mater.scichina.com link.springer.com

Published online 19 March 2024 | https://doi.org/10.1007/s40843-024-2851-9
Sci China Mater 2024, 67(4): 1042–1081

SPECIAL TOPIC: Computation-assisted Materials Screening and Design

Methods and applications of machine learning in computational design of optoelectronic semiconductors

Xiaoyu Yang, Kun Zhou, Xin He* and Lijun Zhang*

ABSTRACT **The development of high-throughput computation and materials databases has laid the foundation for the emergence of data-driven machine learning methods in recent years. Machine learning has become a crucial methodology propelling researches in computational materials. It has demonstrated tremendous potential in analyzing materials data, expediting materials calculations, predicting material properties, advancing the discovery, screening, and design of new materials. Consequently, an increasing number of methodologies, models, and frameworks of machine learning have emerged. This review provides a comprehensive overview of the latest advancements and applications of machine learning in computational design of optoelectronic semiconductors. We introduce the workflow and strategies of machine learning shallow models, ensemble models, and deep neural networks based on various material representation methods. The associated material databases and toolkits are also discussed. Furthermore, we delve into the applications of these models in predicting material stability, optoelectronic properties, materials inverse design, and establishing relationships between material structures and properties. Finally, we summarize and discuss the key challenges existing in current machine learning, with a specific focus on issues related to the size of available data, data quality, material representation, and materials inverse design.**

Keywords: machine learning, computational materials, optoelectronic semiconductor materials

INTRODUCTION

In 2022, the release of chat generative pre-trained transformer (ChatGPT) [1] gained widespread attention and acclaim. Its remarkable capabilities in composing articles, modifying code, and translating languages swiftly captured the imagination of the public, heralding the advent of the artificial intelligence (AI) era. This pivotal moment underscored the visionary concept of "data-intensive scientific discovery", posited by Jim Gray of the Microsoft Research Institute as early as 2007 [2]. Depicted in Fig. 1a, this paradigm shift augments the traditional trio of scientific research methodologies—experimentation, theory, and computational simulation—by accentuating the pivotal role of big data and advanced analytics. In this paradigm, researchers leverage a myriad of tools, including machine learning (ML) and data mining, to amass, organize, and dissect vast datasets, thereby unearthing novel insights and knowledge [3,4]. This novel approach to scientific inquiry has quietly revolutionized our lives. Noteworthy initiatives, such as the Materials Genome Initiative (MGI) initiated by the USA government in 2011, have epitomized this transformation by harnessing cutting-edge computational, experimental, and data science technologies to propel materials innovation [5,6]. The founding of the OpenAI team in 2015 marked the inception of groundbreaking research in AI models, culminating in the launch of ChatGPT in 2022. Subsequently, milestones in AI, such as the development of AlphaGo by the DeepMind team in 2016, have pushed the boundaries of machine cognition, as exemplified by its victory over the Korean Go champion, Lee [7]. Moreover, the introduction of AlphaFold by DeepMind in 2020 addressed a long-standing conundrum in biology—deciphering the intricate process of protein folding [8]. Notably, recent endeavors by both the Google DeepMind and Microsoft teams have led to the development of material generative models, namely GNoME [9], and MatterGen [10] , facilitating inverse material design. In essence, the past decade has witnessed an exponential surge in data-driven scientific research methodologies, propelling us towards unprecedented frontiers of knowledge and innovation.

The integration of AI into materials science has yielded profound advancements, evident in the accelerated pace of research and discovery of novel materials boasting advanced performance and diverse applications, courtesy of data-driven ML methodologies [11,12]. Prior to this paradigm shift, traditional computational simulation methods served as the primary approach in materials research. Offering cost-effective alternatives to experimental studies, these methods furnish crucial direction and guidance to experimental endeavors. Given the intricate relationship between material properties and their chemical compositions and intrinsic structures, the construction of material models incorporating information on chemical elements and atomic positions has enabled the prediction of a myriad of material properties. These encompass mechanical, thermal, optical, electrical, and magnetic attributes, attainable through theoretical calculations [13–16]. Computational simulation methodologies for materials encompass a spectrum of approaches, including but not limited to, first-principles methods such as density functional theory (DFT). These methods

State Key Laboratory of Integrated Optoelectronics, Key Laboratory of Automobile Materials of MOE, Key Laboratory of Material Simulation Methods & Software of MOE, and School of Materials Science and Engineering, Jilin University, Changchun 130012, China
* Corresponding authors (emails: xin_he@jlu.edu.cn (He X); lijun_zhang@jlu.edu.cn (Zhang L))

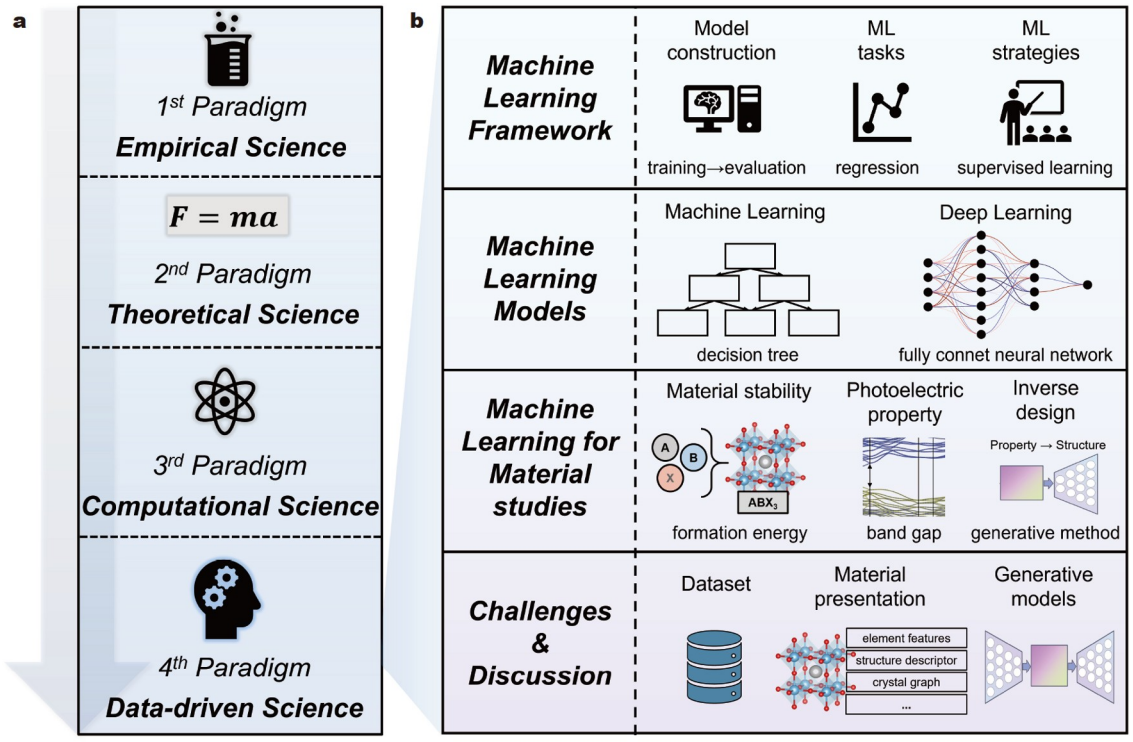

Figure 1 (a) Data-driven science has become the fourth paradigm of scientific research. (b) ML methods and applications. It is also the framework of this review. The content includes the basic framework and processes of ML, popular ML models, and specific applications of ML in predicting material stability and optoelectronic properties. Finally, we discuss some key challenges of ML. Some more detailed examples are given on the right.

simulate the electronic structure and properties of materials by solving the Schrödinger equation [14,17]. Molecular dynamics methods, on the other hand, elucidate the motion and interactions of atoms and molecules within materials by solving classical mechanics equations. Leveraged to study material structure, dynamic behavior, and thermodynamic properties, these methods play a pivotal role in investigating phenomena such as phase transitions, mechanical properties, and heat conduction. [18]. Presently, the practice of predicting material structures and properties using computational methodologies prior to experimental investigations has become ubiquitous in materials research. This approach not only facilitates the planning and optimization of material synthesis processes, encompassing considerations of chemical composition, material structure, and requisite synthesis routes, but also aids in the analysis and interpretation of potential relationships between material structures and properties. Consequently, it furnishes invaluable insights for the advancement of functional materials, chemistry, and biology, thereby guiding the refinement and development of theoretical methodologies.

In recent years, propelled by advancements in computer hardware and software, researchers have leveraged batch, automated calculations to analyze an extensive array of materials swiftly—a research paradigm termed high-throughput computing for materials [16,19,20]. This approach harnesses parallel computing and automation technologies to facilitate efficient and expedited large-scale calculations and data processing. The significance of high-throughput computing in materials science and engineering is multifaceted, encompassing the acceleration of materials discovery and design, reduction of time and costs, provision of expansive datasets for in-depth analysis, and the advancement of materials simulation and theoretical research [21]. The advent of high-throughput computing has led to the emergence of several materials databases, including material project (MP) [22], the open quantum materials database (OQMD) [23], and automatic-FLOW for materials discovery (AFLOW) [24], which harbor detailed structural information and diverse material properties. These repositories serve as invaluable resources for researchers seeking comprehensive datasets to inform their investigations and analyses in the realm of materials science and engineering.

The intertwined advancement of high-throughput computing and materials databases has propelled the development of data-driven methodologies in materials simulation [25–31]. Data-driven approaches in materials research encompass the utilization of techniques such as data mining and ML to aggregate, structure, and analyze extensive datasets derived from experimental, computational, or literature sources. These methodologies are geared towards uncovering latent patterns within materials, swiftly predicting material properties, and expediting the discovery, design, and screening of novel materials. Analogous to how computational simulation enhances the efficacy of material experiments, data-driven ML methodologies have markedly accelerated the pace of materials computational simulation by several orders of magnitude. In the realm of materials research, ML finds diverse applications, encompassing but not limited to the following areas: (1) utilizing interpretable models to evaluate material descriptors and establish correlations between material composition, structure, and properties [32,33], (2) employing deep neural networks to develop atomic potential functions that rival the accuracy of DFT for accelerated material structure optimization or MD simulation [34–36],

(3) leveraging supervised learning models for rapid and cost-effective property prediction (e.g., energy, bandgap, defects, phonon, optical, and elastic properties) to facilitate the screening of functional materials [37–46], (4) utilizing generative models to generate chemical compositions, molecules, and crystal structures to facilitate materials inverse design [47–50], and (5) applying unsupervised natural language processing (NLP) for mining literature texts to extract high-quality professional knowledge and data [51–53].

In this review, as shown in Fig. 1b, we provide an in-depth exploration and analysis of the contemporary landscape of ML in computational materials, following a structured progression from fundamental to sophisticated dimensions. Our emphasis is placed on the following pivotal elements: the methodologies and techniques employed in ML, the repertoire of models and tools utilized, the utilization of ML for property prediction in materials, and its applications in materials inverse design. A special focus is directed towards optoelectronic semiconductor materials, notably the widely studied metal halide materials. In the end of the review, we further discuss the key challenges in current ML: augmenting the quantity and quality of training data, refining the precision of crystal representations to augment the model's learning capabilities, and effectuating material inverse design. It is pertinent to acknowledge that our review does not aim to encompass all facets of the field comprehensively. Given the expansive domain of materials science, our discussions are delimited to the purview of computational materials. It is noteworthy that our primary focus lies in the realm of applying ML methodologies to optoelectronic semiconductor materials and solid crystals. This targeted approach allows us to dedicate our efforts to a more thorough analysis of the pertinent subject matter.

FRAMEWORK OF ML BASED STUDIES

The general workflow and methods of ML is depicted in Fig. 2a. It starts from acquiring a dataset and concludes with model evaluation. We categorize ML into four types: classification, regression, clustering, and dimensionality reduction, based on different tasks. We also classify ML into categories such as supervised learning and unsupervised learning according to different learning strategies, accompanied by relevant case studies.

General workflow

Preparing a dataset

The dataset serves as the starting point for data-driven ML methods and is the most crucial link in the ML process. It comprises material information (chemical composition and crystal structure) and the properties or performance of materials that serve as the learning objectives for the model. All the knowledge that an ML model acquires comes from the dataset, hence determining the upper limit of the model's performance [54]. The data size must be sufficient to encompass hidden relationships between features adequately. However, an excessive amount of data can slow down the model training process and pose challenges for parameter tuning. Methods for obtaining material data include manual extraction from literature [55], NLP-based automated literature mining [52], high-throughput computation [56], and extraction from material databases [57]. Organizing literature or conducting high-throughput computation can yield customized and accurate data, but the dataset is often limited in size. Extracting data directly from databases allows for quickly obtaining a large amount of data, but some specific materials may not be present in existing databases.

Data evaluation and cleaning

After obtaining the dataset, the first step is to assess the data quality to determine if it is suitable for constructing an ML model. Evaluation methods include checking if the samples are representative, identifying outliers or erroneous samples, ensuring that sample labels are obtained under consistent parameters, and examining the balanced distribution of label values (close to a normal distribution) [58,59]. Subsequently, data cleaning is performed to eliminate noise, errors, and inconsistencies in the data to ensure its quality and reliability. For missing values in the data, one can choose to delete samples or features with missing values or use imputation methods (such as mean, median, and regression) to fill in the missing values. Normalization and standardization of the data can be applied to eliminate differences in scale between different features, ensuring that the data are comparable and analyzable on the same scale.

Feature engineering

Feature engineering, also known as descriptor design, is a closely integrated stage in the ML process with domain-specific knowledge of materials. Its goal is to transform raw data into a feature representation that ML models can understand and process. Feature engineering includes feature extraction, feature selection, and feature construction. Feature extraction aims to more accurately represent samples as numerical values that a computer can understand, such as extracting information about crystal elements, atomic positions, atomic interactions, and local structures, to help the model understand material characteristics [11,60–62]. Feature selection involves removing irrelevant or redundant features unrelated to the learning objective, reducing samples from a high-dimensional space to a low-dimensional space, allowing the model to focus on features most closely related to material properties and thus improving model performance [63,64]. Feature construction involves combining original features to obtain composite features that are more closely related to the target [65]. In actual ML tasks, one may need to continuously fit ML models to evaluate the effectiveness of the current feature set and iterate through the three tasks in feature engineering.

Selecting ML methods

As shown in Fig. 2b, ML tasks include classification, regression, dimensionality reduction, and clustering. ML strategies include supervised learning, unsupervised learning, semi-supervised learning, reinforcement learning, active learning, and transfer learning. ML models can be categorized based on model complexity into shallow models, ensemble models, and deep models. Different ML tasks, strategies, and models are suitable for addressing completely different problems. We will elaborate further on this in subsequent chapters.

Dataset split

Before model training, it is necessary to split the dataset into training, validation, and test sets. During the model training process, the training set is used for model fitting, and the vali-

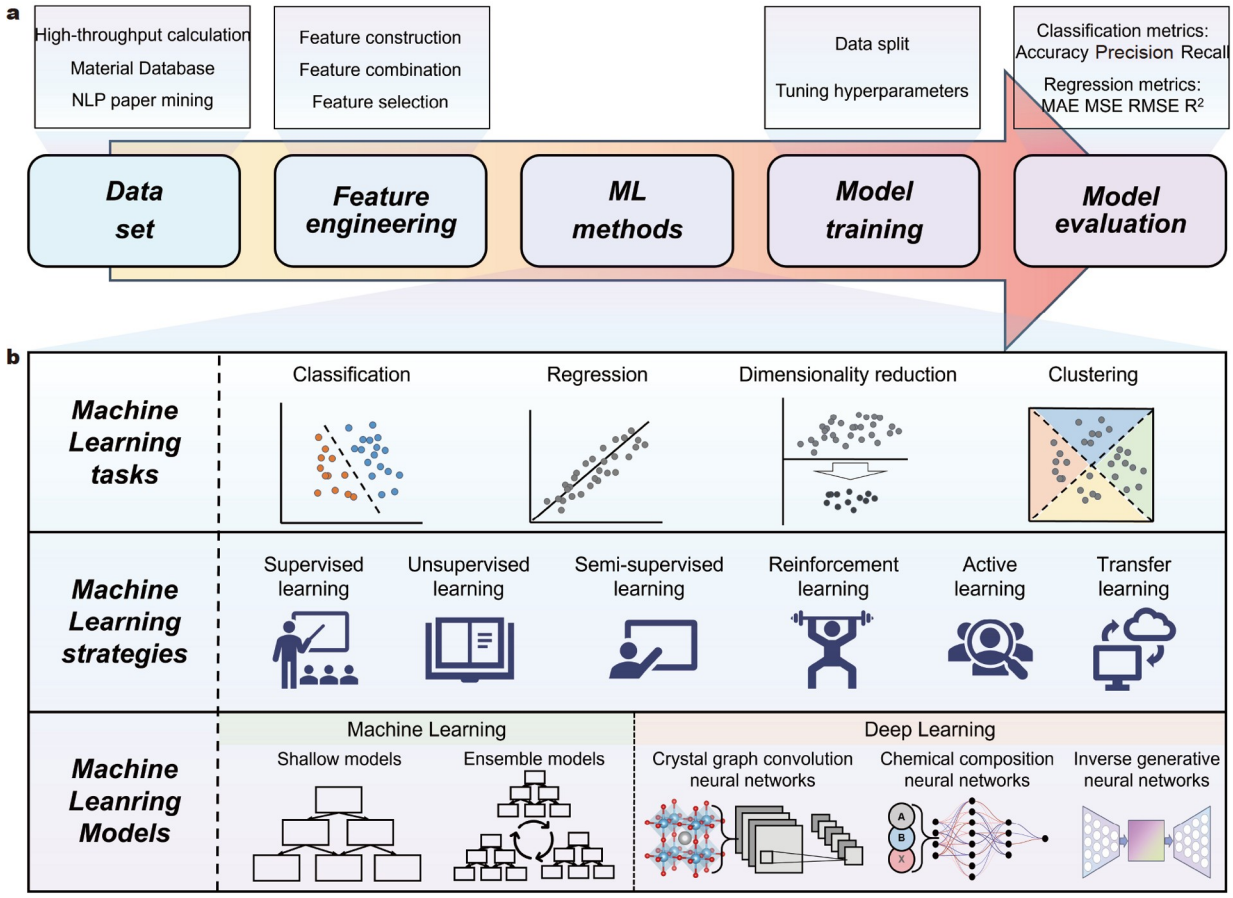

Figure 2 (a) General workflow of ML. It starts from dataset and goes through feature engineering, method selection, and model training, and ends with model evaluation. (b) Classification of ML methods in this review. ML tasks are divided into four categories: classification, regression, dimensionality reduction, and clustering. ML strategies are categorized as supervised learning, unsupervised learning, semi-supervised learning, reinforcement learning, active learning, and transfer learning. ML models are divided into ML and deep learning. ML includes shallow models and ensemble models. Deep learning consists of neural networks based on crystal graphs, neural networks based on chemical compositions, and generative neural networks for inverse materials design.

dation set is used for adjust model hyperparameters. We compare the model's performance on the training and validation sets to assess the model's learning progress, avoiding underfitting and overfitting. After training is completed, the test set is used to evaluate the model's performance on unseen data [66].

Model training

Model training involves letting the model learn from the training set data, adjusting model parameters and structure. The model adjusts its internal parameters (such as node weights in neural network models) by learning on the training set and then evaluating on the validation set. We adjust hyperparameters (such as decision tree maximum depth for decision tree models, the number of hidden layers for neural network models) based on the model's performance on the test set [67–69].

Model evaluation

After model training is complete, it is necessary to evaluate the model based on the test set [70,71]. For supervised models, we measure model performance by comparing predicted values with sample labels. For classification models, evaluation metrics include accuracy, precision, recall, F1 Score, receiver operating characteristic (ROC) curve and area under curve (AUC), confusion matrix. For regression models, commonly used metrics for evaluation are root mean squared error (RMSE), mean absolute error (MAE), mean squared error (MSE), R^2 score, etc. Since the model's performance depends on its performance on the test set, the choice of the test set is particularly crucial. The test set should be entirely different from the training and validation sets, i.e., it should not contain duplicate samples, to ensure the accuracy of the evaluation results. Otherwise, the model may encounter samples during the test that it learned during the training process, creating a false impression of excellent model performance. Additionally, the test set should not be too small, and the sample selection should be entirely random to ensure the non-accidental nature of the model's evaluation results [72–74].

Specific tasks

Specific tasks of ML studies include classification, regression, dimensionality reduction, and clustering. In classification tasks, the model learns the mapping relationship from input features to category labels, enabling the classification prediction of new samples. Classification tasks in ML include binary classification and multiclass classification [75–77]. Binary classification tasks can be used for yes/no decisions, such as an ML classification model determining the stability of an unknown crystal. Multiclass models can predict which category a new sample belongs to, such as a human face recognition model.

Regression tasks are most common, where each sample label is

a continuous value, such as the adsorption energy of two-dimensional (2D) materials, the bandgap of semiconductors, and the formation energy (E_f) of crystals [78,79].

In clustering tasks, the model learns the distance or similarity of different samples, partitioning the sample points in the dataset into different groups, ensuring high similarity among samples within the same group to achieve the goal of classifying samples [80]. Clustering tasks belong to a type of unsupervised learning, where the training dataset's samples do not require predefined labels. Clustering models can be used to discover potential correlations among samples for data analysis and preprocessing. Common clustering models include K-means clustering, hierarchical clustering, and density-based clustering.

Dimensionality reduction aims to map high-dimensional data to a low-dimensional space, reducing the number and complexity of features while preserving key information from the original data. It helps visualize data in two or 3D feature spaces for sample analysis and classification, reduces data complexity to lower model training costs, and addresses the curse of dimensionality [81,82]. Common dimensionality reduction methods include principal component analysis (PCA) [65], linear discriminant analysis (LDA) [83], and t-distributed stochastic neighbor embedding (t-SNE) [68].

State-of-the-art strategies

Supervised learning

In supervised learning, each sample in the dataset consists of features and their corresponding target values (labels). Features are used to describe the information of the sample, while the target variable represents the sample's property. During training, the model learns the relationship between the features and target values in the training set to acquire predictive capabilities. However, supervised learning heavily relies on labeled samples, making it labor-intensive to label samples before constructing a supervised learning model. This challenge particularly affects the construction of supervised learning models in data-scarce domains.

Classification and regression models based on supervised learning are the most common models in materials science. Typically, various features represent the structure and composition of materials, and the properties or performance of materials (such as concrete strength, semiconductor bandgap, defect formation energy of 2D materials, stability of perovskite materials, and thermal expansion coefficient of high-entropy alloys) serve as their target values. Through supervised learning, the model learns the latent relationship between materials and their properties, enabling the rapid prediction of relevant properties for candidates. This accelerates the discovery of new materials. In tasks like predicting the bandgap of perovskites using supervised learning, a well-trained model can quickly predict the bandgaps of new materials, achieving computational speeds several orders of magnitude faster than traditional DFT calculations [33,56,58,63,77,84–92].

Interpretable models can also reveal specific relationships between material composition, structure, and properties or performance. For example, in a supervised learning task related to perovskite crystal stability, analyzing the feature importance output by the ML model can highlight that the octahedral factor has a significant impact on the stability of perovskites [33,77,84–86,88–90,92]. Examples of supervised learning work will be further discussed in subsequent chapters.

Unsupervised learning

In unsupervised learning, each sample in the dataset is represented only by features without corresponding target values. Therefore, unsupervised learning can be used to discover the intrinsic structure or relationships within unlabeled data when obtaining sample labels is challenging. Since there are no target values to guide the model's learning direction, the construction of effective unsupervised learning models requires algorithms to autonomously discover features and patterns in the data, and the accuracy of the model also heavily depends on the knowledge and experience of ML engineers for evaluation [93].

Clustering and dimensionality reduction based on unsupervised learning can be employed to categorize samples, reduce data complexity, and visualize feature spaces, for instance, using algorithms like hierachical density-based spatial clustering of applications with noise (HDBSCAN) and t-SNE to cluster layered homologous groups of 2D flat-band materials for discovering material templates [47]. Xie and Grossman [94] utilized PCA to reduce the dimensionality of metallic elements in perovskite materials, projecting sample points into a 2D visualization space to identify sample distributions. They also employed t-SNE for clustering analysis on a boron dataset, oxygen, and sulfur element coordination environments. Additionally, as shown in Fig. 3a, Bhattacharya *et al.* [46] used t-SNE to visualize the feature space of various graph neural network (GNN) models to compare their learning capabilities regarding crystal structures and chemical compositions.

NLP methods based on unsupervised learning are widely used for literature text mining. This is done to collect data from literature and automatically analyze potential relationships between knowledge by constructing knowledge graphs (KGs) [95,96]. Zhang and He [97] utilized word embeddings to automatically extract information from a material literature database, constructing an unsupervised ML model that successfully established implicit relationships between material chemical formulas and their photovoltaic applications. Zheng *et al.* [51] guided ChatGPT through engineering to perform text mining and extract 26k different synthetic parameters from 800 peer-reviewed articles on MOFs. This dataset was then used for subsequent ML tasks. The ChemDataExtractor software toolkit employs NLP and ML methods to extract chemical data from scientific literature. Huang and Cole [98] used it for data mining in 229k academic papers, resulting in a battery materials database containing 292k entries. Dong and Cole [99] utilized this software to extract 100k semiconductor bandgap records from 128k journal articles, automatically generating a database.

The variational autoencoder (VAE) model is a type of ML model based on unsupervised learning. Vasylenko *et al.* [100] employed an unsupervised learning VAE model to evaluate the synthesizability of unexplored chemical compositions. As shown in Fig. 3b, they utilized the anomaly detection function of the VAE model to rank the new phases formed by random combinations of elements by their reasonability. The reconstruction error of the VAE represents the deviation of the candidate phase space from the chemical systems in the training set, making it suitable for the assessment of synthesizability of new phases.

Generative models based on unsupervised learning are also important models in materials science [49]. Common generative models are generally based on generative adversarial network

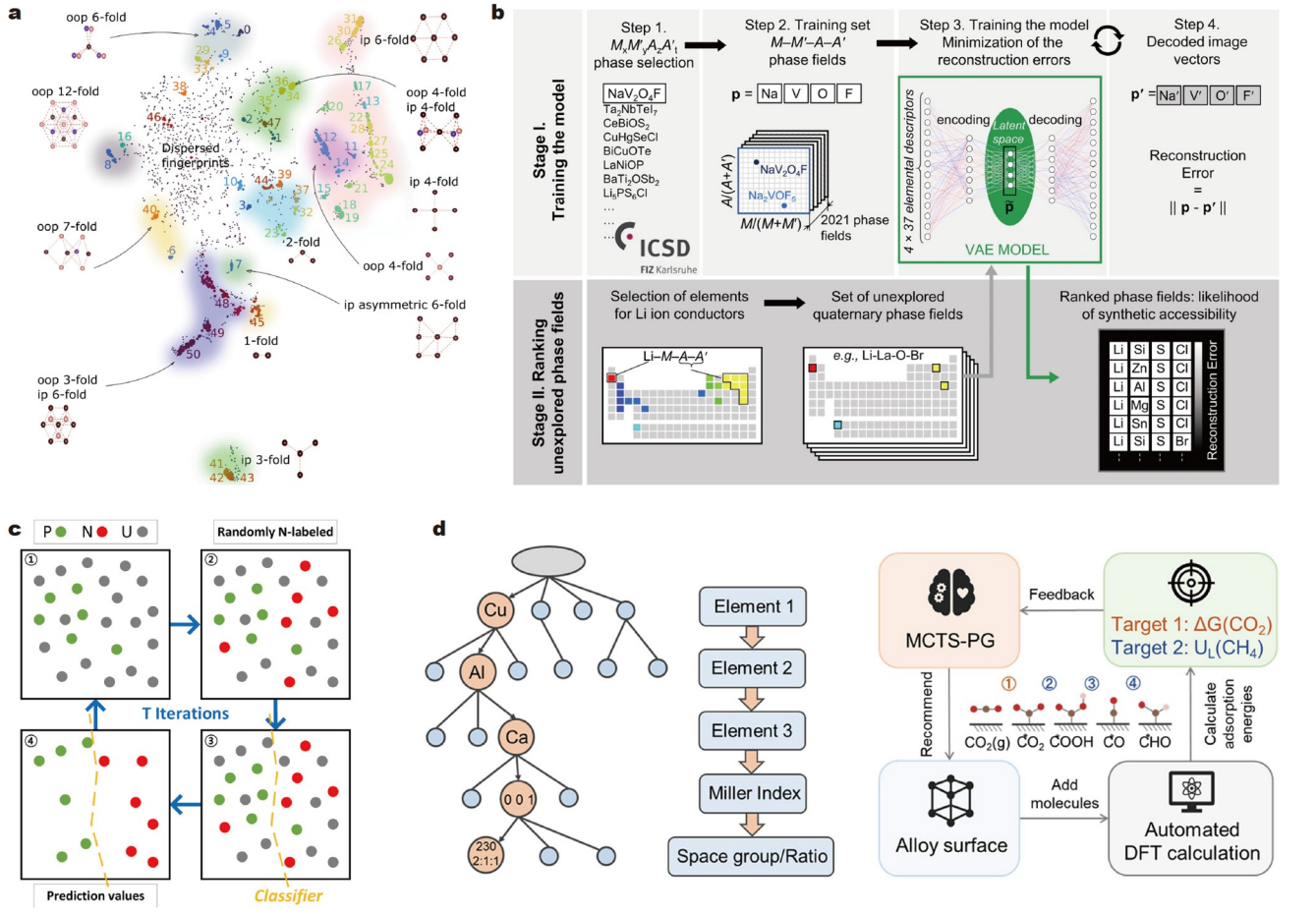

Figure 3 (a) t-SNE 2D visualization of the structure fingerprint space, with different coordination patterns color-coded. Reprinted with permission from Ref. [46]. Copyright 2023, the Author(s). (b) Workflow of the VAE model. Reprinted with permission from Ref. [100]. Copyright 2021, the Author(s). (c) Schematic diagram of PU learning. Green, red, and gray circles express positive, negative, and unlabeled data, respectively. Reprinted with permission from Ref. [107]. Copyright 2020, American Chemical Society. (d) Five-layer tree search structure of MCTS-PG. The light red nodes represent the 001 surface of Cu_2AlCa with the symmetry of a 230 space group. And the general framework of MCTS-PG combined with automated DFT in optimizing the alloy surface for CO_2 activation (target 1) and methanation (target 2). Reprinted with permission from Ref. [121]. Copyright 2023, American Chemical Society.

(GAN) or VAE. They have been widely used for generating entirely new chemical formulas, molecular structures, and crystal structures [50,100–113]. The methods and applications of generative models will be further detailed in subsequent chapters.

Semi-supervised learning

Semi-supervised learning is a learning paradigm that falls between supervised learning and unsupervised learning. In semi-supervised learning, only a small portion of the training dataset is labeled, while most samples are unlabeled. When labeling samples is costly, semi-supervised learning can make efficient use of both labeled and unlabeled samples for training. Common methods include self-training and co-training [114]. Because unlabeled samples lack labels, evaluating the distribution and quality of these samples is challenging. Ensuring that unlabeled samples do not contain erroneous information and belong to the same distribution as labeled samples is one of the difficulties in unsupervised learning.

Positive-unlabeled learning (PU learning) is a research direction in semi-supervised learning that involves training a binary classifier in situations where only positive class and unlabeled data are available. In practical scenarios, obtaining negative class samples can be challenging, and negative class data may be highly diverse and dynamically changing. For instance, in the problem of predicting the synthesizability of materials, one may only have data on synthesized materials (positive samples) and unexplored materials (unlabeled samples), without information on materials that cannot be synthesized (negative samples) because it is difficult to know in advance which crystals cannot be synthesized. In such cases, PU learning can effectively utilize these two types of data for model training, thus gaining the ability to classify samples [115,116]. The core idea behind training PU models is iterative, as shown in Fig. 3c. First, a portion of unlabeled samples is selected, assuming them to be negative samples. Then, a binary classification model is trained using labeled positive examples and assumed negative examples. The trained model is used to predict the labels of unlabeled samples. Based on the prediction results, labels of unlabeled samples are adjusted (samples predicted as positive are labeled as positive, while samples predicted as negative remain unlabeled). Iteratively repeating these steps allows the model to gradually learn discriminative features between positive and negative examples, thus completing the construction of the binary clas-

sification model. Jang *et al.* [107] and Gu *et al.* [117] used PU learning methods to predict the synthesizability of crystal materials and perovskite materials, respectively.

Reinforcement learning

Reinforcement learning differs significantly from other types of learning, primarily used to solve sequential decision-making problems. In reinforcement learning, an agent executes specific actions in an environment and receives rewards or feedback from the environment. Subsequently, the environment transitions to a new state. The goal of the agent is to learn a policy through interaction with the environment, such that it selects optimal actions in different states to maximize rewards, achieving the learning objective [118]. Reinforcement learning does not rely on labeled training sets but learns through interaction with the environment. It is a general decision-making framework enabling computers to improve themselves autonomously, holding the potential to achieve general AI a key factor in the success of the Go-playing robot AlphaGo [7].

Monte Carlo tree search (MCTS) can be seen as a specific form of reinforcement learning. It evaluates the value of possible decisions and actions through random simulation and statistical sampling. The search process is guided by statistical information rather than learning and policy updates through interaction with the environment [119,120]. Banik *et al.* [43] found a connection between the intermediate configurations of materials involved in the defect design problem in low-dimensional materials and the concept of "delayed rewards" in reinforcement learning. They constructed a reinforcement learning model based on MCTS with delayed rewards, effectively used for exploring the defect configuration space, determining the optimal arrangement, and evolution of defects in 2D MoS_2 materials. Song *et al.* [121] used the MCTS algorithm combined with policy gradient (PG), forming the MCTS-PG algorithm. It was integrated with DFT calculations to create an iterative search for the optimal material, forming a versatile adaptive reinforcement learning framework. As shown in Fig. 3d, they employed this framework to search for materials with desired properties from initial data. After iterative searches, they successfully filtered out 100 alloy surfaces capable of chemically adsorbing CO_2 and predicted 9 alloy surfaces with high CO_2 methanation activity.

Active learning

Active learning algorithms are a strategy to minimize data collection costs. It actively selects the most valuable unlabeled samples based on limited labeled data through learning, guiding us in labeling or querying [122,123]. As shown in Fig. 4a, the active learning process generally includes model training, sample exploration, sample labeling, and model retraining. This iterative process is repeated to collect sufficient data to improve the model's performance. Common sample selection strategies in active learning include uncertainty sampling, margin sampling, information gain, diversity sampling, model-based uncertainty sampling, and others. Choosing an appropriate sample selection strategy can help improve sample labeling efficiency.

Active learning is used to explore unknown material spaces and discover valuable data [124]. Kim and Min [58], after constructing an ML model to predict the formation energy and bandgap of double halide perovskite materials, further used an active learning strategy to select data from the database to train the model and improve prediction accuracy. They chose the exploration method as a sampling strategy, considering the database with the maximum predicted standard deviation (i.e., uncertainty), and compared it with a random sampling strategy to demonstrate the superiority of the sampling strategy. Newly labeled data were added to the training set by performing DFT calculations. As shown in Fig. 4b, by adding a small amount of data with maximum uncertainty to the training set, the model's predictive performance significantly improved. Kim *et al.* [125] proposed a deep learning framework for exploring the material design space. As shown in Fig. 4c, this framework gradually extends the reliable prediction domain of the deep learning neural network (DNN) model to the desired attribute area through active transfer learning.

In the active learning framework, the development and exploration of the unexplored material search space are balanced by uncertainty, guiding the next best experiment or computation. The results of experiments or computations enhance the training data, and the cycle continues until the ideal ML model is built (Fig. 4d), significantly improving the efficiency of materials development [124].

Transfer learning

Transfer learning involves two ML tasks: the source task and the target task. In the learning of the source task, the model is typically trained on a larger and more easily obtainable dataset to acquire foundational knowledge. Subsequently, the knowledge gained from the source task is transferred to the target task. This transfer can include the model's parameters, feature representations, and representations in intermediate layers. Finally, the model is further fine-tuned on the target dataset to adapt to the specific characteristics of the target task. The advantage of transfer learning lies in its ability to leverage easily obtainable data for pre-training the model. By equipping the model with general knowledge and experience, it reduces the demand for a large sample size on the target task [126,127].

This strategy is well-suited for addressing the issue of data scarcity in specific materials. For ML tasks, pre-training models on numerous data from material open databases allows them to understand fundamental knowledge related to materials science and physical chemistry. Models trained in this way are more likely to exhibit better performance on specific material domain problems. Jha *et al.* [128] used the deep neural network ElemNet that pre-trained on 341k data from the OQMD database to predict formation energy, then applied transfer learning to fine-tune the model parameters on two smaller datasets (joint automated repository for various integrated simulations (JARVIS) and the MP, with 11k and 23k data points, respectively). The results showed that transfer learning significantly improved the model's prediction accuracy. Similar work had done by Goodall and Lee [129]. Additionally, Gu *et al.* [117] pre-trained the MatErials graph network (MEGNet) [130] on the MP and applied a transfer learning strategy by fixing the model weights in the encoding layer and the first graph convolutional layer, then retraining the rest of the model using specific perovskite data (Fig. 4e). The results demonstrated that combining transfer learning with training on limited data in specific material domains significantly improved the model's performance in those domains.

Transfer learning strategies can be applied in a more flexible manner by transferring knowledge between different material domains and diverse sets of material properties. This strategy

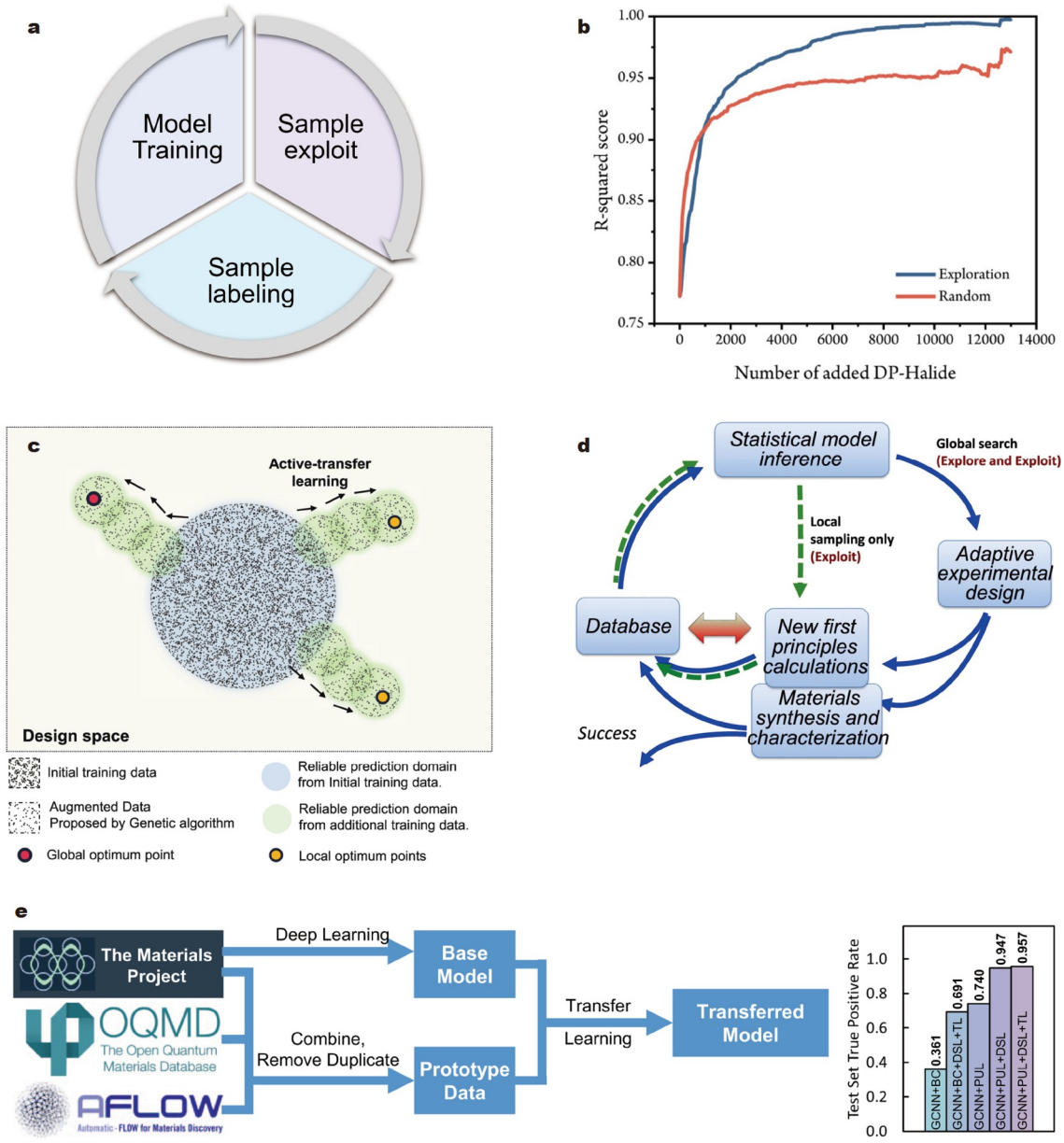

Figure 4 (a) Schematic of active learning. (b) Prediction accuracy changes during optimization for E_{form} from the exploration and random selection techniques with the metric of *R*-squared score. Reprinted with permission from Ref. [58]. Copyright 2022, Wiley-VCH GmbH. (c) Schematic of gradual expansion of reliable prediction domain of DNN based on the addition of data generated from the hyper-heuristic genetic algorithm and active transfer learning. Reprinted with permission from Ref. [125]. Copyright 2021, the Author(s). (d) Adaptive design paradigm to iteratively learn a surrogate model and use uncertainties to trade-off exploitation and exploration of the search space of unexplored materials to select the next best experiment or calculation. Reprinted with permission from Ref. [124]. Copyright 2019, the Author(s). (e) Domain-specific transfer learning workflow and the out-of-sample true positive rate for perovskites for various tested models. TL indicates transfer learning. Reprinted with permission from Ref. [117]. Copyright 2022, the Author(s).

provides a solution to the challenge of ML modeling in material domains with scarce data. For instance, even though low-dimensional materials have been extensively studied for various properties in diverse devices, the scarcity of relevant data, such as the well-known 2D materials database computational 2D materials database (C2DB) containing only 4k entries, hinders deep learning tasks. Frey *et al.* [44] aimed to overcome this limitation during the design of point defects in 2D materials. To obtain an adequate pool of 2D candidates, they trained MEGNet on a dataset with abundant bulk crystal data. Subsequently, they transferred this model to a 2D material dataset to predict the formation energy, Fermi energy, and bandgap of 2D materials. This approach enabled the identification of the most promising 2D material candidates. Later, they generated nearly ten thousand defect structures in transition metal dichalcogenide (TMD), h-BN, and over 150 2D wide-bandgap materials for the subsequent construction of defect models.

Moreover, Chen and Ong [131] further developed a transfer learning framework called AtomSets based on MEGNet. Although MEGNet was only pre-trained on 130k samples with formation energies from the MP, the testing results demonstrated that the transfer learning strategy performed well in predicting other material properties such as bulk modulus, bandgap, and metallic attributes. This not only proves the

effectiveness of the transfer learning strategy but also affirms that some different material properties stem from the unified fundamental principles of physical and chemical laws.

ML BASED MODELS AND TOOLKITS

Shallow learning models

Shallow learning models refer to relatively simple ML models with shallow frames and learning capabilities. These models are typically based on traditional statistical learning methods. Although they may struggle with overly complex problems, due to their lower computational complexity and friendliness with small datasets, they remain highly effective in many practical applications [132,133]. Popular shallow models include decision tree [78] and support vector machine (SVM) [134,135]. Decision tree predicts the target value by learning simple decision rules inferred from data features. Each internal node in the tree represents a feature decision, and based on the decision results, samples are assigned to different child nodes. Leaf nodes represent the final classification label or regression value. Decision tree is easy to understand and interpret, and the tree structure can be visualized. However, it can be unstable, as small changes in the data may lead to the generation of entirely different trees. Overfitting issues may arise if a decision tree creates too many trees during training, necessitating tree pruning and limiting the maximum depth of the tree. SVM works by mapping data to a high-dimensional feature space and finding an optimal hyperplane in this space to maximally separate samples of different classes. For problems that are not linearly separable, SVM can use a Kernel Function to map data to a high-dimensional feature space. Unlike SVM applied to classification tasks, the goal of support vector regression (SVR) is to fit the data as closely as possible within a tolerance, making the distance between samples and the hyperplane as small as possible.

SVM and SVR have consistently been popular shallow models in materials science. For instance, Shen *et al.* [88] constructed an SVR model to predict the bandgap of 2D organic–inorganic halide perovskites materials. The model was used for high-throughput screening and successfully identified 18 candidates with appropriate bandgaps, environmental friendliness, and stability from a vast chemical space containing 1017k virtual samples. Kumar *et al.* [91] built decision tree and SVM to determining whether transition metal chalcogenides and oxides are semi-metals (zero bandgap) or semiconductors (non-zero bandgap). Yang *et al.* [86] constructed an SVM classifier to screen out 3098 perovskites from 6529 virtual samples. They also used an SVR model for predicting bandgaps, ultimately selecting 60 oxide double perovskites with bandgap between 1.00 and 1.60 eV. Chen *et al.* [64] developed an SVR model to predict the energy above the convex hull (E_{hull}) of ABO_3-type compounds, enabling the assessment of material stability.

For other shallow models, Wang *et al.* [39] built a Gaussian regression process (GRP) model to predict bandgaps and band edge positions, revealing layer-dependent electronic properties in heterostructures of transition metal chalcogenides. Maddah *et al.* [136] constructed a decision tree to determine the stability of Ti-based perovskites and investigated the influential factors.

Ensemble learning models

We can obtain a more powerful and robust ensemble model [137,138] by combining shallow models through certain strategies. These strategies typically involve weighting the predictions of multiple shallow models to derive the final prediction. Common ensemble models mainly include three types: bagging, boosting, and stacking [139]. In the bagging method, multiple sub-training sets are generated by randomly sampling with replacement from the original training set. Then, a basic model is independently trained based on each sub-training set. The final prediction of the ensemble model is obtained by averaging the predictions of these basic models. For example, random forest (RF) [140,141] predicts by averaging the predictions of multiple decision trees. It has been used in some material prediction works [136,142]. Boosting involves sequentially training a series of basic models, each attempting to correct the output of the previous model. The predictions of multiple basic models are then combined through weighted voting or weighted averaging. Common boosting algorithms include AdaBoost tree (AdaBT) [143], gradient boosting tree (GBT) [144], XGBoost tree (XGBT) [145], LightGBM (LGBM) [146]. The core idea of stacking is to independently train a series of models of different types and then integrate the output results of each model using a meta-model to form the output of the ensemble model. For example, using logistic regression and naive Bayes as basic models and a decision tree as the meta-model creates a stacking ensemble model.

Shallow models and ensemble models heavily rely on manually crafted material descriptors. The ability of descriptors to accurately capture the underlying relationships between materials and the properties to be predicted directly determines the model's performance. Crystal descriptors include elemental descriptors (atomic radius, electronegativity, electron affinity, electron count, etc.), local structure descriptors (average bond length, bond angle, octahedral factors O_f etc.) [63,87,92,147], crystal global structure descriptors (crystal density, packing fraction, crystal complexity, symmetry, etc.), as well as band structure descriptors [148], electron density descriptors. There are also many structure descriptors that satisfy translational and rotational invariance, including but not limited to Coulomb matrix, atom-centered symmetry functions (ACSF), and smooth overlap of atomic positions (SOAP) [149], etc. [60,62,150,151]. However, these manually constructed descriptors have significant uncertainty, meaning it is not known a priori whether these descriptors are effective before model training. It is also challenging to ascertain whether these descriptors possess uniqueness. Therefore, relying on expert knowledge to find effective descriptors for materials may be challenging.

In materials science, ensemble models are widely recognized and used due to their higher accuracy and interpretability compared with shallow models, and less dependency on data size compared with deep learning models. This has made some ensemble models such as RF, GBT, XGBoost, LGBM, widely accepted and utilized [32,33,37,38,44,58,63,65,77,84,88–90,92,136,147,152–155]. They find broad applications in predicting material properties based on small datasets obtained from high-throughput computations, including assessing crystal stability (predicting E_f or E_{hull}) [32,58,63,64,147,153], exploring the optoelectronic properties of crystals (predicting bandgaps) [33,58,63,77,84–86,88–92], and conducting high-throughput virtual screening of materials to accelerate the discovery or design of new materials. Specific examples will be elaborated in subsequent chapters.

Deep learning models

Deep learning models primarily refer to neural networks based on deep learning [3,156]. Neural networks are a type of ML models that mimics the structure and functioning of the human brain's neural system. They consist of multiple artificial neurons connected by weighted connections, and information is transmitted and processed through activation functions. Neural network models typically consist of an input layer, multiple hidden layers, and an output layer (Fig. 5a). The input layer receives raw data as the model's input, hidden layers process and extract features from the input, and the output layer generates the final prediction. While both the input and output layers have only one layer each, the hidden layers can have many layers, and each hidden layer can have many nodes (neurons). All nodes in each layer can communicate information with nodes in adjacent layers. The non-linear features of neural networks enhance the model's complexity, allowing it to handle more information and perform more challenging prediction tasks than shallow and ensemble models [157].

Compared with shallow models and ensemble models, neural networks not only have the ability to learn more complex knowledge but also come with additional advantages. Firstly, they are well-suited for transfer learning [158]. The training process of neural networks involves finding an appropriate network architecture and node weights. By retaining the framework and some weights, the model can be transferred to a target domain for retraining, further adjusting model parameters without the need to train a neural network model from scratch. Secondly, the training of neural networks involves a large number of matrix multiplications and convolution operations, which can be efficiently executed on graphic process units (GPUs) through parallel computing [159]. This capability facilitates the processing of large-scale data and the construction of large models, as exemplified by the massive 175 billion parameters in large models like ChatGPT (GPT-3.5) [1].

However, neural networks also have some drawbacks. Firstly, they exhibit low interpretability (black-box model) [160]. Due to the presence of hidden layers and numerous nodes, neural networks transform inputs into a high-dimensional feature space for learning relationships between features. This abstract mechanism is challenging to interpret and understand. Secondly, neural networks have a dependency on data size [3,68,129]. Compared with traditional ML methods, neural networks have parameters that are orders of magnitude higher. As shown in Fig. 5b, the performance of neural networks surpasses that of shallow models and ensemble models only when the data size reaches a certain level [133,161]. For example, for GNN, a data size on the order of 10^3 to 10^4 is necessary for the model to achieve sufficient accuracy [68]. As depicted in Fig. 5c, an increase in the number of training set samples significantly reduces the model's error. However, an excess of data size and increased model complexity lead to longer training times and greater computational requirements for neural networks. Omee *et al.* [162] found that due to the workload of model training, some GNN models suffer from undertraining. They demonstrated that 500 epochs of training are required to reach a point where the model no longer improves, yet some GNN models were trained for only 200 epochs. Finally, neural networks are sensitive to hyperparameters [163]. Hyperparameters are parameters set before training begins, such as learning rate, batch size, regularization parameters, the choice of optimizer and activation function, the number of layers in the network, and the number of neurons in each layer. Any change in these parameters may significantly impact the final results of the model. For example, the number of layers and nodes in hidden layers has a non-linear relationship with model performance. Unfortunately, due to workload constraints, it is not feasible to perform an exhaustive search for optimal hyperparameters.

We primarily divided the popular neural networks applied in computational materials into two parts: supervised learning neural networks for property prediction and unsupervised learning neural networks for material generation. Further classified based on material representation methods, property prediction models can be categorized into models based on crystal structure graphs and models based on chemical compositions. Generative models can be categorized into chemical compositions generation models and crystal structures generation models. In the last five years, material representation methods and the architecture of neural network models have been continuously improved, leading to enhanced model performance. We will introduce the core ideas of these models chronologically and compare them to demonstrate the development process of these models.

Neural networks based on crystal graph

GNNs are popular methods in material science because provide a more automated, standardized, and accurate representation for crystals, compared with the feature engineering in shallow or ensemble models. In GNNs, a graph is represented as a collection of nodes and edges, where nodes represent entities, and edges represent relationships between nodes. In a graph structure, each node is defined by its own features and the features of the nodes connected to it. The core idea of GNN is to update the representation of nodes by iteratively aggregating information from the nodes and their neighbors, capturing information within the graph [164]. This definition of graphs is highly applicable to describing crystal structures [61,165–167]. In periodic crystals, the crystal is composed of regularly arranged atoms, and interactions between atoms through atomic bonds. The interactions between atoms significantly influence the structure and properties of the crystal. The representation of an atom depends on the types of coordinating atoms, and coordinating with different atoms significantly alters the meaning of an atom in the crystal. Therefore, we can analogize crystals to graphs in GNNs, where atoms are nodes, and atomic bonds (interactions between atoms and their coordinating atoms) are edges. The collection of nodes and edges characterizes the entire structure of the crystal. Based on this idea, as shown in Table 1, GNNs have been proven to effectively learn the structures of molecules and crystals and accomplish the prediction of material properties [57,107,117,129,130,162,168–172].

Xie and Grossman [170] first proposed the crystal graph convolutional neural network (CGCNN) for crystal structure representation and material property prediction. As shown in Fig. 6a, the core of CGCNN includes graph convolutional layers and pooling layers [175]. The convolutional layer iteratively updates the representation of atoms by performing convolution operations on atoms and their neighbors. This process extracts information about the local structure in which an atom is located and utilizes a non-linear graph convolution function for bonding. During the model training, the weights of the convolutional kernel are updated to distinguish the strength of

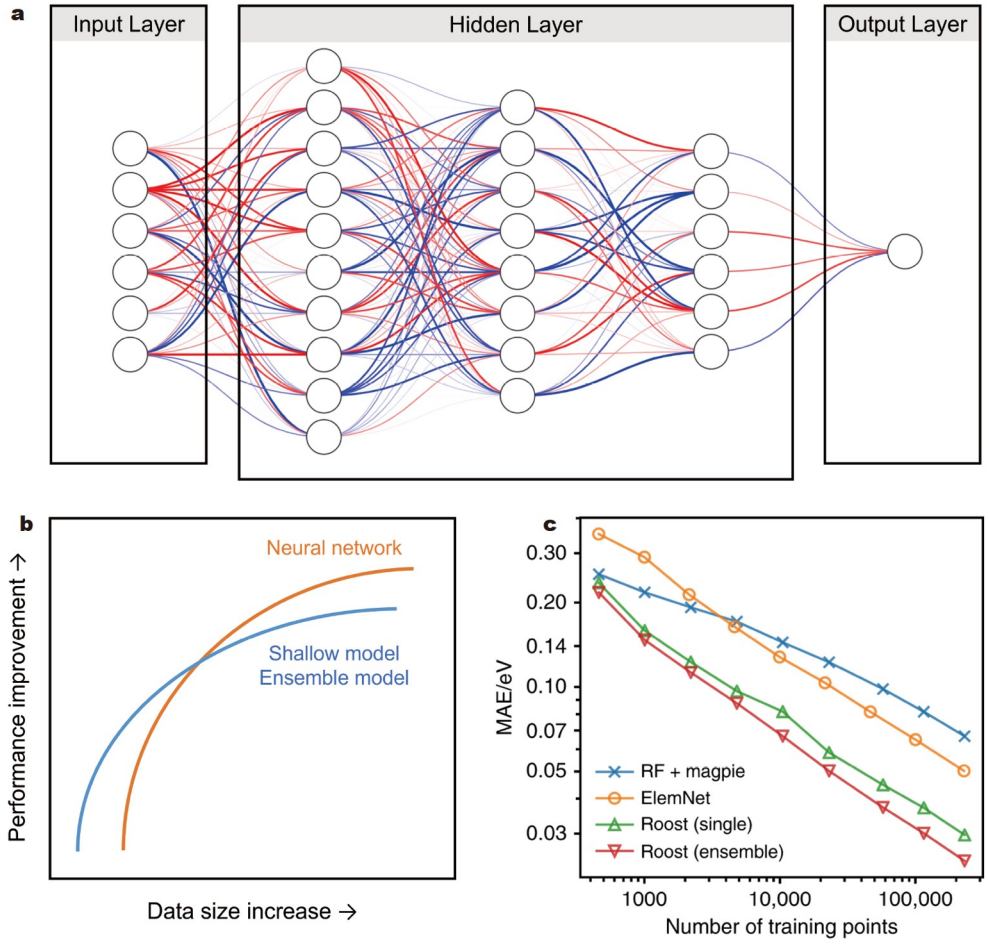

Figure 5 (a) Neural network architecture. (b) Comparison of data dependency between neural networks and non-neural networks. After training with a larger data size, the performance of neural networks surpasses that of non-neural networks. (c) Increasing data size significantly reduces the error of neural networks, compared with that of Random Forest. Reprinted with permission from Ref. [129]. Copyright 2020, the Author(s).

Table 1 Neural networks based on crystal graph (GNNs) with the authors, year, training data, and MAE of predicting formation energy

GNN models	Author	Year	Training data	MAE (eV/atom)
CGCNN [170]	Xie and Grossman	2018	MP 28k	0.039
MEGNet [130]	Chen *et al.*	2019	MP 60k	0.028
OGCNN [173]	Karamad *et al.*	2020	MP 60k	0.030
iCGCNN [174]	Park and Wolverton	2020	OQMD 180k	0.031
GATGNN [172]	Louis *et al.*	2020	MP 60k	0.048
ALIGNN [168]	Choudhary and DeCost	2021	MP 60k	0.022
DeeperGATGNN [162]	Omee *et al.*	2022	MP 36k × 80%	0.0296

bonding between the central atom and its neighbors. After convolution, the pooling layer performs normalized summation to aggregate atomic information, generating the overall feature vector for the crystal.

Although CGCNN provides a highly flexible framework for representing crystal structures and demonstrates excellent performance in predicting various material properties, there is still room for improvement. For instance, the pooling method used by CGCNN, known as normalized summation, while simple and computationally efficient, may lose some important feature information because all atomic information is assigned the same weight. Building on the core idea similar to CGCNN, several improved crystal graph network models have been proposed.

Chen *et al.* [130] found that CGCNN does not consider the influence of global states (such as temperature) on the system. They introduced MEGNet, which takes temperature, pressure, entropy, and other state variables into account as global states. As shown in Fig. 6b, during the training, the bond attributes, atom attributes, and state attributes of crystals are successively updated through convolution operations. In the MEGNet architecture, a single block captures interactions between each atom and its local environment. By stacking more graph

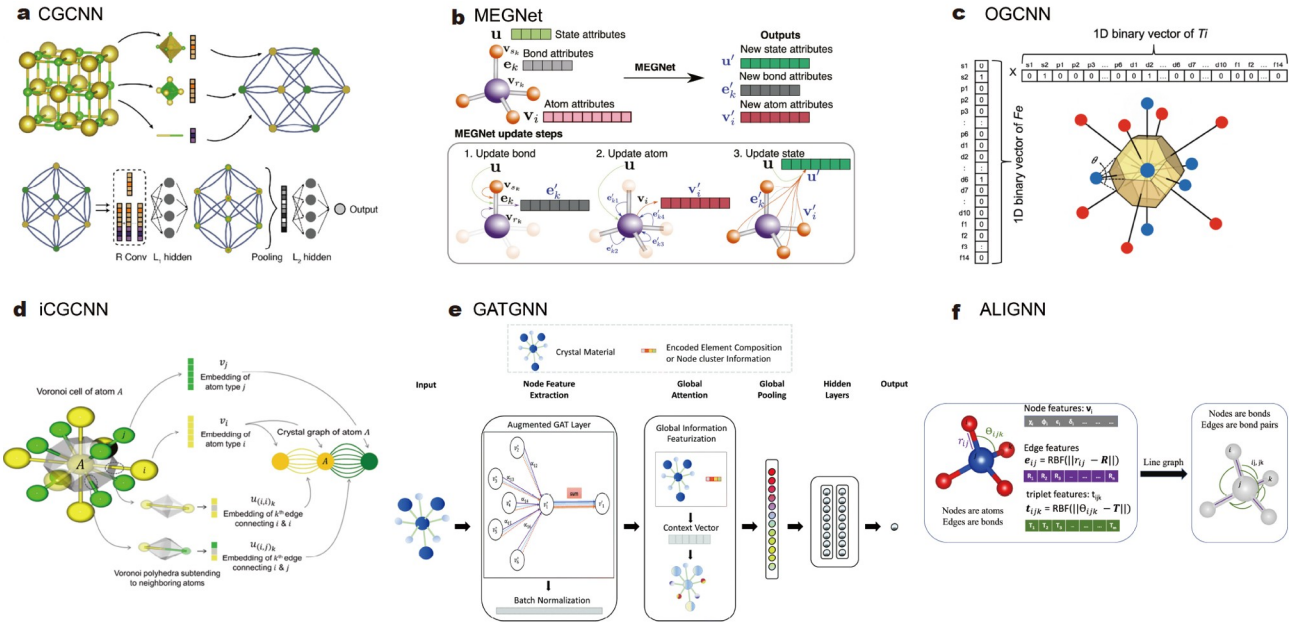

Figure 6 Neural networks based on crystal graph. (a) Illustration of CGCNN. Reprinted with permission from Ref. [170]. Copyright 2018, American Physical Society. (b) Overview of an MEGNet module. The initial graph is represented by the set of atomic attributes ***v***, bond attributes ***e*** and global state attributes ***u***. The bond, atomic, and global state attributes are updated in turn. Reprinted with permission from Ref. [130]. Copyright 2019, American Chemical Society. (c) OFM representation for the FeTi alloy. Blue and red atoms are Fe and Ti, respectively. The inset shows the Voronoi polyhedron for the center Fe atom forming a truncated octahedron. The 1D binary vectors for the Fe and Ti atoms are shown as well. Reprinted with permission from Ref. [173]. Copyright 2020, American Physical Society. (d) Illustration of iCGCNN crystal graph. The crystal graph shown on the far right represents the local environment of atom A. Multiple edges connect A to neighboring nodes to show the number of Voronoi neighbors. Reprinted with permission from Ref. [174]. Copyright 2020, American Physical Society. (e) Architecture of our global attention graph CNN model GATGNN. Reprinted with permission from Ref. [172]. Copyright 2020, Royal Society of Chemistry. (f) Schematic showing undirected crystal graph representation and the corresponding line graph construction for a SiO_4 polyhedron. The ALIGNN convolution layer alternates between message passing on the bond graph (left) and its line graph (or bond adjacency graph, right). Reprinted with permission from Ref. [168]. Copyright 2021, Springer Nature.

network modules, atoms and bonds can capture longer-range interactions. This clearly addresses the limitation in CGCNN, which only considers neighboring atoms within a fixed 6 Å range. This is particularly meaningful for predicting properties related to short and long-range interactions, such as zero-point vibrational energy (ZPVE) and electronic spatial extent. Based on MEGNet, Chen and Ong [131,176] further developed a transfer learning framework called AtomSets and an interatomic potential (IAP) model named M3GNet.

Karamad *et al.* [173] found that CGCNN did not account for features involving orbital-orbital interactions. Consequently, they introduced the orbital graph convolutional neural network (OGCNN). As illustrated in Fig. 6c, in addition to elemental features, the orbital-orbital interactions between atoms and their neighbors are represented as an orbital field matrix (OFM), with weights assigned based on the area of the Voronoi polyhedra for coordinating atoms. Furthermore, the inclusion of an encoder-decoder network and a convolutional network in OGCNN allows it to learn crucial features from basic atoms (elemental features), orbital-orbital interactions, and topological features.

Park and Wolverton [174] proposed improved crystal graph convolutional neural network (iCGCNN) by addressing three shortcomings of CGCNN. Firstly, they found that CGCNN rigidly considers the 12 nearest neighboring atoms around each atom, learning the strength of atomic bonds through updates to convolutional kernel weights during model training. This representation is not always accurate. Therefore, as shown in Fig. 6d, iCGCNN directly specifies the strength of interactions between the central atom and neighbors by using the solid angle, area, and volume of Voronoi polyhedra as edges for graph network nodes. Secondly, iCGCNN also considers three-body interactions that CGCNN overlooks. Lastly, the edge vectors representing atomic bonds in CGCNN remain unchanged during the training process. Therefore, iCGCNN updates atomic bond vectors by designing a new convolution function.

Louis *et al.* [172] introduced graph-attention graph neural network (GATGNN). This model incorporates an attention mechanism, originating from neural networks in NLP, used to learn the contributions of different context vector components [177]. As shown in Fig. 6e, GATGNN uses the attention graph attention layer (AGAT) to capture the attributes of the local atomic environment and employs a global attention layer to replace CGCNN's inaccurate normalized summation pooling aggregation strategy. The global attention layer performs weighted aggregation on all these atomic environment vectors to create a global representation of the entire crystal structure. This strategy considers the weights of different atomic information relative to global information, allowing the model to better capture the fact that different atoms contribute differently to global material properties in a crystal. GATGNN outperforms CGCNN and MEGNet in predicting bandgap, shear modulus, and bulk modulus.

Choudhary and DeCost [168] proposed atomistic line graph neural network (ALIGNN) to consider local structural features, as the electronic characteristics such as bandgaps are sensitive to the changes in local structural features. The specific idea, as

depicted in Fig. 6f, involves deriving a line graph $L(g)$ from a graph g, which describes the connectivity of edges in g. In the original graph g, nodes correspond to atoms, and edges correspond to bonds, while in the atomic line graph $L(g)$, nodes correspond to inter-atomic bonds, and edges correspond to bond angles. Employing an edge-gated graph convolution strategy, the model alternates between graph convolutions on these two graphs, propagating information about bond angles to atomic representations through the representation of interatomic bonds and *vice versa*. Since the bond distances and bond angles in the line graph capture finer details of the local structure in a crystal, further characterization of the local structure enhances the model's performance.

The aforementioned GNNs focus on improving model performance through more accurate representations of material graphs. However, Omee *et al.* [162] found that several existing GNN models have fewer than nine convolutional layers. They discovered that merely increasing the number of layers in GNN models can better capture the underlying relationships between structure and properties, thereby further enhancing model performance. They designed DeeperGATGNN by increasing the number of convolutional layers from 5 to 15 based on GATGNN. The predictive performance on crystal formation energy significantly improved compared with the original GATGNN model.

Neural networks based on chemical compositions

While GNN models based on crystal structures may have advantages in predictive accuracy, they also exhibit some drawbacks. In many cases, material property datasets lack appropriate structural information. For instance, datasets like the experimental bandgap data collected by Zhuo *et al.* [178] may not provide structural information that GNN models can use. Furthermore, GNN models are sensitive to changes in crystal structures, making them reliant on relaxed crystal structures. These prerequisites somewhat restrict the applicability of GNN models. This limitation is particularly evident in the discovery of new materials, where it is impossible to have a priori knowledge of the specific structural information for unexplored compounds. As a result, some neural network models have been proposed that are not based on crystal structures but instead rely on chemical compositions (Table 2). Constructing models solely based on elemental composition allows for faster exploration of unknown material spaces, facilitating the discovery of valuable candidates.

Jha *et al.* [179] introduced ElemNet (Fig. 7a). It takes only the chemical composition as input and employs a deep neural network with 17 hidden layers to automatically capture the physical and chemical interactions and similarities between different elements, thereby predicting material properties in a straightforward manner. The model was trained on 275k compounds from OQMD to predict the formation energy of materials. The results indicate that this model is at least 30% more accurate (MAE = 0.055 eV/atom) than previous shallow models.

Although the predictive accuracy is not as high as the later GNN models mentioned, it accomplishes the prediction task without utilizing any domain knowledge about material stability, relying solely on the information-capturing capability of deep learning models. This has opened new directions for material representation and prediction, leading to the emergence of many deep learning models constructed solely based on elemental composition.

Similar to ElemNet, atom table convolutional neural networks (ATCNN) proposed by Zeng *et al.* [180] automatically mines information about elements in the elemental composition based on deep learning. As shown in Fig. 7b, in the framework of ATCNN, compounds are treated as 10 × 10-pixel images referred to as an atom table (AT). Each pixel in the AT represents an element, and its value is the proportion of that element in the compound. Unlike ElemNet, ATCNN performs feature extraction of elements through convolution operations rather than an extensive number of hidden layers in neural networks.

In ElemNet and ATCNN, the relationships between elements in the chemical composition, as well as the importance of different elements to the compound, are not explicitly considered. A notable drawback of this approach is that the proportion of different elements in the dataset is approximately equal to their abundance, resulting in the importance of elements to the compound being determined by their stoichiometry. However, this does not reflect the real-world scenario accurately. For instance, dopants in materials may have a critical role in controlling the properties of the material, even though they constitute a very small proportion. ElemNet and ATCNN cannot capture the importance of dopant elements with low proportions. To address this limitation, Goodall and Lee [129] introduced the representation learning from stoichiometry (Roost). As shown in Fig. 7c, they reframe the chemical formula of a material as a dense, weighted graph of its elements and then directly learn from it using a message-passing neural network. The advantage of this approach is that, with an increase in data size, the model's overall learning ability improves across different samples with diverse chemical formulas. Secondly, the model employs a weighted soft attention mechanism to update the representations of element nodes in the graph. This process allows the model to learn features of each constituent element, capturing some prior knowledge about the correlations between elements.

Wang *et al.* [181] made improvements upon the Roost. They introduced the compositionally restricted attention-based network (CrabNet) (Fig. 7d). Like Roost, CrabNet employs the mat2vec method to represent a material's chemical formula as a set of element vectors. However, CrabNet differs in that it introduces a self-attention mechanism based on the transformer architecture [177] in material performance prediction tasks. It treats the chemical composition as a system, elements as items within that system, dynamically learns and updates the representation of individual elements based on their chemical environment, and shares information between elements. This representation allows CrabNet to learn interactions between elements within compounds and directly predict the contribution of each element's vector to property predictions. Fig. 7e illustrates the average contributions of each element to the bulk modulus predicted by the CrabNet model trained on AFLOW volume modulus data. CrabNet predicts a small contribution of lithium to the overall bulk modulus, while tungsten has a significant contribution. This suggests that visualization results based on attention mechanisms can display correlations between elements or relationships between elements and target properties.

Neural networks based on generative methods

Generative models are a type of deep learning model that

Table 2 Neural networks based on chemical compositions with the authors, year, training data, and MAE of predicting formation energy

Composition-based models	Author	Year	Training data	MAE (eV/atom)
ElemNet [179]	Jha *et al.*	2018	OQMD 275k	0.055
ATCNN [180]	Zeng *et al.*	2019	OQMD 5886 × 80%	0.078
Roost [129]	Goodall and Lee	2020	OQMD 256k × 90%	0.024
CrabNet [181]	Wang *et al.*	2021	OQMD 341k × 70%	0.031

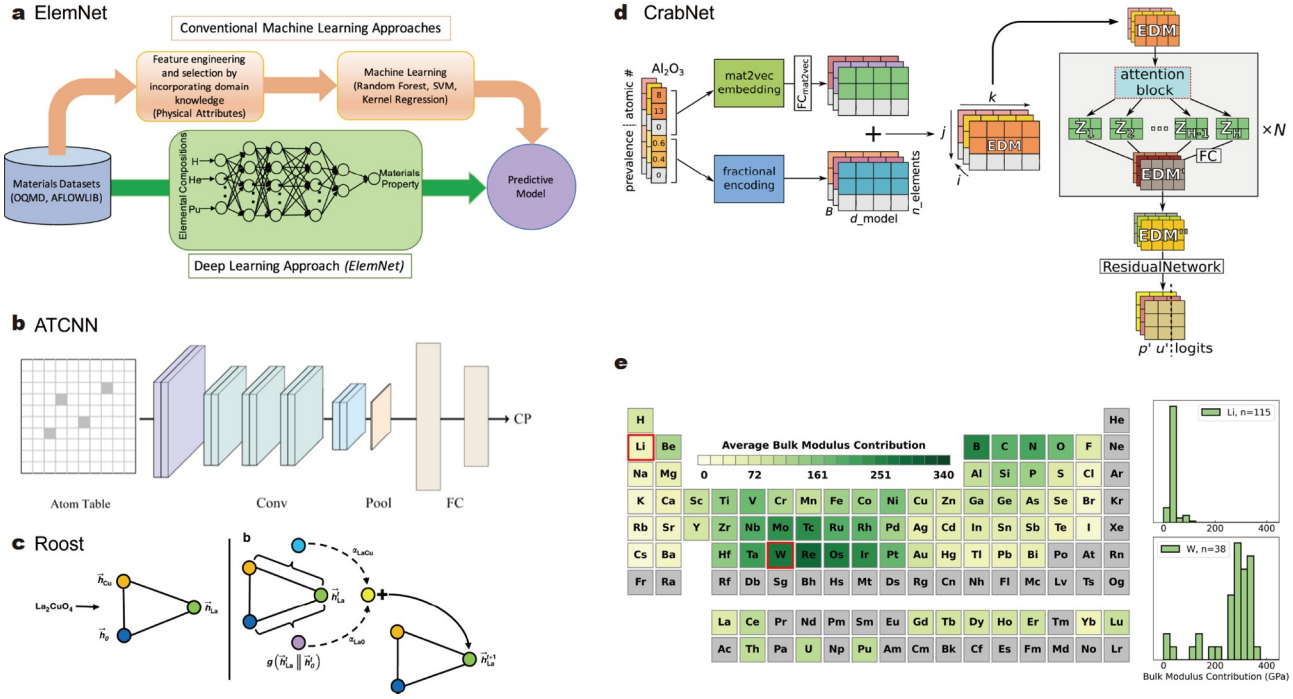

Figure 7 (a) ElemNet based predictive approach directly learns to predict properties of materials. Reprinted with permission from Ref. [179]. Copyright 2018, the Author(s). (b) Schematic diagram of the ATCNN model for superconducting critical temperature (T_c) prediction. Reprinted with permission from Ref. [180]. Copyright 2019, the Author(s). (c) An example stoichiometry graph for La_2CuO_4. And a graphical representation of the the update function for the La representation. Reprinted with permission from Ref. [129]. Copyright 2020, the Author(s). (d) Schematic illustration of the element-derived matrix (EDM) representation for Al_2O_3. And the schematic of the CrabNet architecture including the input EDM, the self-attention layers (repeated N times), the updated and final element representations (EDM and EDM″), the residual network, and the final model output. Reprinted with permission from Ref. [181]. Copyright 2021, the Author(s). (e) Average contribution of all elements to bulk modulus predictions. Reprinted with permission from Ref. [181]. Copyright 2021, the Author(s).

creatively generates new samples similar to input samples by unsupervised learning of the latent distribution and features of the samples [182]. In recent years, well-known generative models include the stable diffusion model applied to image generation [183] and ChatGPT used for text generation [1]. Generative models for generating organic molecular structures or protein structures have found widespread applications in life sciences, exemplified by AlphaFold [8]. In materials science, generative models are employed to create crystal structures or chemical compositions [48,49,184]. When considering conditional constraints, the generated materials can possess specified properties, aiming at the inverse design of materials [185]. Compared with traditional methods such as element substitution or structure search, data-driven generative models significantly enhance the efficiency of exploring new materials.

Based on different model architectures, generative models applied in materials science can be broadly classified into two main categories: probabilistic VAE models (Fig. 8a) and adversarial GAN models (Fig. 8b).

VAE consists of an encoder and a decoder. The encoder encodes samples into a latent space, from which samples are then sampled and decoded into new samples. The latent space is a multidimensional continuous vector space where each point or vector corresponds to a state, or feature of the data automatically extracted during the model's learning process. By manipulating vectors in the latent space, we can influence the attributes of the generated samples. Typically, the sum of the reconstruction error and the Kullback-Leibler (KL) divergence is used as the loss function for VAE. They characterize the difference between the data generated by the decoder and the original data, as well as the difference between the distribution of the latent variables generated by the encoder and the prior distribution (usually assumed to be a standard normal distribution) [186].

GAN consists of a generator and a discriminator. The generator is a neural network whose goal is to learn a mapping that can transform noise vectors in the latent space into new samples similar to the training data distribution. The discriminator's role is to judge whether a sample is a real sample from the training

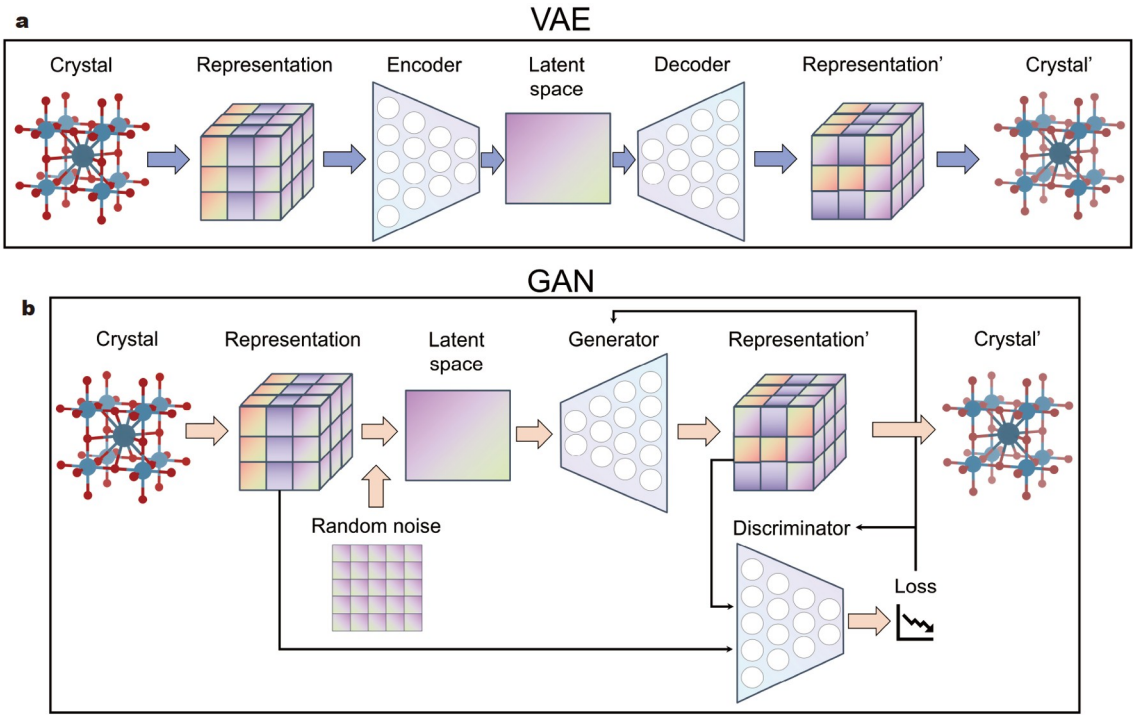

Figure 8 Frame of generative models (a) VAE and (b) GAN.

set or a new sample generated by the generator. In the training process of GAN models, the generator and discriminator continuously adjust their parameters to outperform each other. This parameter optimization is achieved by minimizing the loss functions of the generator and discriminator, which respectively characterize the ability of the generator and discriminator to generate realistic new samples and distinguish between real and fake samples [187].

Generative models can be used to generate chemical compositions or crystal structures (Table 3). The former is more flexible and efficient compared with the latter, enabling rapid exploration of feasible compositional spaces. However, one compound may correspond to multiple phases, and after generating chemical formulas, crystal structure exploration methods are still needed to find reasonable crystal structures. For instance, methods like USPEX [188] or CALYPSO [189] can be used to determine the crystal structure for a given set of elements and their stoichiometry. Furthermore, crystal structure generative models can directly generate the crystal structures, including lattice constants, space groups, atomic coordinates. However, their construction is typically more challenging because it requires a reversible representation of 3D crystals. This means establishing a one-to-one mapping between the crystal entity and the material representation, enabling accurate reconstruction of cell parameters and atomic positions.

Noh *et al.* [110] first proposed the inverse design framework for crystals called iMatGen, which was used to predict new crystal structures of vanadium oxides. To generate crystal structures with continuous representations, they decomposed the crystal structure into cell images (lattice constants and angles) and basis images (atomic positions). In the crystal reconstruction process, lattice parameters and atomic positions were reconstructed separately from these two images. As shown in Fig. 9a, iMatGen consists of two parts: an AE image compressor used to reduce the size of the two aforementioned images, and a VAE generator used to encode the elemental information from the first step. Additionally, to ensure the model could reconstruct stable crystal structures, the VAE training process included a classification task to categorize stable and unstable crystals ($E_f > 0.5$ eV/atom). The specific method for generating new materials involved sampling the material vector (z) from the material latent space, and then applying the two decoders successively to obtain the two types of material images in grid space. These grid space images were then inversely transformed into real space atomic positions and unit parameters, completing the crystal construction.

On the other hand, there have been attempts to efficiently sample the design space of inorganic materials by generating valuable chemical compositions. Dan *et al.* [108] proposed MatGAN based on a GAN architecture. Each material is represented by an 85 (number of elemental types) × 8 (maximum number of atoms) one-hot encoded matrix. As shown in Fig. 9b, in MatGAN, both the discriminator (D) and the generator (G) are deep neural networks containing convolutional and deconvolutional layers. After constructing MatGAN, Dan *et al.* [108] conducted a comprehensive evaluation of the model-generated new samples, including charge balance, thermodynamic stability, uniqueness, and novelty. Additionally, Pathak *et al.* [190] also proposed a model for generating material chemical compositions, but with a different approach based on the VAE architecture. As shown in Fig. 9c, deep learning based inorganic material generator (DING) consists of two parts: a generator network based on conditional VAE (CVAE) and an attribute prediction network. CVAE provides a continuous latent space along with control over attributes. The attributes of the material to be generated (formation energy, volume of each atom, and energy of each atom) are represented as a conditional vector, which is directly passed as input to CVAE. Thus, it can

Table 3 Generative models with the authors, model frame, year, generative target, material representation, and training data

Generative models	Author	Model frame	Year	Generative target	Material representation	Training data
iMatGen [110]	Noh *et al.*	VAE	2019	V_xO_y crystal structures	Lattice atomic coordinates	MP 10k V_xO_y crystal
MatGAN [108]	Dan *et al.*	GAN	2020	Chemical composition	One-hot matrix of 85 elements × 8 atoms	QMD 291k
DING [190]	Pathak *et al.*	VAE	2020	Chemical composition	One-hot matrix of 89 elements × 11 atoms	OQMD 272k × 72%
Cond-DFC-VAE [109,191]	Callum J. Court	VAE	2020	Crystal structures	256-dimensional electron-density map	MP 78k
CubicGAN [103]	Zhao *et al.*	GAN	2021	Ternary cubic crystal structure	Lattice Atomic coordinate 3 elements × 23 features vector	375k Ternary cubic crystal
FTCP [50]	Ren *et al.*	VAE	2022	Crystal structures	Real-space CIF-like features Reciprocal-space Fourier-transformed features	MP
CDVAE [101]	Xie *et al.*	VAE	2022	Crystal structures	Crystal multi-graph	Perov-5 Carbon-24 MP-20

selectively sample the latent space during the decoding process as needed to generate samples with the expected attributes. After generating new samples, three predictors trained on the OQMD database are used to predict the three attributes of the generated samples, thereby evaluating and filtering the candidate objects.

After Noh *et al.* [110] introduced iMatGen, more exploration into crystal generation has been undertaken. As shown in Fig. 9d, Court *et al.* [109] proposed the crystal generation model called conditional deep-feature-consistent VAE (Cond-DFC-VAE), characterized by utilizing the U-net for crystal structure construction. The crystal structure is initially represented as a voxelized 256-dimensional electron-density map. Similar to the strategy in the iMatGen model, they further enhance this latent space by training a binary classification formation energy model on the latent vector of input crystals to distinguish stable structures from unstable ones. Sampling the latent space based on formation energy generates new electron density maps. The density map is then reconstructed into atomic positions through a combination of UNet semantic segmentation network and morphological transformations. Visual evaluations of the latent space show it to be sufficiently smooth (continuous latent space), capable of generating realistically novel (similar but not identical to training samples) new samples.

Zhao *et al.* [103] from the same research group as the developers of MatGAN, later proposed the CubicGAN for generating cubic crystal structures. As shown in Fig. 9e, a crystal is represented by its lattice parameters, atomic coordinates, element features (3 elements × 23 element features), and space group. The generator takes randomly chosen space group, element combinations, and random noise Z as input, and then generates a material structure with the specified space group and element composition. The discriminator's input includes non-equivalent atomic coordinates, element attributes, cell parameters, and space group. Information extraction and implicit relationships between these four components are carried out through convolutional operations to determine the authenticity of the crystal.

Based on the aforementioned exploration, some have started considering conditional constraints for crystal structure generation, enabling inverse design of crystals. For instance, Ren *et al.* [50] proposed an inverse design framework for inorganic crystal structures based on the VAE model called Fourier-transformed crystal properties (FTCP). As shown in Fig. 9f, it includes a general crystallographic representation (varied in composition and structure) and a VAE model. The reversible material representation consists of real-space CIF-like features (element matrix, lattice matrix, atomic coordinate matrix, atomic occupancy matrix, and element property matrix) combined with reciprocal-space Fourier-transformed features. The VAE's latent space is connected to a target learning branch that maps points in the latent space to certain properties. Hence, the loss function of the VAE model includes an additional attribute mapping loss, a strategy that differs from CVAE by introducing target values into the VAE. The inverse design of new crystals is achieved by sampling points in the latent space of structural attributes outside existing crystals but within a local perturbation (Lp) strategy that satisfies user-defined design objectives.

Several of the earlier-generation generative models did not satisfy crystal invariance in the encoding and decoding of crystal structures. In other words, they lacked translational, rotational, permutation, and supercell invariance. Xie *et al.* [101], the creators of the CGCNN, proposed a crystal structure generative model named crystal diffusion variational autoencoder (CDVAE) that meets the requirements of crystal invariance. CDVAE is based on the diffusion mechanism and the VAE model. Property constraints are implemented based on the diffusion mechanism: adding noise (random values following a normal distribution) to a stable structure and denoising it might increase stability. They represent a crystal as $M = (A, X, L) \in A^N \times R^{N\times 3} \times R^{3\times 3}$ (A: atomic types, X: atomic coordinates, L: periodic

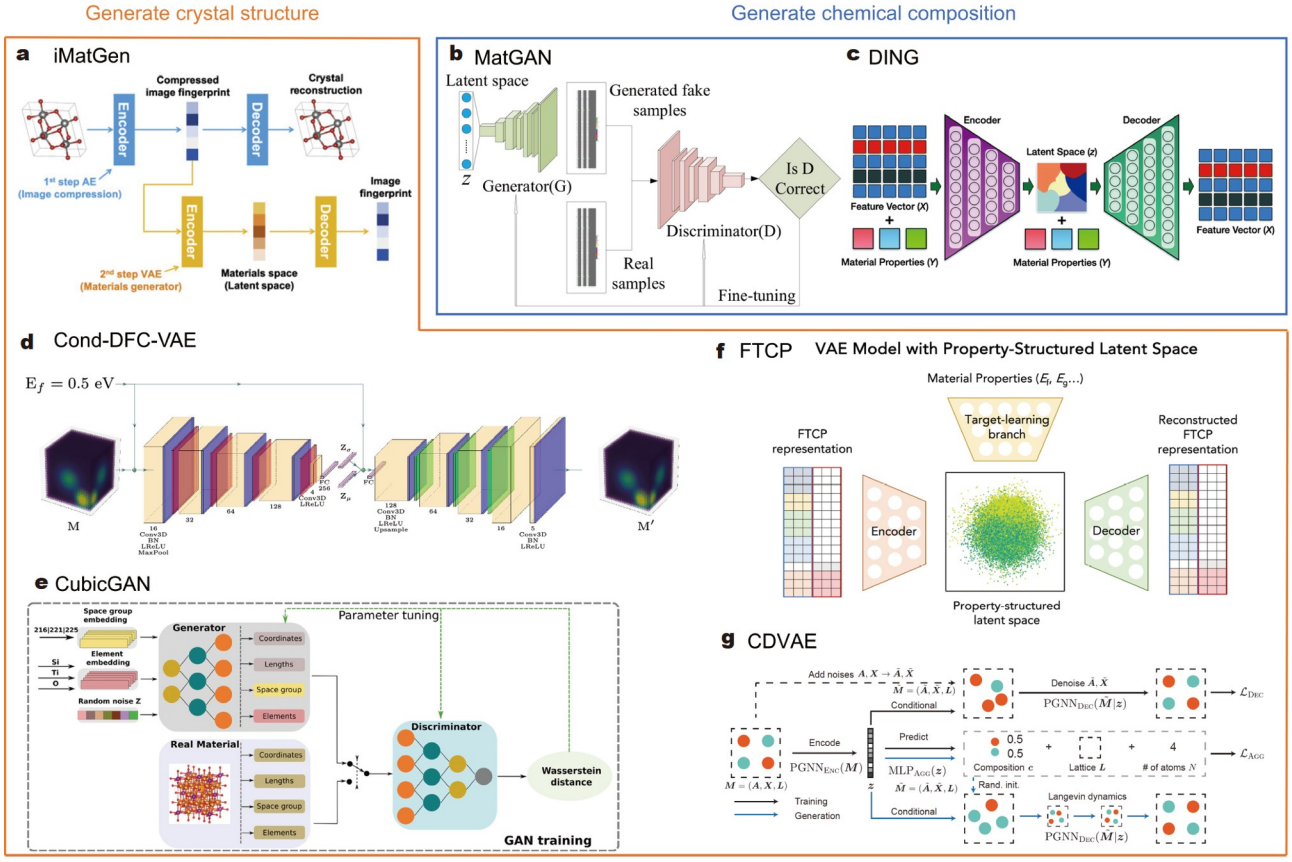

Figure 9 (a) Proposed hierarchical two-step image-based materials generator. Reprinted with permission from Ref. [110]. Copyright 2019, Elsevier. (b) MatGAN is composed of a generator, which maps random vectors into generated samples and a discriminator, and tries to differentiate real materials and generated ones. Reprinted with permission from Ref. [108]. Copyright 2020, the Author(s). (c) In model DING, CVAE model is used for generation of materials. The encoder networks encode 979-dimensional one-hot feature vector into a vector in the latent space ($\boldsymbol{z}$). The decoder networks take the property of the material along with its latent vector and regenerate the material. Reprinted with permission from Ref. [190]. Copyright 2020, Royal Society of Chemistry. (d) The Cond-DFC-VAE takes electron-density maps (M) with a corresponding property and produces reconstructed maps (M′). Reprinted with permission from Ref. [109]. Copyright 2020, American Chemical Society. (e) Fworkflow of the CubicGAN framework. Reprinted with permission from Ref. [103]. Copyright 2021, the Author(s). (f) VAE architecture using the invertible FTCP representation for inverse design. On top of the encoder + decoder architecture of a normal VAE, the latent space is also connected to a target-learning branch for property mapping, reflecting a property gradient(s) (property-structured latent space). Reprinted with permission from Ref. [50]. Copyright 2022, Elsevier. (g) Overview of the proposed CDVAE approach. Reprinted with permission from Ref. [116]. Copyright 2022, the Author(s).

lattice, N: number of atoms, R: Euclidean space). As shown in Fig. 9g, CDVAE consists of three neural networks: (1) periodic GNN encoder: encodes the crystal M into the latent space vector $\boldsymbol{z}$ (adding noise), (2) property predictor: sampling from the latent space vector $\boldsymbol{z}$ and predicting the composition c, lattice L, and the number of atoms N for crystal M, and (3) periodic GNN decoder: a diffusion model that denoises X' and A' conditioned on $\boldsymbol{z}$. Crystal generation involves two pathways: (1) sampling $\boldsymbol{z}$ from the latent space with added noise. The property predictor is used to predict the three properties of the crystal: composition c, lattice L, and the number of atoms N. These values are used to randomly initialize the new crystal structure. (2) Using the decoder to denoise X' and A' conditioned on $\boldsymbol{z}$ to enhance the stability of the new structure. Through testing, they demonstrated that the CDVAE model surpasses previous generative models in both crystal generation capabilities and optimizing crystal properties.

Toolkits for ML based materials research

Data collection, material representation, model construction is crucial for ML tasks. These processes heavily rely on various toolkits.

Data collection

Material databases provide us with convenient access to a vast amount of material data. Large neural networks are typically trained based on extensive material databases. Computational material databases include crystal structure files, related properties, and computational parameters. Table 4 provides detailed information about these databases, including their names, data size, links, and constructed structures. MP [22], OQMD [23], and JARVIS [193] are popular computational material databases and are often used as training sets for models. Constructing models based on the same database allows for the possibility of cross-model comparisons. Additionally, the experimental database inorganic crystal structure database (ICSD) [202] is widely used for high-throughput calculations to further generate datasets due to its detailed crystal information and convenient file types.

The recent development of low-dimensional materials has

Table 4 Computational material databases with description, data size, link, and institution

Databases	Description	Data size	Link	Institution
AFLOW [24]	Computational database of materials	3.5m	http://aflowlib.org	Duke University
Materials Project [22]	Computational database of materials	154k	https://materialsproject.org	U.S. Department of Energy
OQMD [23]	Computational database of materials	1m	http://oqmd.org	Northwestern University
CSD [192]	Database of organic and inorganic materials searched from previous journal publications	504k	http://crystallography.net	University of Cambridge
JARVIS-DFT [193]	A materials property repository focused on DFT predictions of material properties	56k	https://www.nist.gov/programs-projects/jarvis-dft	National Institute of Standards and Technology
C2DB [194,195]	Computational 2D materials database	4k	https://cmr.fysik.dtu.dk/c2db/c2db.html	Aalborg University
2DMatPedia [21]	Computational 2D materials database	6k	http://www.2dmatpedia.org	National University of Singapore
C1DB [196]	Computational database for 1D materials	820	https://cmr.fysik.dtu.dk/c1db/c1db.html	Technical University of Denmark
NOMAD [197]	Novel materials discovery project	12m	https://nomad-lab.eu/prod/rae/gui/search	Humboldt-Universität zu Berlin
Materials Cloud	A platform for open computational science	29m	https://www.materialscloud.org	cole Polytechnique Fédérale de Lausanne
MOFX-DB [198]	Computational adsorption data for nanoporous materials	168k	https://mof.tech.northwestern.edu	Northwestern University
CEP [199]	Harvard clean energy project	2m	http://cleanenergy.harvard.edu	Harvard University
OMDB [200]	An electronic structure database for various organic and organometallic materials	12.5k	https://omdb.mathub.io	KTH Royal Institute of Technology and Stockholm University
PubChem [201]	An open chemistry database for small molecules	115m	https://pubchem.ncbi.nlm.nih.gov	National Institutes of Health (NIH)
QM	Quantum chemistry structures and properties of molecules	134k	http://quantum machine.org/datasets	Argonne National Laboratory
NREL MatDB	Computational materials database for renewable energy applications	20k	https://materials.nrel.gov	National Renewable Energy Laboratory
aNANt	A functional 2D materials database	23k	http://anant.mrc.iisc.ac.in	Indian Institute of Science

prompted the construction of related material databases. The C2DB, constructed by Haastrup *et al.* [194] and Gjerding *et al.* [195] is applied in high-throughput calculations or ML work on 2D materials. However, it is evident that low-dimensional material databases face the challenge of data scarcity. For example, C2DB has only around 4k data, and the data size of C1DB [196] is even less than 1k, significantly smaller than the million-level data points in 3D databases, making it difficult to support the training of deep neural networks. The development of ML research of low-dimensional materials urgently demands the construction of relevant databases, which is expected to accelerate with time.

On the other hand, scientific literature indexed in databases like science citation index (SCI) can provide trustworthy knowledge and high-quality data. In recent years, an increasing number of researchers have been constructing material databases by extracting information from literature. Jacobsson *et al.* [55] manually reviewed every paper retrieved by searching "perovskite solar cells" on the Web of Science. This exhaustive process involved more than 15k papers and the manual extraction of data from over 42k devices. A more efficient approach involves the use of unsupervised NLP to automatically mine data from literature. Court and Cole [203] created a database with 20k records of magnetic and superconducting phase transition temperatures and their associated compound names. They achieved this by using the ChemDataExtractor toolkit [204] on a corpus of 74k scientific articles crawled from publishers such as Elsevier, Springer, and the Royal Society of Chemistry. Moreover, Pyzer-Knapp *et al.* [31] designed the IBM DeepSearch platform based on NLP. It is used to extract unstructured data from documents, enabling the construction of document-centric KGs and supporting sophisticated queries and data extraction for downstream applications.

However, it is essential to note that from published experimental or computational results, we can only extract positive data. Unsuccessful experiments and computational results containing negative data are unlikely to be published. Using such data for analysis or building ML models may introduce biases towards successful cases and may miss valuable knowledge and experiences from failures.

Features engineering

Feature engineering for materials is a laborious task, besides the GNNs mentioned earlier that can automatically extract features,

there are tools dedicated to assisting in material feature engineering. The functionalities of these relevant toolkits are outlined in Table 5. The Mendeleev package [205] provides an application programming interface (API) for accessing various properties of elements in the periodic table. Python materials genomics (Pymatgen) [206] is a robust open-source Python library for material analysis. It includes highly flexible classes for the representation of element, site, molecule, and structure objects. Matminer [207] is used for data mining of material properties. It can convert material objects into features such as average electronegativity or differences in ionic radii. Matminer also encompasses features for complex material data, such as band structures and electron density. The DScribe [149,208] is used to build descriptors for materials ML, including Coulomb matrices, sine matrices, Ewald matrices, ACSF, SOAP, many-body tensor representation, and local many-body tensor representation. Robocrystallographer [209] can construct semi-local structure descriptors (connectivity, tilt angle) for crystals. It can represent them as readable JavaScript object notation (JSON) and text files. Sure independence screening and sparsifying operator (SISSO) [210] can automatically extract effective feature combinations through mathematical operations in an expanded, vast candidate feature space. Additionally, it serves as a symbolic regression algorithm, used to fit linear mathematical formulas between material features and targets.

Model construction

The common underlying tools for model construction are Scikit-learn [214], PyTorch [215], and TensorFlow [216]. Scikit-learn is an open-source Python ML library that provides a range of supervised and unsupervised learning algorithms. It offers a wealth of ML tools, including classification, regression, clustering, dimensionality reduction, model selection, and preprocessing. Both shallow ML models and ensemble models mentioned earlier are included in Scikit-learn. It enables tasks such as data cleaning, model selection, model fitting, model evaluation, and model optimization. PyTorch is an open-source ML library primarily developed by Facebook's AI Research team. It offers a modular way to build and train deep learning models, supporting various types of neural network architectures. All the deep learning models mentioned earlier can be constructed and executed using PyTorch, for tasks like image and video processing, NLP. TensorFlow is an open-source library developed by the Google Brain team, used for building and training ML and deep learning models. It has overlapping functionalities with PyTorch. It features a robust visualization tool called TensorBoard, which helps users visualize the model training process in real-time.

The aforementioned three underlying tools are powerful, providing support for the development of general ML models and offering developers maximum flexibility. However, this implies that each research group needs to develop a complete workflow based on these tools to accomplish specific ML tasks in materials science. Obviously, on the one hand, this increases the threshold for researchers not majoring in computer science to engage in ML work. On the other hand, the lack of standardized

Table 5 Tools for data acquisition or processing, feature engineering, model construction or usage in ML process

Tools	Data acquisition or processing	Feature engineering	Model construction or usage
ChemDataExtractor [204]	●		
IBM DeepSearch [31]	●		
Matminer [207]	●	●	
Pymatgen [206]	●	●	
DScribe [149,208]		●	
Mendeleev [205]		●	
Robocrystallographer [209]		●	
SISSO [210]		●	
Gplearn [211]		●	
MagPie [212]		●	
Atom2vec [213]		●	
Scikit-learn [214]	●	●	●
Pytorch [215]			●
Tensorflow [216]			●
AFLOW-ML [217]	●	●	●
DeePMD-kit [218]		●	●
JARVIS-tools [193]	●	●	●
JAMIP [219]	●	●	●
ALKEMIE [220]	●	●	●
MatDeepLearn [68]			●
MAGUS [221]		●	●
MAST-ML [222]	●	●	●
MatLearn [223]			●

procedures results in low user-friendliness and reusability. These issues have led to a demand for ML workflows applicable to materials science, prompting the development of relevant tools for users to conveniently build their own models or use developers' semi-finished products. These tools include related software, programs, frameworks, and more.

Among the tools listed in Table 5 for model construction or usage, some are representative. For instance, DeePMD-kit [218] is an open-source software package for building deep learning models for molecular dynamics simulations. Its primary applications lie in materials science and chemistry, where it can be used to simulate large-scale material systems such as proteins and solid-state materials. It provides a range of tools to assist users in building, training, and using deep potential models from scratch. Fung *et al.* [68] proposed a workflow and testing platform named MatDeepLearn. It is designed for the rapid, reproducible evaluation and comparison of GNNs (SchNet, MPNN, CGCNN, MEGNet, GCN) and other models (SOAP, SM) in predicting various properties on different datasets (bulk crystals, alloy surfaces, MOFs, 2D materials, Pt clusters). AFLOW-ML [217] overcomes the high threshold of ML by simplifying the ML methods developed by the AFLOW consortium. This framework provides an open RESTful API that allows direct access to continually updated algorithms, which can be seamlessly integrated into any workflow to predict electronic, thermal, and mechanical properties. JAMIP, developed by Zhao *et al.* [219], includes high-throughput materials calculation as its core, along with data generation, management tools, data storage, ML, and data mining modules. It provides toolkits for the discovery and design of new materials based on functional materials big data and AI ML algorithms. The ML module encompasses data cleaning, feature engineering, model construction, and model evaluation for common ML algorithms. ALKEMIE, developed by Wang *et al.* [220], achieves data generation through high-throughput calculations, data management, and data mining using ML models. Additionally, it includes a module for ML of cross-scale molecular dynamics potentials and a user-friendly interface, enhancing the operability of workflows.

APPLICATIONS OF ML METHODS IN MATERIALS DISCOVERY

Optoelectronic semiconductors refer to the materials that can respond to light and generate electron-hole pairs (charge carriers), thereby achieving photoelectric conversion. They are widely used in energy conversion, information, and electronic devices. Traditional optoelectronic semiconductors include Si [224] and GaAs [225], which are used for efficient solar cells and photodetectors, as well as CdTe used in thin-film solar cells [226]. New types of optoelectronic semiconductor materials, such as 2D materials and metal halide materials, have gained more attention. 2D materials, including graphene [227], black phosphorus [228], and 2D TMDs [229] represented by MoS_2, are of particular interest. These layered 2D materials can function as single-layer structures or be stacked together to form van der Waals (vdW) homogeneous structures or heterostructures [230]. The selectivity, interlayer distance, coupling strength, and interlayer twist angle of stacked materials provide rich tunability of the optoelectronic properties of 2D vdW materials. Suitable bandgaps, outstanding conductivity, and high optical absorption make them important for applications in photodetectors and photocatalysis [231]. Metal halide materials, especially the perovskite with the chemical formula ABX_3 [232], where A represents monovalent cations like $CH_3NH_3^+$ (MA^+), $CH(NH_2)_2^+$ (FA^+) and Cs^+; B for divalent metal cations like Pb^{2+} and Sn^{2+}; and X for halide ions: I^-, Br^-, and Cl^-, have also attracted significant attention. The replaceable components in the chemical composition of ABX_3 provide abundant adjustability to their optoelectronic properties [232–234]. In addition to ABX_3-type perovskites, other metal halides or chalcogenides containing octahedral or tetrahedral motifs also exhibit excellent optoelectronic properties [235], such as $Cs_2AgBiBr_6$ [236], CuAgSe [237], $BaZrS_3$ [238], and $MnGeO_3$ [239]. Due to their high light harvesting ability, long and balanced carrier diffusion length, high defect tolerance, high photoluminescence quantum yield, and readily tunable bandgap, they have broad application prospects in solar cells [240], light-emitting diodes (LEDs) [241], photodetectors [242], lasers [243].

The thermodynamic and kinetic stability, electronic bandgap, carrier effective mass, optical properties such as optical absorption and dielectric constant, of optoelectronic semiconductor materials are crucial for evaluating the performance of optoelectronic energy materials. In the past, the study of these properties was mainly discovered through experimental trial and error methods. However, experimental methods are often inefficient and difficult to quickly discover new materials or modulate material properties. DFT methods can theoretically simulate and predict the above material properties, and have been widely used in the study of optoelectronic semiconductor materials, as summarized by Luo *et al.* [19]. Moreover, by harnessing computational simulation methods to gather data and subsequently leveraging ML techniques, we are now able to rapidly predict the aforementioned material properties, thereby achieving the discovery of new materials or the modulation of their properties. We will discuss in detail the use of ML methods for the design of optoelectronic semiconductor materials in subsequent contents. For example, Cai *et al.* [63] predicted the formation energy of perovskites using ML to assess their thermodynamic stability, and then screened for stable semiconductor materials by predicting bandgaps. Based on DFT methods, further calculations were conducted on the band structures, excitonic effects, and molecular dynamics simulations of candidates. Finally, experimental synthesis or literature retrieval was used to validate the stability and optoelectronic properties of newly discovered materials. This is the general process of discovering or modulating optoelectronic semiconductor materials by combining DFT, ML, and experimentation. The accuracy of ML models, the number of candidates, the completeness of screening criteria, and whether they are experimentally validated are criteria that determine the excellence of such studies. However, most of the current studies using DFT + ML for virtual screening of materials only provide potentially synthesizable new materials and their DFT properties, without experimental validation, due to the expensive nature of experimental methods.

Some optoelectronic semiconductor materials that are designed using various ML techniques and then validated by experiment synthesis are summarized in Table 6. These materials include inorganic perovskites, hybrid organic-inorganic perovskites (HOIPs), metal halides, metal sulfides, and metal oxides, and used as photocatalytic materials, mid-infrared (IR) nonlinear optical (NLO) materials, ultraviolet (UV)-light emit-

Table 6 Optoelectronic semiconductor materials that are designed using ML techniques and then validated by experiment synthesis

Type of material	Composition	Application	ML method	The role of ML
Inorganic perovskites [246]	$Cs_2AgBiBr_6$	Photocatalytic materials	EXT	Predicting effective mass
Inorganic perovskite oxides [244]	$AgTaO_3$ $RbInO_3$ $NaOsO_3$	Multiple applications	Transfer learning DNN-CE	Predicting energy
HOIDPs [63]	$(CH_3NH_3)_2AgSbI_6$ $(CH_3NH_3)_2AgBiBr_6$ $(CH_3NH_3)_2TlBiBr_6$	Photovoltaic materials	GBRT	Predicting energy and bandgap
Mixed-cation HOIPs [154]	$MA_xDMA_{1-x}PbI_3$ $(1-x) = 0.0$ to 0.15	Photovoltaic materials	Deep learning XGBoost	Predicting the structural properties
Ternary metal sulfides [41]	$LiGaSe_2$ $KAlSe_2$	Mid-IR NLO materials	ATCNN	Predicting bandgap
Ternary metal halides [247]	K_2CuCl_3 K_2CuBr_3	UV-light emitting materials	Evolutionary algorithm Neural network	Predicting bandgap
Quaternary metal oxides [245]	Li_2MnSiO_5	Photovoltaic materials	GBRT	Predicting bandgap

ting materials, and so on. This highlights how ML methods can be applied to various types of optoelectronic semiconductor materials, and serves different purposes, being used to predict energies, bandgaps, effective masses, and structural properties. In detail, Cai *et al.* [63] utilized gradient boosting regression tree (GBRT) to predict the formation energy of HOIDPs. Then they screened the candidates with the chemical space of 25.9k by formation energy less than −0.2 eV/atom to obtain 17k candidates. They found that three compounds ($(CH_3NH_3)_2AgSbI_6$, $(CH_3NH_3)_2AgBiBr_6$, and $(CH_3NH_3)_2TlBiBr_6$) have been experimentally synthesized, demonstrating the effectiveness of ML screening process. Li *et al.* [244] employed transfer learning and neural networks to predict the formation energy of 5329 inorganic perovskite oxides, subsequently identifying 1314 thermodynamically stable candidates. Among these, 144 oxides were reported to be synthesized experimentally, including but not limited to the semiconductors $AgTaO_3$, $RbInO_3$, and $NaOsO_3$. There are four studies in Table 6 employed ML to predict the bandgap of materials. Determining bandgap is often a crucial step in screening for optoelectronic semiconductors. Both Cai *et al.* [63] and Davies *et al.* [245] utilized GBRT models for the screening of photovoltaic materials. The other two studies predicting the bandgap employed neural networks. Consistent with our previous discussion, ensemble models and neural networks each have their own advantages. Depending on the sample types and sample sizes, different models can be adopted. Although effective mass is more challenging to predict compared with energy and bandgap, Li *et al.* [246] used a relatively simple extra tree model to predict effective mass due to their limited dataset (only 31 data points). They found Zn^{2+} with a fully occupied d orbital as the optimal candidate for enhancing the electronic structures of $Cs_2AgBiBr_6$.

Furthermore, there are more cases of material design and discovery based on ML methods, and these approaches can be applied to semiconductor optoelectronic materials research. From a computational perspective, they provide insights and guidance into the modulation of chemical composition, crystal structure, and properties of functional materials. We further introduce and analyze the specific applications of ML methods in materials discovery. It includes predicting the stability and optoelectronic properties of materials, and using generative models for material inverse design.

Prediction of new stable materials by investigating stability

The stability of materials refers to the ability of a material to maintain its physical and chemical properties under specific conditions, representing the most fundamental property that materials should possess. For instance, in applications such as batteries, catalysts, electronic devices, the stability of materials directly influences the performance and lifespan of the devices [248]. We typically use thermodynamic stability and kinetic stability to measure whether a crystal is stable. The thermodynamic stability of a crystal can be assessed by the formation energy (E_f) and convex hull energy (E_{hull}), while the kinetic stability can be measured by whether the phonon spectrum has imaginary frequencies. We focus on ML for material stability learning, primarily including ML for predicting the thermodynamic stability, kinetic stability, synthesizability of materials, and the application of ML atomic potentials.

Thermodynamic stability

Formation energy is defined as the difference between the total energy of the crystal and the total energy of the corresponding elemental substance [249]. Generally, if $E_f > 0$ eV/atom for a crystal, it means that the crystal cannot be formed from its constituent elements. However, it is important to note that $E_f < 0$ eV/atom is only a necessary condition for the thermodynamic stability of a crystal, not a sufficient condition. Based on the predicted E_f, we can further construct the phase diagram of the material [179,223] and obtain E_{hull} of the material.

Ye *et al.* [250] were among the early pioneers attempting to predict E_f using neural networks. They used only two descriptors (electronegativity and ionic radius) to construct a relatively simple fully connected neural network (FCNN), predicting the formation energy of $C_3A_2D_3O_{12}$ garnets and ABO_3 perovskites with MAEs of 7–10 and 20–34 meV/atom, respectively. Because both garnets and perovskites have fixed structural prototypes, it is easy to build elemental features for atoms at different sites, ensuring the uniformity of feature lengths. Consequently, there is no need for convolution and pooling operations in the neural

network, allowing an FCNN to achieve good performance. ML for predicting E_f can not only operate independently but can also be combined with structural optimization algorithms (OAs). As shown in Fig. 10a, Cheng *et al.* [57] constructed the MEGNet for predicting E_f based on OQMD and Matbench. This model was then combined with OAs such as random search (RAS), particle swarm optimization (PSO), and Bayesian optimization (BO) to search for crystal structures with the minimum E_f. Comparative studies indicate that the GN (MatB)-BO model, trained by combining BO, can predict crystal structures with the best accuracy and extremely low computational cost.

Currently, state-of-the-art crystal GNN models perform well in predicting formation energy, with MAE ranging from 0.02 to 0.04 eV/atom after training on large databases [168]. However, they are dependent on the atomic coordinates of the structure. This dependence may lead them to bias towards learning and predicting stable ground state (GS) structures from the training set, significantly affecting the prediction accuracy of high-energy structures deviating from their relaxed states. To address this issue, Pandey *et al.* [251] balanced the training of GNN models on a combined dataset of GS and high-energy structures to accurately predict their total energy. The results show that the model achieved an overall MAE of 0.04 eV/atom on the combined dataset, which is comparable to models trained solely on ICSD. Models trained on the combined dataset improved the prediction accuracy for both ICSD structures and hypothetical structures, overcoming biases observed when training total energy models separately on each dataset. Similarly, Gibson *et al.* [67] addressed this issue through a data augmentation strategy. As shown in Fig. 10b, they perturbed the atomic coordinates of relaxed structures to generate additional training samples that describe the region around the minimum of the potential energy surface (PES). These perturbed structures were then mapped to the energy of the relaxed structures. Compared with models trained only on relaxed structures, models trained on the augmented dataset comprising both relaxed and perturbed structures significantly improved the prediction accuracy for unrelaxed structures.

While E_f is one of the indicators for assessing the thermodynamic stability of materials, as shown in Fig. 10c, even crystals with $E_f < 0$ eV/atom may be unstable because there might be more stable competing phases with lower E_f [252]. Therefore, E_{hull} is an assessment indicator for the thermodynamic stability. As illustrated in Fig. 10d, the convex hull is the lowest energy

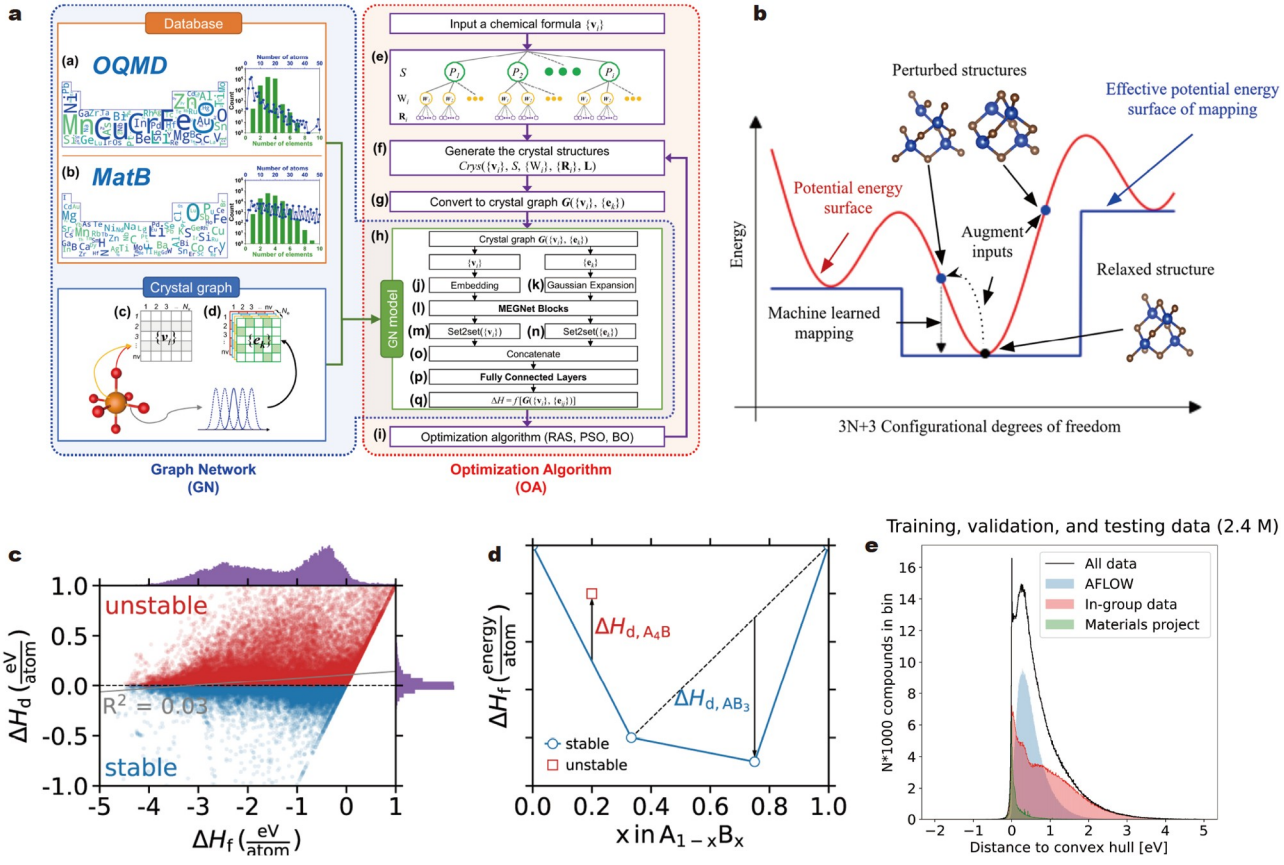

Figure 10 (a) Flowchart of GN-OA approach. Reprinted with permission from Ref. [57]. Copyright 2022, the Author(s). (b) The red line denotes a 2D representation of the continuous PES of materials. The blue line illustrates the effective PES, which describes the energy of a relaxed structure for a given unrelaxed input structure. The black circle indicates the relaxed structures contained in the dataset, and the blue circles symbolize artificially generated structures for the data augmentation. Reprinted with permission from Ref. [67]. Copyright 2022, the Author(s). (c) Illustration of the convex hull construction to obtain the decomposition enthalpy (ΔH_d), from the formation enthalpy (ΔH_f). Reprinted with permission from Ref. [252]. Copyright 2020, the Author(s). (d) ΔH_d shown against ΔH_f, for 85k ground-state entries in MP, indicating effectively no correlation between the two quantities. Reprinted with permission from Ref. [252]. Copyright 2020, the Author(s). (e) A total of 2.7 million calculations from AFLOW, MP and authors' group were accumulated and curated, leaving in the end 2.09 million data points. The histogram depicts the distance to the convex hull of the dataset. Reprinted with permission from Ref. [169]. Copyright 2021, the Author(s).

line formed by the E_f of all stable compounds (or phases). ΔE_{hull} refers to the difference between the formation energy of a compound and the formation energy of its corresponding point on the convex hull line. If ΔE_{hull} = 0 eV/atom (decomposition enthalpy ΔH_d < 0 eV/atom), then the compound lies on the convex hull line and is considered thermodynamically stable. However, it is important to note that, in practice, materials with E_{hull} < 36 meV/atom may be stable or metastable and may have synthesizability [253].

For ML tasks aiming to predict E_{hull}, the issue of dataset availability is often a challenge. On one hand, E_{hull} is relatively more challenging to obtain compared with E_f because it requires finding the decomposition phases of materials and obtaining their formation energies. On the other hand, as shown in Fig. 10e, the E_{hull} in the dataset often exhibits a non-normal distribution. We can easily obtain stable structures from material databases where their E_{hull} = 0 eV/atom, while it is difficult to obtain the E_{hull} of unstable structures. Because positive samples are readily added to the database, while negative samples are not considered for addition. Therefore, this can lead to an uneven numerical distribution of samples, with an abundance of samples with E_{hull} = 0 eV/atom in the dataset, which is clearly detrimental to building ML models [58,169,247].

Due to the uneven distribution of E_{hull}, Kim and Min [58] adopted a classification model to predict whether the E_{hull} of an $A_2BB'X_6$-type double perovskite halides is greater than 0 eV/atom. They used 145 elemental features and the space group number as descriptors. The model achieved an accuracy of 0.65, and tends to classify materials as unstable or metastable.

Moreover, some researchers have utilized regression models combined with special strategies to predict E_{hull}. For example, Schmidt *et al.* [169] used a dataset comprising 2.09 million entries to pretrain the crystal graph attention network, achieving a model MAE of 30 meV/atom. Subsequently, they transferred the model to mixed quaternary compounds for further training, and the test results demonstrated a significant improvement in the model's accuracy in predicting E_{hull}. Similarly, Choubisa *et al.* [247] also utilized a GNN along with a transfer learning strategy to achieve precise predictions of E_{hull}. They found that the model employing the transfer learning strategy performed best in predicting E_{hull} for unrelaxed crystals (MAE = 34 meV/atom). This model was fine-tuned by pretraining a GNN based on 500k E_f data from OQMD. Chen *et al.* [64] built a regression model PSO-SVR [254] to predict E_{hull}. They selected 2031 ABO_3-type compounds from MP and WebElements database as their dataset. Seven multi-scale descriptors, including 118 features, were established, with E_f also considered as a feature. The PSO algorithm was used for the initial parameter optimization of the SVR model to avoid the randomness of initial parameters. The PSO-SVR model achieved R^2 = 0.957 and RMSE = 0.087 eV.

Kinetic stability

Determining the dynamical stability is more challenging compared with thermodynamic stability. This is because the phonon spectrum involves the vibration of atoms, relative displacements, and non-harmonic effects. Manti *et al.* [37] built an XGBoost classifier to predict the dynamic stability of 2D materials. The dataset used was a subset of C2DB, consisting of 3212 materials. In addition to the radial distribution of the projected density of states (RAD-PDOS) fingerprint map, they considered a low-dimensional fingerprint composed of five features: Perdew-Burke-Ernzerhof (PET) electronic bandgap, E_f, DOS at the Fermi level (DOS at EF), E_{hull}, and the total energy of each atom in the unit cell. The commonality among these features is that they are all obtained from a single DFT calculation, making them much faster than computing phonon frequencies. The test results of the model showed that the AUC for 10-fold cross-validation was 0.9 ± 0.01. As shown in Fig. 11a, the model can be used to avoid performing expensive phonon calculations on materials that can be labeled as unstable by ML models, thereby accelerating the screening of dynamically stable materials.

Differing from Manti *et al.* [37], who used ML to predict phonon frequencies at the Γ point, Chen *et al.* [255] employed ML to predict the phonon DOS, a continuous attribute. Clearly, predicting continuous attributes from limited input information is more challenging than predicting low-dimensional outputs consisting of one or a few discrete points. As shown in Fig. 11b, they utilized the Euclidean neural network E(3)NN, which transforms the input 3D structural data into coefficients of a spherical harmonic function expansion. Subsequently, the data are processed and learned through a multi-layer neural network. The structural information of crystals is converted into a periodic graph, and atomic types are encoded as mass-weighted one-hot encoding. The target quantity learned and predicted by the model is the phonon DOS containing 51 scalars. The model predictions displayed excellent consistency with the real values on the test set. For 70% of the test samples, the relative error was below 10%. As depicted in Fig. 11b, the model-predicted spectra were consistent with the actual spectra. By using the model to predict the phonon DOS of 4346 new crystal structures, they further identified some high heat capacity new materials.

Synthesizability

The synthesizability of a material measures whether a material is likely to exist from the perspective of experimental synthesis. This evaluation metric was proposed because assessing the possibility of a material's existence based solely on thermodynamic stability might not be accurate. A compelling piece of evidence is the existence of metastable crystals in the ICSD, which have E_{hull} > 0 eV/atom but have been successfully synthesized experimentally [253]. Due to the complexity of the material synthesis process, factors other than thermodynamic stability, such as kinetic conditions, precursors, environmental conditions, and experimental parameters, also play a role in determining whether a material can be synthesized. This leads to situations where the thermodynamic stability and synthesizability of materials are not directly correlated in certain cases [256,257]. Therefore, some research endeavors focus on directly predicting the synthesizability of materials by ML models.

Jang *et al.* [107] employed a partially supervised classification model (PU learning) to predict the synthesis probability of crystal structures. The dataset was sourced from MP database, comprising 46k materials experimentally synthesized and cataloged in the ICSD treated as positive samples. Additionally, 77k virtual crystals were generated through DFT calculations. Since the synthesizability of these virtual crystals was undetermined, they were considered "unlabeled data". The training involved 100 rounds, resulting in the construction of 100 models. The final prediction was obtained by averaging the results of these 100 models, defining the average as the crystal similarity score (CLscore) ranging between 0 and 1, quantifying the synthesizability of a given crystal structure. Model testing revealed that

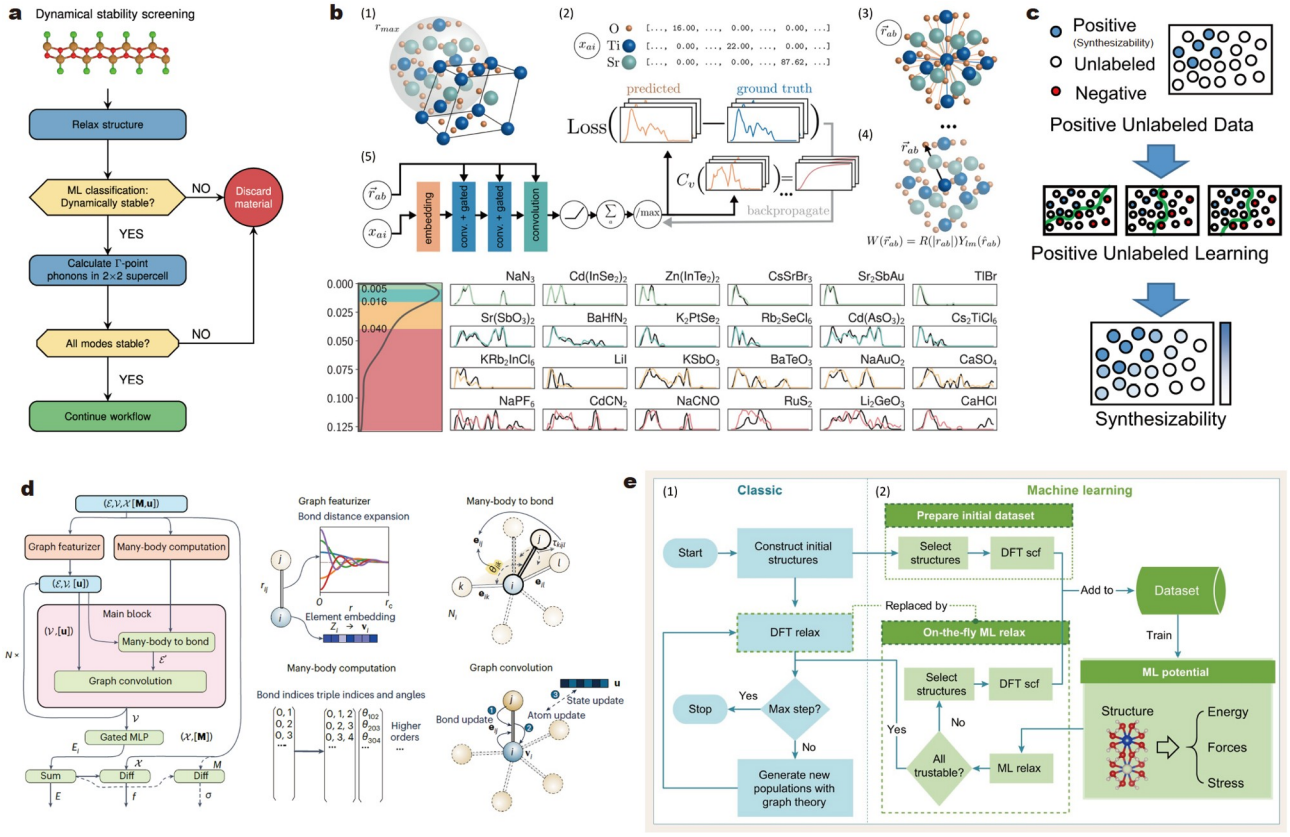

Figure 11 (a) ML classification algorithm can be used to filter out unstable crystals at a minimal computational cost. Reprinted with permission from Ref. [37]. Copyright 2023, the Author(s). (b) Overview of the E(3)NN architecture for phonon DOS prediction. And randomly selected examples in the test set within each error quartile. MSE distribution showing that it is heavily peaked in the 1st and 2nd quartiles with lower error. Reprinted with permission from Ref. [255]. Copyright 2021, the Author(s). (c) Positive and unlabeled learning (PU-learning) procedure overview. Reprinted with permission from Ref. [117]. Copyright 2022, the Author(s). (d) Schematic of the many-body graph potential and the major computational blocks of M3GNet. Reprinted with permission from Ref. [176]. Copyright 2022, the Author(s). (e) Workflow of MAGUS, which includes (1) classic evolutional algorithm and (2) machine-learning crystal structure predictions. Reprinted with permission from Ref. [221]. Copyright 2023, the Author(s).

87.4% of the experimentally synthesized materials were predicted as synthesizable by the model (CLscore > 0.5). To validate the predictions, they conducted a literature search for the top 100 materials in MP based on CLscore and found that 71 of them had been successfully synthesized and reported. The uniqueness of this PU learning method, relying solely on labeled data from ICSD, lies in its complete avoidance of DFT calculations for constructing the dataset, significantly saving time in data collection.

Gu *et al.* [117], after completing the aforementioned work, subsequently combined PU learning and transfer learning strategies. As shown in Fig. 11c, they further narrowed the focus of the material system to perovskite materials. The complete dataset for inorganic crystals, identical to the previous work, was used to train the source model. The perovskite dataset, consisting of 943 synthesizable positive samples and 11.9k virtual samples, was derived from MP, OQMD, and AFLOW, and it was used to train the transfer model. The evaluation results of the transfer model showed a true positive rate of 0.957, significantly higher than the previous work, highlighting the effectiveness of transfer learning in a specific domain.

IAP

It is deserved to point out that IAP is a model used to calculate the potential energy of interactions between atoms. In the prediction of material stability, IAPs play a crucial role and are employed in tasks such as crystal structure optimization and material dynamics simulations. The selection and accuracy of IAP functions directly impact the shape and precision of the PES. [258] While DFT calculations are often used to obtain accurate PES, they can be computationally expensive. ML IAPs constitute a category of models for atomic interactions constructed using ML methods. Compared with traditional empirical potentials or DFT, ML IAPs are more efficient, providing an accurate description of atomic interactions and enabling simulations of large-scale atomic systems. Therefore, developing ML IAP models that approach the accuracy of DFT calculations is an important direction in the application of ML methods in materials science [36,259,260].

Currently, the M3GNet, a universal IAP model proposed by Chen and Ong [176] for accurate assessment in crystal structure optimization, holds a leading position in this domain. In this work, the dataset comprises 187k energy data points, 16 million force data points, and 1.6 million stress data points from the MP. As depicted in Fig. 11d, the main modules of the GNN model M3GNet for three-body interactions include two crucial steps: the many-body bonding module and the standard graph convolution. The many-body bonding step computes new bonding information by considering the bonding environment and bond lengths of atoms. Similar to the ALIGNN model, this bonding

environment also incorporates information about bond angles between atoms. The standard graph convolution iteratively updates information about bonds, atoms, and optional states. Based on M3GNet, they trained 89 universal IAPs for elements in the periodic table with low errors in energy, force, and stress. In the case of IAP fitting, atomic information is mapped to atomic energy E_i, summed for total energy E, and then forces (f) and stresses (σ) are computed *via* automatic differentiation. Compared with DFT crystal structure optimization, the crystal structure optimization error for M3GNet IAPs is 0.035 eV/atom, and relaxation times approximately one-third of DFT. M3GNet is capable of accurately and rapidly relaxing arbitrary crystal structures.

Wang *et al.* [221] developed a crystal structure prediction framework called MAGUS, which combines ML potentials with structure search methods. The workflow includes a graph-theory-based classical evolutionary algorithm and a machine-learning algorithm based on ML potentials. MAGUS can be used to predict stable chemical compositions in chemical composition space or explore metastable structures with desired properties. The workflow is illustrated in Fig. 11e. After generating random structures, some of them are randomly selected for DFT single-point energy calculations to obtain energy, forces, and stresses, constructing a training set. The initial ML force field is then trained. In the subsequent search process, ML force field structure optimization is used to replace the most time-consuming DFT structure optimization. During optimization, structures that extrapolate are recorded, and if the extrapolation exceeds a specified threshold, DFT self-consistent calculations are performed, and the structures are added to the training set for retraining to correct the original ML potential. This iterative process continues to train the ML potential model until no more extrapolated structures are encountered.

Accelerating discovery of new semiconductors by optimizing optoelectronic properties

In addition to stability, functional optoelectronic semiconductors need to possess a range of properties applicable to their use in specific scenarios, such as appropriate bandgaps, small effective carrier masses, high optical absorption, and large dielectric constants [261,262]. ML methods could predict these properties to accelerate the discovery of new optoelectronic semiconductors.

Electronic bandgap

Bandgap determines the wavelength range of light that a material can absorb and convert. Semiconductors for different applications require different bandgap ranges. ML could be used to accelerate the screening of materials with suitable bandgaps [263,264]. Lu *et al.* [92] conducted an early study utilizing ML to predict the bandgap of perovskites. Employing high-throughput calculations, they compiled a dataset containing the bandgap information for 212 organic perovskites. Considering 14 material features, they developed a GBRT model. Subsequently, they used the trained model to predict the bandgap for 5158 candidates, ultimately selecting 218 ideal materials with bandgap in the range of 0.9–1.6 eV. Cai *et al.* [41] accelerated the discovery of nonlinear optical crystals using the ATCNN model based on chemical compositions. As shown in Fig. 12a, they predicted bandgap for 3887 ternary selenide chemical compositions, selecting 1620 materials with bandgap greater than 2.5 eV. Subsequently, crystal structure prediction methods were employed to determine stable crystal structures. Finally, high-throughput calculations were used to accurately screen bandgap and calculate the nonlinear-optical coefficients of candidates.

However, some characteristics of the bandgap contribute to the difficulty in its prediction:

Firstly, bandgap is not solely determined by the elemental composition of the material but is also influenced by subtle variations in the crystal structure. During the feature construction process, it is crucial to comprehensively consider both the elemental and structural features. For instance, Im *et al.* [153] constructed a GBRT model to predict the bandgap of $ABB'X_3$ perovskites. They found that the spacegroup of the crystal exhibited the highest feature importance, indicating its significant influence on the bandgap. Wang *et al.* [77], using a GBDT model, conducted a classification prediction for the bandgap of double perovskites. They discovered that the dipole polarization between B and B′ site cations had the most substantial impact on the bandgap.

Secondly, there are a significant number of materials with zero bandgap. The numerical distribution of the dataset often deviates from a normal distribution [58]. This deviation can hinder the model's learning and prediction capabilities. To address this issue, Saidi *et al.* [87] employed a hierarchical CNN (HCNN) approach when predict the bandgap of metal halide perovskites. This model consists of a classifier and a regression model, initially categorizing the predicted samples into six intervals before predicting the specific values within each interval. The model evaluation results showed that the HCNN method reduced the error of the standard CNN by three times, with a mean bandgap error of 0.14 eV. Similarly, Wang *et al.* [77] chose a classification model to handle the non-uniform distribution of bandgap. As shown in Fig. 12b, they collected bandgap for 1747 double perovskite materials from the MP as the dataset. They labeled the bandgap of materials in the dataset as 0 (0–1 eV), 1 (1–2 eV), 2 (>2 eV) to train a GBDT three-classification model. Using the model to predict 23k candidates, they filtered out 2711 ideal candidates with bandgap in the range of 1–2 eV. Similarly, Talapatra *et al.* [142] adopted a strategy of first classification and then regression when predicting the bandgap of double perovskite oxides. Using the model to predict 23k candidates, they filtered out 2711 ideal candidates with bandgap in the range of 1–2 eV.

Finally, we need to consider the differences between the PBE-bandgap, Heyd-Scuseria-Ernzerhof (HSE)-bandgap, and experimentally measured bandgap [263]. As shown in Fig. 12c, Chen *et al.* [84] used a Δ-ML method to predict the HSE bandgap of double HOIPs. They used HSE functional to calculate accurate bandgap for 1923 structures as the training set. They used the difference between PBE-bandgap and HSE-bandgap as the target value to train a Δ-GBRT model. This allows the model to predict the HSE-bandgap based on the PBE-bandgap of new materials, which is evidently easier than directly predicting the more complex HSE-bandgap. The validation results of the model showed that considering the anisotropy of organic cations could improve the model's accuracy. Using the model, they conducted high-throughput virtual screening on 78.4k DHOIPs, resulting in 19 promising DHOIPs. Mannodi-Kanakkithodi and Chan [56] also considered the importance of the HSE functional. As shown in Fig. 12d, they performed calculations using the HSE functional for 229 perovskites. Using

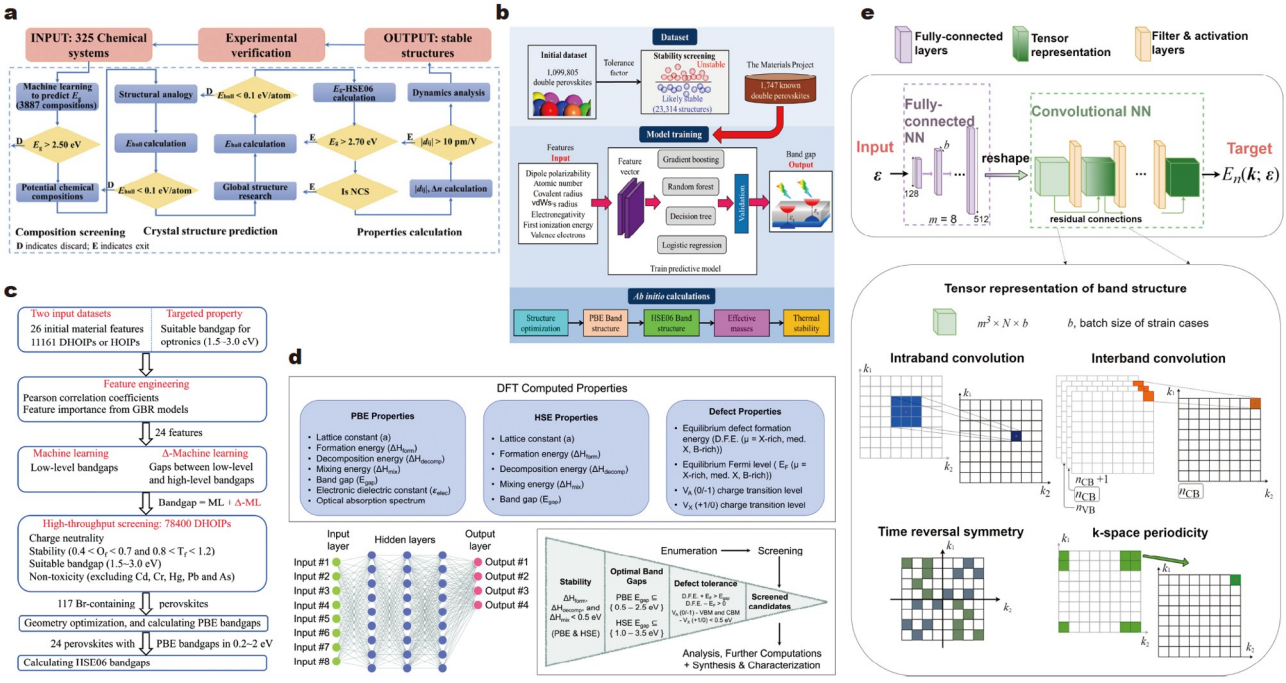

Figure 12 (a) Well-designed target-driven materials discovery workflow used for exploring novel promising mid-IR NLO materials. Reprinted with permission from Ref. [41]. Copyrtight 2022, Wiley-VCH GmbH. (b) ML workflow for identifying double perovskites. Reprinted with permission from Ref. [77]. Copyright 2022, American Chemical Society. (c) Framework for screening DHOIPs with combining ML models with DFT calculations. Reprinted with permission from Ref. [84]. Copyrtight 2022, Royal Society of Chemistry. (d) DFT properties computed for 229 perovskite compounds at the PBE and HSE06 levels of theory and screening performed on ML predicted dataset of 17k perovskite compounds in terms of their stability, bandgaps, and defect tolerance. Reprinted with permission from Ref. [56]. Copyrtight 2022, Royal Society of Chemistry. (e) CNN architecture for band structure prediction. The strain components are passed through fully connected layers, with the last layer reshaped into a rank-5 tensor. After a few convolutional layers with residual connections that improve convergence, the network produces the band structure as the output, which is fitted against the targeted DFT-computed band structure. A mesh comprising 8 × 8 × 8 *k*-points is used. And the tensor representation and physical insights incorporated into the CNN model: time-reversal symmetry, K-space periodicity, and inter-band and intra-band convolution. Reprinted with permission from Ref. [148]. Copyright 2021, the Author(s).

the average elemental properties of A, B, and X-site atoms or molecules as input descriptors, they constructed a neural network to predict HSE bandgap. Subsequently, they conducted high-throughput virtual screening on approximately 18k materials using the trained models, resulting in 392 stable candidates with appropriate bandgaps, defect tolerance, and photovoltaic quality factors.

Carriers effective mass

The effective mass of charge carriers in semiconductor materials is also an important optical and electronic property. It describes the inertia effect exhibited by charge carriers when accelerated in an electric or magnetic field. A lower effective mass is a screening criterion for optoelectronic materials, as it allows the material to achieve higher carrier mobility and greater carrier diffusion distance [265]. Typical materials such as graphene, with an effective mass of zero, exhibit giant intrinsic mobility [266]. However, to our knowledge, there has been scant exploration in ML concerning the prediction of effective mass of charge carriers.

As mentioned earlier, even though predicting bandgaps based on crystal composition and structure does not require an accurate reconstruction of the band structure, accurately predicting bandgaps remains challenging. Furthermore, the complexity of effective mass is significantly higher than that of bandgaps, whether obtained by constructing band structures to calculate band edge curvature for effective mass determination or by solving the semiclassical Boltzmann transport equation to obtain electronic transport properties [267]. This complexity results in very limited data available in existing databases, making ML predictions for effective mass more difficult than predictions for bandgaps.

Tsymbalov *et al.* [148] utilized ML methods to reconstruct the band structure and subsequently predict effective mass. They employed a CNN as the main framework of their model, taking strain tensors and band structures as input to predict the band structure changes under strain (Fig. 12e). The CNN, by convolving matrices associated with intraband for neighboring *K*-points and interband transitions for the same *K*-points, learned the intricate correlation between the energy eigenvalues related to intraband and interband transitions. This allowed the model to effectively capture the information within the band structure. The model also considered crucial physical properties such as crystal periodicity and time-reversal symmetry. Using the impact of strain on the diamond band structure as an example, they demonstrated that the model could accurately predict the band structure of diamond under elastic strain, thereby obtaining bandgap and effective mass.

Optical properties

Some studies have used shallow and ensemble models to predict the optical properties of materials, accelerating the screening and research of new optoelectronic semiconductor materials. For instance, Choudhary *et al.* [268] constructed a GBT model using

classical force field descriptors, totaling 1557 dimensions, to predict materials' infrared intensities, Born-effective charges, piezoelectric, and dielectric tensors (Fig. 13a). They opted for a classification model for a qualitative assessment of dielectric tensors, achieving an ROC AUC of 0.93. Using the established ML model, the authors rapidly screened materials in a large database, identifying 32k materials with high dielectric constants (>20). Takahashi *et al.* [269] performed DFT calculations for 1226 metal oxides to obtain their dielectric constants. Subsequently, they developed two RF models to predict electronic and ionic dielectric constants (Fig. 13b). They found that structural information was not crucial for electronic dielectric constants. On the other hand, in the model predicting ionic dielectric constants, structural descriptors significantly enhanced the prediction accuracy more than electronic contributions.

The complexity of optical properties and the scarcity of samples collectively present challenges for predicting optical properties. As a result, some studies have adopted transfer learning strategies. Kong *et al.* [105] extracted 84k metal oxides from the materials experiment and analysis database (MEAD) as their dataset, and used only the chemical composition of materials as descriptors to predict optical absorption. They constructed a model named H-CLMP (hierarchical correlation learning for multi-property prediction) that combines VAE, attention mechanisms, and transfer learning strategies (Fig. 13c). They divided the optical absorption of materials into 10 segments based on photon energy range (1.39–3.11 eV). Comparing with previous chemistry-based models like ElemNet, CrabNet, and Roost, they found that the H-CLMP(T) model, leveraging transfer learning strategies, achieved the lowest average MAE (0.428). They employed this model to predict the optical absorption of 129k new three-cation metal oxide compositions. Dong *et al.* [270] proposed an optical material composition inverse design framework based on transfer learning and global optimization. The framework (Fig. 13d) aims to predict the optical absorption spectra of metal oxide materials based on their chemical formulas. The database comprises a total of 178k samples composed of 42 different elements, along with their corresponding optical absorbance values at 220 energies ranging from 1.32 to 3.2 eV. Materials are represented using 132-dimensional Magpie features. The neural network model achieved a performance with $R^2 > 0.99$. The inverse design of materials, based on genetic algorithms and Bayesian OAs, allows the determination of element categories and stoichiometry by inversely analyzing the absorption spectrum.

Structure-properties relationship

During the process of predicting material properties using ML methods, we also usually construct material structure-property relationships and uncover the underlying physical-chemical principles that govern materials by the approaches shown in Fig. 14.

Feature importance is the most popular method to construct material structure-property relationships [271]. It is obtained

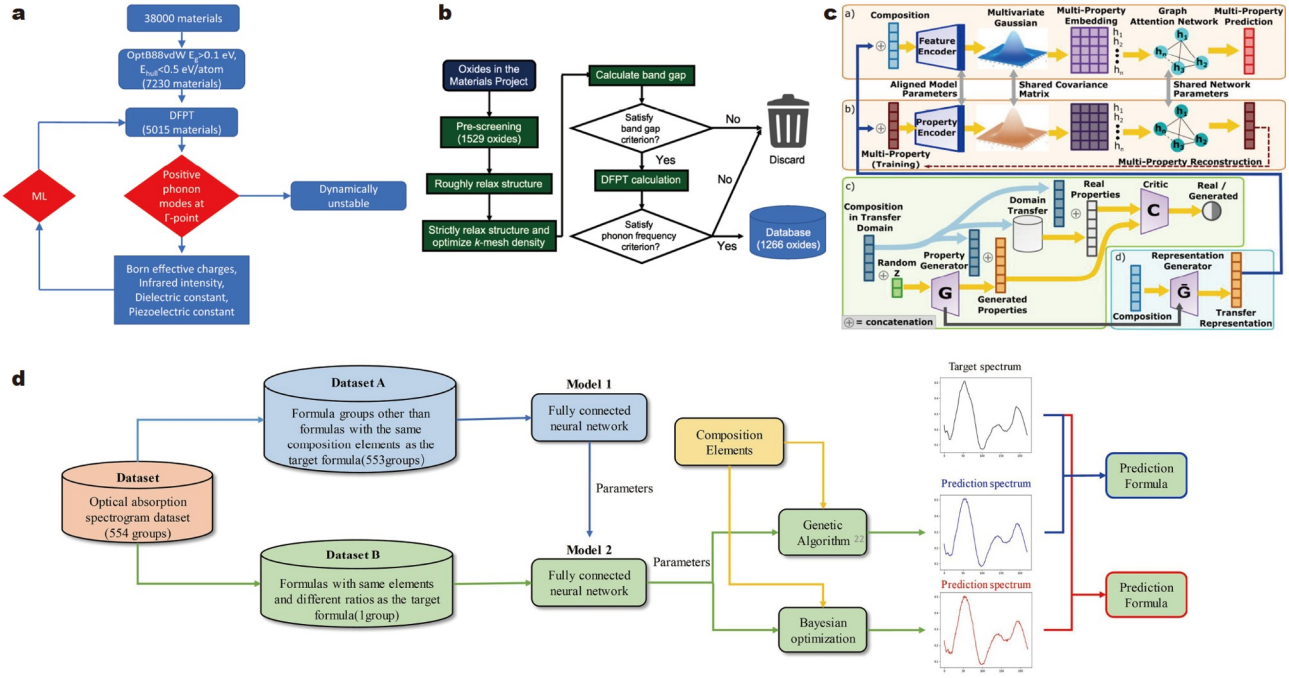

Figure 13 (a) Flow-chart portraying different steps for the DFT and ML methods. Reprinted with permission from Ref. [268]. Copyright 2020, the Author(s). (b) Workflow for constructing a computational database of the dielectric constants of oxides. Reprinted with permission from Ref. [269]. Copyright 2020, American Physical Society. (c) The H-CLMP(T) framework. Components a) and b) are jointly trained parallel models for multi-property prediction and multi-property reconstruction, respectively, where component b) is a variational auto-encoder. Component a) performs the desired multi-property prediction task, while the multi-property reconstruction of component b) facilitates training of component a). Training and deployment of transfer learning are achieved by components c) and d), respectively. Reprinted with permission from Ref. [105]. Copyright 2021, AIP Publishing. (d) The framework composed of an FCNN-based transfer learning model trained with Magpie features and global optimization based search model including a genetic algorithm and a BO. Firstly, they use large amounts of known data (Dataset A) for initial training of the model 1. Then the authors transfer parameters of model 1 to model 2 (with the same type as model 1), and use a small amount of sample data (Dataset B) to fine-tune the model. Reprinted with permission from Ref. [270]. Copyright 2020, Elsevier.

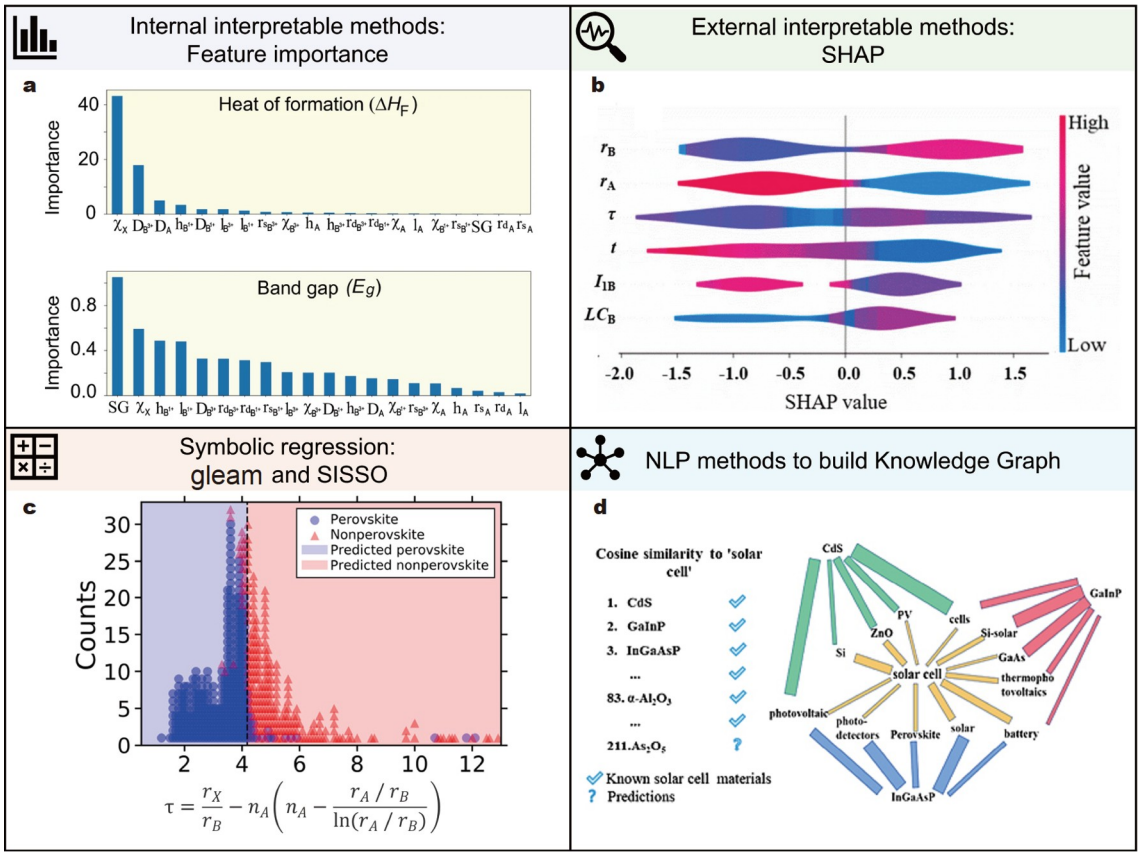

Figure 14 Four ML methods for constructing relationships between material components (structures) and properties. (a) Feature importance from GBRT for c heat of formation and d bandgap of halide double perovskite. Reprinted with permission from Ref. [153]. Copyright 2019, the Author(s). (b) SHAP value distribution. Reprinted with permission from Ref. [273]. Copyright 2021, American Chemical Society. (c) Assessing the performance of the improved tolerance factor. τ achieves a classification accuracy of 92% on the set of 576 ABX_3 solids based on perovskite classification for $t < 4.18$, with this decision boundary identified using a one-node decision tree. Reprinted with permission from Ref. [279]. Copyright 2019, the Author(s). (d) Output list of solar cell materials predicted by the ML method. The ranking is based on the absolute value of cosine similarity between the word vectors of the chemical formula and the solar cell. Reprinted with permission from Ref. [97]. Copyright 2022, AIP Publishing.

through specific algorithms associated with the model, such as the tree-based models that measure the importance of features by the number of splits or the reduction in split criteria in the tree. By constructing a GBRT model to predict E_f and bandgap of $ABB'X_3$ perovskites, Im *et al.* [153] discovered that the average Pauling electronegativity of materials has the highest feature importance (Fig. 14a). On the other hand, bandgap model shows that the spacegroup has the highest feature importance. In most cases, transitioning from a cubic to a monoclinic space group leads to an increase in bandgap. This indicates that E_f is mainly influenced by element features, while bandgap is more sensitive to the structure. In addition, Liu *et al.* [147] found that the tolerance factor has the highest feature importance when constructing a classification model for perovskite materials, proving it to be a primary factor controlling the formability of perovskites. Wang *et al.* [77] used a GBDT model for the classification prediction of bandgap for double perovskites. They discovered that the differences in dipole polarization between B and B′ site cations, covalent radius, valence electron count, and the variance in Pauling electronegativity have a greater impact on the model's accuracy.

Similar to feature importance, the Shapley Additive exPlanations (SHAP) method also constructs structure-properties relationships by revealing the extent to which features influence the target [33,63,88–90]. It is based on cooperative game theory, treating features in ML as participants and the model's predictive results as cooperative outcomes [272]. By calculating the average contribution of each feature to different subsets of features, Shapley values for each feature are obtained, measuring the relative importance of features for model predictions. The advantage of SHAP is that it can visualize the positive and negative impacts of feature changes on predictions through forms like bar charts and waterfall plots. Zhang *et al.* [273] used a dataset of 102 samples (44 HOIPs and 58 non-HOIPs) to build an XGBoost classification model to differentiate the formability of HOIPs. As shown in Fig. 14b, SHAP analysis indicated that the radius and lattice constant of the B site are positively correlated with the formability of HOIPs, while the ionic radius of the A site, tolerance factor (t), and the first ionization energy of the B site are negatively correlated with formability.

Through symbolic regression method, we can construct mathematical equations describing the relationship between material features and target properties, allowing for a quantitative analysis of their relationship. The basic idea of symbolic regression methods involves performing mathematical operations on features to construct descriptors and mathematical expressions used to calculate the target quantity. Symbolic regression tools applied in materials science include gplearn

[40,274,275] and SISSO [210,276,277]. Gplearn is based on genetic programming. It evolves and optimizes mathematical expressions represented as tree structures through genetic algorithm operations like crossover, mutation, and selection to find the most accurate mathematical expression. Weng *et al.* [278] used a gplearn method to propose descriptors for describing the catalytic activity of perovskite oxide catalysts. Among numerous results, they found that μ/t (μ and t are the octahedral and tolerance factors) is the optimal compromise between complexity and accuracy. This concise descriptor indicates that smaller μ and larger t should lead to higher OER activity. SISSO is also a popular symbolic regression method that, by introducing the concepts of sparsity and operator selection, can select the most important features and operators from a large number of candidate mathematical expressions. Bartel *et al.* [279] used SISSO to improve the well-known Goldschmidt tolerance factor (t) for predicting the stability of perovskite structures, significantly enhancing the accuracy of predicting perovskite stability (Fig. 14c). The new descriptor τ, in addition to geometric constraints such as the ionic radii of atoms at the three sites (A, B, and X) in perovskites, also includes chemical information, specifically the oxidation state of the A-site atom. The false positive rates for τ and t are 11% and 51%, respectively, indicating that the main advantage of τ over t is a significant reduction in predicting compounds as perovskites that have not been experimentally confirmed as stable perovskites.

Constructing a KG of materials using ML is another way to build the structure-property relationships of materials. A materials KG is a representation method that organizes knowledge into a graph structure [280,281]. It models and connects information about material components, structures, properties, relationships, or researchers in a graphical way, forming a structured knowledge network. For instance, Zhang and He [97] conducted data mining on 50k materials science papers using NLP method (Fig. 14d). They used ChemDataExtractor for tokenization in the abstract database, identified material names in the named entity recognition step, and built a model using word2vec to establish relationships between material names and their applications. In text mining centered around the term "solar cell", the model unsupervisedly output well-known materials for solar cells. In addition to commonly reported solar cell materials in the literature, the model also predicted several uncommon materials. By performing first-principles calculations on the optical and electronic properties of candidates, they discovered a new solar cell material, As_2O_5.

Inverse design of new materials by generative models

In general, the inverse design of materials can be broadly defined as the process of designing new materials with desired properties starting from the expected material performance. Traditional materials design often involves iterative improvement of material properties through trial and error and experimentation. In contrast, the inverse design approach starts with the desired material properties and employs computational and simulation methods to identify materials that exhibit these properties. Under this broad definition, we can consider high-throughput computational methods for crystal materials, structure search methods, and the previously mentioned use of ML models for high-throughput virtual screening of materials as inverse design methods for materials (Fig. 15a) [282–284]. This is because these methods also begin by specifying the desired material properties and then identify new materials that meet these properties. However, if we adopt a more stringent and narrower definition, we might consider materials inverse design as the direct deduction of specific materials from their properties, without the need for calculations or predictions to determine their properties, effectively avoiding the challenges present in the forward design process of materials. Therefore, under this definition, we can confine direct inverse design of materials to using generative models to obtain new materials that satisfy specified properties [49,110,185].

Using generative models for the inverse design of materials has advantages in terms of speed compared with traditional structure search methods, high-throughput computational methods, and ML screening methods. Structure search methods do not eliminate the need for property evaluation (calculation) of generated structures, involving a cumbersome and time-consuming process of structure generation, structure evaluation, and structure optimization updates. High-throughput computational methods require substantial computational support. ML screening methods typically involve predicting the properties of unexplored materials obtained through element substitutions. This necessitates a rational assessment of the new materials, and the discovered materials are often constrained by inherent crystal structure prototypes. In contrast, for generative models, although the generated crystals may only partially meet expectations due to the accuracy of the current material generation models, the generation of structures and the determination of properties occur simultaneously through sampling in latent space. This is due to the data-driven nature, resulting in a fundamental difference from traditional materials inverse design methods.

Some generative models have been introduced in the previous chapter. iMatGen and FTCP both consider conditional constraints by incorporating the loss function of a target property learning model outside of the VAE into the loss function. Cond-DFC-VAE further enhances the latent space by training a classification model on the latent vectors of input crystals. CDVAE introduces conditional constraints in the denoising process of the sampled element types and atomic quantities from the latent space, aiming to make the sampled results meet initial expectations. In summary, these models consider conditional constraints through various strategies, allowing VAE models to generate new samples from the latent space that closely meet the properties possessed by the materials in the training set.

Some studies have already employed generative models for the inverse design of materials with specified properties [285–287]. For instance, Lyngby and Thygesen [285] used a dataset of 2615 2D materials from C2DB with convex hull $\Delta H_{\text{hull}} < 0.3$ eV/atom for training the CDVAE. As shown in Fig. 15b, they generated 5k new materials, followed by DFT structure relaxation to remove duplicate and non-2D structures, resulting in 3k new 2D structures. The probability of successful relaxation for these structures was 69%. Among the successfully relaxed materials, 73.8% had $\Delta H_{\text{hull}} < 0.3$ eV/atom. Wines *et al.* [287] set the optimization target property for CDVAE as the superconducting transition temperature (T_c). As shown in Fig. 15c, they trained the CDVAE model using 1058 superconducting materials from the JARVIS-SC database. The model was then used to generate 3k new superconductors with unique structures and chemical compositions. High-throughput virtual screening of the mate-

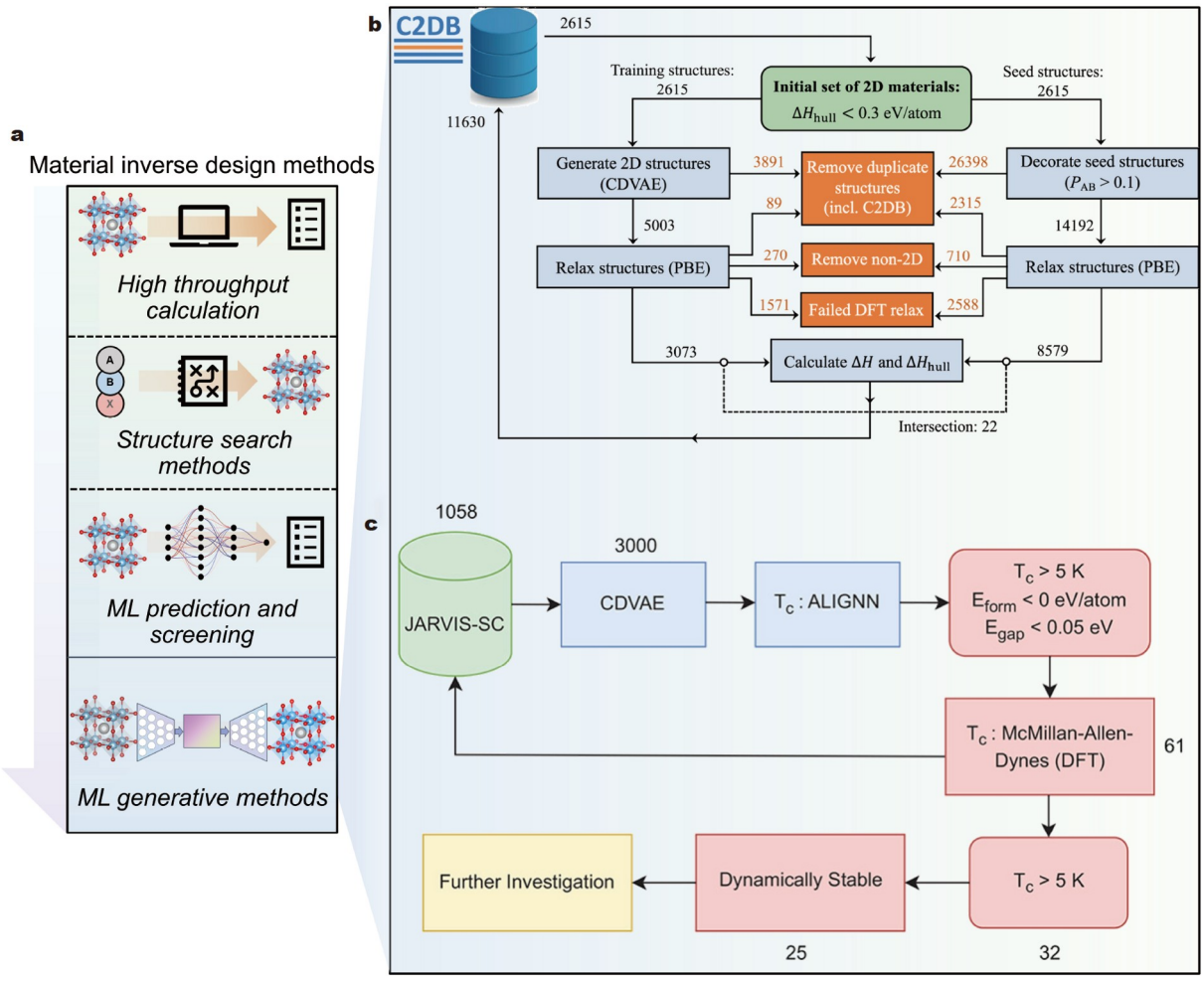

Figure 15 (a) Development of material inverse design methods. Using data-driven generative models to generate desired materials is the current advanced method. (b) Workflow to generate 2D candidates using the CDVAE generative model (left branch) and lattice decoration (right branch). The same set of 2615 materials is used to train the CDVAE model and as seed structures for lattice decoration, respectively. Black numbers indicate the number of materials present at a given step of the workflow while orange numbers indicate the number of materials discarded. Reprinted with permission from Ref. [285]. Copyright 2022, the Author(s). (c) Full inverse design workflow for new superconductors using DFT, ALIGNN, and the CDVAE generative model. Reprinted with permission from Ref. [287]. Copyright 2023, American Chemical Society.

rials was performed using the ALIGNN, with properties filtered based on the thresholds $T_c > 5$ K, $E_f < 0$ eV/atom, $E_{gap} < 0.05$ eV. This led to 32 candidates with $T_c > 5$ K. These materials were not present in the Supercon database, demonstrating that the CDVAE method can generate unique materials with specific desired properties, covering previously undiscovered regions in the phase space.

Clearly, although these works utilize generative models for the inverse design of materials, it still has not eliminated the need for additional forward screening work. This is because, on the one hand, a qualified functional material needs to satisfy multiple properties, and current generative models can only consider a single condition constraint. After obtaining new structures generated by the model, various property screenings are still required to meet practical purposes. On the other hand, similar to ML models used for target value prediction, generative models also need to be assessed for accuracy because the new materials produced by current generative models may have a certain rate of substandard quality. This necessitates additional efforts to exclude substandard materials from the generated set.

SUMMARY AND DISCUSSION

In this review, we have elucidated and examined the methodologies of ML models within computational materials science, delineating the capabilities and functionalities of these models. Certain methodologies have exhibited remarkable efficacy and displayed promising prospects, notably crystal GNNs and generative models.

The construction of ML models typically commences with data collection, involving the extraction of material data from databases or employing high-throughput computing techniques. Depending on factors such as data size, the complexity of prediction targets, and the necessity for result interpretation, researchers may opt for shallow models, ensemble models, or delve into deep neural networks. These models facilitate predictions concerning material stability (formation energy, convex hull energy, and synthesizability), as well as optoelectronic properties (bandgap, effective mass, optical absorption, and dielectric constant), and enable materials inverse design. Moreover, the interpretability of these models facilitates exploration into the underlying relationships between material composition,

structure, and properties, thereby offering valuable insights into physical chemistry. In summary, data-driven ML methodologies have emerged as pivotal tools in computational materials science, significantly expediting the development of novel materials.

Nevertheless, it is imperative to recognize that challenges persist in employing ML methodologies for the design and discovery of semiconductor optoelectronic materials. These challenges encompass various facets, including augmenting the quantity and quality of training data, refining the precision of crystal representations to augment the model's learning capabilities, and effectuating material inverse design based on properties such as bandgap and effective mass. In summary, ongoing investigations into optoelectronic semiconductors utilizing ML methodologies confront significant hurdles, yet they also present abundant opportunities for delving deeper into the latent potential of ML in expediting correlational research within materials science.

Size of available data

Firstly, there is the widely discussed issue of data size. As ML methods continue to evolve, researchers are increasingly recognizing the importance of data [54,288]. Some attempts have shown that the error of deep learning models significantly decreases with the increase in the volume of training data [34,68,129]. Although we can extract data in the order of millions from computational materials databases, for some challenging-to-calculate properties such as HSE functional bandgaps, exciton binding energies, absorption coefficients, phonon spectra, high-order force constants, and nonlinear optical coefficients, it is still difficult to obtain large amounts of data from databases. Additionally, for special material families such as low-dimensional materials, cluster materials, hybrid organic-inorganic halide perovskites, and metal-organic frameworks (MOFs), there is currently no database of the same volume as 3D inorganic crystal materials, even though they are crucial [289,290]. These challenges urgently require researchers from different organizations to reach collaborative agreements, and share experimentally or computationally obtained results with consistent parameters, to collectively advance the construction of databases for special materials, while adhering to basic norms.

Another solution to address the data issue is to employ transfer learning and active learning strategies. Regarding transfer learning, as demonstrated by Frey *et al.* [44] and Chen and Ong [131], since all material properties originate from chemical composition and structure, pretraining a model on a sufficiently large dataset (such as the formation energy of 3D crystals) allows the model to learn universal basic physical and chemical knowledge. Subsequently, transferring the model to specific material families or prediction tasks for specific properties facilitates fine-tuning, alleviating the problem of small data set. Active learning, on the other hand, helps alleviate the pressure of data labeling by allowing us to focus limited resources on labeling the most valuable samples [124,148].

Data quality and data cleaning

Samples within computational materials databases exhibit varying levels of quality, stemming from factors such as adherence to precise or coarse computational standards, utilization of inappropriate calculation parameters, or inadvertent editing errors during database curation. For example, based on our observations, materials projects and OQMD contain numerous erroneous energy entries. As previously highlighted, low-quality data resemble noise and can severely impede the efficacy of ML models. Thus, it is paramount to mitigate the risk of drawing erroneous conclusions stemming from such noise.

On one hand, addressing the presence of incorrect samples within the database necessitates their identification and subsequent removal through meticulous data cleaning procedures. We contend that data cleaning includes cleaning feature set, cleaning materials themselves, and cleaning target values. Cleaning the materials themselves entails scrutinizing the accuracy of material composition and structure, while also assessing whether the samples align with the research objectives. Taking ICSD-2022.2 as a case in point, based on our observations, approximately 72% of the total 204k crystals were deemed unsuitable for our study, characterized by issues such as fractional occupation, duplication with other samples, or the inclusion of isotopes. The exclusion of these samples from the dataset contributes to its enhanced rationality. Cleaning the feature set entails techniques such as filling, deletion, normalization, and encoding. However, our primary focus lies on cleaning the target values (labels). A common approach is to visualize and statistically analyze the target values. This involves using visualization tools such as scatter plots, histograms, as well as statistical analyses like mean, standard deviation, and box plots to identify outliers and inconsistencies in the target values, and to check whether the data conforms to the anticipated physical and chemical laws. Outliers may signify the presence of noise and warrant careful further examination. Most ML endeavors scrutinize the distribution of chemical compositions, space groups, feature values, and target values to underscore the absence of significant noise in the dataset [58,63,84,87,92,153,169,291]. For example, as demonstrated by Kim and Min [58], the formation energy of double perovskite halides exhibits a bimodal distribution resembling a normal distribution, while the bandgap shows a step-like distribution with a higher frequency of samples possessing smaller bandgaps. Such anticipated numerical distributions tend to exhibit reduced levels of noise.

On the other hand, for samples lacking evident errors, a delicate balance must be struck between data quality and quantity. This stems from the fact that high-quality samples are frequently scarce. For instance, databases frequently abound with a multitude of unreliable PBE bandgaps, whereas scarce are the expensive HSE bandgaps or DFT + U energies. Such circumstances often necessitate a tailored analysis in accordance with the specific ML task at hand. Accordingly, depending on the intricacy of the composition, structure, and target values, appropriate ML models and quantities must be selected. Moreover, endeavors should be made to enhance data quality while ensuring an ample quantity of data.

Moreover, the matter of data quality underscores the necessity for transparency in databases. Database creators ought to ensure transparency regarding data sources and computational methodologies, while meticulously documenting the procedures for data collection and processing. This approach enables users to evaluate the quality and reliability of the data more effectively. In turn, users should endeavor to utilize multiple databases whenever feasible to mitigate the potential for biased conclusions stemming from reliance on a single database.

Data imbalance

Samples within databases often exhibit a bias towards positive outcomes or materials with exceptional performance. This departure from reality can result in flawed learning by models. The reason lies in the model's prediction range, which hinges on the range of target values within the training dataset. For example, when training a model to forecast energy levels using materials from a database, the model may incline towards underestimating the energy of novel samples, and consequently misclassifying them as stable. As previously noted, databases tend to prioritize stable materials with lower energies and those with documented superior performance, often overlooking others.

We contend that the following strategies can alleviate bias or overfitting stemming from an overly favorable training set in the model: (1) generating artificial negative samples and incorporating them into the dataset to facilitate data augmentation. For example, as demonstrated by Gibson *et al.* [67], perturb stable crystal structures to obtain crystals with slightly higher energies. This enhances the robustness of the model. (2) Replicating negative or underperforming samples within the dataset to amplify their significance, thereby mimicking real-world scenarios as closely as feasible. For instance, in the context of predicting bandgaps, if the dataset contains an insufficient number of samples with bandgaps equal to 0 eV, the model may struggle to learn the difference between metals and semiconductors. By duplicating these samples, the model may enhance its predictive capabilities. (3) Making full use of negative samples and promoting the establishment of a negative sample database. Large numbers of negative samples are often overlooked or discarded, even though significant resources are invested in obtaining them. For example, Shen *et al.* [292] computed 400k compounds, but ultimately only discovered fewer than 8k stable materials. Gan *et al.* [239] computed 21k ABC_3 chalcogenide compounds but only showcased 93 stable materials among them. Unstable materials, considered as negative samples, can be incorporated into the dataset to facilitate a more comprehensive understanding of material properties by the model. Furthermore, they serve the purpose of filtering out unsuitable candidates, thereby averting redundant verification efforts by different researchers.

Representation of materials with descriptors

Accurate representation of materials is also an active topic. It is addressing how to provide a detailed description of a material's chemical composition and crystal structure, ensuring both accessibility and invariance. This is particularly important for predicting the optical and electronic properties of complex semiconductor materials such as HSE bandgap and effective mass. As mentioned earlier, Ye *et al.* [250] were able to effortlessly enable the ML model to capture the relationship between perovskites and formation energies using only two elemental features. In contrast, Chen *et al.* [84] required complex feature engineering to enable the model to learn the difference between PBE bandgap and HSE bandgap. Obtaining more accurate material representations helps to further elucidate the structure-property relationships of optoelectronic semiconductor materials.

On one hand, manually constructed material descriptors can target specific material structures or properties, such as descriptors measuring the stability [58,63,92,147,293] and distortion [294–296] of perovskites, algorithms for the coordination number and motify of local crystal structures [297,298], differences in local structure properties [32], descriptors applicable to d-band centers for metal catalysts [299], descriptors suitable for defect and surface systems [300], and descriptors capturing the orientation and volume of organic molecules in HOIPs [84]. Additionally, Li *et al.* [244] focused on the central atom and the coordinating atoms in the local structure to construct crystal features. They referred to this strategy as the "center-environment" method. However, the drawback is the heavy reliance on domain experts' knowledge, and the development process involves significant uncertainty.

On the other hand, describing materials based on their chemical composition, while bypassing crystal structure, has the clear limitation of being unable to capture phase transitions in materials.

Feature extraction methods based on crystal structure graphs, though enabling automation and competitive model accuracy [68,166,301], also have some issues. Firstly, they are highly sensitive to crystal structures, requiring precise crystal structures as input. Gibson *et al.* [67] and Choubisa *et al.* [247] treated unrelaxed crystals as noise added to the training set for data augmentation, thus improving the model's accuracy in predicting unrelaxed crystals. Secondly, as observed by Gong *et al.* [165], when the length of the periodicity of crystal structures exceeds the length of the receptive fields of atoms, GNN models may fail to capture long periodicities. They attempted to enhance the model's performance by combining artificially designed material descriptors with graph networks.

Finally, when predicting complex material properties such as carrier effective mass, one can follow the approach of Tsymbalov *et al.* [148] by using underlying physics as input, such as band structures. In such cases, ML models may struggle to directly capture the relationship between crystal structures and target properties.

In conclusion, there is a need for more flexible and accurate material representations or strategies that circumvent existing drawbacks in material representation methods to further strengthen the model's understanding of materials.

Inverse design by generative models

Inverse design of materials represents an idealized methodology within the realm of new materials discovery and materials research fields. While the development and application of some generative models have made achieving this goal possible, numerous challenges still exist.

Firstly, the bottleneck is achieving accurate equivariant and reversible representations of materials. Invariance refers to the property that the representation of a crystal remains unchanged after operations like translation, rotation, and permutation. Reversibility implies a one-to-one mapping between the material and its representation, enabling the precise reconstruction of atomic positions from the representation. Generative models for organic molecules in drug development have achieved good results using SMILES and graph representations, gradually integrated with the industry [302–304]. In contrast to organic molecules, crystals exhibit periodicity, symmetry, and constraints imposed by the crystal cell, making generative models for crystal materials more challenging.

Secondly, there is controversy regarding model frame. VAE is simple and easy to build, while GAN is challenging to train but

generates more realistic samples. There are also diffusion models, which are popular in image generation and molecular design, simulating random diffusion processes to generate new samples. Xie *et al.* [101] were among the first to attempt this, drawing inspiration from NCSN [305] and proposing the CDVAE.

Finally, unlike supervised learning, where we can easily compare the predictive accuracy of different crystal graph network models through the model's MAE, the evaluation of unsupervised generative models is more complex and awaits standardization. We need to assess the quality of the latent space, including its continuity, which characterizes the richness of new samples that the model can generate, and its completeness, which indicates whether the latent space covers all the information in the training set. We also need to evaluate the quality of generated samples, including the efficiency of new samples (the proportion of samples where the chemical composition is electrically neutral, and the crystal is stable), novelty (the proportion of samples with components and structural prototypes different from known materials), repetition rate (the proportion of repeated samples generated), and conditional compliance rate (the proportion of samples that meet the specified constraints). Additionally, as mentioned by Türk *et al.* [184], the model's generalization ability, i.e., how well the model generalizes to new samples that are unlike the training examples.

Furthermore, just as we are about to finalize the manuscript, the material generative models have witnessed two major achievements. These are GNoME (Graph Networks for Materials Exploration) [9] developed by the Google DeepMind team and MatterGen [10] (diffusion-based generative model for designing stable inorganic materials) developed by the Microsoft team. As illustrated in Fig. 16a, GNoME consists of two frameworks: the structural pipeline creates candidates with similar structures to known crystal structures through crystal modifications, while compositional models predict stability without structural information. In both frameworks, models provide a prediction of energy, and a threshold is chosen based on the relative stability (decomposition energy) concerning competing phases. Through GNoME, DeepMind has discovered 2.2 million new materials. Among them, 736 have been independently synthesized in experiments, and approximately 380k estimated to be relatively stable new materials will be made openly accessible in the future. MatterGen, like CDVAE, is based on diffusion principles, generating samples by reversing a fixed destructive process through learning score networks (Fig. 16b). To imbue the generated stable materials with desired properties,

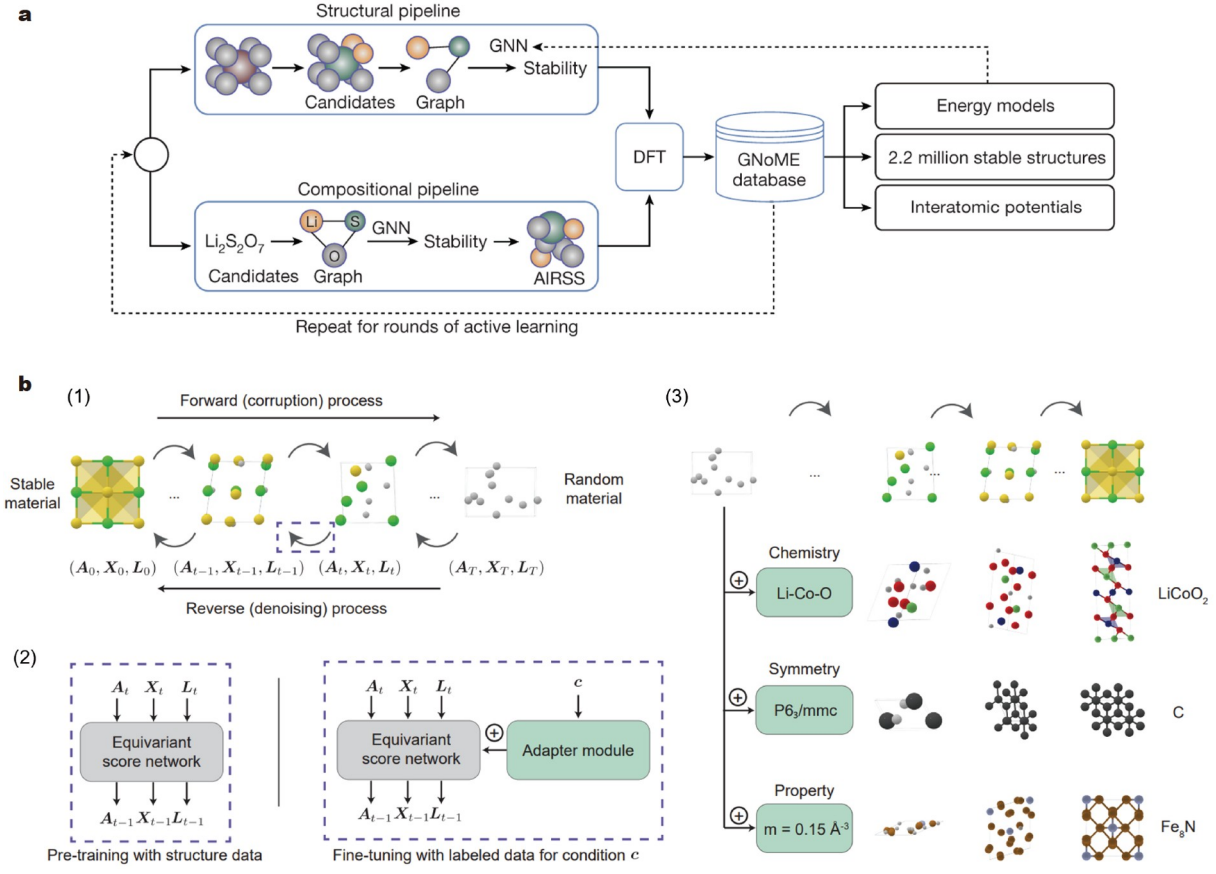

Figure 16 (a) A summary of the GNoME-based discovery shows how model-based filtration and DFT serve as a data flywheel to improve predictions. Reprinted with permission from Ref. [9]. Copyright 2023, the Author(s). (b) Inorganic materials design with MatterGen. (1) MatterGen generates stable materials by reversing a corruption process through iteratively denoising an initially random structure. The forward diffusion process is designed to independently corrupt atom types A, coordinates X, and the lattice L to approach a physically motivated distribution of random materials. (2) An equivariant score network is pretrained on a large dataset of stable material structures to jointly denoise atom types, coordinates, and the lattice. The score network is then fine-tuned with a labeled dataset through an adapter module that alters the model using the encoded property c. (3) MatterGen can be fine-tuned to steer the generation towards materials with desired chemistry, symmetry, and scalar property constraints. Reprinted with permission from Ref. [10]. Copyright 2023, the Author(s).

they have introduced adapter modules, which can be used to fine-tune the underlying model on an additional dataset with attribute labels, thereby achieving property constraints. MatterGen-MP has shown 1.8 times increase in the percentage of S.U.N. (stable, unique, and novel) structures and 3.1 times decrease in average root mean squared displacement compared with the previous state-of-the-art CDVAE. With these continuous developments, we are optimistic about achieving the true sense of material reverse design in the near future.

Received 15 January 2024; accepted 1 March 2024; published online 19 March 2024

1 OpenAI: Optimizing language models for dialogue. 2023. https://openai.com/blog/chatgpt/
2 Hey T, Trefethen A. The fourth paradigm 10 years on. Informatik Spektrum, 2020, 42: 441–447
3 LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521: 436–444
4 Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. Science, 2015, 349: 255–260
5 de Pablo JJ, Jackson NE, Webb MA, *et al.* New frontiers for the materials genome initiative. npj Comput Mater, 2019, 5: 41
6 de Pablo JJ, Jones B, Kovacs CL, *et al.* The materials genome initiative, the interplay of experiment, theory and computation. Curr Opin Solid State Mater Sci, 2014, 18: 99–117
7 Silver D, Huang A, Maddison CJ, *et al.* Mastering the game of Go with deep neural networks and tree search. Nature, 2016, 529: 484–489
8 Jumper J, Evans R, Pritzel A, *et al.* Highly accurate protein structure prediction with AlphaFold. Nature, 2021, 596: 583–589
9 Merchant A, Batzner S, Schoenholz SS, *et al.* Scaling deep learning for materials discovery. Nature, 2023, 624: 80–85
10 Zeni C, Pinsler R, Zügner D *et al.* MatterGen: A generative model for inorganic materials design. 2023. http://arxiv.org/abs/2312.03687
11 Schmidt J, Marques MRG, Botti S, *et al.* Recent advances and applications of machine learning in solid-state materials science. npj Comput Mater, 2019, 5: 83
12 Butler KT, Davies DW, Cartwright H, *et al.* Machine learning for molecular and materials science. Nature, 2018, 559: 547–555
13 Lejaeghere K, Van Speybroeck V, Van Oost G, *et al.* Error estimates for solid-state density-functional theory predictions: An overview by means of the ground-state elemental crystals. Crit Rev Solid State Mater Sci, 2014, 39: 1–24
14 Kresse G, Furthmüller J. Efficient iterative schemes for *ab initio* total-energy calculations using a plane-wave basis set. Phys Rev B, 1996, 54: 11169–11186
15 Giannozzi P, Baroni S, Bonini N, *et al.* QUANTUM ESPRESSO: A modular and open-source software project for quantum simulations of materials. J Phys-Condens Matter, 2009, 21: 395502
16 Curtarolo S, Hart GLW, Nardelli MB, *et al.* The high-throughput highway to computational materials design. Nat Mater, 2013, 12: 191–201
17 Perdew JP, Burke K, Ernzerhof M. Generalized gradient approximation made simple. Phys Rev Lett, 1996, 77: 3865–3868
18 Phillips JC, Braun R, Wang W, *et al.* Scalable molecular dynamics with NAMD. J Comput Chem, 2005, 26: 1781–1802
19 Luo S, Li T, Wang X, *et al.* High-throughput computational materials screening and discovery of optoelectronic semiconductors. WIREs Comput Mol Sci, 2021, 11: e1489
20 Bordonhos M, Galvão TLP, Gomes JRB, *et al.* Multiscale computational approaches toward the understanding of materials. Advcd Theor Sims, 2023, 6: 2200628
21 Shen L, Zhou J, Yang T, *et al.* High-throughput computational discovery and intelligent design of two-dimensional functional materials for various applications. Acc Mater Res, 2022, 3: 572–583
22 Jain A, Ong SP, Hautier G, *et al.* Commentary: The materials project: A materials genome approach to accelerating materials innovation. APL Mater, 2013, 1: 011002
23 Kirklin S, Saal JE, Meredig B, *et al.* The open quantum materials database (OQMD): Assessing the accuracy of DFT formation energies. npj Comput Mater, 2015, 1: 15010
24 Curtarolo S, Setyawan W, Hart GLW, *et al.* AFLOW: An automatic framework for high-throughput materials discovery. Comput Mater Sci, 2012, 58: 218–226
25 Himanen L, Geurts A, Foster AS, *et al.* Data-driven materials science: Status, challenges, and perspectives. Adv Sci, 2019, 6: 1900808
26 Ramprasad R, Batra R, Pilania G, *et al.* Machine learning in materials informatics: Recent applications and prospects. npj Comput Mater, 2017, 3: 54
27 Shen SC, Khare E, Lee NA, *et al.* Computational design and manufacturing of sustainable materials through first-principles and materiomics. Chem Rev, 2023, 123: 2242–2275
28 Bishara D, Xie Y, Liu WK, *et al.* A state-of-the-art review on machine learning-based multiscale modeling, simulation, homogenization and design of materials. Arch Computat Methods Eng, 2023, 30: 191–222
29 Bhat V, Callaway CP, Risko C. Computational approaches for organic semiconductors: From chemical and physical understanding to predicting new materials. Chem Rev, 2023, 123: 7498–7547
30 Singh V, Patra S, Murugan NA, *et al.* Recent trends in computational tools and data-driven modeling for advanced materials. Mater Adv, 2022, 3: 4069–4087
31 Pyzer-Knapp EO, Pitera JW, Staar PWJ, *et al.* Accelerating materials discovery using artificial intelligence, high performance computing and robotics. npj Comput Mater, 2022, 8: 84
32 Ward L, Liu R, Krishna A, *et al.* Including crystal structure attributes in machine learning models of formation energies *via* Voronoi tessellations. Phys Rev B, 2017, 96: 024104
33 Obada DO, Okafor E, Abolade SA, *et al.* Explainable machine learning for predicting the band gaps of ABX3 perovskites. Mater Sci Semiconductor Processing, 2023, 161: 107427
34 Schütt KT, Sauceda HE, Kindermans PJ, *et al.* SchNet—A deep learning architecture for molecules and materials. J Chem Phys, 2018, 148: 241722
35 Deringer VL, Csányi G. Machine learning based interatomic potential for amorphous carbon. Phys Rev B, 2017, 95: 094203
36 Zuo Y, Chen C, Li X, *et al.* Performance and cost assessment of machine learning interatomic potentials. J Phys Chem A, 2020, 124: 731–745
37 Manti S, Svendsen MK, Knøsgaard NR, *et al.* Exploring and machine learning structural instabilities in 2D materials. npj Comput Mater, 2023, 9: 33
38 Willhelm D, Wilson N, Arroyave R, *et al.* Predicting van der Waals heterostructures by a combined machine learning and density functional theory approach. ACS Appl Mater Interfaces, 2022, 14: 25907–25919
39 Wang T, Tan X, Wei Y, *et al.* Unveiling the layer-dependent electronic properties in transition-metal dichalcogenide heterostructures assisted by machine learning. Nanoscale, 2022, 14: 2511–2520
40 Loftis C, Yuan K, Zhao Y, *et al.* Lattice thermal conductivity prediction using symbolic regression and machine learning. J Phys Chem A, 2021, 125: 435–450
41 Cai W, Abudurusuli A, Xie C, *et al.* Toward the rational design of mid-infrared nonlinear optical materials with targeted properties *via* a multi-level data-driven approach. Adv Funct Mater, 2022, 32: 2200231
42 Dong SS, Govoni M, Galli G. Machine learning dielectric screening for the simulation of excited state properties of molecules and materials. Chem Sci, 2021, 12: 4970–4980
43 Banik S, Loeffler TD, Batra R, *et al.* Learning with delayed rewards—A case study on inverse defect design in 2D materials. ACS Appl Mater Interfaces, 2021, 13: 36455–36464
44 Frey NC, Akinwande D, Jariwala D, *et al.* Machine learning-enabled design of point defects in 2D materials for quantum and neuromorphic information processing. ACS Nano, 2020, 14: 13406–13417
45 Huang P, Lukin R, Faleev M, *et al.* Unveiling the complex structure-property correlation of defects in 2D materials based on high throughput datasets. npj 2D Mater Appl, 2023, 7: 6
46 Bhattacharya A, Timokhin I, Chatterjee R, *et al.* Deep learning

approach to genome of two-dimensional materials with flat electronic bands. npj Comput Mater, 2023, 9: 101
47 Zhao Y, Siriwardane EMD, Wu Z, *et al.* Physics guided deep learning for generative design of crystal materials with symmetry constraints. npj Comput Mater, 2023, 9: 38
48 Yan D, Smith AD, Chen CC. Structure prediction and materials design with generative neural networks. Nat Comput Sci, 2023, 3: 572–574
49 Anstine DM, Isayev O. Generative models as an emerging paradigm in the chemical sciences. J Am Chem Soc, 2023, 145: 8736–8750
50 Ren Z, Tian SIP, Noh J, *et al.* An invertible crystallographic representation for general inverse design of inorganic crystals with targeted properties. Matter, 2022, 5: 314–335
51 Zheng Z, Zhang O, Borgs C, *et al.* ChatGPT chemistry assistant for text mining and the prediction of MOF synthesis. J Am Chem Soc, 2023, 145: 18048–18062
52 Shon YJ, Min K. Extracting chemical information from scientific literature using text mining: Building an ionic conductivity database for solid-state electrolytes. ACS Omega, 2023, 8: 18122–18127
53 Smith A, Bhat V, Ai Q, *et al.* Challenges in information-mining the materials literature: A case study and perspective. Chem Mater, 2022, 34: 4821–4827
54 Xu P, Ji X, Li M, *et al.* Small data machine learning in materials science. npj Comput Mater, 2023, 9: 42
55 Jacobsson TJ, Hultqvist A, García-Fernández A, *et al.* An open-access database and analysis tool for perovskite solar cells based on the FAIR data principles. Nat Energy, 2022, 7: 107–115
56 Mannodi-Kanakkithodi A, Chan MKY. Data-driven design of novel halide perovskite alloys. Energy Environ Sci, 2022, 15: 1930–1949
57 Cheng G, Gong XG, Yin WJ. Crystal structure prediction by combining graph network and optimization algorithm. Nat Commun, 2022, 13: 1492
58 Kim J, Min K. Data-driven investigation of the synthesizability and bandgap of double perovskite halides. Advcd Theor Sims, 2022, 5: 2200068
59 Li XG, Blaiszik B, Schwarting ME, *et al.* Graph network based deep learning of bandgaps. J Chem Phys, 2021, 155: 154702
60 Damewood J, Karaguesian J, Lunger JR, *et al.* Representations of materials for machine learning. Annu Rev Mater Res, 2023, 53: 399–426
61 Gong W, Yan Q. Graph-based deep learning frameworks for molecules and solid-state materials. Comput Mater Sci, 2021, 195: 110332
62 Li S, Liu Y, Chen D, *et al.* Encoding the atomic structure for machine learning in materials science. WIREs Comput Mol Sci, 2022, 12: e1558
63 Cai X, Zhang Y, Shi Z, *et al.* Discovery of lead-free perovskites for high-performance solar cells *via* machine learning: Ultrabroadband absorption, low radiative combination, and enhanced thermal conductivities. Adv Sci, 2022, 9: 2103648
64 Chen L, Wang X, Xia W, *et al.* PSO-SVR predicting for the Ehull of ABO_3-type compounds to screen the thermodynamic stable perovskite candidates based on multi-scale descriptors. Comput Mater Sci, 2022, 211: 111435
65 Ma XY, Lewis JP, Yan QB, *et al.* Accelerated discovery of two-dimensional optoelectronic octahedral oxyhalides *via* high-throughput *ab initio* calculations and machine learning. J Phys Chem Lett, 2019, 10: 6734–6740
66 Zhu JJ, Yang M, Ren ZJ. Machine learning in environmental research: Common pitfalls and best practices. Environ Sci Technol, 2023, 57: 17671–17689
67 Gibson J, Hire A, Hennig RG. Data-augmentation for graph neural network learning of the relaxed energies of unrelaxed structures. npj Comput Mater, 2022, 8: 211
68 Fung V, Zhang J, Juarez E, *et al.* Benchmarking graph neural networks for materials chemistry. npj Comput Mater, 2021, 7: 84
69 Bischl B, Binder M, Lang M, *et al.* Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. WIREs Data Min Knowl, 2023, 13: e1484
70 Li Z, Yoon J, Zhang R, *et al.* Machine learning in concrete science: Applications, challenges, and best practices. npj Comput Mater, 2022, 8: 127
71 Artrith N, Butler KT, Coudert FX, *et al.* Best practices in machine learning for chemistry. Nat Chem, 2021, 13: 505–508
72 Ho SY, Phua K, Wong L, *et al.* Extensions of the external validation for checking learned model interpretability and generalizability. Patterns, 2020, 1: 100129
73 Baştanlar Y, Özuysal M. Introduction to Machine Learning. miRNomics: MicroRNA Biology and Computational Analysis. Totowa: Humana, 2013. 1107
74 Hoffmann F, Bertram T, Mikut R, *et al.* Benchmarking in classification and regression. WIREs Data Min Knowl, 2019, 9: e1318
75 Palanivinayagam A, El-Bayeh CZ, Damaševičius R. Twenty years of machine-learning-based text classification: A systematic review. Algorithms, 2023, 16: 236
76 Sebastiani F. Machine learning in automated text categorization. ACM Comput Surv, 2002, 34: 1–47
77 Wang Z, Han Y, Lin X, *et al.* An ensemble learning platform for the large-scale exploration of new double perovskites. ACS Appl Mater Interfaces, 2022, 14: 717–725
78 Loh W. Classification and regression trees. WIREs Data Min Knowl, 2011, 1: 14–23
79 Smola AJ, Schölkopf B. A tutorial on support vector regression. Stat Computing, 2004, 14: 199–222
80 Xu R, WunschII D. Survey of clustering algorithms. IEEE Trans Neural Netw, 2005, 16: 645–678
81 Yan S, Xu D, Zhang B, *et al.* Graph embedding and extensions: A general framework for dimensionality reduction. IEEE Trans Pattern Anal Mach Intell, 2007, 29: 40–51
82 Anowar F, Sadaoui S, Selim B. Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE). Comput Sci Rev, 2021, 40: 100378
83 Martinez AM, Kak AC. PCA *versus* LDA. IEEE Trans Pattern Anal Machine Intell, 2001, 23: 228–233
84 Chen J, Xu W, Zhang R. Δ-Machine learning-driven discovery of double hybrid organic–inorganic perovskites. J Mater Chem A, 2022, 10: 1402–1413
85 Venkatraman V. The utility of composition-based machine learning models for band gap prediction. Comput Mater Sci, 2021, 197: 110637
86 Yang X, Li L, Tao Q, *et al.* Rapid discovery of narrow bandgap oxide double perovskites using machine learning. Comput Mater Sci, 2021, 196: 110528
87 Saidi WA, Shadid W, Castelli IE. Machine-learning structural and electronic properties of metal halide perovskites using a hierarchical convolutional neural network. npj Comput Mater, 2020, 6: 36
88 Shen Y, Wang J, Ji X, *et al.* Machine learning-assisted discovery of 2D perovskites with tailored bandgap for solar cells. Advcd Theor Sims, 2023, 6: 2200922
89 Liu Y, Yan W, Han S, *et al.* How machine learning predicts and explains the performance of perovskite solar cells. Sol RRL, 2022, 6: 2101100
90 Rath S, Sudha Priyanga G, Nagappan N, *et al.* Discovery of direct band gap perovskites for light harvesting by using machine learning. Comput Mater Sci, 2022, 210: 111476
91 Kumar U, Mishra KA, Kushwaha AK, *et al.* Bandgap analysis of transition-metal dichalcogenide and oxide *via* machine learning approach. J Phys Chem Solids, 2022, 171: 110973
92 Lu S, Zhou Q, Ouyang Y, *et al.* Accelerated discovery of stable lead-free hybrid organic-inorganic perovskites *via* machine learning. Nat Commun, 2018, 9: 3405
93 Längkvist M, Karlsson L, Loutfi A. A review of unsupervised feature learning and deep learning for time-series modeling. Pattern Recognition Lett, 2014, 42: 11–24
94 Xie T, Grossman JC. Hierarchical visualization of materials space with graph convolutional neural networks. J Chem Phys, 2018, 149: 174111
95 Young T, Hazarika D, Poria S, *et al.* Recent trends in deep learning based natural language processing. IEEE Comput Intell Mag, 2018, 13: 55–75
96 Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language

processing: An introduction. J Am Med Inform Assoc, 2011, 18: 544–551
97 Zhang L, He M. Unsupervised machine learning for solar cell materials from the literature. J Appl Phys, 2022, 131: 064902
98 Huang S, Cole JM. A database of battery materials auto-generated using ChemDataExtractor. Sci Data, 2020, 7: 260
99 Dong Q, Cole JM. Auto-generated database of semiconductor band gaps using ChemDataExtractor. Sci Data, 2022, 9: 193
100 Vasylenko A, Gamon J, Duff BB, *et al.* Element selection for crystalline inorganic solid discovery guided by unsupervised machine learning of experimentally explored chemistry. Nat Commun, 2021, 12: 5561
101 Xie T, Fu X, Ganea OE, *et al.* Crystal diffusion variational autoencoder for periodic material generation. 2022, https://doi.org/10.48550/arXiv.2110.06197
102 Binks DJ, Dawson P, Oliver RA, *et al.* Cubic GaN and InGaN/GaN quantum wells. Appl Phys Rev, 2022, 9: 041309
103 Zhao Y, Al-Fahdi M, Hu M, *et al.* High-throughput discovery of novel cubic crystal materials using deep generative neural networks. Adv Sci, 2021, 8: 2100566
104 Lee IH, Chang KJ. Crystal structure prediction in a continuous representative space. Comput Mater Sci, 2021, 194: 110436
105 Kong S, Guevarra D, Gomes CP, *et al.* Materials representation and transfer learning for multi-property prediction. Appl Phys Rev, 2021, 8: 021409
106 Rigoni D, Navarin N, Sperduti A. Conditional constrained graph variational autoencoders for molecule design. 2020. http://arxiv.org/abs/2009.00725
107 Jang J, Gu GH, Noh J, *et al.* Structure-based synthesizability prediction of crystals using partially supervised learning. J Am Chem Soc, 2020, 142: 18836–18843
108 Dan Y, Zhao Y, Li X, *et al.* Generative adversarial networks (GAN) based efficient sampling of chemical composition space for inverse design of inorganic materials. npj Comput Mater, 2020, 6: 84
109 Court CJ, Yildirim B, Jain A, *et al.* 3-D inorganic crystal structure generation and property prediction *via* representation learning. J Chem Inf Model, 2020, 60: 4518–4535
110 Noh J, Kim J, Stein HS, *et al.* Inverse design of solid-state materials *via* a continuous representation. Matter, 2019, 1: 1370–1384
111 Jin W, Barzilay R, Jaakkola T. Junction tree variational autoencoder for molecular graph generation. 2019. http://arxiv.org/abs/1802.04364
112 Kipf TN, Welling M. Variational graph auto-encoders. 2016. http://arxiv.org/abs/1611.07308
113 Hu J, Li M, Gao P. MATGANIP: Learning to discover the structure-property relationship in perovskites with generative adversarial networks. 2019, https://doi.org/10.48550/arXiv.1910.09003
114 Zhou ZH, Li M. Semi-supervised learning by disagreement. Knowl Inf Syst, 2010, 24: 415–439
115 Han K, Chen W, Xu M. Investigating active positive-unlabeled learning with deep networks. In: Proceedings of AI 2021: Advances in Artificial Intelligence: 34th Australasian Joint Conference. Sydney: Springer-Verlag, 2022. 13151
116 Bekker J, Davis J. Learning from positive and unlabeled data: A survey. Mach Learn, 2020, 109: 719–760
117 Gu GH, Jang J, Noh J, *et al.* Perovskite synthesizability using graph neural networks. npj Comput Mater, 2022, 8: 71
118 Arulkumaran K, Deisenroth MP, Brundage M, *et al.* Deep reinforcement learning: A brief survey. IEEE Signal Process Mag, 2017, 34: 26–38
119 Świechowski M, Godlewski K, Sawicki B, *et al.* Monte Carlo tree search: A review of recent modifications and applications. Artif Intell Rev, 2023, 56: 2497–2562
120 Jensen JH. A graph-based genetic algorithm and generative model/Monte Carlo tree search for the exploration of chemical space. Chem Sci, 2019, 10: 3567–3572
121 Song Z, Zhou Q, Lu S, *et al.* Adaptive design of alloys for CO_2 activation and methanation *via* reinforcement learning Monte Carlo tree search algorithm. J Phys Chem Lett, 2023, 14: 3594–3601
122 Ureel Y, Dobbelaere MR, Ouyang Y, *et al.* Active machine learning for chemical engineers: A bright future lies ahead! Engineering, 2023, doi: 10.1016/j.eng.2023.02.019
123 Wen Y, Li Z, Xiang Y, *et al.* Improving molecular machine learning through adaptive subsampling with active learning. Digital Discov, 2023, 2: 1134–1142
124 Lookman T, Balachandran PV, Xue D, *et al.* Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. npj Comput Mater, 2019, 5: 21
125 Kim Y, Kim Y, Yang C, *et al.* Deep learning framework for material design space exploration using active transfer learning and data augmentation. npj Comput Mater, 2021, 7: 140
126 Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng, 2010, 22: 1345–1359
127 Zhuang F, Qi Z, Duan K, *et al.* A comprehensive survey on transfer learning. Proc IEEE, 2021, 109: 43–76
128 Jha D, Choudhary K, Tavazza F, *et al.* Enhancing materials property prediction by leveraging computational and experimental data using deep transfer learning. Nat Commun, 2019, 10: 5316
129 Goodall REA, Lee AA. Predicting materials properties without crystal structure: Deep representation learning from stoichiometry. Nat Commun, 2020, 11: 6280
130 Chen C, Ye W, Zuo Y, *et al.* Graph networks as a universal machine learning framework for molecules and crystals. Chem Mater, 2019, 31: 3564–3572
131 Chen C, Ong SP. AtomSets as a hierarchical transfer learning framework for small and large materials datasets. npj Comput Mater, 2021, 7: 173
132 Pasupa K, Sunhem W. A comparison between shallow and deep architecture classifiers on small dataset. In: 2016 8th International Conference on Information Technology and Electrical Engineering (ICITEE). Yogyakarta: IEEE, 2016. 1–6
133 Janiesch C, Zschech P, Heinrich K. Machine learning and deep learning. Electron Markets, 2021, 31: 685–695
134 Cortes C, Vapnik V. Machine learning and deep learning. Machine Learn, 1995, 20: 273–297
135 Shawe-Taylor J, Sun S. A review of optimization methodologies in support vector machines. Neurocomputing, 2011, 74: 3609–3618
136 Maddah HA, Berry V, Behura SK. Cuboctahedral stability in titanium halide perovskites *via* machine learning. Comput Mater Sci, 2020, 173: 109415
137 Sagi O, Rokach L. Ensemble learning: A survey. WIREs Data Min Knowl, 2018, 8: e1249
138 Dietterich TG. Ensemble methods in machine learning. In: Multiple Classifier Systems. Berlin, Heidelberg: Springer Berlin Heidelberg, 2000: 1–15
139 Bauer E, Kohavi R. An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. Machine Learn, 1999, 36: 105–139
140 Biau G, Scornet E. A random forest guided tour. TEST, 2016, 25: 197–227
141 Breiman L. Random forests. Machine Learn, 2001, 45: 5–32
142 Talapatra A, Uberuaga BP, Stanek CR, *et al.* Band gap predictions of double perovskite oxides using machine learning. Commun Mater, 2023, 4: 46
143 Freund Y, Schapire RE. Experiments with a new boosting algorithm. In: Proceedings of the Thirteenth International Conference on International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers Inc., 1996. 148–156
144 Friedman JH. Greedy function approximation: A gradient boosting machine.. Ann Statist, 2001, 29
145 Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: Association for Computing Machinery, 2016. 785–794
146 Ke G, Meng Q, Finley T, *et al.* LightGBM: A highly efficient gradient boosting decision tree. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2017. 3149–3157
147 Liu H, Cheng J, Dong H, *et al.* Screening stable and metastable ABO_3 perovskites using machine learning and the materials project. Comput

Mater Sci, 2020, 177: 109614
148 Tsymbalov E, Shi Z, Dao M, *et al.* Machine learning for deep elastic strain engineering of semiconductor electronic band structure and effective mass. npj Comput Mater, 2021, 7: 76
149 Himanen L, Jäger MOJ, Morooka EV, *et al.* DScribe: Library of descriptors for machine learning in materials science. Comput Phys Commun, 2020, 247: 106949
150 Batra R, Song L, Ramprasad R. Emerging materials intelligence ecosystems propelled by machine learning. Nat Rev Mater, 2021, 6: 655–678
151 Wang Z, Sun Z, Yin H, *et al.* Data-driven materials innovation and applications. Adv Mater, 2022, 34: 2104113
152 Hu W, Zhang L, Pan Z. Designing two-dimensional halide perovskites based on high-throughput calculations and machine learning. ACS Appl Mater Interfaces, 2022, 14: 21596–21604
153 Im J, Lee S, Ko TW, *et al.* Identifying Pb-free perovskites for solar cells by machine learning. npj Comput Mater, 2019, 5: 37
154 Park H, Ali A, Mall R, *et al.* Data-driven enhancement of cubic phase stability in mixed-cation perovskites. Mach Learn-Sci Technol, 2021, 2: 025030
155 Balachandran PV, Emery AA, Gubernatis JE, *et al.* Predictions of new ABO_3 perovskite compounds by combining machine learning and density functional theory. Phys Rev Mater, 2018, 2: 043802
156 Choudhary K, DeCost B, Chen C, *et al.* Recent advances and applications of deep learning methods in materials science. npj Comput Mater, 2022, 8: 59
157 Alom MZ, Taha TM, Yakopcic C, *et al.* A state-of-the-art survey on deep learning theory and architectures. Electronics, 2019, 8: 292
158 Tan C, Sun F, Kong T, *et al.* A survey on deep transfer learning. In: Kůrková V, Manolopoulos Y, Hammer B, et al. (eds). Artificial Neural Networks and Machine Learning–ICANN 2018. Cham: Springer, 2018
159 Qian J, Kim T, Jeon M. Reliability of large scale GPU clusters for deep learning workloads. In: Companion Proceedings of the Web Conference. Now York: Association for Computing Machinery, 2021. 179–181
160 Zhang Q, Zhu S. Visual interpretability for deep learning: A survey. Front Inf Technol Electron Eng, 2018, 19: 27–39
161 Bailly A, Blanc C, Francis É, *et al.* Effects of dataset size and interactions on the prediction performance of logistic regression and deep learning models. Comput Methods Programs Biomed, 2022, 213: 106504
162 Omee SS, Louis SY, Fu N, *et al.* Scalable deeper graph neural networks for high-performance materials property prediction. Patterns, 2022, 3: 100491
163 Domhan T, Springenberg JT, Hutter F. Speeding up automatic hyperparameter optimization of deep neural networks by extrapolation of learning curves. In: Yang Q, Wooldridge M (eds). Proceedings of the 24th International Conference on Artificial Intelligence. Buenos Aires: AAAI Press, 2015. 3460–3468
164 Scarselli F, Gori M, Ah Chung Tsoi M, *et al.* The graph neural network model. IEEE Trans Neural Netw, 2009, 20: 61–80
165 Gong S, Yan K, Xie T, *et al.* Examining graph neural networks for crystal structures: Limitations and opportunities for capturing periodicity. Sci Adv, 2023, 9: eadi3245
166 Reiser P, Neubert M, Eberhard A, *et al.* Graph neural networks for materials science and chemistry. Commun Mater, 2022, 3: 93
167 Goswami L, Deka MK, Roy M. Artificial intelligence in material engineering: A review on applications of artificial intelligence in material engineering. Adv Eng Mater, 2023, 25: 2300104
168 Choudhary K, DeCost B. Atomistic line graph neural network for improved materials property predictions. npj Comput Mater, 2021, 7: 185
169 Schmidt J, Pettersson L, Verdozzi C, *et al.* Crystal graph attention networks for the prediction of stable materials. Sci Adv, 2021, 7: eabi7948
170 Xie T, Grossman JC. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. Phys Rev Lett, 2018, 120: 145301
171 Antunes LM, Grau-Crespo R, Butler KT. Distributed representations of atoms and materials for machine learning. npj Comput Mater, 2022, 8: 44
172 Louis SY, Zhao Y, Nasiri A, *et al.* Graph convolutional neural networks with global attention for improved materials property prediction. Phys Chem Chem Phys, 2020, 22: 18141–18148
173 Karamad M, Magar R, Shi Y, *et al.* Orbital graph convolutional neural network for material property prediction. Phys Rev Mater, 2020, 4: 093801
174 Park CW, Wolverton C. Developing an improved crystal graph convolutional neural network framework for accelerated materials discovery. Phys Rev Mater, 2020, 4: 063801
175 Li Z, Liu F, Yang W, *et al.* A survey of convolutional neural networks: Analysis, applications, and prospects. IEEE Trans Neural Netw Learn Syst, 2022, 33: 6999–7019
176 Chen C, Ong SP. A universal graph deep learning interatomic potential for the periodic table. Nat Comput Sci, 2022, 2: 718–728
177 Vaswani A, Shazeer N, Parmar N, *et al.* Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010
178 Zhuo Y, Mansouri Tehrani A, Brgoch J. Predicting the band gaps of inorganic solids by machine learning. J Phys Chem Lett, 2018, 9: 1668–1673
179 Jha D, Ward L, Paul A, *et al.* ElemNet: Deep learning the chemistry of materials from only elemental composition. Sci Rep, 2018, 8: 17593
180 Zeng S, Zhao Y, Li G, *et al.* Atom table convolutional neural networks for an accurate prediction of compounds properties. npj Comput Mater, 2019, 5: 84
181 Wang AYT, Kauwe SK, Murdock RJ, *et al.* Compositionally restricted attention-based network for materials property predictions. npj Comput Mater, 2021, 7: 77
182 Gm H, Gourisaria MK, Pandey M, *et al.* A comprehensive survey and analysis of generative models in machine learning. Comput Sci Rev, 2020, 38: 100285
183 Wang B, Vastola JJ. Diffusion models generate images like painters: An analytical theory of outline first, details later. 2023, https://doi.org/10.48550/arXiv.2303.02490
184 Türk H, Landini E, Kunkel C, *et al.* Assessing deep generative models in chemical composition space. Chem Mater, 2022, 34: 9455–9467
185 Sanchez-Lengeling B, Aspuru-Guzik A. Inverse molecular design using machine learning: Generative models for matter engineering. Science, 2018, 361: 360–365
186 Kingma DP, Welling M. An introduction to variational autoencoders. FNT Machine Learn, 2019, 12: 307–392
187 Goodfellow I, Pouget-Abadie J, Mirza M, *et al.* Generative adversarial networks. Commun ACM, 2020, 63: 139–144
188 Glass CW, Oganov AR, Hansen N. USPEX—Evolutionary crystal structure prediction. Comput Phys Commun, 2006, 175: 713–720
189 Wang Y, Lv J, Zhu L, *et al.* CALYPSO: A method for crystal structure prediction. Comput Phys Commun, 2012, 183: 2063–2070
190 Pathak Y, Juneja KS, Varma G, *et al.* Deep learning enabled inorganic material generator. Phys Chem Chem Phys, 2020, 22: 26935–26943
191 Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells W, *et al.* (eds). Medical Image Computing and Computer-assisted Intervention–MICCAI 2015. Cham: Springer, 2015
192 Groom CR, Bruno IJ, Lightfoot MP, *et al.* The Cambridge structural database. Acta Crystlogr B Struct Sci Cryst Eng Mater, 2016, 72: 171–179
193 Choudhary K, Garrity KF, Reid ACE, *et al.* The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design. npj Comput Mater, 2020, 6: 173
194 Haastrup S, Strange M, Pandey M, *et al.* The computational 2D materials database: High-throughput modeling and discovery of atomically thin crystals. 2D Mater, 2018, 5: 042002
195 Gjerding MN, Taghizadeh A, Rasmussen A, *et al.* Recent progress of the computational 2D materials database (C2DB). 2D Mater, 2021, 8: 044002
196 Moustafa H, Larsen PM, Gjerding MN, *et al.* Computational exfolia-

tion of atomically thin one-dimensional materials with application to Majorana bound states. Phys Rev Mater, 2022, 6: 064202
197 Draxl C, Scheffler M. The NOMAD laboratory: From data sharing to artificial intelligence. J Phys Mater, 2019, 2: 036001
198 Bobbitt NS, Shi K, Bucior BJ, *et al.* MOFX-DB: An online database of computational adsorption data for nanoporous materials. J Chem Eng Data, 2023, 68: 483–498
199 Hachmann J, Olivares-Amaya R, Atahan-Evrenk S, *et al.* The Harvard clean energy project: Large-scale computational screening and design of organic photovoltaics on the World Community Grid. J Phys Chem Lett, 2011, 2: 2241–2251
200 Borysov SS, Geilhufe RM, Balatsky AV. Organic materials database: An open-access online database for data mining. PLoS ONE, 2017, 12: e0171501
201 Kim S, Thiessen PA, Bolton EE, *et al.* PubChem substance and compound databases. Nucleic Acids Res, 2016, 44: D1202–D1213
202 Bergerhoff G, Hundt R, Sievers R, *et al.* The inorganic crystal structure data base. J Chem Inf Comput Sci, 1983, 23: 66–69
203 Court CJ, Cole JM. Magnetic and superconducting phase diagrams and transition temperatures predicted using text mining and machine learning. npj Comput Mater, 2020, 6: 18
204 Swain MC, Cole JM. ChemDataExtractor: A toolkit for automated extraction of chemical information from the scientific literature. J Chem Inf Model, 2016, 56: 1894–1904
205 Mentel LM. Mendeleev—A Python resource for properties of chemical elements, ions and isotopes. 2014. https://github.com/lmmentel/mendeleev
206 Ong SP, Richards WD, Jain A, *et al.* Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. Comput Mater Sci, 2013, 68: 314–319
207 Ward L, Dunn A, Faghaninia A, *et al.* Matminer: An open source toolkit for materials data mining. Comput Mater Sci, 2018, 152: 60–69
208 Laakso J, Himanen L, Homm H, *et al.* Updates to the DScribe library: New descriptors and derivatives. J Chem Phys, 2023, 158: 234802
209 Ganose AM, Jain A. Robocrystallographer: Automated crystal structure text descriptions and analysis. MRS Commun, 2019, 9: 874–881
210 Ouyang R, Curtarolo S, Ahmetcik E, *et al.* SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates. Phys Rev Mater, 2018, 2: 083802
211 gplearn. https://gplearn.readthedocs.io/en/latest/intro.html
212 Ward L, Agrawal A, Choudhary A, *et al.* A general-purpose machine learning framework for predicting properties of inorganic materials. npj Comput Mater, 2016, 2: 16028
213 Zhou Q, Tang P, Liu S, *et al.* Learning atoms for materials discovery. Proc Natl Acad Sci USA, 2018, 115: E6411–E6417
214 Pedregosa F, Varoquaux G, Gramfort A, *et al.* Scikit-learn: Machine learning in Python. J Mach Learn Res, 2011, 12: 2825–2830 ,
215 Paszke A, Gross S, Massa F, *et al.* PyTorch: An imperative style, high-performance deep learning library. 2019, https://doi.org/10.48550/arXiv.1912.01703
216 Abadi M, Barham P, Chen J, *et al.* TensorFlow: A system for large-scale machine learning. 2016, https://doi.org/10.48550/arXiv.1605.08695
217 Gossett E, Toher C, Oses C, *et al.* AFLOW-ML: A RESTful API for machine-learning predictions of materials properties. Comput Mater Sci, 2018, 152: 134–145
218 Wang H, Zhang L, Han J, *et al.* DeePMD-kit: A deep learning package for many-body potential energy representation and molecular dynamics. Comput Phys Commun, 2018, 228: 178–184
219 Zhao XG, Zhou K, Xing B, *et al.* JAMIP: An artificial-intelligence aided data-driven infrastructure for computational materials informatics. Sci Bull, 2021, 66: 1973–1985
220 Wang G, Peng L, Li K, *et al.* ALKEMIE: An intelligent computational platform for accelerating materials discovery and design. Comput Mater Sci, 2021, 186: 110064
221 Wang J, Gao H, Han Y, *et al.* MAGUS: Machine learning and graph theory assisted universal structure searcher. Natl Sci Rev, 2023, 10: nwad128
222 Jacobs R, Mayeshiba T, Afflerbach B, *et al.* The materials simulation toolkit for machine learning (MAST-ML): An automated open source toolkit to accelerate data-driven materials research. Comput Mater Sci, 2020, 176: 109544
223 Peterson GGC, Brgoch J. Materials discovery through machine learning formation energy. J Phys Energy, 2021, 3: 022002
224 Ballif C, Haug FJ, Boccard M, *et al.* Status and perspectives of crystalline silicon photovoltaics in research and industry. Nat Rev Mater, 2022, 7: 597–616
225 Barrigón E, Heurlin M, Bi Z, *et al.* Synthesis and applications of III–V nanowires. Chem Rev, 2019, 119: 9170–9220
226 Lee TD, Ebong AU. A review of thin film solar cell technologies and challenges. Renew Sustain Energy Rev, 2017, 70: 1286–1297
227 Castro Neto AH, Guinea F, Peres NMR, *et al.* The electronic properties of graphene. Rev Mod Phys, 2009, 81: 109–162
228 Li L, Yu Y, Ye GJ, *et al.* Black phosphorus field-effect transistors. Nat Nanotech, 2014, 9: 372–377
229 Aftab S, Hegazy HH. Emerging trends in 2D TMDs photodetectors and piezo-phototronic devices. Small, 2023, 19: 2205778
230 Liang SJ, Cheng B, Cui X, *et al.* Van der Waals heterostructures for high-performance device applications: Challenges and opportunities. Adv Mater, 2020, 32: 1903800
231 Liu Y, Duan X, Shin HJ, *et al.* Promises and prospects of two-dimensional transistors. Nature, 2021, 591: 43–53
232 Zhang L, Mei L, Wang K, *et al.* Advances in the application of perovskite materials. Nano-Micro Lett, 2023, 15: 177
233 Chen X, Wang C, Li Z, *et al.* Bayesian optimization based on a unified figure of merit for accelerated materials screening: A case study of halide perovskites. Sci China Mater, 2020, 63: 1024–1035
234 Wang JT, Wang SZ, Zhou YH, *et al.* Flexible perovskite light-emitting diodes: Progress, challenges and perspective. Sci China Mater, 2023, 66: 1–21
235 Zhang B, Sun B, Liu F, *et al.* TiO_2-based S-scheme photocatalysts for solar energy conversion and environmental remediation. Sci China Mater, 2024, 67: 424–443
236 Zhou L, Xu Y, Chen B, *et al.* Synthesis and photocatalytic application of stable lead-free $Cs_2AgBiBr_6$ perovskite nanocrystals. Small, 2018, 14: 1703762
237 Qiu P, Shi X, Chen L. Cu-based thermoelectric materials. Energy Storage Mater, 2016, 3: 85–97
238 Wu X, Gao W, Chai J, *et al.* Defect tolerance in chalcogenide perovskite photovoltaic material $BaZrS_3$. Sci China Mater, 2021, 64: 2976–2986
239 Gan Y, Miao N, Lan P, *et al.* Robust design of high-performance optoelectronic chalcogenide crystals from high-throughput computation. J Am Chem Soc, 2022, 144: 5878–5886
240 Min H, Lee DY, Kim J, *et al.* Perovskite solar cells with atomically coherent interlayers on SnO_2 electrodes. Nature, 2021, 598: 444–450
241 Zhu L, Cao H, Xue C, *et al.* Unveiling the additive-assisted oriented growth of perovskite crystallite for high performance light-emitting diodes. Nat Commun, 2021, 12: 5081
242 Dou L, Yang YM, You J, *et al.* Solution-processed hybrid perovskite photodetectors with high detectivity. Nat Commun, 2014, 5: 5404
243 Qin C, Sandanayaka ASD, Zhao C, *et al.* Stable room-temperature continuous-wave lasing in quasi-2D perovskite films. Nature, 2020, 585: 53–57
244 Li Y, Zhu R, Wang Y, *et al.* Center-environment deep transfer machine learning across crystal structures: From spinel oxides to perovskite oxides. npj Comput Mater, 2023, 9: 109
245 Davies DW, Butler KT, Walsh A. Data-driven discovery of photoactive quaternary oxides using first-principles machine learning. Chem Mater, 2019, 31: 7221–7230
246 Li X, Mai H, Lu J, *et al.* Rational atom substitution to obtain efficient, lead-free photocatalytic perovskites assisted by machine learning and DFT calculations. Angew Chem Int Ed, 2023, 62: e202315002
247 Choubisa H, Todorović P, Pina JM, *et al.* Interpretable discovery of semiconductors with machine learning. npj Comput Mater, 2023, 9: 117
248 Cho H, Kim YH, Wolf C, *et al.* Improving the stability of metal halide perovskite materials and light-emitting diodes. Adv Mater, 2018, 30:

1704587
249 Bartel CJ. Review of computational approaches to predict the thermodynamic stability of inorganic solids. J Mater Sci, 2022, 57: 10475–10498
250 Ye W, Chen C, Wang Z, *et al.* Deep neural networks for accurate predictions of crystal stability. Nat Commun, 2018, 9: 3800
251 Pandey S, Qu J, Stevanović V, *et al.* Predicting energy and stability of known and hypothetical crystals using graph neural network. Patterns, 2021, 2: 100361
252 Bartel CJ, Trewartha A, Wang Q, *et al.* A critical examination of compound stability predictions from machine-learned formation energies. npj Comput Mater, 2020, 6: 97
253 Sun W, Dacek ST, Ong SP, *et al.* The thermodynamic scale of inorganic crystalline metastability. Sci Adv, 2016, 2: e1600225
254 Li X, Xie Y, Guo Q. A new intelligent prediction method for grade estimation. In: Zhang L, Lu BL, Kwok J. (eds). Advances in Neural Networks–ISNN 2010. Berlin, Heidelberg: Springer, 2010
255 Chen Z, Andrejevic N, Smidt T, *et al.* Direct prediction of phonon density of states with euclidean neural networks. Adv Sci, 2021, 8: 2004214
256 Noh J, Kim S, Gu G, *et al.* Unveiling new stable manganese based photoanode materials *via* theoretical high-throughput screening and experiments. Chem Commun, 2019, 55: 13418–13421
257 De Yoreo JJ, Gilbert PUPA, Sommerdijk NAJM, *et al.* Crystallization by particle attachment in synthetic, biogenic, and geologic environments. Science, 2015, 349: aaa6760
258 Rappe AK, Casewit CJ, Colwell KS, *et al.* UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. J Am Chem Soc, 1992, 114: 10024–10035
259 Yu W, Ji C, Wan X, *et al.* Machine-learning-based interatomic potentials for advanced manufacturing. Int J Mech Sys Dyn, 2021, 1: 159–172
260 Haghighatlari M, Li J, Guan X, *et al.* NewtonNet: A Newtonian message passing network for deep learning of interatomic potentials and forces. Digital Discov, 2022, 1: 333–343
261 Wang QH, Kalantar-Zadeh K, Kis A, *et al.* Electronics and optoelectronics of two-dimensional transition metal dichalcogenides. Nat Nanotech, 2012, 7: 699–712
262 Saparov B, Mitzi DB. Organic-inorganic perovskites: Structural versatility for functional materials design. Chem Rev, 2016, 116: 4558–4596
263 Yang J, Mannodi-Kanakkithodi A. High-throughput computations and machine learning for halide perovskite discovery. MRS Bull, 2022, 47: 940–948
264 Liu Y, Tan X, Liang J, *et al.* Machine learning for perovskite solar cells and component materials: Key technologies and prospects. Adv Funct Mater, 2023, 33: 2214271
265 Miyata A, Mitioglu A, Plochocka P, *et al.* Direct measurement of the exciton binding energy and effective masses for charge carriers in organic–inorganic tri-halide perovskites. Nat Phys, 2015, 11: 582–587
266 Geim AK. Graphene: Status and prospects. Science, 2009, 324: 1530–1534
267 Madsen GKH, Singh DJ. BoltzTraP. A code for calculating band-structure dependent quantities. Comput Phys Commun, 2006, 175: 67–71
268 Choudhary K, Garrity KF, Sharma V, *et al.* High-throughput density functional perturbation theory and machine learning predictions of infrared, piezoelectric, and dielectric responses. npj Comput Mater, 2020, 6: 64
269 Takahashi A, Kumagai Y, Miyamoto J, *et al.* Machine learning models for predicting the dielectric constants of oxides based on high-throughput first-principles calculations. Phys Rev Mater, 2020, 4: 103801
270 Dong R, Dan Y, Li X, *et al.* Inverse design of composite metal oxide optical materials based on deep transfer learning and global optimization. Comput Mater Sci, 2021, 188: 110166
271 Mi JX, Li AD, Zhou LF. Review study of interpretation methods for future interpretable machine learning. IEEE Access, 2020, 8: 191969–191985
272 Lundberg S, Lee SI. A unified approach to interpreting model predictions. 2017. http://arxiv.org/abs/1705.07874
273 Zhang S, Lu T, Xu P, *et al.* Predicting the formability of hybrid organic–inorganic perovskites *via* an interpretable machine learning strategy. J Phys Chem Lett, 2021, 12: 7423–7430
274 Stephens T. gpleran. https://gplearn.readthedocs.io/en/latest/ intro.html
275 Liu S, Wang J, Duan Z, *et al.* Simple structural descriptor obtained from symbolic classification for predicting the oxygen vacancy defect formation of perovskites. ACS Appl Mater Interfaces, 2022, 14: 11758–11767
276 Guo Z, Hu S, Han ZK, *et al.* Improving symbolic regression for predicting materials properties with iterative variable selection. J Chem Theor Comput, 2022, 18: 4945–4951
277 Song Z, Wang X, Liu F, *et al.* Distilling universal activity descriptors for perovskite catalysts from multiple data sources *via* multi-task symbolic regression. Mater Horiz, 2023, 10: 1651–1660
278 Weng B, Song Z, Zhu R, *et al.* Simple descriptor derived from symbolic regression accelerating the discovery of new perovskite catalysts. Nat Commun, 2020, 11: 3513
279 Bartel CJ, Sutton C, Goldsmith BR, *et al.* New tolerance factor to predict the stability of perovskite oxides and halides. Sci Adv, 2019, 5: eaav0693
280 Aggour KS, Detor A, Gabaldon A, *et al.* Compound knowledge graph-enabled AI assistant for accelerated materials discovery. Integr Mater Manuf Innov, 2022, 11: 467–478
281 Xie C, Pan Z, Shu C. Microstructure representation knowledge graph to explore the twinning formation. Crystals, 2022, 12: 466
282 Zunger A. Inverse design in search of materials with target functionalities. Nat Rev Chem, 2018, 2: 0121
283 Wang J, Wang Y, Chen Y. Inverse design of materials by machine learning. Materials, 2022, 15: 1811
284 Mroz AM, Posligua V, Tarzia A, *et al.* Into the unknown: How computation can help explore uncharted material space. J Am Chem Soc, 2022, 144: 18730–18743
285 Lyngby P, Thygesen KS. Data-driven discovery of 2D materials by deep generative models. npj Comput Mater, 2022, 8: 232
286 Moustafa H, Lyngby PM, Mortensen JJ, *et al.* Hundreds of new, stable, one-dimensional materials from a generative machine learning model. Phys Rev Mater, 2023, 7: 014007
287 Wines D, Xie T, Choudhary K. Inverse design of next-generation superconductors using data-driven deep generative models. J Phys Chem Lett, 2023, 14: 6630–6638
288 Zhu L, Zhou J, Sun Z. Materials data toward machine learning: Advances and challenges. J Phys Chem Lett, 2022, 13: 3965–3977
289 Zhang Y, Ling C. A strategy to apply machine learning to small datasets in materials science. npj Comput Mater, 2018, 4: 25
290 Acar P. Recent progress of uncertainty quantification in small-scale materials science. Prog Mater Sci, 2021, 117: 100723
291 Emery AA, Wolverton C. High-throughput DFT calculations of formation energy, stability and oxygen vacancy formation energy of ABO_3 perovskites. Sci Data, 2017, 4: 170153
292 Shen C, Li T, Zhang Y, *et al.* Accelerated screening of ternary chalcogenides for potential photovoltaic applications. J Am Chem Soc, 2023, 145: 21925–21936
293 Goldschmidt VM. Die gesetze der krystallochemie. Naturwissenschaften, 1926, 14: 477–485
294 Robinson K, Gibbs GV, Ribbe PH. Quadratic elongation: A quantitative measure of distortion in coordination polyhedra. Science, 1971, 172: 567–570
295 Stoumpos CC, Frazer L, Clark DJ, *et al.* Hybrid germanium iodide perovskite semiconductors: Active lone pairs, structural distortions, direct and indirect energy gaps, and strong nonlinear optical properties. J Am Chem Soc, 2015, 137: 6804–6819
296 Baur WH. The geometry of polyhedral distortions. Predictive relationships for the phosphate group. Acta Crystlogr B Struct Sci, 1974, 30: 1195–1215
297 Pan H, Ganose AM, Horton M, *et al.* Benchmarking coordination number prediction algorithms on inorganic crystal structures. Inorg

Chem, 2021, 60: 1590–1603
298 Zimmermann NER, Jain A. Local structure order parameters and site fingerprints for quantification of coordination environment and crystal structure similarity. RSC Adv, 2020, 10: 6063–6081
299 Tamtaji M, Gao H, Hossain MD, *et al.* Machine learning for design principles for single atom catalysts towards electrochemical reactions. J Mater Chem A, 2022, 10: 15309–15331
300 Birschitzky VC, Ellinger F, Diebold U, *et al.* Machine learning for exploring small polaron configurational space. npj Comput Mater, 2022, 8: 125
301 Wu X, Wang H, Gong Y, *et al.* Graph neural networks for molecular and materials representation. J Mater Inf, 2023, 3: 12
302 Bilodeau C, Jin W, Jaakkola T, *et al.* Generative models for molecular discovery: Recent advances and challenges. WIREs Comput Mol Sci, 2022, 12: e1608
303 Peña-Guerrero J, Nguewa PA, García-Sosa AT. Machine learning, artificial intelligence, and data science breaking into drug design and neglected diseases. WIREs Comput Mol Sci, 2021, 11: e1513
304 Bagal V, Aggarwal R, Vinod PK, *et al.* MolGPT: Molecular generation using a transformer-decoder model. J Chem Inf Model, 2022, 62: 2064–2076
305 Song Y, Ermon S. Generative modeling by estimating gradients of the data distribution. 2020. http://arxiv.org/abs/1907.05600

Acknowledgements This work was supported by the National Natural Science Foundation of China (62125402 and 62321166653).

Author contributions Yang X prepared the manuscript under the direction of Zhang L; Zhou K helped prepare the figures and tables; Zhang L and He X revised the manuscript. All authors contributed to the general discussion.

Conflict of interest The authors declare that they have no conflict of interest.

Xiaoyu Yang is a doctoral student at the School of Materials Science and Engineering, Jilin University. He received a Bachelor degree in materials physics from Jilin University in 2020. His research interests focus on promoting the study of new optoelectronic semiconductor materials through high-throughput computation and machine learning.

Xin He obtained her PhD degree at Jilin University (2019), and now is an associate professor and Tang Aoqing Young Scholars of Jilin University. She was awarded the Postdoctoral Innovative Talents Supporting Program in 2019. Her current interests focus on designing novel semiconductor materials for optoelectronic applications.

Lijun Zhang is the Tang Aoqing Distinguished Professor, Dean of the School of Materials Science and Engineering, Jilin University, China. He obtained his BS degree from the Northeast Normal University (2003), and completed his PhD degree at Jilin University (2008), China. He then worked as a postdoctoral researcher at Oak Ridge National Laboratory (2008–2010) and National Renewable Energy Laboratory (2010–2013), and became a research assistant professor at the University of Colorado at Boulder (2013–2014). In September 2014, he became a permanent faculty of the School of Materials Science and Engineering, Jilin University. His current interest focuses on the design of new materials and the enhancement of semiconductor performance tailored for diverse optoelectronic applications.

机器学习方法及应用: 光电半导体材料计算设计

杨晓雨, 周琨, 贺欣*, 张立军*

摘要 高通量计算与材料数据库推动了数据驱动的机器学习方法的发展. 机器学习已经成为材料计算研究的重要方法, 在分析材料数据、加速材料计算、预测材料性质、推进新材料发现、筛选和设计等方面展现出极大的潜力. 众多与材料计算相交叉的机器学习方法、模型以及框架不断涌现. 本文综述了近年来光电半导体材料计算设计领域内机器学习方法的最新进展与应用. 介绍了机器学习的流程与类型, 基于不同材料表示方法的浅层模型、集成模型和深度神经网络, 以及相关材料数据库和相关工具. 我们还讨论了这些模型在预测材料稳定性与光电性质、材料逆向设计、构建材料构效关系等方面的应用. 最后, 本文对目前机器学习方法存在的机遇与挑战, 即数据数量与质量、材料的表示、材料逆向设计做了进一步总结与讨论.

