75] developed DDSE (Dynamic Database for SSEs), a special dynamic database for all-solid-state batteries. Using DDSE, Xiang et al. [76] integrated their experimental data to develop a predictive and analytical framework for the ionic conductivity of antiperovskite (AP) SSEs through multiple model training, achieving a classification accuracy of up to 94%. Hargreaves et al. [28] also published an experimentally validated database of lithium solid conductors, containing 820 ionic conductivity data points.

In the realm of synthesis method databases, the Text-mined Data set of Inorganic Materials Synthesis Recipes from the University of Cambridge extracts inorganic material synthesis methods from scientific literature using text mining [77]. It enables access to process parameters for electrode material synthesis, facilitates the translation from laboratory research to industrial production, and guides the design of material synthesis routes. Another relevant dataset, the Dataset of Solution-based Inorganic Materials Synthesis Procedures Extracted from the Scientific Literature [78], uses NLP technology to extract solution-based inorganic material synthesis procedures.

In catalysis, the Digital Catalysis Platform (DigCat) developed by Zhang and Li [79] represents an innovative integration of big data and advanced computational tools. It compiles over 400000 experimental performance datasets covering electrocatalysts, thermocatalysts, and photocatalysts, along with over 300000 catalyst structure records. Thanks to such fundamental efforts, many of these material datasets hold significant potential for application in "AI for Energy Research" research. Some typical databases for energy materials can be found in Electronic Supplementary Material.

7.3 Perspectives

In recent years, numerous scholars have published review articles on the application of AI in energy materials from different perspectives, application scenarios, and research methods [84–88]. Wang et al. [84] focused on the effects of external fields (mechanical, electrical, and magnetic) on electrocatalysis and ML-guided design, highlighting the applications of ML interatomic potentials (MLIPs) and data-driven methods in electrocatalysis optimization, emphasizing that the combining ML with external fields is an effective strategy for improving electrocatalytic performance. Hu et al. [86] reviewed AI application in the research and development of rechargeable battery materials, covering the design optimization of core components such as electrodes and electrolytes.

Mortazavi [85] systematically summarized advances in ML interatomic potentials (MLIPs), generative model-based material design, the application of DL in material continuum modeling, as well as the applications of these technologies in crystal structure prediction, thermal conductivity calculation, interface modeling, and other area. Yao et al. [88] looked ahead to the broader field of sustainable energy, proposing Acc(X)eleration Performance Indicators (XPIs) to quantify the efficiency of ML-driven energy material discovery, covering key dimensions such as synthesis speed and the number of materials meeting performance thresholds. In contrast, the present review specifically focuses on battery and electrocatalytic materials, providing targeted insights for researchers in this core energy domain. By tracing the evolution from classical ML to the latest language models, it aims to systematically organize AI application scenarios along the trajectory of technological development. For researchers seeking to bridge AI technologies with energy materials research, this review is intended as a valuable comprehensive knowledge supplement.

8 Conclusions

This review systematically traces the evolution and application of AI technology in battery and electrocatalytic materials research. Classical ML lays a solid foundation for small-data scenarios and interpretable analysis, while advanced representation methods such as GNNs and Transformers realize precise property prediction and efficient material design. Generative models facilitate a shift toward inverse design, and LLMs break barriers in multimodal information integration. High-quality domain-specific databases provide essential support for AI model training and validation.

Looking ahead, two priority directions are critical the robust integration of AI into energy materials research and industrial applications: materials design and experimental integration. In materials design, it is essential to develop more versatile material representation approaches to overcome the inherent limitations of existing methods, which often struggle with diverse material systems and generative design tasks. In experimental integration, multimodal analytical frameworks should be built using LLMs to efficiently process complex, multi-dimensional characterization and performance data. The ultimate goal is to create a universal AI toolkit that spans the entire materials research and development workflow.

Acknowledgements This work was supported by the National Natural Science Foundation of China (Grant Nos. 52302302 and 52572257), and the Fundamental Research Funds for Central Universities.

Electronic supplementary material Supplementary material is available in the online version of this article at https://doi.org/10.1007/s11708-026-1053-5 and is accessible for authorized users.

Competing Interests The authors declare no competing interests.

References

1. Xiao P, Yun X, Chen Y, et al. Insights into the solvation chemistry in liquid electrolytes for lithium-based rechargeable batteries. Chemical Society Reviews, 2023, 52(15): 5255–5316
2. Xiao J, Shi F, Glossmann T, et al. From laboratory innovations to materials manufacturing for lithium-based batteries. Nature Energy, 2023, 8(4): 329–339
3. Fan Z, Chen X, Shi J, et al. Functionalized separators boosting electrochemical performances for lithium batteries. Nano-Micro Letters, 2025, 17(1): 128
4. Wu D, Chen L, Li H, et al. Solid-state lithium batteries-from fundamental research to industrial progress. Progress in Materials Science, 2023, 139: 101182
5. Alsaç E P, Nelson D L, Yoon S G, et al. Characterizing electrode materials and interfaces in solid-state batteries. Chemical Reviews, 2025, 125(4): 2009–2119
6. Huang X Y, Zhao C Z, Kong W J, et al. Tailoring polymer electrolyte solvation for 600 Wh kg^{-1} lithium batteries. Nature, 2025, 646(8084): 343–350
7. Xie C, Chen W, Wang Y, et al. Dynamic evolution processes in electrocatalysis: Structure evolution, characterization and regulation. Chemical Society Reviews, 2024, 53(22): 10852–10877
8. Ren J T, Chen L, Wang H Y, et al. Water electrolysis for hydrogen production: From hybrid systems to self-powered/catalyzed devices. Energy & Environmental Science, 2024, 17(1): 49–113
9. Wang J, Ma J, Cheng H. Nanomaterials-based enzymatic biofuel cells for wearable and implantable bioelectronics. Frontiers in Energy, 2025, 19(3): 283–299
10. Leung K. DFT modelling of explicit solid–solid interfaces in batteries: Methods and challenges. Physical Chemistry Chemical Physics, 2020, 22(19): 10412–10425
11. Wan H, Zhang B, Liu S, et al. Interface design for high-performance all-solid-state lithium batteries. Advanced Energy Materials, 2024, 14(19): 2303046
12. Xu G L, Liu X, Zhou X, et al. Native lattice strain induced structural earthquake in sodium layered oxide cathodes. Nature Communications, 2022, 13(1): 436
13. Kong Z, Wu J, Liu Z, et al. Advanced electrocatalysts for fuel cells: Evolution of active sites and synergistic properties of catalysts and carrier materials. Exploration, 2025, 5(1): 20230052
14. Bi T, Liu Y, Wei Y, et al. Deep learning-based structural characterization and mass transport analysis of CO_2 reduction catalyst layers. Frontiers in Energy, 2025, 19(5): 681–693
15. Zhang X, Feng J, Cai F, et al. A novel state of health estimation model for lithium-ion batteries incorporating signal processing and optimized machine learning methods. Frontiers in Energy, 2025, 19(3): 348–364
16. Han Z, Tao S, Jia Y, et al. Data-driven insight into the universal structure–property relationship of catalysts in lithium–sulfur batteries. Journal of the American Chemical Society, 2025, 147(26): 22851–22863
17. Han Z, Chen A, Li Z, et al. Machine learning-based design of electrocatalytic materials towards high-energy lithium||sulfur batteries development. Nature Communications, 2024, 15(1): 8433
18. Cui K, Wang T, Zhang Q, et al. Multi-functional descriptor design of V-based double atomic catalysts for room temperature sodium-sulfur batteries. Small, 2025, 21(4): 2409866
19. Liu B, Yang J, Yang H, et al. Rationalizing the interphase stability of Li|doped-$Li_7La_3Zr_2O_{12}$ via automated reaction screening and machine learning. Journal of Materials Chemistry. A, Materials for Energy and Sustainability, 2019, 7(34): 19961–19969
20. Jiang M, Zhang Y, Yang Z, et al. A data-driven interpretable method to predict capacities of metal ion doped TiO_2 anode materials for lithium-ion batteries using machine learning classifiers. Inorganic Chemistry Frontiers, 2023, 10(22): 6646–6654
21. Jiang M, Yang Z, Lu T, et al. Machine learning accelerated study for predicting the lattice constant and substitution energy of metal doped titanium dioxide. Ceramics International, 2024, 50(1): 1079–1086
22. Ling N, Wang Y, Song S, et al. Experimentally validated screening strategy for alloys as anode in Mg-air battery with multi-target machine learning predictions. Chemical Engineering Journal, 2024, 496: 153824
23. Madani M, Lacivita V, Shin Y, et al. Accelerating materials property prediction via a hybrid Transformer Graph framework that leverages four body interactions. npj Computational Materials, 2025, 11(1): 15
24. Huang S, Cole J M. BatteryBERT: A pretrained language model for battery database enhancement. Journal of Chemical Information and Modeling, 2022, 62(24): 6365–6377
25. Xu C, Wang Y, Barati Farimani A. TransPolymer: A transformer-based language model for polymer property predictions. npj Computational Materials, 2023, 9(1): 64
26. Choi E, Jo J, Kim W, et al. Searching for mechanically superior solid-state electrolytes in Li-ion batteries via data-driven approaches. ACS Applied Materials & Interfaces, 2021, 13(36):

42590–42597
27. Sendek A D, Yang Q, Cubuk E D, et al. Holistic computational structure screening of more than 12000 candidates for solid lithium-ion conductor materials. Energy & Environmental Science, 2017, 10(1): 306–320
28. Hargreaves C J, Gaultois M W, Daniels L M, et al. A database of experimentally measured lithium solid electrolyte conductivities evaluated with machine learning. npj Computational Materials, 2023, 9(1): 9
29. Hu Q, Chen K, Li J, et al. Speeding up the development of solid state electrolyte by machine learning. Next Energy, 2024, 5: 100159
30. Bzdok D, Krzywinski M, Altman N. Machine learning: Supervised methods. Nature Methods, 2018, 15(1): 5–6
31. Mishra A K, Rajput S, Karamta M, et al. Exploring the possibility of machine learning for predicting ionic conductivity of solid-state electrolytes. ACS Omega, 2023, 8(18): 16419–16427
32. Pilania G, Wang C, Jiang X, et al. Accelerating materials property predictions using machine learning. Scientific Reports, 2013, 3(1): 2810
33. Klein Moberg H, Abbondanza G, Nedrygailov I, et al. Deep-learning-enabled online mass spectrometry of the reaction product of a single catalyst nanoparticle. Nature Communications, 2025, 16(1): 7203
34. Li G, Zhu Y, Guo Y, et al. Deep learning to reveal the distribution and diffusion of water molecules in fuel cell catalyst layers. ACS Applied Materials & Interfaces, 2023, 15(4): 5099–5108
35. Bi T, Liu Y, Wei Y, et al. Deep learning-based structural characterization and mass transport analysis of CO_2 reduction catalyst layers. Frontiers in Energy, 2025, 19(5): 681–693
36. Himanen L, Jäger M O J, Morooka E V, et al. DScribe: Library of descriptors for machine learning in materials science. Computer Physics Communications, 2020, 247: 106949
37. Chen D, Wang B, Li S, et al. Superionic ionic conductor discovery via multiscale topological learning. Journal of the American Chemical Society, 2025, 147(24): 20888–20898
38. Zhang Y, He X, Chen Z, et al. Unsupervised discovery of solid-state lithium ion conductors. Nature Communications, 2019, 10(1): 5260
39. Park J H, Hwang S K, Ji S G, et al. Characterization of various tandem solar cells: Protocols, issues, and precautions. Exploration, 2023, 3(2): 20220029
40. Cheng W, Zhao M, Lai Y, et al. Recent advances in battery characterization using *in situ* XAFS, SAXS, XRD, and their combining techniques: From single scale to multiscale structure detection. Exploration, 2024, 4(1): 20230056
41. Xie T, Grossman J C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. Physical Review Letters, 2018, 120(14): 145301
42. Schütt K, Kindermans P J, Sauceda Felix H E, et al. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. In: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, USA, 2017
43. Choudhary K, DeCost B. Atomistic line graph neural network for improved materials property predictions. npj Computational Materials, 2021, 7(1): 185
44. Louis S Y, Siriwardane E M D, Joshi R P, et al. Accurate prediction of voltage of battery electrode materials using attention-based graph neural networks. ACS Applied Materials & Interfaces, 2022, 14(23): 26587–26594
45. He X, Chen Y, Wang S, et al. Employing graph neural networks for predicting electrode average voltages and screening high-voltage sodium cathode materials. ACS Applied Materials & Interfaces, 2024, 16(19): 24494–24501
46. Law J N, Pandey S, Gorai P, et al. Upper-bound energy minimization to search for stable functional materials with graph neural networks. JACS Au, 2023, 3(1): 113–123
47. Wang Y, Liu J, Du P H, et al. Screening topological quantum cathode materials for K-ion batteries by graph neural network and first-principles calculations. ACS Applied Energy Materials, 2023, 6(9): 4503–4510
48. Kuenneth C, Ramprasad R. PolyBERT: A chemical language model to enable fully machine-driven ultrafast polymer informatics. Nature Communications, 2023, 14(1): 4099
49. Antunes L M, Butler K T, Grau-Crespo R. Crystal structure generation with autoregressive large language modeling. Nature Communications, 2024, 15(1): 10570
50. Li H, Hao J, Qiao S. AI-driven electrolyte additive selection to boost aqueous Zn-ion batteries stability. Advanced Materials, 2024, 36(49): 2411991
51. Nandishwara K M, Cheng S, Liu P, et al. Data-driven microstructural optimization of Ag-Bi-I perovskite-inspired materials. npj Computational Materials, 2025, 11(1): 210
52. Tao S, Liu H, Sun C, et al. Collaborative and privacy-preserving retired battery sorting for profitable direct recycling via federated machine learning. Nature Communications, 2023, 14(1): 8032
53. Schneider G, Fechner U. Computer-based de novo design of drug-like molecules. Nature Reviews. Drug Discovery, 2005, 4(8): 649–663
54. Zeni C, Pinsler R, Zügner D, et al. A generative model for inorganic materials design. Nature, 2025, 639(8055): 624–632
55. Parida C, Roy D, Lastra J M G, Bhowmik A. Mining chemical space with generative models for battery materials. Batteries & Supercaps, 2025: e202500309
56. Gao X, Yang A Q, Yu W B, et al. Generative artificial intelligence navigated development of solvents for next generation high-performance magnesium batteries. Advanced Materials, 2025: e10083
57. Chen Y, Wang X, Deng X, et al. MatterGPT: A generative transformer for multi-property inverse design of solid-state materials. arXiv preprint, arXiv:2408.07608, 2024
58. Wang Z, You F. Leveraging generative models with periodicity-aware, invertible and invariant representations for crystalline materials design. Nature Computational Science, 2025, 5(5): 365–376
59. Gayon-Lombardo A, Mosser L, Brandon N P, et al. Pores for thought: Generative adversarial networks for stochastic

reconstruction of 3D multi-phase electrode microstructures with periodic boundaries. npj Computational Materials, 2020, 6(1): 82
60. Kench S, Squires I, Dahari A, et al. Li-ion battery design through microstructural optimization using generative AI. Matter, 2024, 7(12): 4260–4269
61. Müller S, Sauter C, Shunmugasundaram R, et al. Deep learning-based segmentation of lithium-ion battery microstructures enhanced by artificially generated electrodes. Nature Communications, 2021, 12(1): 6205
62. Liu D, Shadike Z, Lin R, et al. Review of recent development of *in situ*/operando characterization techniques for lithium battery research. Advanced Materials, 2019, 31(28): 1806620
63. Cheng G, Gong X G, Yin W J. Crystal structure prediction by combining graph network and optimization algorithm. Nature Communications, 2022, 13(1): 1492
64. Cheng G, Yin W J. De novo inverse design superhard C–N compounds via global machine learning interatomic potentials and multiobjective optimization algorithm. Journal of Physical Chemistry Letters, 2025, 16(18): 4392–4400
65. Yin J, Chen H, Qiu J, et al. SurFF: A foundation model for surface exposure and morphology across intermetallic crystals. Nature Computational Science, 2025, 5(9): 782–792
66. Chen H, Liu H, Tew Y, et al. Distilling knowledge from catalysis literature with long-context large language model agents. ACS Catalysis, 2025, 15(21): 18244–18254
67. Wein S, Schneider N. Assessing the cross-linguistic utility of abstract meaning representation. Computational Linguistics, 2024, 50(2): 419–473
68. Huo H, Rong Z, Kononova O, et al. Semi-supervised machine-learning classification of materials synthesis procedures. npj Computational Materials, 2019, 5(1): 62
69. Kim E, Jensen Z, Van Grootel A, et al. Inorganic materials synthesis planning with literature-trained neural networks. Journal of Chemical Information and Modeling, 2020, 60(3): 1194–1201
70. Dagdelen J, Dunn A, Lee S, et al. Structured information extraction from scientific text with large language models. Nature Communications, 2024, 15(1): 1418
71. Wang Q, Yang F, Wang Y, et al. Unraveling the complexity of divalent hydride electrolytes in solid-state batteries via a data-driven framework with large language model. Angewandte Chemie International Edition, 2025, 64(25): e202506573
72. Zhang Z, Ren Z, Hsu C W, et al. A multimodal robotic platform for multi-element electrocatalyst discovery. Nature, 2025, 647(8089): 390–396
73. Alampara N, Schilling-Wilhelmi M, Ríos-García M, et al. Probing the limitations of multimodal language models for chemistry and materials research. Nature Computational Science, 2025, 5(10): 952–961
74. Priessner M, Lewis R J, Lemurell I, et al. Advancing structure elucidation with a flexible multi-spectral AI model. Angewandte Chemie, 2025: e17611
75. Yang F, Campos dos Santos E, Jia X, et al. A dynamic database of solid-state electrolyte (DDSE) picturing all-solid-state batteries. Nano Materials Science, 2024, 6(2): 256–262
76. Xiang S, Lu S, Li J, et al. Ionic conductivity study of antiperovskite solid-state electrolytes based on interpretable machine learning. ACS Applied Energy Materials, 2025, 8(3): 1620–1628
77. Kononova O, Huo H, He T, et al. Text-mined dataset of inorganic materials synthesis recipes. Scientific Data, 2019, 6(1): 203
78. Wang Z, Kononova O, Cruse K, et al. Dataset of solution-based inorganic materials synthesis procedures extracted from the scientific literature. Scientific Data, 2022, 9(1): 231
79. Zhang D, Li H. Digital catalysis platform (DigCat): A gateway to big data and AI-powered innovations in catalysis. Chemistry, Preprint, 2024, https://doi.org/10.26434/chemrxiv-2024–9lpb9
80. Tran R, Lan J, Shuaibi M, et al. 2022 (OC22) Dataset and challenges for oxide electrocatalysts. ACS Catalysis, 2023, 13(5): 3066–3084
81. Rosen A S, Iyer S M, Ray D, et al. Machine learning the quantum-chemical properties of metal–organic frameworks for accelerated materials discovery. Matter, 2021, 4(5): 1578–1597
82. Venugopal V, Olivetti E. MatKG: An autonomously generated knowledge graph in material science. Scientific Data, 2024, 11(1): 217
83. Barroso-Luque L, Shuaibi M, Fu X, et al. Open materials 2024 (omat24) inorganic materials dataset and models. arXiv preprint, arXiv: 2410.12771, 2024
84. Wang L, Zhou X, Luo Z, et al. Review of external field effects on electrocatalysis: Machine learning guided design. Advanced Functional Materials, 2024, 34(49): 2408870
85. Mortazavi B. Recent advances in machine learning-assisted multiscale design of energy materials. Advanced Energy Materials, 2025, 15(9): 2403876
86. Hu Q, Lu J, Hui J, et al. Artificial intelligence-driven development in rechargeable battery materials: Progress, challenges, and future perspectives. Advanced Functional Materials, 2025, 35(52): e08438
87. Kim K S. Machine learning for accelerating energy materials discovery: Bridging quantum accuracy with computational efficiency. Advanced Energy Materials, 2025: e03356
88. Yao Z, Lum Y, Johnston A, et al. Machine learning for a sustainable energy future. Nature Reviews. Materials, 2022, 8(3): 202–215

Author Biography

Menghao Yang, a doctoral supervisor at Tongji University, is dedicated to constructing artificial intelligence models to advance fundamental research on novel solid-state energy storage materials such as solid-state batteries and fuel cells. In the past five years, he has published over 40 papers as the first or corresponding author (including co-contributions) in journals such as Nat. Mater. (2 papers), Nat. Catal. (2 papers), Nat. Commun. (3 papers), and Adv. Mater. (2 papers). He has led projects including the National Natural Science Foundation of China (General/Youth Program) and the 2025 Shanghai Fundamental Research Explorer Program. He was selected for the 2021 Shanghai Magnolia Talent Program. He has been granted 3 national patents and 10 software copyrights. He serves as a youth editorial board member or guest editor for 12 journals, including Nano-Micro Letters, Chinese Chemical Letters, Exploration, Innovation, and MGE Advances.