Review | Received 29 August 2025; Revised 20 November 2025; Accepted 1 December 2025; Published 6 January 2026
https://doi.org/10.55092/aimat20060001

Application of artificial intelligence combined with density functional theory in materials

Xiemeng Zhu[1], Juhong Yu[2,*], Yong Liu[3], Yangyang Song[4], Liang Zhang[4] and Shiyu Du[1,2,5,*]

[1] Qingdao Institute of Software College of Computer Science and Technology, China University of Petroleum (East China), Qingdao 266580, China
[2] School of Materials Science and Engineering, China University of Petroleum (East China), Qingdao 266580, China
[3] Department of Chemistry, University of Colorado Denver, Denver, Colorado 80217-3364, USA
[4] Putuo People's Hospital, Frontier Science Center for Stem Cell Research, School of Life Sciences and Technology, Tongji University, Shanghai 200092, China
[5] Milky-Way Sustainable Energy Ltd., Zhuhai 519000, China

* Correspondence authors: E-mails: 20230031@upc.edu.cn (J.Y.); dushiyu@nimte.ac.cn (S.D.).

Highlights:

- AI accelerates materials discovery by integrating with DFT calculations.
- Details key algorithms and workflows, with applications in semiconductors, perovskites, and 2D materials.
- Outlines the evolution from computational assistance to intelligent autonomous discovery and future R&D directions.

Abstract: Artificial intelligence (AI), as an important driving force of the technological revolution, is changing the way humans produce, live, and learn. In recent years, massive training data, advanced algorithms, and efficient computational power have advanced the widespread application of AI. The cross-integration of AI and materials science is currently an important scenario and technological frontier for the application of AI. The application in the field of new materials has shown great potential and value. Compared to traditional experimental and Density Functional Theory (DFT)-based approaches, which are time-consuming and inefficient for studying material properties, the rapid advancement of AI technology is dramatically accelerating the exploration, design, synthesis, and optimization of novel materials. This review introduces the application of AI techniques combined with DFT theoretical computation in materials innovation, such as research on new materials, materials design, property prediction and synthesis. It highlights the advantages of AI techniques combined with DFT theoretical computation over traditional methods in the field of materials, as well as the future directions and unknown challenges of materials science.

Keywords: artificial intelligence; machine learning; density functional theory; deep learning; semiconductor material; perovskite material; two-dimensional material

1. Introduction

Materials science focuses on the relationship between material structure, processing, properties, and application. After centuries of development, it has accumulated a vast amount of data. However, due to cognitive limitations, it is difficult for humans to quickly extract useful information from massive literature and data [1,2]. Materials informatics, on the other hand, employs AI techniques to mine and utilize existing materials data to obtain information on material composition, processing, and properties. This allows for the rapid prediction of new compounds, providing a high-speed and effective approach for discovering new materials [3]. In the past, the research and development of new materials primarily relied on traditional experimental trial-and-error methods. After clarifying the material structure and properties, the preparation process and methods were determined based on experience. The physical and chemical properties of the developed experimental samples were analyzed by analyzing their performance. This process requires abundant synthesis and characterization of the material to determine its crystal structure, which is highly dependent on the experimental samples and equipment and requires repeated work, resulting in time-consuming and costly results. Recent technological advances have enabled numerous methods for calculating material structures and properties, accelerating the discovery of new electronic materials. Such as Molecular Dynamics (MD), Finite Element Method (FEM), DFT and so on [4–6]. Among these, DFT calculations have found widespread application in materials science, mainly used to research the electronic structure, magnetism, mechanical properties, and other characteristics of materials. For example, it is used to study and optimize the electronic structure and magnetic properties of nanoparticles in nanomaterials; to predict the energy band structure of semiconductor materials to optimize their electrical properties; and to study the reaction mechanism and active sites on the surface of catalysts, to predict the catalytic activity and selectivity of catalysts, and to guide the design of new catalysts [7–10]. By studying these properties to predict and optimize their physical and chemical properties, it can also be applied to high-throughput screening to discover new materials [11–13]. Although DFT is used to calculate the electronic structure of material compositions to avoid additional experimental costs, the electronic structure of materials is often complex, which leads to longer computation times and less accuracy [14,15]. Despite advances in algorithms and computational conditions, computation with DFT is still relatively expensive when the number of computational tasks is in the thousands and the scalability of the system scale is poor, which is the main factor limiting its development in the materials field. However, these problems are gradually being solved by rapid advances in AI and its cross-disciplinary application in materials science.

Computers play an indispensable role in modern life, and the speed of their development is astonishing. In multiple fields, computers are gradually replacing humans. Their growing computing power and storage capacity attract people and make it possible to use computers to handle complex tasks and systems. Recent advances in computer science and technology have given modern computers human-like ‘self-learning’ and ‘self-adaptive’ capabilities [16–18]. The emergence of AI has brought new hope to the development of materials, becoming one of the most important topics in computer science. The research and development (R&D) of materials is no longer limited to traditional trial-and-error methods and theoretical calculations. To accelerate materials research, and to solve the problem of long

cycle time and large consumption of human and material resources with traditional materials, the strategy of combining AI and theoretical calculation (DFT) can be adopted. Which can greatly improve the calculation accuracy and efficiency, as well as enhance the efficiency of model design and development. Compared to traditional DFT calculations, which require computationally expensive, one-by-one analysis of a pre-defined materials space and struggle to scale to complex systems, the integration of AI and DFT has enabled a paradigm shift from "passive screening" to "active design". This data-driven approach allows for the rapid generation and optimization of new materials based on existing knowledge, establishing a closed-loop system of "prediction-verification-learning" [19]. Machine Learning (ML) and Deep Learning (DL), which are important components of AI to reveal relationships embedded in data and predict new viable materials [20–22], are rapidly becoming central to a variety of modern technologies [23,24]. AI techniques are now widely used in a variety of fields such as genomics [25,26], drug discovery [27,28], and automation [29].

In this review, we discuss the progress of the application of AI techniques combined with DFT in different types of materials. The focus is that the addition of AI techniques solves the disadvantage of traditional theory in terms of costly and time-consuming calculations, mainly by combining ML and DFT. This involves using ML algorithms to extract effective information and performance data, establishing suitable models, and outputting the final prediction results. Finally, we summarize and look forward to the advancements in the materials field driven by the combination of these two methods, while also revealing existing issues and challenges and offering reasonable recommendations.

2. Introduction to ML techniques and algorithms

2.1. The introduction of ML

ML is the science of studying how to simulate or realize human learning activities using computers. It is currently one of the most cutting-edge research areas in AI [30]. As an important branch of AI, ML can effectively identify high-dimensional data, quickly extract valuable information, uncover hidden patterns, reduce computational costs, and shorten development cycles. Since its inception by Samuel [31] in 1959, ML has found applications in computer vision (CV), gaming, data mining, and bioinformatics [32–35]. After the 1980s, it attracted widespread attention as a major pathway for realizing AI. Currently, ML can be divided into two research directions: traditional ML and ML in the big data environment. Traditional ML research primarily focuses on learning mechanisms, exploring and simulating human learning processes. In contrast, ML research in the big data environment emphasizes how to effectively utilize information, extracting hidden, useful, and interpretable knowledge from massive datasets [36]. Nowadays, ML is widely used to solve numerous problems in materials science. In the past century, it was used to detect the solubility of C_{60} in material science. Currently, it is commonly used to discover new materials, predict material and molecular properties, engage in molecular inverse design, and design drugs [37–39]. Over the past decade, AI and ML have gradually matured, achieving significant progress in various fields.

DL as a branch of ML, is a ML method based on artificial neural networks (ANN). Its core idea is to mimic the brain's inter-neuronal connections via multi-layer networks [40]. DL systems utilize gradient-based optimization to optimize multi-layer network parameters from output errors [41]. The characteristic of DL is that it can train models using large-scale data and automatically learn the feature

representations of the data. However, DL techniques have certain drawbacks, such as being time-consuming and demanding in terms of data requirements [42]. Moreover, DL shows great promise in fields such as video games, image recognition, structural engineering, chemoinformatics, and materials science [43–46].

2.2. Basic workflow of ML

The primary workflow of ML in materials research consists of five components: data preparation, feature engineering, model selection, model evaluation, and model application. The workflow is depicted in Figure 1 below.

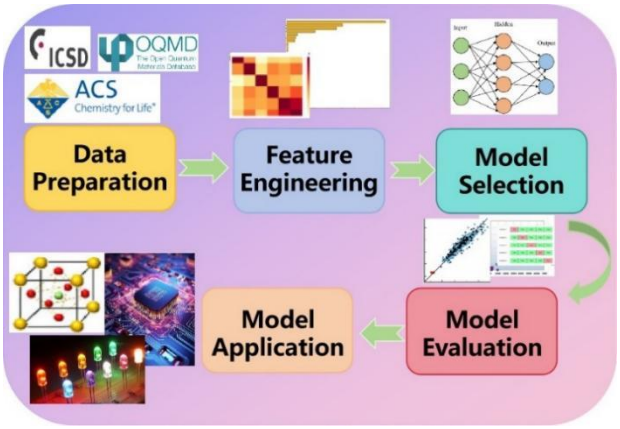

Figure 1. Machine Learning flowchart.

2.2.1. Data preparation

Datasets used in ML typically contain features or descriptors related to materials, primarily information related to the structure and properties of materials, including the physicochemical properties of molecules and atoms, the structural properties of compounds, and the process conditions during synthesis. The quantity and quality of data critically affect material predictions. Insufficient data or significant experimental and calculation errors can compromise data quality, directly impacting the final model effectiveness. Generally, obtaining high-quality data requires preliminary data processing, including methods such as deleting missing and repeated values, data normalization, and data standardization [47,48]. Then, samples in the dataset are classified and screened based on elements and structures to remove interfering samples.

Relevant data can be obtained through experiments, literature, simulation calculations (such as DFT, MDs), and databases. The database has collected a large volume of diverse data types generated from experiments, simulation calculations, and ML. However, the origin data obtained from experimental measurements or computational simulations often suffer from incompleteness, noise, and inconsistency. Therefore, data preprocessing is necessary to ensure their integrity and consistency [49]. Specific steps include integrating scattered data, imputing missing values, remove erroneous data. Additionally, data normalization can improve model accuracy and convergence efficiency [50]. To ensure data quality, it should be collected from authoritative databases (some common databases are listed in Table 1). When required data is not available in databases, datasets can be generated through theoretical calculations, with common calculation software including Materials Studio (MS), Vienna Ab initio Simulation Package (VASP), and so on.

Table 1. Public database of various materials.

Database	Description
Materials Project (MP)	A large number of computational and experimental data of inorganic materials
Open Quantum Materials Database (OQMD)	Material property data (e.g., crystal structure, energy, electronic structure) from experiments and simulations
The Inorganic Crystal Structure Database (ICSD)	Experimentally characterized inorganic crystal data
Cambridge Structural Database (CSD)	Structure database of small molecule and metal-organic molecule crystals from X-ray and neutron diffraction
Crystallography Open Database (COD)	Structures data of organic, inorganic, and metal-organic compounds and minerals
Materials Platform for Data Science (MPDS)	A web-based platform that provides curated data and calculations for materials science research

2.2.2. Feature engineering

Feature engineering constructs effective predictive features from raw data by selecting and transforming physicochemical properties that are closely related to the target property [51]. To effectively train the model and avoid overfitting, it is essential to have fewer features than samples. Therefore, feature selection ensures model accuracy and efficiency by reducing input dimensionality and removing redundant features while preserving critical information [52,53].

2.2.3. Model selection

ML algorithms construct models based on sample data to make predictions and decisions for target tasks [54]. Depending on whether the training data is labeled, ML is generally divided into supervised learning (classification and regression) and unsupervised learning (clustering) [55]. If the target is continuous, the task is regression; otherwise, it's classification. A suitable model is selected for predicting the performance of the target property based on the target task. Algorithm selection criteria primarily rely on cross-validation and independent testing results. Common evaluation metrics include Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Coefficient of Determination (R^2), accuracy, recall, F1-score, *etc.* [56,57].

2.2.4. Model evaluation and application

Model evaluation is the process of assessing a model's generalization capability (performance). Evaluating the trained model is an indispensable part of the model development process, helping to select the best model for representing the data and ensuring accurate prediction of unknown data, with the goal of determining the effectiveness of the model in practical applications. Common model evaluation methods include independent testing, cross-validation, and bootstrap [58]. A model's generalization error is typically assessed using a test set to evaluate its predictive performance on unseen samples. Generally, the smaller the error in independent testing, the stronger the model's generalization ability. Cross-validation is a statistical technique for assessing model performance by repeatedly holding out one of several data subsets (usually k-fold cross-validation) for validation while using the rest for training [59]. Bootstrap is a statistical method that performs simulated sampling based on the original data to assess the confidence intervals of certain parameters [60]. In ML and statistics, we typically use bootstrap for model evaluation, and subsequently apply the evaluated model to relevant fields.

2.3. Algorithms introduction

The common algorithms for ML currently include K-Nearest Neighbors (KNN), Decision Trees, ensemble methods (e.g. Random Forest (RF) and Gradient Boosting), Neural Networks, and DL algorithms (e.g. Convolutional Neural Networks (CNN)). Next, these algorithms will be briefly introduced.

(1) KNN algorithm

The KNN algorithm [61,62] is a basic classification and regression algorithm. As a supervised method, KNN relies on labeled data for training. The core idea of this algorithm is to assign a sample to the category most common among its K nearest neighbors in the feature space. For classification problems, given an unknown sample, the KNN algorithm searches for the K training samples closest to this sample in the training set and predicts the category of the unknown sample based on the most common category among these K samples. For regression problems, the KNN algorithm finds the K training samples most similar to the target sample and predicts the value of the target sample based on the average of these samples.

The logic of the KNN algorithm is very simple, easy to understand and implement. It can directly use training data for prediction, so it is fast when processing new data. The KNN algorithm does not assume any prior distribution of the data and is suitable for various types of feature data. It performs well on simple problems but may be inefficient on large-scale datasets. When applying the KNN algorithm, it is necessary to pay attention to selecting appropriate K values and distance metrics. The KNN algorithm can only provide classification results and cannot give the probability of the prediction results, which may require additional processing in some cases.

(2) Decision tree

The Decision Tree (DT) algorithm [63,64] is a common supervised learning algorithm used to address classification and regression problems. The DT model is presented in a tree structure, where internal nodes represent features, branches represent feature values, and leaf nodes represent categories or numerical values. For classification problems, DT constructs a tree-like model based on training data by learning the relationships between sample features, allowing classification of unknown data through this model. In regression, DT can predict a target value according to feature values. Additionally, to avoid overfitting, pruning can be applied to the generated DT to remove some branches or leaf nodes.

DT are algorithms that are straightforward and easy to interpret, performing well in processing various types of data. They can handle both numerical and categorical data without requiring preprocessing. Additionally, they can manage missing values without causing the algorithm to fail. DT also have the capability to automatically select important features, eliminating the need for parameter tuning and simplifying the model training process. However, they also have some drawbacks, such as being prone to overfitting, sensitive to small changes, and difficult to handle high-dimensional data [65]. When utilizing the DT algorithm, it is crucial to weigh its advantages and disadvantages based on specific problems and take corresponding measures to mitigate these issues.

(3) RF

RF [66] comprises abundant individual decision trees, known as Classification and Regression Trees (CART). As an ensemble algorithm, RF determines its output through majority voting of the

individual trees. It is a variation of Bagging [67], where the final result is produced by voting based on the principle of "majority rules". When selecting samples for Random Forest, a random sampling method with replacement is adopted, where data is randomly drawn as the training data for one of the decision tree models [68]. Unlike traditional decision tree training, a subset of attributes is first selected, and then the optimal splitting attribute is chosen from this subset, which can speed up the training process. To improve prediction accuracy and reduce overfitting, Random Forest employs the Bagging method to combine multiple decision tree models. The basic idea is to draw multiple datasets from the original data using the Bootstrap resampling method, and then construct decision tree models for each of these datasets. The final output is produced by voting or averaging the predictions of the multiple decision trees. This ensemble method can enhance prediction accuracy and mitigate the overfitting problem of a single decision tree.

RF can effectively handle high-dimensional data, tolerate missing values, and can assess feature importance. It has achieved widespread success in fields such as medical diagnosis, financial risk assessment, and image classification. However, it also has some issues, such as model complexity, high memory consumption, and sensitivity to imbalanced class data.

(4) Neural networks

Neural Networks (NNs) [69–71] are ML models inspired by the structure of biological neurons, designed to solve various complex problems, including classification, regression, clustering, and more. NNs consist of multiple neurons (or nodes) organized into multiple layers, with each neuron connected to all neurons in the next layer, processing input data by learning appropriate weights. The advantages of NNs include their ability to handle complex nonlinear relationships, suitability for large-scale datasets, and good generalization capability. It has achieved great success in natural language processing, image recognition, speech recognition, and other fields. However, NNs face challenges like requiring numerous labeled data, high computational resource consumption, and poor model interpretability. In recent years, with the advancement of DL technologies, NNs have become increasingly critical in AI.

(5) CNN

CNN is a feedforward neural network that incorporates convolutional computations and possesses a deep structure [72,73]. CNNs are capable of representation learning and enable translation-invariant classification of inputs through their hierarchical architecture. Therefore, they are also known as "Shift-Invariant Artificial Neural Networks (SIANN)" [74]. CNNs have achieved tremendous success in the field of CV, inspired by the biological visual system and designed to simulate the way human vision processes information [75]. It has now become an integral part of CV and DL research.

CNNs have strong feature extraction, generalization ability, visualization analysis, and large-scale data processing capabilities. They can automatically extract features from speech, graphics, and other types of data, enabling classification and recognition of unseen data. However, they require abundant training data to train the model effectively, otherwise the results may be unsatisfactory. They also consume significant computational resources. Additionally, CNN models often lack interpretability and are prone to overfitting.

(6) GAN

GAN [76–78] are a type of DL model whose core idea is to generate realistic data through the competition of two neural networks. GAN consists of a generator model G and a discriminator model D.

The goal of the generator model G is to generate samples that are as close to real data as possible, while the discriminator model D is designed to differentiate authentic samples from generated ones. These two networks compete against each other continuously, and through this competitive mechanism, GAN enables train data distribution learning and high-quality sample generation.

GAN is an unsupervised learning method that does not require labeled data. It can generate highly realistic images, audio, and text data, making it applicable to various fields [79]. However, GANs are prone to instability during training, which can lead to model collapse. They may also generate repeated patterns, resulting in a lack of diversity in the generated samples. Furthermore, evaluating the performance of GANs is challenging because there is no clear metric to measure the quality of the fake data generated by the generator.

3. Application of AI in materials

Due to the huge space of material combinations, it is difficult to explore various structures within a short time using only DFT methods, and researchers also aim to optimize the desired properties of materials. However, the results based on simulation calculations or experimental trial-and-error are often unsatisfactory, time-consuming and laborious, and the results are average. Recently, with the growing application of AI in materials, particularly the rapid advancement of ML in exploring synthesis and predicting material features and structures, the accuracy and speed of predicting material properties and structures have been greatly improved. Furthermore, the combination of AI and DFT has shown significant advantages in improving computational accuracy and efficiency, overcoming the limitations of traditional methods, promoting interdisciplinary research, fostering theoretical innovation, and increasing the efficiency of model design and development.

The integration of AI and DFT has evolved through three key phases: Starting around 2010, the "computational assistance" phase saw AI primarily used to enhance efficiency, exemplified by Machine Learning Potentials (MLPs) that fitted DFT data to enable large-scale atomic simulations [80]. After 2015, the field entered the stage of augmented intelligence, shifting its focus to learning structure-property relationships from massive DFT data to enable rapid prediction of material properties, significantly expanding the scope of screening. Since 2020, AI has entered the autonomous discovery phase, leveraging technologies such as generative models and Bayesian optimization to proactively design new materials [81–82]. These designs are then automatically validated by DFT, gradually forming a closed-loop intelligent R&D system centered on "design–verification–learning". Next, we will introduce the applications of AI combined with DFT in semiconductor materials, perovskite materials and two-dimensional (2D) materials in recent years.

3.1. Application of AI combined with DFT on semiconductor materials

Semiconductors are important components of modern devices such as transistors [83], light-emitting diodes [84,85], integrated circuits [86], photovoltaics [87], solar cells [88–90]. Figure 2 contains applications of semiconductors in various fields. For example, excellent thermal conductivity, high breakdown strength, and a wide band gap make silicon carbide suitable for demanding conditions like high temperature, frequency, and power [91]. Therefore, the exploration of computational methods for semiconductors is crucial for the improvement of future technologies.

Figure 2. Semiconductor applications in various fields.

Edirisuriya *et al*. [92] built a GAN-classifier-calculation pipeline to identify stable semiconductors. A GAN-based algorithm named CubicGAN was employed to generate cubic materials, and a classifier was developed to screen semiconductors, followed by first-principles studies to investigate their stability. The main framework of CubicGAN is shown in Figure 3. It is shown that $AA'MnH_6$ and $NaYRuH_6$ semiconductors have significantly different properties compared to other $AA'MH_6$ semiconductors. Gao *et al*. [93] to accelerate the development of new quaternary semiconductor materials with excellent performance, combined ML with DFT to predict the band gaps of 2180 quaternary semiconductors, most are unexploited and eco-friendly materials. Four new direct band gap quaternary semiconductors (Ag_2InGaS_4, $AgZn_2InS_4$, Ag_2ZnSnS_4, $AgZn_2GaS_4$) identified by the ML model, and were further validated by DFT calculations. The calculation results are shown in Figure 4. They have feature direct bandgaps, small effective mass, large exciton binding energy and substantial Stokes shifts. Weston *et al*. [94] combined DFT with ML to predict kesterite I_2-II-IV-V_4 semiconductors, achieving MSE as low as 283 meV by using multiple ML models, with support vector regression giving the best results. The trained model identified 717 potential solar absorbers (bandgap 0.5–2.5 eV) from 1,568 screened kesterite I_2-II-IV-V_4 compounds, 242 of which were in the optimal 1.2–1.8 eV range. Screening 242 compounds for stability using MP data yielded 25 promising synthesizable candidates with band gaps of 1.2–1.8 eV. Most of these have been previously explored and have potential in materials for high efficiency photovoltaic devices.

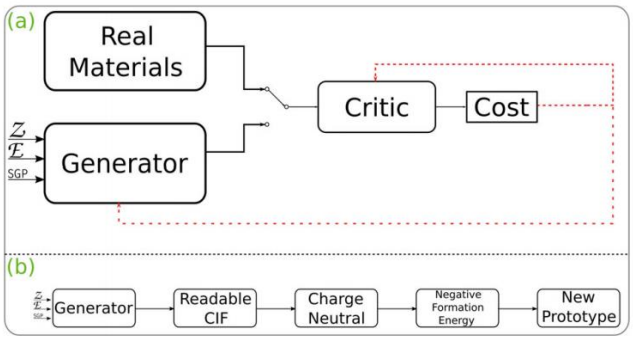

Figure 3. The main framework of CubicGAN [92]. Reprinted with permission. Copyright 2022 Nature.

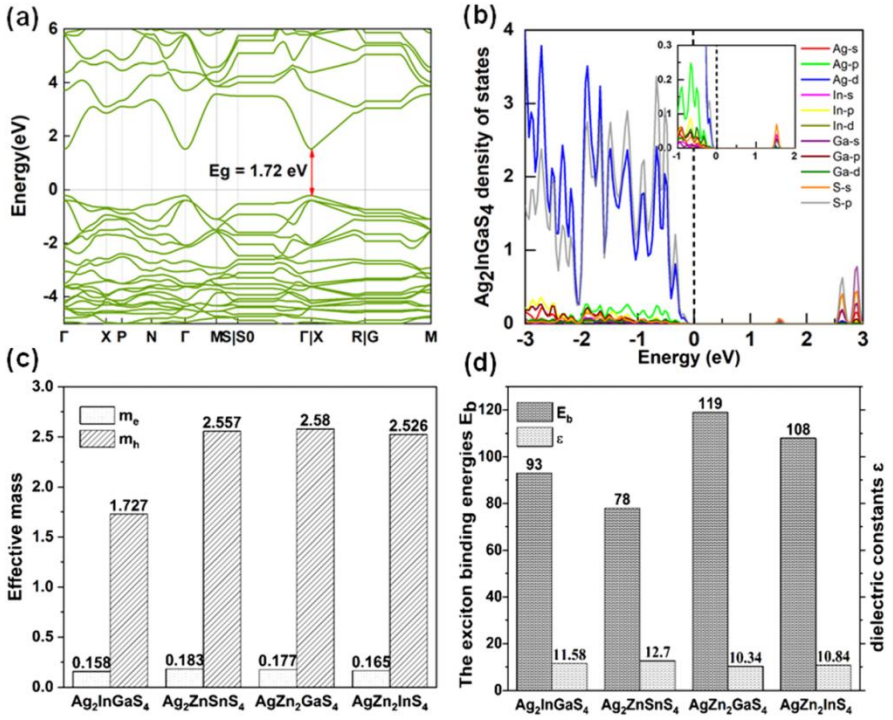

Figure 4. DFT calculation results: **(a)** Electronic band structures calculated with HSE; **(b)** State-density projection diagram; **(c)** Effective masses; **(d)** Exciton binding energies (E_b) and dielectric constants (ε) [93]. Reprinted with permission. Copyright 2023 PCCP.

Min, Gege *et al*. [95] developed a computational pipeline combining DL and DFT to generate initial boron nitride polymorphs using a stochastic strategy combining group and graph theory, and built a graph convolutional neural network (GCN) classifier for semiconductor screening and stability assessment. The proposed model attention mechanism-based orbital crystal graph convolutional neural network (A-OCGCN) classifies the metals and semiconductors during the screening process and predicts the band gap values through a regression task. Finally, 26 new stable boron nitride crystalline forms were successfully discovered in the Pc phase, and the predicted values from the A-OCGCN model were compared with the calculated values from DFT, of which 3 are direct band gap semiconductors, and 10 are quasi-direct band gap semiconductors. This model streamlines the high-throughput screening process and reduces computational costs, thereby enriching the semiconductor materials library and opening new avenues for high-performance optoelectronic devices.

Therefore, AI can not only combine with DFT to rapidly and efficiently discover and screen existing perovskite materials with stable performance, but also make outstanding contributions to discover novel semiconductor materials that are previously unknown, possess excellent properties (suitable band gap and high stability), and are environmentally reliable.

3.2. Application of AI combined with DFT on perovskite materials

Perovskite has a cubic crystal structure, which gives it some unique physical and chemical properties. The molecular formula of perovskite is ABX_3, with A and B as metals and X as an anion. Figure 5 depicts its crystal structure and elemental composition. Its optical and electrical properties can be adjusted by modifying its chemical composition to meet the needs of different applications. Perovskite

materials have attracted widespread attention from the outside world because of excellent optoelectronic properties, and have important prospects in the fields of energy, information technology, and other optoelectronic fields [96–100]. To improve efficiency, ML methods are used to accelerate the screening of stable and suitable band gap perovskites.

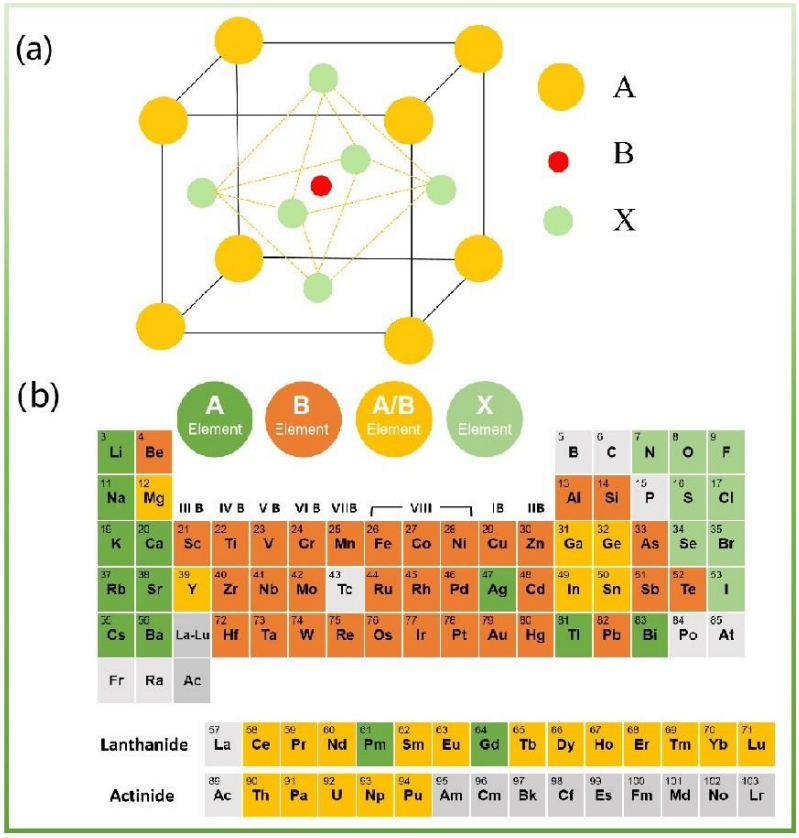

Figure 5. Structure and primary constituent elements of perovskite: **(a)** Diagram of the crystal structure of cubic ABX3 perovskite; **(b)** Elements of the periodic table that form perovskite.

Guo *et al*. [101] proposed an ML framework using different ML algorithms: RF, ridge regression, support vector machine (SVM), and XGBoost, to accelerate lead-free halide perovskite research in stability and band gap. A dataset of 540 halide double perovskites was selected for validation, and their stability and band gaps were predicted using the original DFT-computed materials dataset from Jino Im *et al*. [102]. Among the models, XGBoost achieves the highest predictive performance for thermodynamic stability (R^2: 0.9935, MAE: 0.0126), while RF performs best for band gap prediction (R^2: 0.9410, MAE: 0.1492). The importance of the selected features is analysed and directions are provided for the discovery of potential lead-free perovskite. Li *et al*. [103] proposed a strategy of combining ML with DFT to study stable halide double perovskites. Extract 354 decomposition energies of halide perovskites from DFT as the training set, and explore the mapping relationship between the stability of perovskites and their constituent ion radii based on ML models. The F_1 score for validating 246 $A_2B(I)B(III)X_6$ compounds not found in the training set was 95.9%. The performance of this model is better than the tolerance factor and description factor (F_1 score, 77.5%), and the predictions are confirmed by experimental measurements. Schmidt *et al*. [104] conducted a study on 104 all inorganic perovskite ABX3 samples and predicted the thermodynamic stability of these materials using ML models, with E_{hull} as the reference index. And constructed a DFT calculation containing approximately 250,000 cubic

perovskites, which was trained and tested on a large scale. Ridge regression, extremely randomized trees, RF, and NNs were selected to predict performance. It was found that among 230,000 test samples, after training with 20,000 samples, the extremely randomized trees provided a minimum average absolute error of 121 meV·atom^{-1} for E_{hull}. Furthermore, compounds with first-row elements show higher errors with increasing electronegativity (as shown in Figure 6). The results show that ML accelerates high-throughput DFT calculations by at least fivefold by limiting chemical composition features. Additionally, room-temperature stability remains a significant concern for perovskite devices. Zhao *et al*. [105] revealed a weak correlation between high/low-temperature stability and stability reversal by integrating ML with DFT. Cations significantly influence perovskite stability: under accelerated aging at elevated temperatures, increasing methylammonium (MA) or reducing cesium (Cs) inorganic cations reduces photo/thermal stability. However, at practical operating temperatures (< 100 °C), the situation differs. By calculating the activation energy for decomposition kinetics, the study reveals that compositions containing at least 10 mol% MA and no more than 5 mol% Cs/Rb maximize device stability at operating temperatures. Experiments confirm that such MA-containing perovskite solar cells exhibit negligible efficiency loss after 1800 hours of operation under 30 °C illumination.

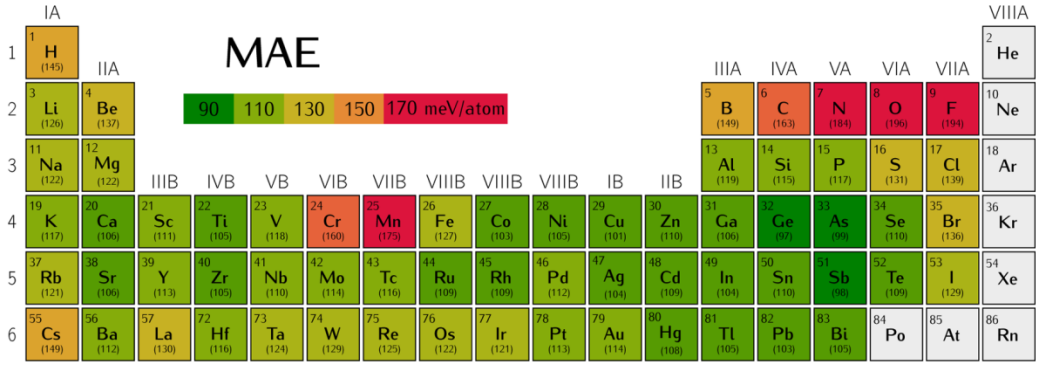

Figure 6. Test set MSE (meV/atom) of AdaBoost and extreme randomized tree models calculated based on elemental averages [104]. Reprinted with permission. Copyright 2017 ACS.

In addition, the discovery of novel lead-free perovskite materials is also an important research direction. S. Alidoust *et al*. [106] used a thorough computational screening of mixed cationic halide perovskites with the general formula AA'BX$_3$ using DFT to screen for promising lead-free candidates. Using 23 A/A'-cations, 29 B-ions, and 4 X-anions yielded 29,000 possible perovskite combinations. Based on this, Fatemeh Jamalinabijan *et al*. [107] combined with ML method can capture the complexity of the space more efficiently and the workflow is shown in Figure 7a. Following an initial screening with octahedral and tolerance factors that reduced the candidates from 29,000 to 2700 for DFT calculation, ML models were trained on this data to predict the properties of the remaining 26,300 perovskite configurations. The GBR model more accurately predicted decomposition energy and band gap (MSEs: 0.063 and 0.113), and identified 930 candidate perovskites that meet the criteria. The compositional distributions of these candidates are shown in Figure 7b. Finally, selected 20 perovskites for DFT to validate the prediction accuracy (shown in Figure 7c). These results underscore the robustness of the ML method and its promising role in discovering novel eco-friendly perovskite photovoltaics.

In summary, the integration of AI with DFT enables rapid prediction of key properties of perovskite materials, such as band gap and stability. It also help analyze the impact of temperature, composition and characteristics on performance, accelerating the discovery of novel eco-friendly lead-free perovskite materials. This is crucial for future advancements in identifying high-performance, eco-friendly perovskite materials.

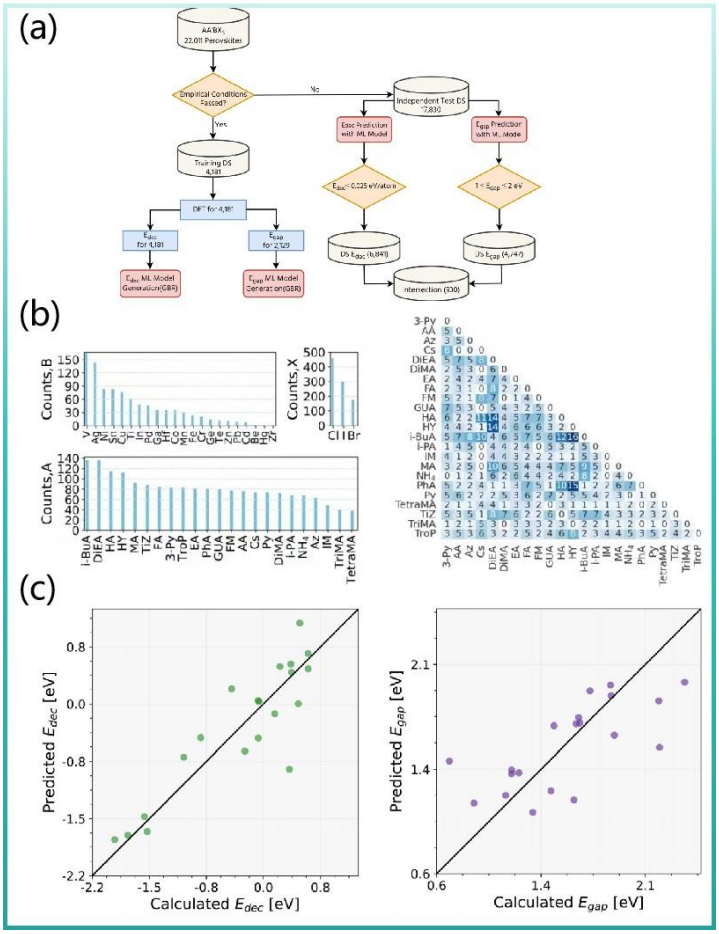

Figure 7. Primary processes and calculation results: **(a)** Data preparation and utilization workflow for ML models; **(b)** Frequency of A, B, X components (left) and A/A'-cation heat map (right) across the 930 compounds; **(c)** Predict and compare decomposition energies (left) and band gap (right) for 20 novel cubic perovskites [106,107]. Reprinted with permission. Copyright 2024 ACS; Reprinted with permission. Copyright 2025 RSC.

3.3. Application of AI combined with DFT on 2D materials

With unique photovoltaic properties, 2D materials [108–111] have become promising for applications in semiconductors and photovoltaics, and also show potential in catalytic and biomedical applications. This section reviews common ML algorithms and their applications across four domains: band gap, magnetism, catalysts, and materials discovery in Figure 8. Although existing databases have screened thousands of 2D material, the search for novel 2D materials with unexplored potentials is still ongoing and the current discovery of new 2D materials remains challenging. Traditional experimental synthesis methods rely heavily on chemical direct and accidental discoveries, and require a long time. These questions have sparked the enthusiasm of researchers and promoted the application of ML in 2D materials. The

penetration of ML in the domain of 2D materials is becoming increasingly deep, and Figure 9 summarizes the application work of different teams in 2D materials.

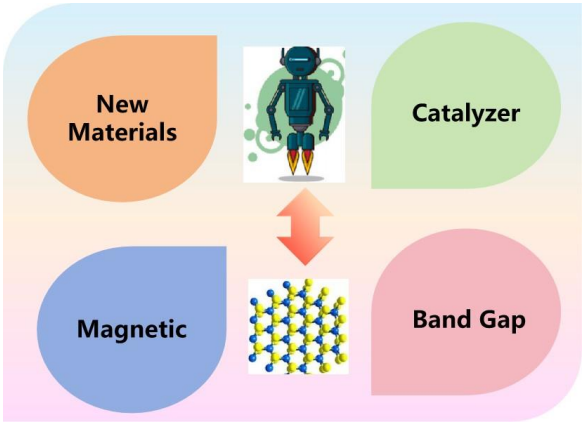

Figure 8. Application of ML in 2D materials.

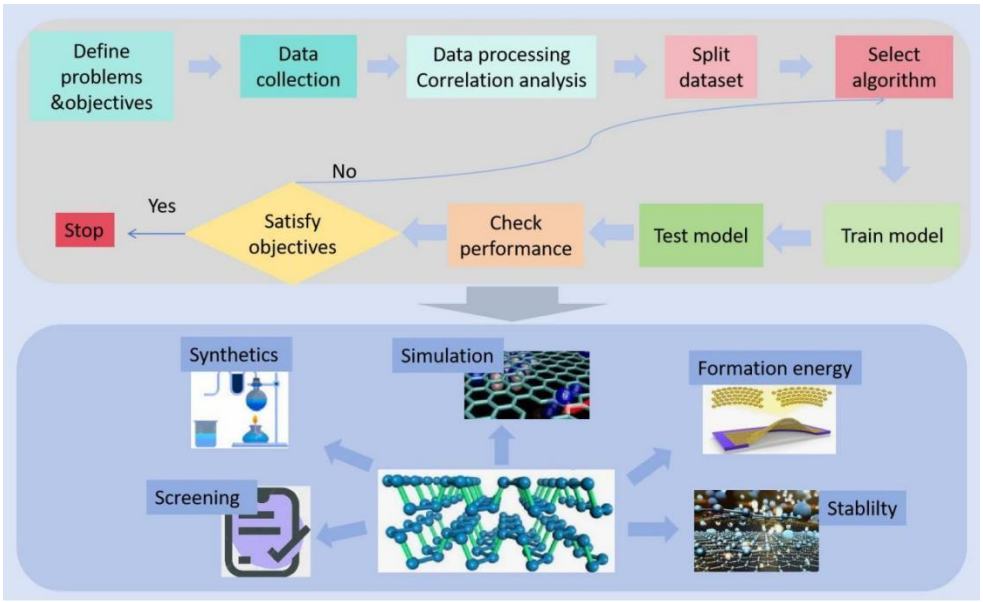

Figure 9. ML processes and applications in the synthesis, screening, identification, and discovery of 2D Materials.

To expedite 2D material development, Song *et al*. [112] developed a generative DL method for exploring new materials in uncharted compositional spaces. And integrated with a RF–based classifier for 2D materials to identify novel candidates, the performance is shown in Figure 10a. The modified classifier was used to screen 2.65 million generated samples using MatGAN, a GAN-based model [113] capable of generating chemically plausible hypothetical materials, and its architecture is shown in Figure 10b. In addition, crystal structures were predicted using elemental substitution and then their structural stability was confirmed using DFT. So far, 267,489 new potential 2D material compositions have been identified (Figure 10c depicts the distribution of both novel and existing 2D materials), of which 1485 have probability scores greater than 0.95, 101 crystal structures have also been predicted by DFT formation energy calculations, and 92 2D/layered materials have been confirmed. Rio *et al*. [114] proposed a DL framework for predicting the electronic structure of materials and molecules based on DFT by training a neural network to predict the electronic density of states (DOS) of the total electrons of

various graphene-derived isomers, including different types of carbon nanotubes, fullerene molecules, graphene and graphite. The R^2 value for the test set is 0.996. It predicts the DOS with exceptional speed and chemical accuracy by systematically learning the input-output mapping of the Kohn-Sham equation (Standard deviation: 0.15 eV). This algorithm significantly accelerates the prediction of DOS and charge density, enabling a high-fidelity, ultra-fast DFT simulator.

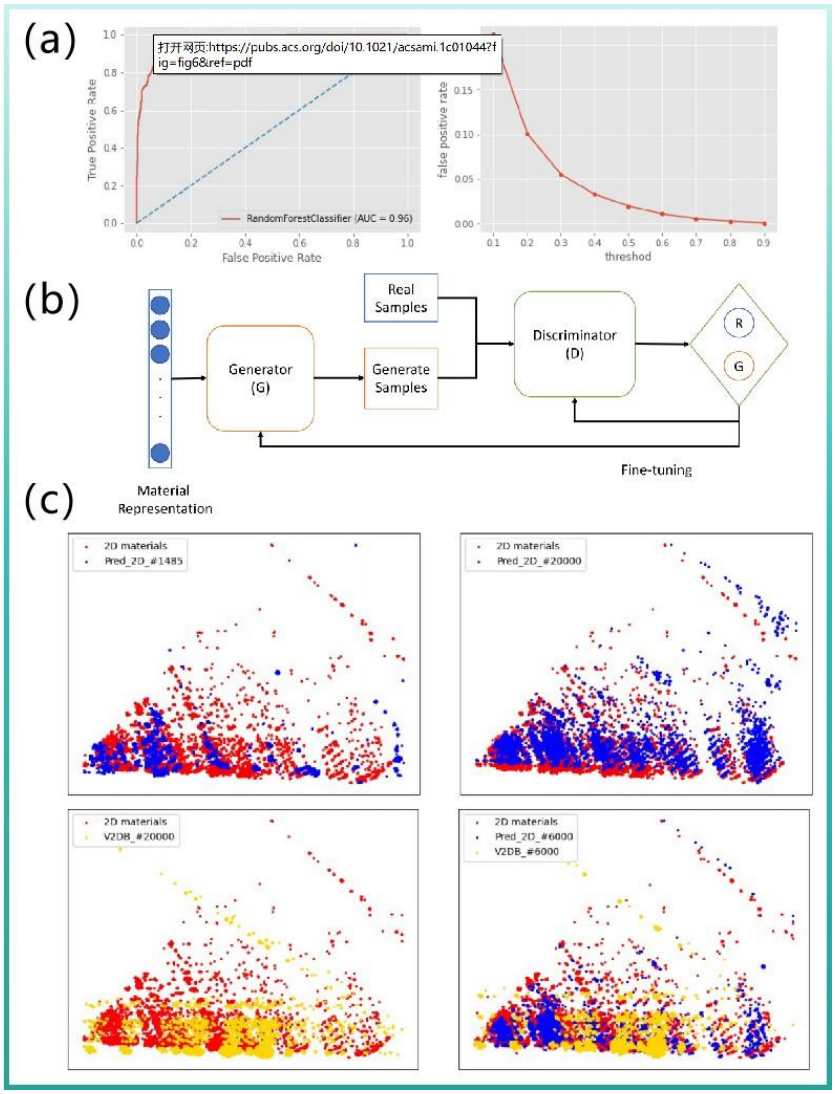

Figure 10. AI-driven approaches for 2D materials discovery and visualization: **(a)** Performance of our RF random forest classifier; **(b)** Architecture of MatGAN; **(c)** T-SNE visualization of new and existing 2D materials based on Magpie features [112]. Reprinted with permission. Copyright 2021 ACS.

As a fundamental material property, the band gap governs electronic structure and optical properties. The band gap of 2D materials is used in semiconductor devices, photodetectors, field-effect transistors, heterostructures, *etc*., as shown in Figure 11. Therefore, predicting the band gap is crucial, and ML models are widely applicable for screening materials with target functions. Zhu's team [115] cleverly combined DFT and ML to explore the fundamental band gap of 2D isoelectronic phosphazene semiconductors and the alignment. Calculations were first performed using the Perdew-Burke-Ernzerhof exchange-correlation generalization and the Heyd-Scuseria-Ernzerhof hybrid generalization (HSE) in the gradient density approximation (GGA-PBE) as a reference. The band gaps, valence band maxima (VBMs) and

conduction band minima (CBMs) are calculated for different material types in Figure 12a, where the blue region indicates phase I materials, while the green region indicates phase II materials. It was found that such materials have a similar crystalline structure, but band gap values roughly distributed 0–8 eV. Then three ML methods—linear regression (LR), RFR, and support vector regression (SVR)—were employed to predict the electronic properties. The model performance is shown in Figure 12b. Among them, SVR delivered the best performance (RMSE < 0.15 eV) on band gap, VBM, and CBM predictions using PBE results and elemental information.

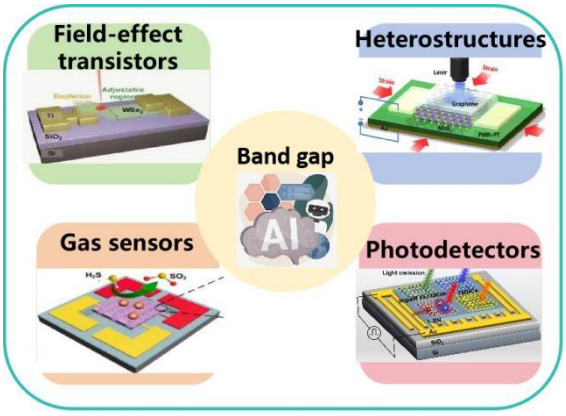

Figure 11. Application scenarios of 2D material band gap.

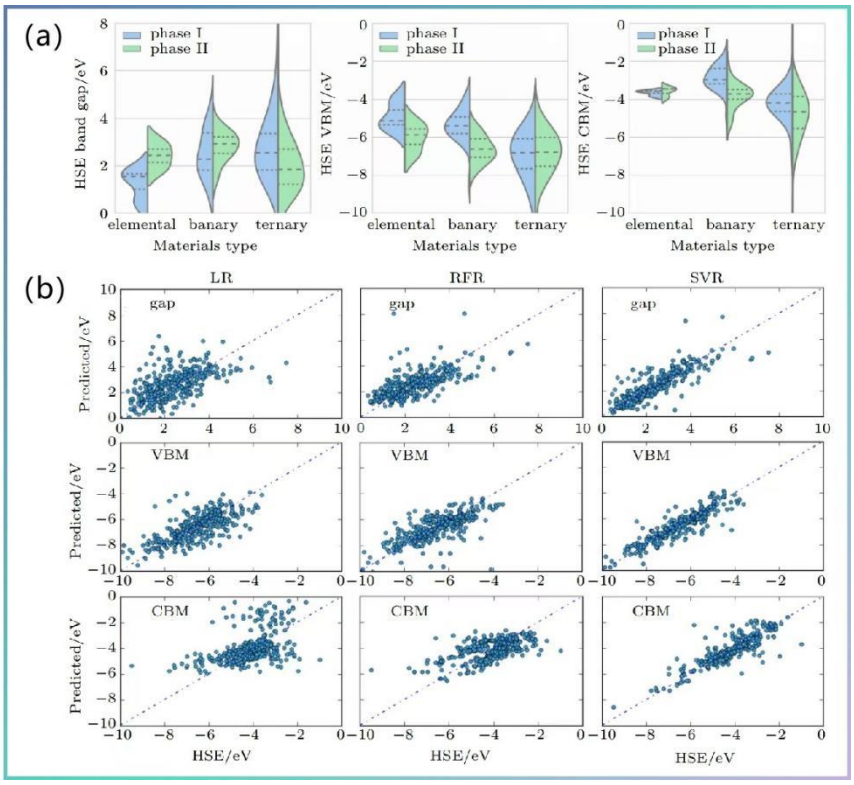

Figure 12. Band structure calculations and predictive model evaluation: **(a)** Calculation of band gap, VBMs and CBMs for different material types; **(b)** The performance of model LR, RFR and SVR [115]. Reprinted with permission. Copyright 2020 IOP.

In addition, 2D materials are also a good catalytic material, where graphene-based materials have become a key component in the study of electrocatalytic reactions [116]. ML has demonstrated tremendous potential in discovering high-performance catalytic materials, with various research groups

employing different ML algorithms to accelerate discovery of catalytic materials. In addition, highly active catalyst screening can be synergistically enhanced by integrating DFT [117]. Deng *et al*. [118] investigated the structure-property correlation and the origin of catalytic activity of biatomic catalysts (BACs) using DFT simulations combined with ML techniques. ML techniques were used to identify the origin of the oxidation-reduction reaction (ORR) activity on BACs. RFR identified a strong correlation between the DFT-calculated limiting potential (UL) and the input features (pearson correlation coefficient: 0.98; MSE for the test set and training set: 0.004 and 0.022, respectively), linking them to catalytic activity. This work not only identifies promising BACs that have not been explored experimentally, but also offers valuable guidance for designing new and efficient ORR catalysts. Lin *et al*. [119] developed ML models based on sacDFT calculations of 104 graphene supports to describe readily accessible physical properties and potential modes of the limiting potentials (the MSE for the ORR, oxygen precipitation reaction (OER), hydrogen precipitation reaction (HER) is 0.027/0.021/0.035 respectively), and applied the models to evaluate the catalytic performance of 260 alternative graphene supports. The ORR/OER/HER of the top ML-recommended catalysts were recalculated by DFT, which confirmed the model's reliability and identified two OER catalysts superior to the noble metal oxides RuO_2 and IrO_2.

In summary, the integration of AI and DFT plays a crucial role in accelerating the development cycle of 2D materials, discovering new 2D materials, predicting their electronic structures and band gaps, and identifying high-performance catalytic materials. It provides researchers with essential theoretical guidance for advancing the field of 2D materials.

4. Summary and outlook

This paper reviews the application of AI techniques combined with DFT in the research of new materials, structural stabilisation and performance prediction. Some basic ML and DL algorithms are briefly outlined with their advantages and disadvantages. Then, the research progress of ML and DL algorithms for semiconductor materials, perovskite materials and 2D materials respectively in recent years is presented. Although ML has been widely used in property prediction, and quantum chemistry owing to its predictive power and computational efficiency. However, the application of materials in ML still face many challenges: Heavy reliance on data: High-quality material-related data is scarce, making it difficult to perfectly characterize material properties, resulting in prediction accuracy lower than DFT calculations. Lack of strict physical constraints: May produce results violating physical laws and struggles to provide interpretable chemical insights, thus currently serving primarily as a powerful auxiliary tool. Insufficient reliability: Most models are only applicable to specific environments and problems. They perform well within the scope of trained knowledge (interpolation) but exhibit severe prediction inaccuracies when encountering new, unknown chemical structures (extrapolation), lacking true physical reasoning capabilities. Future research will focus on developing smarter, more reliable AI models. This includes leveraging techniques like active learning to build more efficient models with less data, integrating physical principles to enhance extrapolation capabilities and ensure predictions comply with physical laws, and designing explainable AI architectures to open the "black box". This approach aims to enable models not only to make predictions but also to provide meaningful chemical insights.

Although AI as an important tool to accelerate the process of materials will not completely replace human expertise and traditional computational methods, materials researchers can master this technology to solve more problems in materials. Next, we can improve the existing problems and strive

to promote the formation of the "component-structure-property-application" chain as soon as possible to accelerate materials research.

Acknowledgments

The authors acknowledge the support of the National Key R&D Program of China (No. 2024YFB3817300), the National Natural Science Foundations of China (Grant Nos. 52250005, 21875271, U20B2021), the support of the Key R & D Projects of Zhejiang Province (No. 2022C01236, 2019C01060), the National Key Laboratory of Nuclear Reactor Technology (Grant No. STRFML-2023-06), Nuclear Power Institute of China, the Entrepreneurship Program of Foshan National Hi-tech Industrial Development Zone, the Major Project of the Ministry of Science and Technology of China (Grant No. 2015ZX06004-001), Natural Science Foundation of Shanghai Science and Technology Commission (25ZR1401353), the Fundamental Research Funds for the Central Universities (22120250339), the Fundamental Research Funds for the Central Universities (22120250374), Peak Disciplines (Type IV) of Institutions of Higher Learning in Shanghai, Ningbo Natural Science Foundations (Grant Nos. 2014A610006, 2016A610273, and 2019A610106).

Authors' contribution

Conceptualization, Shiyu Du and Xiemeng Zhu; validation, Juhong Yu and Yong Liu; investigation, Xiemeng Zhu; resources, Juhong Yu and Liang Zhang; data curation, Yong Liu and Yangyang Song; writing—original draft preparation, Xiemeng Zhu; writing—review and editing, Shiyu Du and Xiemeng Zhu; project administration, Yangyang Song and Liang Zhang; funding acquisition, Shiyu Du. All authors have read and agreed to the published version of the manuscript.

Conflicts of interests

Shiyu Du holds the position of Editor-in-chief for *AI & Materials* and has not peer reviewed or made any editorial decisions for this paper.

References

[1] Mittemeijer EJ. *Fundamentals of Materials Science*, 1st ed. Berlin: Springer, 2010.
[2] Callister WD, Rethwisch DG. *Materials Science and Engineering: An Introduction*, 10th ed. Hoboken: John Wiley & Sons, 2020.
[3] Deringer VL, Caro MA, Csányi G. Machine learning interatomic potentials as emerging tools for materials science. *Adv. Mater.* 2019, 31(46):1902765.
[4] Wolf D, Yamakov V, Phillpot SR, Mukherjee A, Gleiter H. Deformation of nanocrystalline materials by molecular-dynamics simulation: relationship to experiments? *Acta Mater*. 2005, 53(1):1–40.
[5] Jagota V, Sethi APS, Kumar K. Finite element method: an overview. *Walailak J. Sci. Technol.* 2013, 10(1):1–8.
[6] Orio M, Pantazis DA, Neese F. Density functional theory. *Photosynth. Res.* 2009, 102(2):443–453.
[7] Makkar P, Ghosh NN. A review on the use of DFT for the prediction of the properties of nanomaterials. *RSC Adv.* 2021, 11(45):27897–27924.

[8] Xiao H, Tahir-Kheli J, Goddard III WA. Accurate band gaps for semiconductors from density functional theory. *J. Phys. Chem. Lett.* 2011, 2(3):212–217.

[9] Xu T, Liu M, Wu K, Liu C. Density functional theory calculations to increase the efficiency of oxygen electrode catalysts from ytterbium single atom catalysts using nitrogen solid supports. ACS Appl. *Nano Mater.* 2024, 7(13):15526–15534.

[10] Pang Y, Ding Z, Ma A, Fan G, Xu H. Electroreduction of nitrate to ammonia on graphyne-based single-atom catalysts by combined density functional theory and machine learning study. *Sep. Purif. Technol.* 2025, 354:129422.

[11] Landers J, Gor GY, Neimark AV. Density functional theory methods for characterization of porous materials. *Colloids Surf., A* 2013, 437:3–32.

[12] Jain A, Shin Y, Persson KA. Computational predictions of energy materials using density functional theory. *Nat. Rev. Mater.* 2016, 1(1):1–13.

[13] Maurer RJ, Freysoldt C, Reilly AM, Brandenburg JG, Hofmann OT, *et al*. Advances in density-functional calculations for materials modeling. *Annu. Rev. Mater. Res.* 2019, 49:1–30.

[14] Kim C, Huan TD, Krishnan S, Ramprasad R. A hybrid organic-inorganic perovskite dataset. *Sci. Data* 2017, 4(1):1–11.

[15] Mounet N, Gibertini M, Schwaller P, Campi D, Merkys A, *et al*. Two-dimensional materials from high-throughput computational exfoliation of experimentally known compounds. *Nat. Nanotechnol.* 2018, 13(3):246–252.

[16] Hilbert M. Big data for development: a review of promises and challenges. *Dev. Policy Rev.* 2016 34(1):135–174.

[17] Zhang X, Pérez-Stable EJ, Bourne PE, Peprah E, Duru OK, *et al*. Big data science: opportunities and challenges to address minority health and health disparities in the 21st century. *Ethn. Dis.* 2017, 27(2):95.

[18] Reed DA, Dongarra J. Exascale computing and big data. *Commun. ACM* 2015, 58(7):56–68.

[19] Pollice R, Gomes GP, Aldeghi M, Hickman RJ, Krenn M, *et al*. Data-driven strategies for accelerated materials design. *Acc. Chem. Res.* 2021, 54(4):849–860.

[20] Hu Z, Huang C, Xie L, Hua L, Yuan Y, *et al*. Machine learning assisted quality control in metal additive manufacturing: a review. *Adv. Powder Mater.* 2025:100342.

[21] Ye Y, Li R, Qu B, Wang H, Liu Y, *et al*. Machine learning for energy band prediction of halide perovskites. *Mater. Futures* 2025, 4(3):035601.

[22] Gou F, Ma Z, Yang Q, Du H, Li Y, *et al.* Machine learning-assisted prediction and control of bandgap for organic–inorganic metal halide perovskites. *ACS Appl. Mater. Interfaces* 2025, 17(12):18383–18393.

[23] Groumpos PP. Artificial intelligence: issues, challenges, opportunities and threats. In *Conference on Creativity in Intelligent Technologies and Data Science*, Volgograd, Russia, September 16–19, 2019, pp. 19–33.

[24] Wang Y, Widrow B, Zadeh LA, Howard N, Wood S, *et al.* Cognitive intelligence: deep learning, thinking, and reasoning by brain-inspired systems. *Int. J. Cogn. Inform. Nat. Intell.* 2016, 10(4):1–20.

[25] Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. *Nat. Rev. Genet.* 2015, 16(6):321–332.

[26] Fan W, Bifet A. Mining big data: current status, and forecast to the future. *ACM SIGKDD Explor. Newsl.* 2013, 14(2):1–5.
[27] Zhang L, Tan J, Han D, Zhu H. From machine learning to deep learning: progress in machine intelligence for rational drug discovery. *Drug Discovery Today* 2017, 22(11):1680–1685.
[28] Zhu H. Big data and artificial intelligence modeling for drug discovery. *Annu. Rev. Pharmacol. Toxicol.* 2020, 60(1):573–589.
[29] Li B, Hou B, Yu W, Lu X, Yang C, *et al.* Applications of artificial intelligence in intelligent manufacturing: a review. *Front. Inf. Technol. Electron. Eng.* 2017, 18(1):86–96.
[30] Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. *Science* 2015, 349(6245):255–260.
[31] Kohavi R, Provost F. Special issue on applications of machine learning and the knowledge discovery process. *Mach. Learn.* 1998, 30: 271–274.
[32] Silver D, Huang A, Maddison CJ, Guez A, Sifre L, *et al.* Mastering the game of Go with deep neural networks and tree search. *Nature* 2016, 529(7587):484–489.
[33] Cambria E, White B. Jumping NLP curves: a review of natural language processing research. *IEEE Comput. Intell. Mag.* 2014, 9(2):48–57.
[34] Tsai C, Lai C, Chiang M, Yang L. Data mining for internet of things: a survey. *IEEE Commun. Surv. Tutor.* 2013, 16(1):77–97.
[35] Kononenko I. Machine learning for medical diagnosis: history, state of the art and perspective. *Artif. Intell. Med.* 2001, 23(1):89–109.
[36] Fan W, Bifet A. Mining big data: current status, and forecast to the future. *ACM SIGKDD Explor. Newsl.* 2013, 14(2):1–5.
[37] Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A. Machine learning for molecular and materials science. *Nature* 2018, 559(7715):547–555.
[38] Wang M, Wang T, Cai P, Chen X. Nanomaterials discovery and design through machine learning. *Small Methods* 2019, 3(5):1900025.
[39] Nash W, Drummond T, Birbilis N. A review of deep learning in the study of materials degradation. *npj Mater. Degrad.* 2018, 2(1):37.
[40] LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015, 521(7553):436–444.
[41] Zhang Q, Yang LT, Chen Z, Li P. A survey on deep learning for big data. *Inf. Fusion* 2018, 42:146–157.
[42] Chen X, Lin X. Big data deep learning: challenges and perspectives. *IEEE Access* 2014, 2:514–525.
[43] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, Las Vegas, USA, June 27–30, 2016, pp. 770–778.
[44] Lee S, Ha J, Zokhirova M, Moon H, Lee J. Background information of deep learning for structural engineering. *Arch. Comput. Methods Eng.* 2018, 25(1):121–129.
[45] Azimi SM, Britz D, Engstler M, Fritz M, Mücklich F. Advanced steel microstructural classification by deep learning methods. *Sci. Rep.* 2018, 8(1):2128.
[46] Lusci A, Pollastri G, Baldi P. Deep architectures and deep learning in chemoinformatics: the prediction of aqueous solubility for drug-like molecules. *J. Chem. Inf. Model.* 2013, 53(7):1563–1575.
[47] Abdallah ZS, Du L, Webb GI. *Data Preparation*, 2nd ed. Boston: Springer, 2017. pp. 318–327.

[48] Brownlee J. *Data Preparation for Machine Learning: Data Cleaning, Feature Selection, and Data Transforms in Python*, 1st ed. London: Machine Learning Mastery, 2020.

[49] García S, Ramírez-Gallego S, Luengo J, Benítez JM, Herrera F. Big data preprocessing: methods and prospects. *Big Data Anal.* 2016, 1(1):9.

[50] Cabello-Solorzano K, Ortigosa AI, Peña M, Correia L, Tallón-Ballesteros AJ. The impact of data normalization on the accuracy of machine learning algorithms: a comparative analysis. In *International Conference on Soft Computing Models in Industrial and Environmental Applications*, Salamanca, Spain, September 5–7, 2023, pp. 344–353.

[51] Ramprasad R, Batra R, Pilania G, Mannodi-Kanakkithodi A, Kim C. Machine learning in materials informatics: recent applications and prospects. *npj Comput. Mater.* 2017, 3(1):54.

[52] Chen C, Zuo Y, Ye W, Li X, Deng Z, *et al.* A critical review of machine learning of energy materials. *Adv. Energy Mater.* 2020, 10(8):1903242.

[53] Braham EJ, Cho J, Forlano KM, Watson DF, Arròyave R, *et al.* Machine learning-directed navigation of synthetic design space: a statistical learning approach to controlling the synthesis of perovskite halide nanoplatelets in the quantum-confined regime. *Chem. Mater.* 2019, 31(9):3281–3292.

[54] Mohammed M, Khan MB, Bashier EBM. *Machine Learning: Algorithms and Applications*, 1st ed. Boca Raton: CRC Press, 2016.

[55] Ang JC, Mirzal A, Haron H, Hamed HNA. Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 2015, 13(5):971–989.

[56] Naidu G, Zuva T, Sibanda EM. A review of evaluation metrics in machine learning algorithms. *In Computer Science On-line Conference*, April 3–5, 2023, pp. 15–25.

[57] Rainio O, Teuho J, Klén R. Evaluation metrics and statistical tests for machine learning. *Sci. Rep.* 2024, 14(1):6086.

[58] Raschka S. Model evaluation, model selection, and algorithm selection in machine learning. *arXiv* 2018, arXiv:1811.12808.

[59] Browne MW. Cross-validation methods. *J. Math. Psychol.* 2000, 44(1):108–132.

[60] Accuracy S. Bootstrap methods for standard errors, confidence lntervals, and other measures of statistical accuracy. *Stat. Sci.* 1986, 1(1):54–77.

[61] Sun S, Huang R. An adaptive k-nearest neighbor algorithm. In *2010 Seventh International Conference On Fuzzy Systems and Knowledge Discovery*, Yantai, China, August 10–12, 2010, pp. 91–94.

[62] Taunk K, De S, Verma S, Swetapadma A. A brief review of nearest neighbor algorithm for learning and classification. In *2019 International Conference on Intelligent Computing and Control Systems*, Madurai, India, May 15–17, 2019, pp. 1255–1260.

[63] Kotsiantis SB. Decision trees: a recent overview. *Artif. Intell. Rev.* 2013, 39(4):261–283.

[64] Charbuty B, Abdulazeez A. Classification based on decision tree algorithm for machine learning. *J. Appl. Sci. Technol. Trends* 2021, 2(01):20–28.

[65] Rokach L, Maimon O. Decision trees, In *Data Mining and Knowledge Discovery Handbook*, 1st ed. New York: Springer, 2005. pp. 165–192.

[66] Rigatti SJ. Random forest. *J. Insur. Med.* 2017, 47(1):31–39.

[67] Zekić-Sušac M, Has A, Knežević M. Predicting energy cost of public buildings by artificial neural networks, CART, and random forest. *Neurocomputing* 2021, 439:223–233.

[68] Huynh-Thu VA, Geurts P. Unsupervised gene network inference with decision trees and random forests. In *Methods in Molecular Biology*, 1st ed. New York: Humana Press, 2019. pp. 195–215.

[69] Miikkulainen R, Liang J, Meyerson E, Rawal A, Fink D, *et al.* Evolving deep neural networks. In *Artificial Intelligence in the Age of Neural Networks and Brain Computing*, 2nd ed. New York: Academic Press, 2024. pp. 269–287.

[70] Yamazaki K, Vo-Ho VK, Bulsara D, Le N. Spiking neural networks and their applications: a review. *Brain Sci.* 2022, 12(7):863.

[71] Khemani B, Patil S, Kotecha K, Tanwar S. A review of graph neural networks: concepts, architectures, techniques, challenges, datasets, applications, and future directions. *J. Big Data* 2024, 11(1):18.

[72] Zhao Q, Shang Z. Deep learning and its development. In *2021 2nd International Conference on Internet of Things, Artificial Intelligence and Mechanical Automation*, Hangzhou, China, May 14–16, 2021, pp. 012023.

[73] Gu J, Wang Z, Kuen J, Ma L, Shahroudy A, *et al.* Recent advances in convolutional neural networks. *Pattern Recognit.* 2018, 77:354–377.

[74] Zhang W, Tanida J, Itoh K, Ichioka Y. Shift-invariant pattern recognition neural network and its optical architecture. In *Proceedings of annual conference of the Japan Society of Applied Physics*, Japan, October 4–7, 1988, pp. 564.

[75] Zhao X, Wang L, Zhang Y, Han X, Deveci M, *et al.* A review of convolutional neural networks in computer vision. *Artif. Intell. Rev.* 2024, 57(4):99.

[76] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, *et al.* Generative adversarial networks. *Commun. ACM* 2020, 63(11):139–144.

[77] Gui J, Sun Z, Wen Y, Tao D, Ye J. A review on generative adversarial networks: algorithms, theory, and applications. *IEEE Trans. Knowl. Data Eng.* 2021, 35(4):3313–3332.

[78] Alajaji SA, Khoury ZH, Elgharib M, Saeed M, Ahmed ARH, *et al.* Generative adversarial networks in digital histopathology: current applications, limitations, ethical considerations, and future directions. *Mod. Pathol.* 2024, 37(1):100369.

[79] Sabnam S, Rajagopal S. Application of generative adversarial networks in image, face reconstruction and medical imaging: challenges and the current progress. *Comput. Methods Biomech. Biomed. Eng. Imaging Vis.* 2024, 12(1):2330524.

[80] Behler J, Csányi G. Machine learning potentials for extended systems: a perspective. *Eur. Phys. J. B* 2021, 94(7):142.

[81] Harshvardhan GM, Gourisaria MK, Pandey M, Rautaray SS. A comprehensive survey and analysis of generative models in machine learning. *Comput. Sci. Rev.* 2020, 38:100285.

[82] Shahriari B, Swersky K, Wang Z, Adams RP, Freitas ND. Taking the human out of the loop: a review of Bayesian optimization. *Proc. IEEE* 2015, 104(1):148–175.

[83] Yin L, Cheng R, Ding J, Jiang J, Hou Y, *et al.* Two-dimensional semiconductors and transistors for future integrated circuits. *ACS Nano* 2024, 18(11):7739–7768.

[84] Chuang RW, Wu RX, Lai LW, Lee CT. Zno-on-Gan heterojunction lightemitting diode grown by vapor cooling condensation technique. *Appl. Phys. Lett.* 2007, 91(23):231113.

[85] Park JH, Kim DY, Schubert EF, Cho J, Kim JK. Fundamental limitations of wide-bandgap semiconductors for light-emitting diodes. *ACS Energy Lett.* 2018, 3(3):655–662.

[86] Yu L, Zubair A, Santos EJG, Zhang X, Lin Y, *et al.* High-performance WSe2 complementary metal oxide semiconductor technology and integrated circuits. *Nano Lett.* 2015, 15(8):4928–4934.

[87] Green MA, Bremner SP. Energy conversion approaches and materials for high-efficiency photovoltaics. *Nat. Mater.* 2017, 16(1):23–34.

[88] Lin Y, Li X, Xie D, Feng T, Chen Y, *et al.* Graphene/semiconductor heterojunction solar cells with modulated antireflection and graphene work function. *Energy Environ. Sci.* 2013, 6(1):108–115.

[89] Bertrandie J, Han J, De Castro CSP, Yengel E, Gorenflot J, *et al.* The energy level conundrum of organic semiconductors in solar cells. *Adv. Mater.* 2022, 34(35):2202575.

[90] Zhang D, Wang B, Gan Y, Zhou J, Sun Z. New layered II–III–VI semiconductors: promising materials for high performance solar cells. *J. Phys. Chem. C* 2024, 128(4):1582–1590.

[91] Casady JB, Johnson RW. Status of silicon carbide (SiC) as a wide-bandgap semiconductor for high-temperature applications: a review. *Solid-State Electron.* 1996, 39(10):1409–1422.

[92] Siriwardane EMD, Zhao Y, Perera I, Hu J. Generative design of stable semiconductor materials using deep learning and density functional theory. *npj Comput. Mater.* 2022, 8(1):164.

[93] Gao M, Cai B, Liu G, Xu L, Zhang S, *et al.* Machine learning and density functional theory simulation of the electronic structural properties for novel quaternary semiconductors. *Phys. Chem. Chem. Phys.* 2023, 25(13):9123–9130.

[94] Weston L, Stampfl C. Machine learning the band gap properties of kesterite I_2-II-IV-V_4 quaternary compounds for photovoltaics applications. *arXiv* 2017, arXiv:1708.08530.

[95] Min G, Wei W, Fan Q, Wan T, Ye M, *et al.* High-throughput exploration of stable semiconductors using deep learning and density functional theory. *Mater. Sci. Semicond. Process.* 2025, 188:109150.

[96] Bailie CD, Christoforo MG, Mailoa JP, Bowring AR, Unger EL, *et al.* Semi-transparent perovskite solar cells for tandems with silicon and CIGS. *Energy Environ. Sci.* 2015, 8(3):956–963.

[97] Tan ZK, Moghaddam RS, Lai ML, Docampo P, Higler R, *et al.* Bright light-emitting diodes based on organometal halide perovskite. *Nat. Nanotechnol.* 2014, 9(9):687–692.

[98] Van Le Q, Jang HW, Kim SY. Recent advances toward high-efficiency halide perovskite light-emitting diodes: review and perspective. *Small Methods* 2018, 2(10):1700419.

[99] Deschler F, Price M, Pathak S, Klintberg LE, Jarausch DD, *et al.* High photoluminescence efficiency and optically pumped lasing in solution-processed mixed halide perovskite semiconductors. *J. Phys. Chem. Lett.* 2014, 5(8):1421–1426.

[100] Ahmadi M, Wu T, Hu B. A review on organic–inorganic halide perovskite photodetectors: device engineering and fundamental physics. *Adv. Mater.* 2017, 29(41):1605242.

[101] Guo Z, Lin B. Machine learning stability and band gap of lead-free halide double perovskite materials for perovskite solar cells. *Sol. Energy* 2021, 228:689–699.

[102] Im J, Lee S, Ko TW, Kim HW, Hyon YK, *et al.* Identifying Pb-free perovskites for solar cells by machine learning. *npj Comput. Mater.* 2019, 5(1):37.

[103] Li Z, Xu Q, Sun Q, Hou Z, Yin W. Thermodynamic stability landscape of halide double perovskites via high-throughput computing and machine learning. *Adv. Funct. Mater.* 2019, 29(9):1807280.

[104] Schmidt J, Shi J, Borlido P, Chen L, Botti S, *et al.* Predicting the thermodynamic stability of solids combining density functional theory and machine learning. *Chem. Mater.* 2017, 29(12):5090–5103.

[105] Zhao Y, Zhang J, Xu Z, Sun S, Langner S, *et al.* Discovery of temperature-induced stability reversal in perovskites using high-throughput robotic learning. *Nat. Commun.* 2021, 12(1):2191.

[106] Alidoust S, Jamalinabijan F, Tekin A. Light-harvesting lead-free mixed cation hybrid halide perovskites: a density functional theory-based computational screening study. *ACS Appl. Energy Mater.* 2024, 7(2):785–798.

[107] Jamalinabijan F, Alidoust S, Demir Gİ, Tekin A. Discovering novel lead-free mixed cation hybrid halide perovskites via machine learning. *Phys. Chem. Chem. Phys.* 2025, 27(14):7389–7398.

[108] Li J, Wang Y, Song H, Guo Y, Hu S, *et al.* Photocatalytic hydrogen under visible light by nitrogen-doped rutile titania graphitic carbon nitride composites: an experimental and theoretical study. *Adv. Compos. Hybrid Mater.* 2023, 6(2):83.

[109] Ahn EC. 2D materials for spintronic devices. *npj 2D Mater. Appl.* 2020, 4(1):17.

[110] Zhang X, Chen A, Zhou Z. High-throughput computational screening of layered and two-dimensional materials. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* 2019, 9(1):e1385.

[111] He H, Wang Y, Qi Y, Xu Z, Li Y, *et al.* From prediction to design: recent advances in machine learning for the study of 2D materials. *Nano Energy* 2023, 118:108965.

[112] Song Y, Siriwardane EMD, Zhao Y, Hu J. Computational discovery of new 2D materials using deep learning generative models. *ACS Appl. Mater. Interfaces* 2021, 13(45):53303–53313.

[113] Dan Y, Zhao Y, Li X, Li S, Hu M, *et al.* Generative adversarial networks (GAN) based efficient sampling of chemical composition space for inverse design of inorganic materials. *npj Comput. Mater.* 2020, 6(1):84.

[114] Del Rio BG, Kuenneth C, Tran HD, Ramprasad R. An efficient deep learning scheme to predict the electronic structure of materials and molecules: the example of graphene-derived allotropes. *J. Phys. Chem. A* 2020, 124(45):9496–9502.

[115] Zhu Z, Dong B, Guo H, Yang T, Zhang Z. Fundamental band gap and alignment of two-dimensional semiconductors explored by machine learning. *Chin. Phys. B* 2020, 29(4):046101.

[116] Akinwande D, Huyghebaert C, Wang CH, Serna MI, Goossens S, *et al.* Graphene and two-dimensional materials for silicon technology. *Nature* 2019, 573(7775):507–518.

[117] Ghosal S, Chowdhury S, Jana D. Impressive thermoelectric figure of merit in two-dimensional tetragonal pnictogens: a combined first-principles and machine-learning approach. *ACS Appl. Mater. Interfaces* 2021, 13(49):59092–59103.

[118] Deng C, Su Y, Li F, Shen W, Chen Z, *et al.* Understanding activity origin for the oxygen reduction reaction on bi-atom catalysts by DFT studies and machine-learning. *J. Mater. Chem. A* 2020, 8(46):24563–24571.

[119] Lin S, Xu H, Wang Y, Zeng X, Chen Z. Directly predicting limiting potentials from easily obtainable physical properties of graphene-supported single-atom electrocatalysts by machine learning. *J. Mater. Chem. A* 2020, 8(11):5663–5670.