

Applications and potentials of machine learning in optoelectronic materials research: An overview and perspectives

Cheng-Zhou Zhang(张城洲) and Xiao-Qian Fu(付小倩)
Citation: Chin. Phys. B, 2023, 32 (**12**): 126103. DOI: 10.1088/1674-1056/ad01a4
Journal homepage: http://cpb.iphy.ac.cn; http://iopscience.iop.org/cpb

What follows is a list of articles you may be interested in

Electronic and thermal properties of Ag-doped single crystal zinc oxide via laser-induced technique

Huan Xing(邢欢), Hui-Qiong Wang(王惠琼), Tinglu Song(宋廷鲁), Chunli Li(李纯莉), Yang Dai(戴扬), Gengming Fu(傅耿明), Junyong Kang(康俊勇), and Jin-Cheng Zheng(郑金成)
Chin. Phys. B, 2023, 32 (**6**): 066107. DOI: 10.1088/1674-1056/acae74

Dislocation mechanism of $Ni_{47}Co_{53}$ alloy during rapid solidification

Yun-Chun Liu(刘云春), Yong-Chao Liang(梁永超), Qian Chen(陈茜), Li Zhang(张利), Jia-Jun Ma(马家君), Bei Wang(王蓓), Ting-Hong Gao(高廷红), and Quan Xie(谢泉)
Chin. Phys. B, 2023, 32 (**6**): 066104. DOI: 10.1088/1674-1056/aca7ea

Evolution of defects and deformation mechanisms in different tensile directions of solidified lamellar Ti-Al alloy

Yutao Liu(刘玉涛), Tinghong Gao(高廷红), Yue Gao(高越), Lianxin Li(李连欣), Min Tan(谭敏), Quan Xie(谢泉), Qian Chen(陈茜), Zean Tian(田泽安), Yongchao Liang(梁永超), and Bei Wang(王蓓)
Chin. Phys. B, 2022, 31 (**4**): 046105. DOI: 10.1088/1674-1056/ac1f0b

CdS/Si nanofilm heterojunctions based on amorphous silicon films: Fabrication, structures, and electrical properties

Yong Li(李勇), Peng-Fei Ji(姬鹏飞), Yue-Li Song(宋月丽), Feng-Qun Zhou(周丰群), Hong-Chun Huang(黄宏春), and Shu-Qing Yuan(袁书卿)
Chin. Phys. B, 2021, 30 (**2**): 026101. DOI: 10.1088/1674-1056/abc15e

Room-temperature strong coupling between dipolar plasmon resonance in single gold nanorod and two-dimensional excitons in monolayer WSe_2

Jinxiu Wen(温锦秀), Hao Wang(汪浩), Huanjun Chen(陈焕君), Shaozhi Deng(邓少芝), Ningsheng Xu(许宁生)
Chin. Phys. B, 2018, 27 (**9**): 096101. DOI: 10.1088/1674-1056/27/9/096101

REVIEW

Applications and potentials of machine learning in optoelectronic materials research: An overview and perspectives

Cheng-Zhou Zhang(张城洲)[1] and Xiao-Qian Fu(付小倩)[1,2,†]

[1] *School of Information Science and Engineering, University of Jinan, Jinan* 250022, *China*

[2] *Shandong Provincial Key Laboratory of Network based Intelligent Computing, University of Jinan, Jinan* 250022, *China*

(Received 7 August 2023; revised manuscript received 2 October 2023; accepted manuscript online 10 October 2023)

Optoelectronic materials are essential for today's scientific and technological development, and machine learning provides new ideas and tools for their research. In this paper, we first summarize the development history of optoelectronic materials and how materials informatics drives the innovation and progress of optoelectronic materials and devices. Then, we introduce the development of machine learning and its general process in optoelectronic materials and describe the specific implementation methods. We focus on the cases of machine learning in several application scenarios of optoelectronic materials and devices, including the methods related to crystal structure, properties (defects, electronic structure) research, materials and devices optimization, material characterization, and process optimization. In summarizing the algorithms and feature representations used in different studies, it is noted that prior knowledge can improve optoelectronic materials design, research, and decision-making processes. Finally, the prospect of machine learning applications in optoelectronic materials is discussed, along with current challenges and future directions. This paper comprehensively describes the application value of machine learning in optoelectronic materials research and aims to provide reference and guidance for the continuous development of this field.

Keywords: optoelectronic materials, devices, machine learning, prior knowledge

PACS: 61.72.–y, 07.05.Mh, 85.60.–q, 71.20.Nr **DOI:** 10.1088/1674-1056/ad01a4

1. Introduction

Optoelectronic materials have a wide range of applications in various fields of modern technology, including fiber optic communication, UV detectors and LEDs, power devices, solar cells, optical sensors, and lasers.[1–8] These materials exhibit excellent performance, and with continuous research and development, they keep offering more efficient and reliable solutions.

The research and development of traditional optoelectronic materials require a lot of trial, error, and improvements.[9] Meanwhile, computational methods such as density functional theory (DFT),[10] molecular dynamics simulation,[11] and Monte Carlo simulation[12] have been widely used in the prediction of material properties and structures.[13–15] The results of theoretical and computational simulations can better guide experiments. For example, Monte Carlo simulations study carrier transport[16] or finite element methods to simulate optical properties.[17] However, although experimental and theoretical research has produced massive databases, their capabilities are still limited. The methodology is time-consuming and costly, not only for optoelectronic materials but for all other materials.

In 2011, the materials genome initiative (MGI) was launched[18] to accelerate the optimization and application of new materials by integrating experimental, theoretical, and computational approaches.[19,20] High-throughput computing is a vital tool and method that provides computational support for the MGI.[21–23] Currently, materials informatics,[24] as an interdisciplinary discipline covering the fields of materials simulation and computer, is intertwined with the MGI to provide more opportunities for materials research.[25] Figure 1 shows the stages of development of materials science, and the integration of knowledge and technology between the locations has contributed to the progress of materials science. The data-driven stage has evolved towards fast, efficient, and sustainable.

Machine learning (ML) empowers computers to learn from experience and continually enhance their performance across various tasks, broadening the scope of computational applications and improving efficiency. Leveraging ML can significantly reduce the number and costs of traditional trial-and-error experiments in optoelectronic materials. By analyzing large amounts of data and patterns, ML can predict the properties and behaviors of materials. This approach significantly reduces the required experiments and expedites the development process.

According to the data from Web of Science, there are a total of 222 papers when searching for the topic of machine learning and optoelectronic materials, with 197 published in the past five years, and the progress is astonishing. These re-

[†]Corresponding author. E-mail: ise_fuxq@ujn.edu.cn

© 2023 Chinese Physical Society and IOP Publishing Ltd

searches apply ML to material design, process optimization, properties prediction, structure analysis, and more. For example, support vector machine (SVM), random forest (RF), and other ensemble models have been important tools in electronic structure properties prediction.[26–29] Deep learning (DL) has been widely used in optoelectronic materials and often necessitates substantial volumes of data. Convolutional neural network (CNN), recurrent neural network (RNN), generative adversarial network (GAN), and graph neural network (GNN) are mainly applied to properties prediction,[30–32] structure prediction,[33–35] image analysis,[36] and optimization.[37] Nevertheless, numerous material or device datasets in practice fall short of achieving such scale. Transfer learning (TL) emerges as a valuable technique, enabling DL models to be effectively trained even when confronted with limited data.[30,31] With the development of ML in materials, advanced ML algorithms have gradually been applied to optoelectronics, providing more intelligent, efficient, and adaptive methods for the research and application of optoelectronic materials. Active learning is used in applications such as property prediction,[38] discovering new materials,[39] and capturing atomic force fields.[40] Reinforcement learning[41] improves the efficiency of material optimization.[42] ML interatomic potentials (ML-IAPs) and force fields (MLFFs) perform nanosecond molecular dynamics simulations.[43,44]

ML enhances research efficiency, improves material properties, and advances optoelectronic materials through these applications. However, choosing proper ML models to achieve desired characteristics is still challenging when relating entirely new material. Review articles could supply some methodologies for beginners. Unfortunately, we notice that most of these articles focus more on a specific material, such as new 2D materials,[45] perovskite materials,[46–48] nanoplasmonics,[49] etc. Seldom covers a wide range of optoelectronic materials and devices.

In this paper, we initially provide a brief development of ML and an overview of the typical workflows in optoelectronic materials ML. Subsequently, we delve into the methodologies and techniques employed to apply ML in predicting and optimizing optoelectronic material structures and device properties. These applications underscore how ML is well-positioned to expedite research, enhance performance, and drive the development of optoelectronic materials. Finally, we discuss the challenges and prospects within this domain, offering insights and recommendations for future research directions and trends.

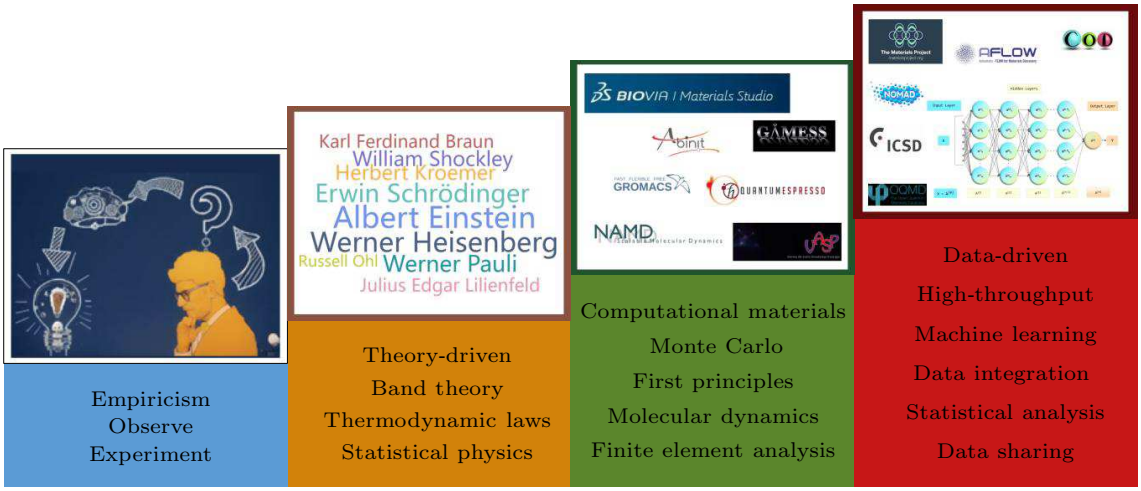


Fig. 1. Advances in materials science.

2. Fundamentals of ML

Research problems in optoelectronic materials are diverse, and different ML algorithms and techniques apply to different problems. Understanding the development of ML can help researchers select appropriate methods to solve specific optoelectronic materials problems. Understanding the general process of ML can help researchers plan better, execute, and improve applications in optoelectronic materials and increase the efficiency and quality of research. This section is organized around two subsections, and we focus on the general process of ML of optoelectronic materials.

2.1. The development of ML

In 1956, John McCarthy *et al.* introduced the concept of artificial intelligence (AI) at the Dartmouth Conference, and various theories and algorithms emerged after decades of development.[50–52] ML, as a branch of AI, has developed from simple to complex and can be summarised in the following stages.

In the early stages of ML, algorithms relied heavily on manual feature engineering, requiring the manual extraction of relevant features as model inputs. Typical algorithms included linear regression and SVM. While these algorithms performed well in many tasks, the reliance on manual feature engineer-

ing limited their applicability. With DL, multilayer neural networks can learn hierarchical representations directly from the raw data, thus eliminating the need for explicit feature engineering. CNN and RNN have demonstrated strong capabilities in feature learning and have made significant progress in natural language processing (NLP).[53] Reinforcement learning focuses on intelligent autonomous decision-making to maximize gains and is suitable for online learning and optimization in control tasks. Current multimodal learning integrates information from different sources, while TL enables knowledge transfer across domains. These models allow ML to solve complex real-world problems.[54]

ML has experienced a leap from weakness to strength, from simple statistical analyses to autonomous learning functions. Its application domains and boundaries expand, providing novel tools and methods for solving multifaceted challenges.

ML opens up new opportunities in various subfields of materials science. Butler *et al.*[55] presented advances in ML in chemical sciences regarding material design, material synthesis, and theoretical guidance. Piccinotti *et al.*[56] summarized AI's research progress in photonics, nanophotonics, and plasmonics. Schleder *et al.*[57] outlined using DFT, high-throughput computing, and ML, modern methods of computing materials, to design novel materials with enhanced properties. Choudhary *et al.*[58] outlined advances in DL methods for atomistic simulations, materials imaging, and spectral analysis. In addition, Golmohammadi and Aryanpour[59] collected and organized the articles containing keywords such as ML and materials discovery from 2011–2022 and predicted the properties of more than thirty materials using ML and proposed four guidelines to help researchers develop better models. Kadilkar *et al.*[60] discussed the progress of using ML to design materials with target properties and how ML can reduce the design space's dimensionality. Liu *et al.*[61] reviewed advances in research for semiconductor materials and manufacturing, where ML has become a powerful tool for semiconductor research, helping to analyze large-scale datasets and solve complex problems in semiconductor materials science and manufacturing. As an emerging strategic material, applying ML to assist the research and development of optoelectronic materials is an important direction of current research.

2.2. ML process for optoelectronic materials

The general ML process for optoelectronic materials is shown in Fig. 2. It typically includes data collection and preprocessing, feature engineering, model selection and training, and model evaluation and optimization. Below, we detail the application of these steps to research in optoelectronics.

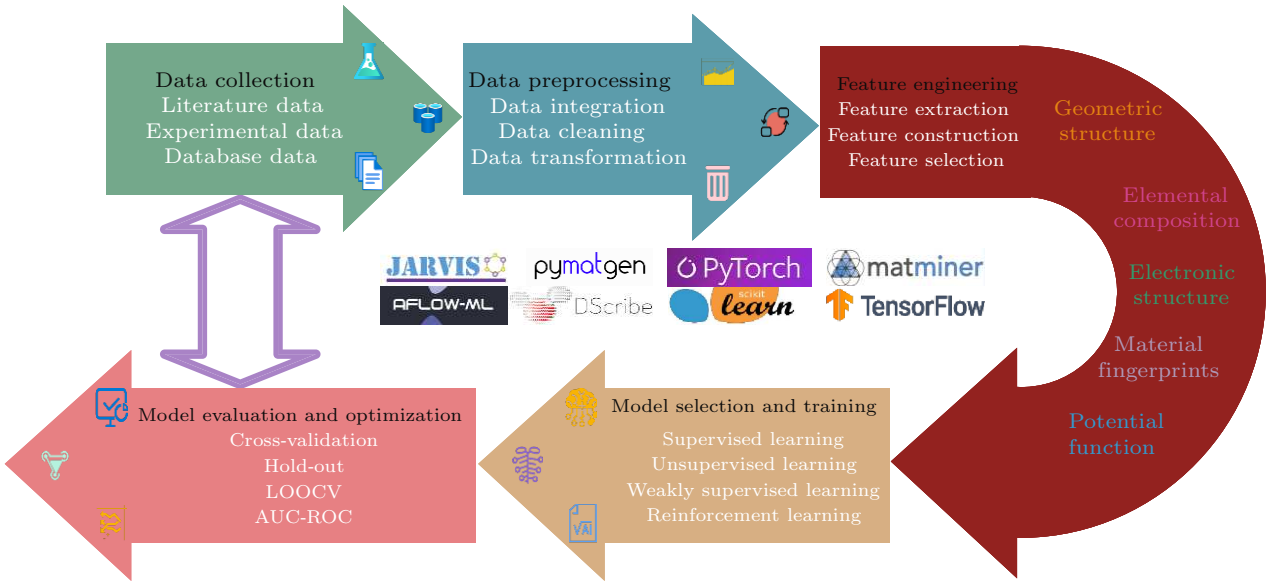


Fig. 2. The general ML process for optoelectronic materials.

2.2.1. Data collection and preprocessing

Data collection and preprocessing are the first steps in ML of optoelectronic materials. Constructing high-quality datasets requires acquiring data from multiple sources, including data obtained from experimental measurements, computational simulation results, and materials reported in the paper. Researchers can collect primary data, such as components and structures of actually synthesized optoelectronic material samples, and property data, such as conductivity and optical properties obtained from measurements from COD,[62] and ICSD[63] databases. Computational material databases include AFLOW,[64] MP,[65] NOMAD,[66] and CMR.[67] Contain property data of optoelectronic materials that have been predicted using first-principle calculations and other methods. Data related to optoelectronic materials can also be extracted from the published literature.

In addition to numerical data, it is possible to create datasets of topographic images of optoelectronic materials obtained by instruments such as scanning electron microscopy (SEM) and atomic force microscopy and datasets of characteristic spectra of optoelectronic materials obtained by testing with devices such as UV-vis spectrometers.

By integrating and processing heterogeneous data from multiple sources, a comprehensive and high-quality optoelectronic material dataset can be constructed for training ML models, material property prediction, and new material discovery. These rich and influential data will significantly improve the learning ability and generalization of the models.

ML in optoelectronic materials science is highly dependent on standardized, high-quality data, and many open initiatives that provide standardized materials data following the FAIR (findable, accessible, interoperable, reusable) principle[68] are of great significance. Table 1 lists some vital standardized initiatives that allow different researchers to compare and reproduce on a uniform database, accelerating research progress. Statistical techniques and data mining can be used to normalize or discretize the raw data to obtain high-quality data. For example, missing data imputation[69] can be used to improve the completeness of datasets; sampling strategy[70] can be used to achieve model performance benchmarks with fewer data; and data augmentation[71] can be used to reduce the reliance on more labeled data. Through these standardization and quality control measures, we can provide high-quality and uniform data sources for ML, reduce the impact of data noise on the model, and promote the development of research in optoelectronic materials.

Table 1. Standardization initiatives.

Name	URL	Description
JCAMP-DX	http://www.jcamp-dx.org/	Electronic data standards for chemical and spectroscopy information.
KIM API	https://openkim.org/kim-api/	API standard for connecting molecular simulation codes with interatomic models.
Nomad meta info	https://nomad-lab.eu/services/metainfo	Schema for storing results of *ab initio* and force-field atomistic simulations.
Open force field toolkit	https://docs.openforcefield.org/	Specification for encoding molecular mechanics force fields.
CIF	https://www.iucr.org/resources/cif	A standard for crystallographic information.
CML	http://www.xml-cml.org/	Molecules, compounds, reactions, spectra, crystals.
ESCF	https://gitlab.com/ElectronicStructureLibrary/escdf/escdf-specifications	An initiative of the electronic structure library contributors.

2.2.2. Feature engineering

Feature engineering is an integral part of data science and ML and can significantly impact the performance and effectiveness of models. Feature engineering mainly includes feature extraction, construction, and selection.

Feature extraction is a critical step in extracting useful information from raw data. This process helps reducing the feature matrix's dimensionality, reducing computational complexity and shortening training time. Among the dimensionality reduction methods, principal component analysis[72] (PCA) and linear discriminant analysis (LDA)[73] are commonly used techniques. Potapov and Lubk[74] applied PCA to denoise spectral images. They believed the data needed to be smoothed, filtered, and weighted. Otherwise, the denoising effect of PCA will be reduced, and human error will be introduced.

Feature construction is constructing new features from material data by combining and transforming them. The optoelectronic material features are constructed as shown in Fig. 3 and include in the following aspects.

(i) **Basic attribute features** Use the basic attributes of the optoelectronic material, such as components, crystal structure parameters, energy gap, electronic density of states, etc., as the initial features.

(ii) **Structural descriptors** Construct various types of descriptors describing the component and structural information of the material, such as atomic symmetry functions, crystal fingerprints, etc., to capture the topological structural features of the material.

(iii) **Graph network features** Use GNN or graph convolutional network (GCN) to extract the connectivity network features of the material to describe the atomic arrangement and crystal symmetry.

(iv) **Spectral features** Extract features from experimental or computational spectra to construct spectral fingerprints and other features to describe the optical properties of materials.

(v) **Image features** Use computer vision technology to process SEM and transmission electron microscope (TEM) images of materials to extract image features that characterize the size, shape, distribution, and other information.

(vi) **Multi-source feature fusion** Synthesize the features obtained from the above channels to construct a comprehensive material descriptor.

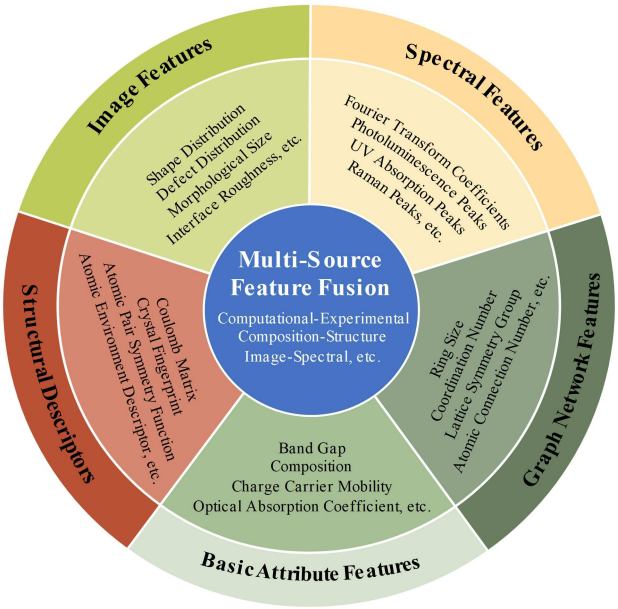


Fig. 3. Feature construction methods for optoelectronic material.

Feature selection plays a crucial role in data analysis and ML, which can improve the performance and interpretability of models. Wang *et al.*[75] introduced standard feature selection methods for perovskite materials. Whether used individually or in combination, the Pearson correlation coefficient in the filtering method, recursive feature elimination in the wrapping method, and tree model in the embedding method appeared more frequently. Choosing appropriate feature selection methods can significantly reduce model complexity and improve model interpretability. Through practical feature engineering, data redundancy can be eliminated, the robustness of model prediction can be improved, and the model's learning of fundamental physical mechanisms can be more explicit. Table 2 lists some of the commonly used features.

If the dataset is small, the complexity of the data distribution is low. In that case, too many features make the model complex overcapacity, resulting in overfitting.[76] When the number of features is close to or exceeds the number of samples, the model can simply memorize each training sample and lose the ability to generalize.

Table 2. Common features.

Electronic structure	Geometric structure	Elemental composition	Material fingerprint	Potential function	Spectrum
band gap	lattice constant	atomic number	Coulomb matrix[a]	ACSF[b]	XRD
electron affinity	dihedral angle	covalent radius	2D diffraction fingerprint[c]	SOAP[d]	XAS
charge density	bond angle	electronegativity	morgan fingerprint	GAP[e]	XES
band center	radial distribution function	first ionization energy	voronoi fingerprint	DEEPmd[f]	UVS

[a]Ref. [77], [b]Ref. [78], [c]Ref. [34], [d]Ref. [79], [e]Ref. [80], [f]Ref. [81]

Randomly omitting half of the feature detectors[82] can prevent the problem. The regularization technique prevents neural network overfitting and improves the model's generalization ability. Its basic idea is to randomly drop some of the hidden layer nodes with a certain probability during the training process to avoid these nodes from over-dependence on the features of other nodes. In testing, all nodes participate in the prediction, but the output is scaled to compensate for the effect of ignoring nodes. It regularizes the network by randomly ignoring nodes during training, encouraging nodes to learn features that do not depend on each other, thus improving generalization.

It is also possible to transform the regression problem into a classification problem by simply determining whether the bandgap is within a specific range without predicting the exact value of the bandgap.[83] To better adapt to the needs of practical applications, we convert continuous bandgap values into binary labels and simplify the task of judging whether the material is usable instead of accurately predicting the bandgap values. This method reduces the difficulty of the prediction task and focuses more on determining the availability of the material.

TL can utilize the information of rich data in the source domain, which is suitable for the case of relatively few samples in the destination domain. Some material representation methods, such as GNN and crystal fingerprints, can be applied to various optoelectronic materials to realize TL. These methods are widely used in materials science and other fields to help understand and interpret complex features. They can reduce the difficulty of the target task and make it easier for ML models to capture essential patterns, thus improving the accuracy and interpretability of predictions.

2.2.3. Model selection and training

Model selection and training is an iterative process that needs to be adjusted and optimized according to the actual situation. The experience of the experts, as well as the types of models and methods used in previous studies, are factors to be considered.

ML algorithms are classified into supervised, unsupervised, weakly supervised, and reinforcement learning. The ML models built based on ML algorithms are shown in Fig. 4. Many papers describe the process of ML algorithms.[84–87] Depending on the nature and characteristics of the problem, we need to choose the appropriate model architecture.

Supervised learning relies on explicit labeling information, unsupervised learning does not use labeling information, and weakly supervised learning uses partial, incomplete, or in-

accurate labeling information. Reinforcement learning learns decision-making strategies by interacting with the environment. The core idea of TL is to improve performance on a new task by utilizing existing knowledge rather than learning from scratch.

Classification algorithms include support vector classification (SVC), logistic regression, decision tree (DT), RF, *k*-nearest neighbor (KNN), etc., and common regression algorithms include linear regression, support vector regression (SVR), gradient boosting regression (GBR), etc. In artificial neural network (ANN), the nature of tasks, whether classification or regression, is typically determined by the network's output layer. Softmax and sigmoid output layers are employed for classification tasks, while linear and single-neuron output layers are used for regression tasks. Furthermore, choosing the neural network's loss function determines whether the network performs classification or regression. For instance, classification tasks typically utilize cross-entropy loss, whereas regression tasks utilize mean squared error loss.

2.2.4. Model evaluation and optimization

Objectively evaluating model performance is necessary to improve the model's generalization ability and better adapt to unknown data. Researchers usually use cross-validation methods to measure model performance.[88] Error assessment in classification tasks can use metrics such as accuracy, F1-score, AUC-ROC curve, etc.; regression tasks can use mean absolute error, root mean square error (RMSE), R^2 score, etc. Different evaluation metrics can measure the performance of the model from different perspectives. By adjusting hyperparameters, regularization techniques, and adopting a stacked ML model,[89] the model's performance can be effectively improved, and more accurate ML models can be constructed.

This subsection summarizes the general ML process. Table 3 lists the optoelectronic material ML tools. These tools can fully support and accelerate the development of optoelectronic materials. Researchers must establish a solid mathematical foundation, deeply understand ML algorithms and principles, and utilize these tools for rigorous experimental design and validation to avoid misleading results.

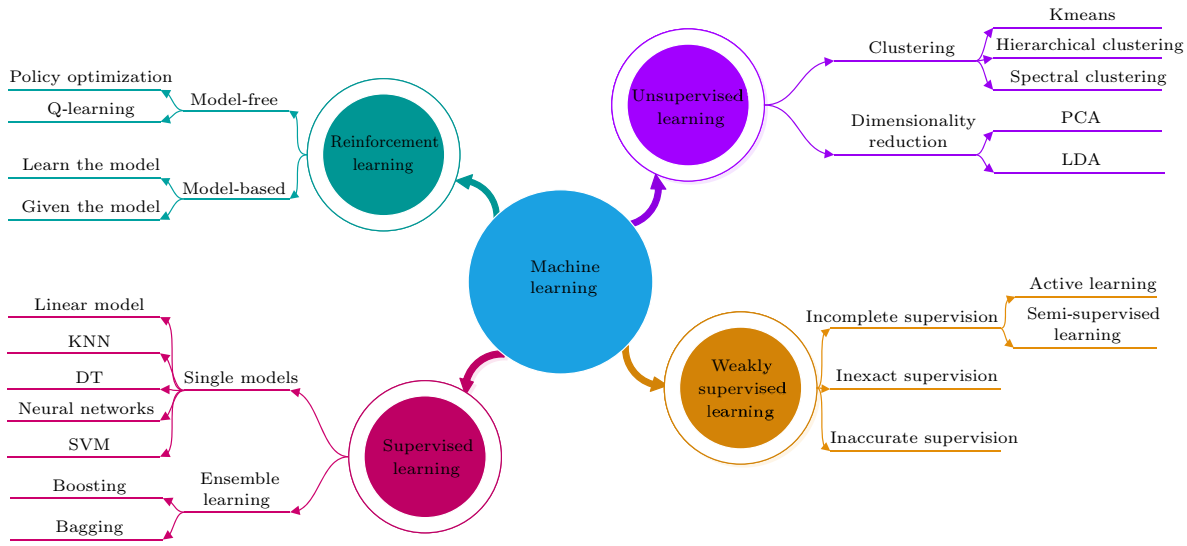


Fig. 4. Common ML models.

Table 3. Resources for materials ML.

Name	URL	Description
ML frameworks		
Pytorch	https://pytorch.org/	Programming deep neural networks (DNN) specifically for GPU acceleration.
Scikit-learn	https://scikit-learn.org/stable/inde-x.html	Encompasses mainstream ML algorithms.
Tensorflow	https://www.tensorflow.org/	Used for building and training various ML models.
Keras	https://keras.io/	Advanced neural network API.
Materials ML tools		
Atomate	https://atomate.org/	Materials workflows based on FireWorks.
Datamol	https://github.com/datamol-io/data-mol	Molecular processing tool.
HTTK	https://httk.openmaterialsdb.se/	High-throughput toolkit.
Matminer	https://github.com/hackingmater-ials/matminer	A library for data mining in materials science.
Pymatgen	https://pymatgen.org/	A library for materials analysis.
ASE	https://wiki.fysik.dtu.dk/ase/	Manipulating, running, visualizing, and analyzing atomistic simulations.
JAMIP	http://jamip-code.com/	An open-source integrated infrastructure for computational materials informatics.
CDVAE	https://github.com/txie93/cdvae	Generating new stable materials.

3. ML applications in optoelectronic materials and devices

This section will introduce the application of ML in optoelectronic materials and devices. The basic information of related literature is summarized in Table 4, including the research content, best ML algorithms (some of the literature used only one

Table 4. Literature analysis.

Research topic	Optimal algorithm	Feature	Contribution
Band gap	RF	Atomic number, atomic orbitals, electronegativity, formation energy, etc.	Prediction of eleven undiscovered materials with desired band gaps and formation energies.[a]
	RF	Electronegativity, molar mass, etc.	Screening of four new quaternary sulfur-based semiconductor materials.[b]
	SVR	As, P, Sb percentages, strain, etc.	Importance of high-quality first principles computational training data for ML model prediction accuracy.[c]
	SVR	Average mass, electronegativity, phase type, etc.	Best results were obtained using PBE results and element information as features.[d]
	SVR	Covalent radius, electronegativity, valence, etc.	Best results obtained using PBE results and element information as features.[e]
	SVR	Atomic radius, atomic number, valence, etc.	ML predictions of band gaps are closer to experimental values than DFT.[f]
	GBR	Tolerance factor, octahedral factor, electronegativity, polarizability, etc.	Establishment of a comprehensive structure-property map of hybrid organic–inorganic perovskites.[g]
	GBR	Crystal space group, electronegativity, energy levels, atomic distance, etc.	Discovery criteria for novel lead-free perovskites through statistical analysis of selected features.[h]
	GBR	Octahedral factors, ionization energy, electron affinity, polarizability, etc.	Feature vectors, including element-related properties and DSOf, provide the best prediction accuracy.[i]
	GPR	Grain size, lattice parameters, etc.	High prediction accuracy and stability using a small sample size (65 samples) in the GPR model.[j]
	ANN	Reciprocal of layer number, ionic radius, oxidation state, etc.	Screening of promising candidate perovskite materials.[k]
	GPR, active learning	Atomic radius, electronegativity, ionization energy, etc.	Discovery of optimal heterostructures with minimal first-principles calculations.[l]
Defect properties	Linear regression	Branch-point energy, charge neutrality level	Prediction of point defect formation energies and transition levels using only branch-point energy from bulk-phase unit cells.[m]
	RF	Element properties, cell defect properties	Optimal features combining impurity atom's element properties with energy and electronic information calculated through low-cost cell defect calculations.[n]
	RF	Structural descriptors, chemical descriptors, comparison of defect structures with host materials	Exploration of two-dimensional materials' defect design space, revealing numerous promising candidate defects.[o]
Defect classification	CNN-TL	SEM image	Introduction of CNN-based TL for automatic defect classification.[p]
	CNN	Photoluminescence spectrum images	Introduction of FCN framework for pixel.[q]
Device characteristic	Logistic regression	Photoresponse current	Higher reconstruction accuracy using an ML algorithm compared to traditional computational methods.[r]
	GPR	*J*–*V* curve	Discovery of 5% Mn-doped CMZTSSe solar cells with an efficiency of 8.9%.[s]
	Ensemble learning	Voltage, current, power, etc.	Tracking the maximum power of PV panel systems.[t]
	Ensemble learning	Resistivity, emission layer resistance, diffusion boat, and tube numbers	Understanding solar cell efficiency and its temporal trends.[u]
	ANN, CNN	Doping concentration, voltage, barrier layer thickness, *C*–*V* curve	Analysis of the *C*–*V* curve of barrier infrared detectors.[v]
	DL	Molecular fingerprints, molecular physics, chemical properties	Efficient additive screening in perovskite LEDs.[w]
	RNN	Solar zenith angle, dry bulb temperature, cloud cover	Comparison of physical and ML models in estimating solar radiation and photovoltaic power accuracy.[x]
	LDA	[Zn]/[Ge] ratio, Raman spectra, photovoltaic parameters	Establishing relationships between Raman spectra, composition, and efficiency.[y]

Table 4. Continued.

Research topic	Optimal algorithm	Feature	Contribution
Crystal structure	RF	Atomic number, atomic radius, electronegativity, etc.	Identification of eight descriptors for crystal structure determination.[z]
	RF	BVVS, atomic number, atomic radius, electronegativity, etc.	Ten feature descriptors to recognize crystal structures.[aa]
	CNN	2D diffraction fingerprint	Introduction of 2D diffraction fingerprint.[ab]
Thin film morphology identification	GAN, ANN	2D standard deviation, similarity index, etc.	Selection of optimal features using a thermal diffusion search algorithm and generation of enhanced morphology images using SinGAN.[ac]
Nanopore identification	CNN	TEM images	Image segmentation methods applied to *in-situ* TEM image analysis.[ad]
Identifying chemical composition	CNN	UV-vis spectra	Effective identification of component information from UV-vis spectra.[ae]
Stability	GPR, active learning	Voronoi tessellation	Generalization of the model for searching binary metal-oxide structures.[af]
DOS	GPR	SOAP	Simulation and explanation of complete phase transitions in disordered silicon.[ag]
Energy, quality	ANN	Effective mass, band gap, lattice parameters	Reduction in solving Schrödinger's equation time using ANN.[ah]
Orbital energies	GCN	Molecular graph	Effective use of DL models for virtual screening of optimal host molecules and emitting molecules for OLEDs.[ai]
Luminous intensity	RNN	Humidity and photoluminescence time series	Framework for designing stable perovskite devices.[aj]
Molecular properties	ANN	Molecular fingerprints	Discovery of new OLED materials with external quantum efficiency up to 22%.[ak]

[a]Ref. [28], [b]Ref. [29], [c]Ref. [26], [d]Ref. [100], [e]Ref. [101], [f]Ref. [102], [g]Ref. [103], [h]Ref. [104], [i]Ref. [105], [j]Ref. [108], [k]Ref. [106], [l]Ref. [38], [m]Ref. [112], [n]Ref. [113], [o]Ref. [114], [p]Ref. [36], [q]Ref. [127], [r]Ref. [126], [s]Ref. [120], [t]Ref. [125], [u]Ref. [119], [v]Ref. [117], [w]Ref. [124], [x]Ref. [118], [y]Ref. [123], [z]Ref. [33], [aa]Ref. [35], [ab]Ref. [34], [ac]Ref. [128], [ad]Ref. [130], [ae]Ref. [129], [af]Ref. [39], [ag]Ref. [43], [ah]Ref. [109], [ai]Ref. [31], [aj]Ref. [32], [ak]Ref. [23].

approach), feature selection, and main contributions. Choosing an appropriate model needs to consider the specifics of the task, and prior knowledge can help researchers identify more effective methods for dealing with specific problems.

3.1. Crystal structure

Studying crystal structures provides essential information for deeply understanding materials' fundamental properties, structural characteristics, and behavior. Takahashi Keisuke and Takahashi Lauren[33] demonstrated that Gaussian mixture models are used to understand the data structure of material databases, while RF is used to predict crystal structures. Ziletti *et al.*[34] constructed a CNN model to classify structures based on crystal symmetry automatically. They proposed a descriptor called 2D diffraction fingerprint. Zhang *et al.*[35] utilized bond-valence vector sum (BVVS) and elemental information to describe the complex geometric shapes of perovskites. It has been demonstrated to adapt well to small-sample datasets for prediction without constructing large datasets, and even with just 10 feature descriptors, it can identify space groups and crystal systems.

By learning the latent representation space of crystal structures, the variational autoencoder (VAE) can generate new structures with features similar to known ones. Banko *et al.*[90] utilized the crystal structure representations learned by VAE to reveal latent information, such as structural similarities in texture diffraction patterns. Ren *et al.*[91] introduced a framework for universal reverse design (not limited to a specific set of elements or crystal structures). It features a universal reversible representation encoding of crystals in both real and reciprocal space and attributes structural latent spaces from the VAE.

The crystal structure can be represented as a crystal structure graph (CSG). GNNs provide a powerful tool for understanding, modeling, and optimizing crystal structures. Xie and Grossman[92] utilized crystal structures for predicting material properties and designing knowledge extraction. The authors proposed a crystal graph convolutional neural network (CGCNN) framework. The crystal structure encodes atomic information and bonds interactions between atoms. Then, a CNN is constructed on the CSG to automatically extract the optimal representation of the target property through training. CGCNN is able to predict eight different properties with high accuracy.

GNNs provide a learning approach for representing material structures as graph structures. Researchers have proposed

GNNs to predict different properties of materials and continuously optimize the structure and performance of the models. Chen *et al.*[93] developed universal materials graph network (MEGNet) models for predicting the properties of molecules and crystals. Karamad *et al.*[94] proposed the orbital graph convolutional neural network (OGCNN), which showed the potential to accurately predict the properties of materials such as metals, minerals, and oxides. Park and Wolverton[95] developed an improved variant of the CGCNN model (iCGCNN), which showed a reduction of about 20% in prediction error compared to the original CGCNN model. Choudhary and DeCost[96] presented an atomistic line graph neural network (ALIGNN), which utilizes bond lengths and bond angles as the edge features of the line graph and explicitly and efficiently incorporates the delicate atomic structure information. Lee and Asahi[97] proposed TL using a crystal graph convolutional neural network (TL-CGCNN) to improve prediction accuracy. Wang *et al.*[98] proposed an attention mechanism-based crystal graph convolutional neural network (A-CGCNN), which enhances the integration of crystal topology and atomic features.

Table 5 compares these models' differences in applications and methods. Most of these models are applied to material property prediction, the inputs are mostly CSG, and the methodology draws on GNNs for modeling, with various improvements based on CGCNN.

Table 5. Process optimization flowchart.

Model	Input data	Method overview
CGCNN	CSG	Encoding atomic information and bonding interactions between atoms.
MEGNet	Crystal chemistry graph	Introducing a unified free energy model with global state variables, extracting interpretable element embeddings, and capable of handling both molecules and crystals.
OGCNN	Crystal atomic orbital graph	Enhancing feature representation by incorporating orbital interactions and using an encoder–decoder structure.
iCGCNN	CSG	Connecting each node to its Voronoi neighbors explicitly introduces 3-body terms in convolution functions and optimizes edge embeddings.
ALIGNN	Lattice-based line graph	Utilizing bond length, bond angle, and other factors as edge embeddings, performing message passing alternatively on atomic bond graphs and corresponding line graphs.
TL-CGCNN	CSG	Improving CGCNN through TL techniques.
A-CGCNN	CSG	Incorporating attention mechanisms into CGCNN.

3.2. Properties research

One of the primary objectives of ML is to predict the physical and chemical properties of optoelectronic materials. We can predict and understand materials' optical and electronic properties more accurately, which is crucial for optoelectronic devices, energy conversion, and optical applications. Additionally, research on electronic structure and defects profoundly impacts optoelectronic materials because they directly relate to the material's optical and electronic properties.[99] By delving into the study of these two aspects of properties, we can highlight the importance of ML in optoelectronic materials research and provide powerful tools and methods for the discovery and performance optimization of new materials. These studies contribute to advancing the forefront of science and provide essential support for innovative technologies and sustainable development.

3.2.1. Electronic structure

Bandgap is an essential concept in electronic structure, which describes the energy difference between the valence band and conduction band in a material. By predicting and controlling the bandgap of semiconductor materials, electronic devices such as transistors, solar cells, and photodetectors can be designed for different applications. When researchers use different models to solve similar problems in different studies, comparing the performance and results of these models can reveal commonalities and differences between them. The SVR and GBR models are strong for bandgap prediction.

Mondal *et al.*[26] used ML to predict the properties of the bandgap for quaternary III–V semiconductors. The author emphasized that the most widely used ML model in semiconductor material bandgap prediction is SVM. They trained an SVR model to predict the bandgap and utilized the trained SVC model to filter out unreliable data. The authors also provided a comprehensive view of the material's properties, constructing bandgap phase diagrams of GaAsPSb, allowing researchers to identify which compositional ranges and strain conditions the material may be suitable for optoelectronic devices and which conditions may not.

Zhu *et al.*,[100] Huang *et al.*,[101] and Zhuo *et al.*[102] used first-principles calculations combined with ML to predict the electronic structure properties of materials. Zhu *et al.*'s research objective was to predict the bandgap, valence band maximum (VBM), and conduction band minimum (CBM) of two-dimensional materials. Huang *et al.* aimed to predict nitride materials' bandgap and band offset, and Zhuo *et al.* preferred to predict the bandgap of inorganic solid materials. We analyze the prediction results of bandgap in these three studies because it is a critical parameter in determining the optoelectronic properties of the materials. Table 6 analyzes the bandgap prediction results of the best models in these three studies, along with their chosen algorithms and features. Missing data indicates that the best model does not apply to this simple feature space.

Zhu *et al.* and Huang *et al.* are very close to each other regarding dataset size. Both of them use the PBE method, which is less computationally expensive, and the HSE method, which is more accurate but more computationally expensive. Still,

there are some differences in the specific content and focus of the studies.

Table 6. Experimental comparison.

Researchers	Data size	Algorithms	Feature	Bandgap RMSE (eV)
Zhu *et al.*[a]	451	LR	PBE results, bandgap, VBM, CBM.	–
		RFR	Element features only.	0.57
		Radial SVR (best)	PBE results and element features.	0.13
Huang *et al.*[b]	300	Linear SVR	18 element properties.	0.298
		Poly SVR	Expanded to 26	
		Radial SVR (best)	dimensions via	0.293
		LR	feature engineering.	
		ANN	With an additional PBE	0.099
		DNN	bandgap feature.	
Zhuo *et al.*[c]	3896	Linear SVR	34 element properties.	0.45
		Poly SVR		
		Radial SVR (best)		
		KNN		
		KRR		

[a]Ref. [100], [b]Ref. [101], [c]Ref. [102].

Zhu *et al.* used three models to predict the bandgap and band positions of 2D semiconductor materials and constructed three different feature sets. Set-I relies on the results of PBE calculations, Set-II is based on the basic properties of the elements only, and Set-III is a combination of Set-I and Set-II. The selection and combination of these feature sets provide diversity in their study. Huang *et al.* focused on predicting nitride semiconductor materials' bandgap and band alignment, employ six models, such as SVM and LR, and construct three different feature sets. Set-I contains 18 elemental attributes, Set-II adds the bandgap obtained from PBE calculations as features to Set-I, and Set-III constructs a new 26-dimensional feature space using feature engineering. By comparison, the relative importance of each attribute of SVR-based bandgap prediction is electronegativity, covalent radius, valence, and first ionization energy in descending order. Zhuo *et al.* trained an ML model using only experimentally measured bandgap and compositional information to quickly and accurately predict the material bandgap. The model predictions are closer to the experimental values than the DFT calculations and can accurately predict the bandgap of inorganic solids.

The three studies explored the applicability of the modeling algorithm in predicting the material properties and provided valuable experience for subsequent studies. The results show that the SVR model can predict the bandgap of materials more efficiently than the DFT calculation with a certain accuracy. In addition, by comparing the RMSEs of bandgaps of the three articles, we can find that the error of predicting bandgaps is still too high compared with the first two studies, although Zhou *et al.* have a larger dataset. So, the accuracy of bandgap prediction can be improved by utilizing PBE calculation results and elemental features.

Similarly, Lu *et al.*,[103] Im *et al.*,[104] and Ma *et al.*[105] had combined ML with first-principles calculations to discover desired materials rapidly. Lu *et al.* collected bandgap data of reported organic–inorganic perovskites and established a nonlinear mapping between material features and bandgaps using the GBR algorithm. They trained six suitable lead-free candidate materials out of unexplored organic–inorganic perovskites. Further first-principles calculations on these six candidates' thermal stability, environmental stability, and electronic structures revealed three lead-free hybrid organic–inorganic perovskites that meet the design requirements for ideal optoelectronic materials. Im *et al.* generated a database of hypothetical lead-free halide double perovskites and obtained their enthalpies of formation and bandgaps using first-principles calculations. Prediction models for enthalpy of formation and bandgap were built using the GBR algorithm, giving importance scores for each feature. Ma *et al.* acquired geometric and electronic structure data of 2D octahedral oxyhalides as training datasets for ML models. After testing SVR, RF, bagging, and GBR algorithms, the GBR algorithm was chosen to build the prediction model.

However, there are some differences in feature engineering and research objectives. Lu *et al.* obtained 14 critical features, including structural characteristics and elemental properties, from 30 initial features, thereby discovering stable lead-free hybrid organic–inorganic perovskites for solar cells. Im *et al.* used 32 features, including octahedral factors and elemental properties, to find lead-free perovskites for solar cells. Ma *et al.* obtained 26 essential features, including stacked octahedral factors and elemental properties from 62 initial features, discovering new two-dimensional oxyhalides for optoelectronic devices. Despite different research goals, the combined computational and GBR for materials discovery is similar, demonstrating the applicability of this approach to discover various types of optoelectronic functional materials.

In addition, several studies utilized different approaches to predict the band gap. Allam *et al.*[106] utilized an ANN model to explore the chemical space of possible layered perovskite structures and to screen suitable material candidates as solar cell materials. They found the number of layers, anion oxidation state, and cation ionic radius to be the most critical factors affecting the bandgap. Choubisa *et al.*[107] proposed the deep adaptive regressive weighted intelligent network (DARWIN) framework to accelerate material discovery. A GCN model was utilized to predict the bandgap. The authors used DARWIN to extract several chemical rules that explain the direct–indirect nature of bandgaps, such as the fact that p-block elements with larger atomic masses are more likely to form direct bandgaps. These rules reconfirmed the

patterns reported in the paper and provide new statistically significant chemical patterns. In designing stable UV-emitting direct bandgap materials, the authors found from halide materials that electronegativity differences within an optimal range are an essential feature for predicting the performance of such materials. Based on this criterion, the authors successfully synthesized K_2CuCl_3 and characterized its structure and luminescence properties. The results showed that the emission wavelength of K_2CuCl_3 is lower than 400 nm, which meets the requirements of the designed UV-emitting materials. First-principles calculations also verified the bandgap and luminescence properties.

Gao *et al.*[29] screened four new quaternary semiconductors (Ag_2InGaS_4, $AgZn_2InS_4$, Ag_2ZnSnS_4, and $AgZn_2GaS_4$) with direct bandgap by RF algorithm. Takahashi *et al.*[28] visualized and analyzed the data and found several implied trends determining the material's band gap. The authors built an RF model to predict the bandgap of perovskite materials, and 11 previously undiscovered perovskite materials with ideal bandgaps and formation energies were discovered.

By adjusting the types and concentrations of different doping elements, the electronic structure of material films can be modified, thereby tuning their bandgap. Zhang and Xu[108] developed a GPR model to predict bandgap based on lattice constants and grain size. The authors collected experimental data on 65 doped ZnO thin films, and the model achieved excellent prediction accuracy on bandgap. It can assist in designing and comprehending these multi-doped materials to make them have a tunable bandgap. The GPR model is a powerful regression method for small sample and uncertainty modeling.

In addition to the bandgap, highest-occupied molecular orbital (HOMO) and lowest-unoccupied molecular orbital (LUMO) also provide crucial information about the material's electronic structure and energy level distribution. Saleh *et al.*[27] predicted the energy levels of organic semiconductor materials in photodetectors to find new building blocks by mining a photovoltaic database. ML methods such as light gradient boosting machine (LGBM) and GBR model were used to train models for predicting HOMO and LUMO energy levels and power conversion efficiency. Jeong *et al.*[31] developed a GCN model to predict the HOMO and LUMO energies of molecules. Using TL, they initialized the model with parameters from a previously developed model for predicting spectral properties and then trained it on the HOMO and LUMO energy databases. Comparing from-scratch training to TL, they found the TL model had lower prediction errors. Using this model to assist in designing and preparing deep-blue OLED devices resulted in an external quantum efficiency of 6.58%. The model enabled rapid and efficient prediction of HOMO and LUMO energies to aid in designing optoelectronic materials.

Studying the properties of quantum wells allows for an in-depth understanding of the electronic structure of semiconductor materials. The ground state energy and effective mass are usually properties closely related to a quantum well's electronic structure. Da Silva Macedo *et al.*[109] used ML to rapidly and accurately predict the electronic properties of quantum well semiconductor heterostructures. The authors obtained the eigenstate energies and effective masses for thousands of heterostructure configurations to train an ANN model. The trained model can accurately predict the properties of new heterostructures in milliseconds, saving approximately 1000 times the time it would take to solve the equations directly. This significantly reduces the computational barriers in designing quantum well devices.

Crystalline materials exhibit excellent electronic and optical properties in optoelectronic devices. Nevertheless, amorphous materials have extensive prospects for specific applications such as flexible electronics and solar cells.[110,111] Deringer *et al.*[43] revealed the amorphous silicon's structural and electronic phase transition at the atomic scale under high pressure. Predicted changes in the electronic structure indicate a semiconductor-to-metal transition in amorphous silicon under high pressure. This transition significantly impacts the optoelectronic response of amorphous silicon. Although optoelectronic applications are primarily focused on crystalline silicon, the methods used in this study can be extended to investigate the structural stability and electronic structure of other optoelectronic materials.

3.2.2. Defects

Defects can alter crucial properties of materials, such as electronic structure and conductivity. Understanding and controlling defects can assist in optimizing the performance of materials to meet the requirements of various applications.

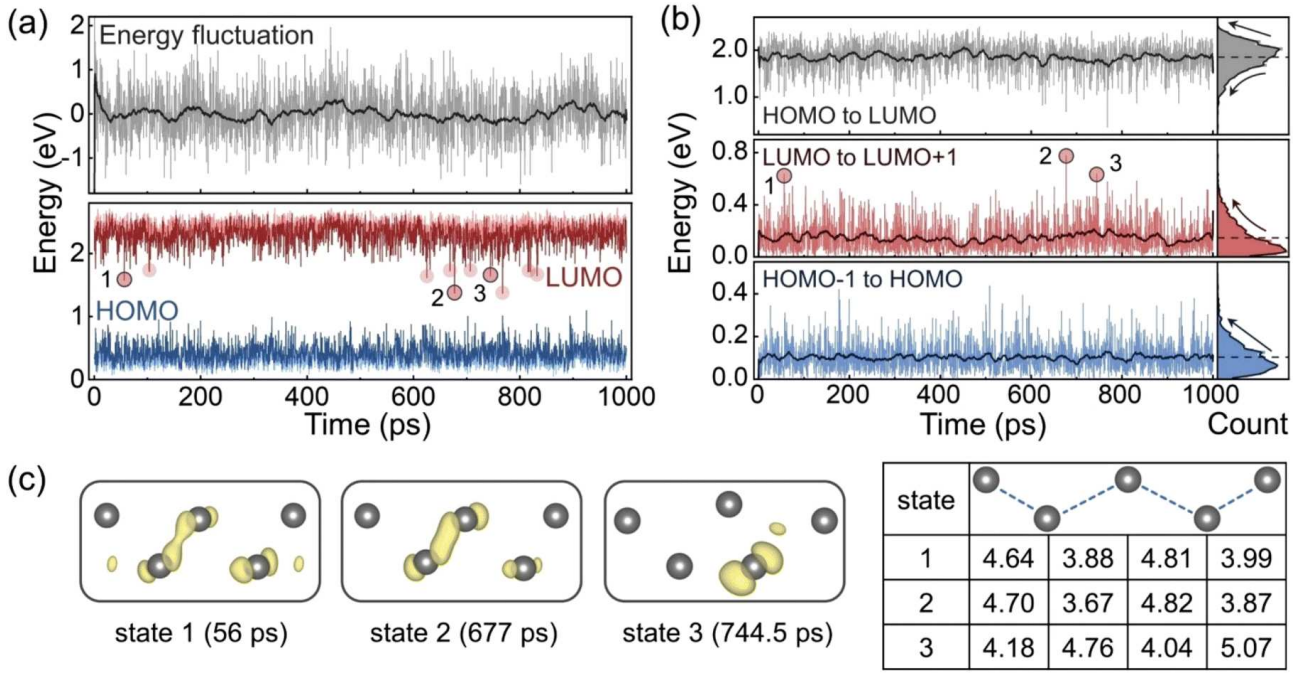
Varley *et al.*[112] showcased a computationally efficient model capable of accurately predicting the formation energies of cation vacancies and pinpointing the positions of their electronic states in an extensive range of II–VI and III–V materials. Importantly, this model eliminates the need for calculations involving supercells composed of $\sim$ 100 atoms, relying exclusively on parameters derived from the primitive unit cell of bulk materials, typically consisting of 2–4 atoms.

By using a dataset generated through DFT, Mannodi-Kanakkithodi *et al.*[113] developed accurate and widely applicable models for predicting impurity properties in cadmium-based chalcogenides. Researchers initially converted the material systems into numerical descriptors, encompassing impurity element attributes, defect site information, and low-cost single-cell calculated defect attributes. Subsequently, modeling was performed using RF, kernel ridge regression (KRR), and LASSO regression, with their predictive performances

compared. The RF model can precisely predict impurity properties in cadmium selenide-sulfide mixtures. Furthermore, including PBE-calculated transition levels as descriptors significantly enhances the accuracy of transition level predictions.

Frey *et al.*[114] utilized a pre-trained MEGNet model for TL to predict critical properties of 2D materials to identify ideal host materials for quantum dot emission. They created nearly ten thousand defect structures and trained an ensemble ML model capable of accurately predicting fundamental defect properties without requiring first-principles calculations. The research team identified 100 promising defect structures, including deep center defects suitable for quantum emission and engineered dopant defects for resistive switching.

Wang *et al.*[115] investigated the fluctuation of defect energy levels in metal halide perovskites (MHPs) and its effect on material properties. They developed an MLFF trained on DFT data to perform nanosecond molecular dynamics (MD) simulations. They demonstrated that the intrinsic softness and anharmonicity of the MHP lattice lead to large defect level fluctuations. The study provided insights into how dynamic defects can produce desirable MHP properties for solar cells and other optoelectronic applications. Liu *et al.*[116] studied the structural dynamics of the $\Sigma5(120)$ grain boundary (GB) in $CsPbBr_3$ on the nanosecond timescale and how the structural changes affect the electronic structure and potentially create detrimental mid-gap trap states using MLFF. The authors first constructed a $CsPbBr_3\Sigma5(120)$ GB model and performed *ab initio* DFT calculations to analyze its initial structure and band structure. Based on the DFT data, they trained an MLFF model and used it to perform longer timescale ML-based MD (MLMD) simulations. The authors established correlations between structural and electronic property changes by analyzing the GB's structural fluctuations and periodic DFT calculations. Figure 5 shows the simulation results based on the MLFF at the nanosecond scale. The article investigated grain boundary defects in $CsPbBr_3$ materials, revealing how the structural dynamics of grain boundaries affected the electronic structure and providing guidance for defect repair.



state				
1	4.64	3.88	4.81	3.99
2	4.70	3.67	4.82	3.87
3	4.18	4.76	4.04	5.07

Fig. 5. Results based on MLMD. (a) Evolutions of the potential energy fluctuation and the frontier energy levels in the MLMD trajectory. (b) Evolutions of the energy level differences in the MLMD trajectory and the distribution histograms. (c) LUMO charge densities and Pb configurations at the labeled states.[116]

3.3. Materials and devices optimization

The optimization of materials and devices in the field of optoelectronics is essential for advancing technology, improving performance, and addressing the growing demand for efficient and sustainable optoelectronic solutions.

By predicting device properties, unnecessary device fabrication or expensive experimentation can be avoided, thus providing a more cost-effective route to device optimization. Using physics-assisted ML, Pratik *et al.*[30] established the relationship between process conditions and device capacitance–voltage (C–V) curves. The models were trained using theoretical C–V data at frequencies of 1 kHz, 3 kHz, 5 kHz, 10 kHz, 50 kHz, and 100 kHz and measured C–V data at 1 kHz and 100 kHz. They were subsequently tested with 3 kHz, 5 kHz, 10 kHz, and 50 kHz frequencies. The most significant advantage of their data-driven model is that it scales smoothly to any number of process parameters. Glasmann *et al.*[117] investigated two ML algorithms for the C–V characteristics of InAsSb-based barrier infrared detectors: an ANN was used to predict the capacitance values of the devices, and

the results showed that the neural network model had a high generalization ability. As shown in Fig. 6, the CNN model is used to predict the doping concentration of the absorber, baffle, and electrical contact layers, as well as the thickness of the baffle layer, using the *C*–*V* curve as input. The results show that the CNN can accurately extract the structural parameter information from the *C*–*V* curves.

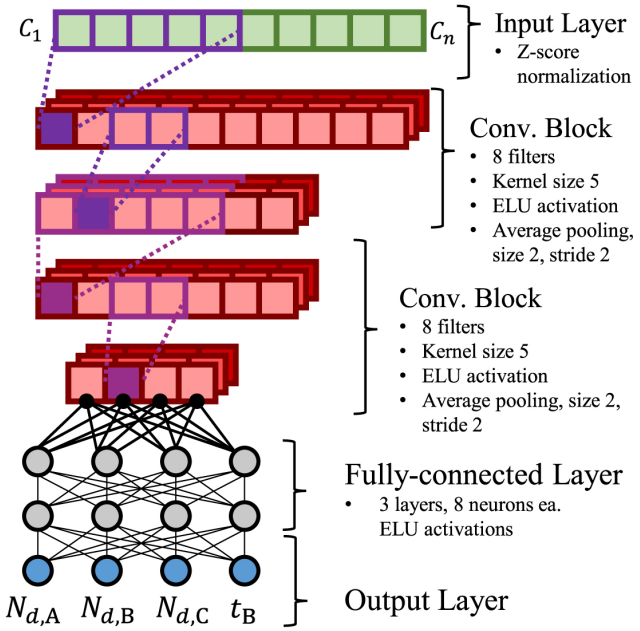


Fig. 6. The CNN framework for predicting capacitance–voltage characteristics.[117]

By carefully designing structures, researchers can customize the optoelectronic properties, advancing developments in fields such as solar cells, lasers, photodetectors, and more. Bassman Oftelie *et al.*[38] constructed a GPR model for predicting material properties of input heterogeneous structures. Additionally, the authors applied an active learning model based on Bayesian optimization (BO), which can discover the optimal heterogeneous structures with minimal first-principles calculations. Through extensive simulation experiments, Epps *et al.*[42] optimized the AI strategy for autonomous materials synthesis and proposed an efficient ensemble neural network framework based on reinforcement learning to enhance the efficiency of synthesizing quantum dot materials. The algorithm comprehensively considers the global context of the synthesis process rather than just local optimization, enabling a more efficient search for global optimality and significantly improving the success rate. Compared to traditional methods, this algorithm can simultaneously optimize multiple objectives, such as emission peak, full width at half maximum, and more, making material optimization more comprehensive and systematic. The research findings are of significant importance in advancing automated synthesis technology, and the framework and methods have broad application prospects and potential for expansion in the field of automated synthesis.

Time-series data typically encompass the variation of device performance parameters, such as photocurrent, open-circuit voltage, short-circuit current, etc. RNNs hold significant application value for handling such time-series data. Howard *et al.*[32] studied the influence of humidity on the luminescent properties of perovskite materials. They conducted dynamic luminescence measurements by controlling relative humidity and established a predictive model using RNN, demonstrating the potential application of ML in predicting the long-term stability of materials. Ramadhan *et al.*[118] systematically compared physical and ML models' performance in estimating solar irradiance and photovoltaic power. The study selected and compared ML models such as SVR, RF, RNN, gated recurrent unit (GRU), and long short-term (LSTM) memory with widely used physical models. LSTM and GRU are two commonly used variants of RNNs. They are both designed to address issues such as vanishing and exploding gradients present in traditional RNNs. Figure 7 provides a comprehensive view of the accuracy of the different models in estimating the global horizontal irradiance (GHI), direct normal irradiance (DNI), and global titled irradiance (GTI). LSTM and GRU are better for estimating solar radiation and photovoltaic power.

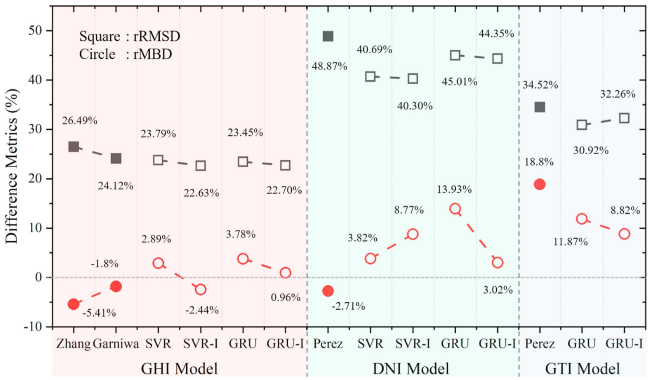


Fig. 7. Comparison of accuracy between ML and physical models on output variables.[118]

Battery conversion efficiency can directly impact the performance and availability of batteries. Wasmer *et al.*[119] used ensemble learning to predict solar cell conversion efficiency and explain the impacts of different features on the predictions. They analyzed the time trends of feature impacts, differentiating which features were more sensitive to short-term efficiency changes versus long-term efficiency levels. For example, the wet bench process significantly influenced the overall efficiency trend, belonging to short-term influencing factors. Wafer base resistivity, emitter sheet resistance, and other in-line measurement data had a persistent impact on efficiency but did not vary with time. This study effectively parsed the time trends of solar cell efficiency and its influencing factors through interpretable ML, providing valuable guidance for production process optimization.

The material's conductivity can be changed in semiconductor devices by adjusting the dopant and doping ratio. Different dopants and doping ratios can lead to other electronic structures and properties.[120,121] Li *et al.*[122] predicted the optimal manganese doping concentration to optimize the performance of Mn-doped CMZTSSe solar cells. The authors used GPR to predict the optimal Mn doping concentration of 5% in CMZTSSe solar cells. The experimental results verified this prediction, and a conversion efficiency of 8.9% was obtained for CMZTSSe solar cells based on 5% Mn doping. Further characterization results showed that compared with pure CZTSSe, the main defect of the 5% Mn-doped CZTSSe solar cell has been changed from CuZn inversion to VCu, which is one of the reasons for the improved efficiency. This study once again underscored the reliability of GPR models for small datasets.

Grau-Luque *et al.*[123] analyzed the influence of off-stoichiometry in $Cu_2ZnGeSe_4$ on defect formation and solar cell performance using ML. A combinatorial sample containing around 200 $Cu_2ZnGeSe_4$ solar cells with different [Zn]/[Ge] compositions was prepared. The samples were characterized by x-ray fluorescence spectroscopy and Raman spectroscopy and correlated with electronic parameters. Furthermore, the Raman spectral data was analyzed using the LDA algorithm. Specifically, all the Raman spectra measured under different excitation conditions for each cell (a total of 588 spectra) were used as inputs. At the same time, efficiency, open-circuit voltage, and [Zn]/[Ge] ratio were set as outputs. Each output target was classified into four groups containing approximately equal data. Figure 8 shows the results of ML analysis of Raman spectral data for different classification targets. The training and test scores were close for the efficiency and open-circuit voltage targets. Solar cells with the highest efficiency (around 6.3%) were obtained when $1.05 < [\mathrm{Zn}]/[\mathrm{Ge}] < 1.15$. This range was considered as the optimal cationic composition. The study demonstrated that solar cell efficiency can be effectively predicted solely based on the Raman spectral and compositional data.

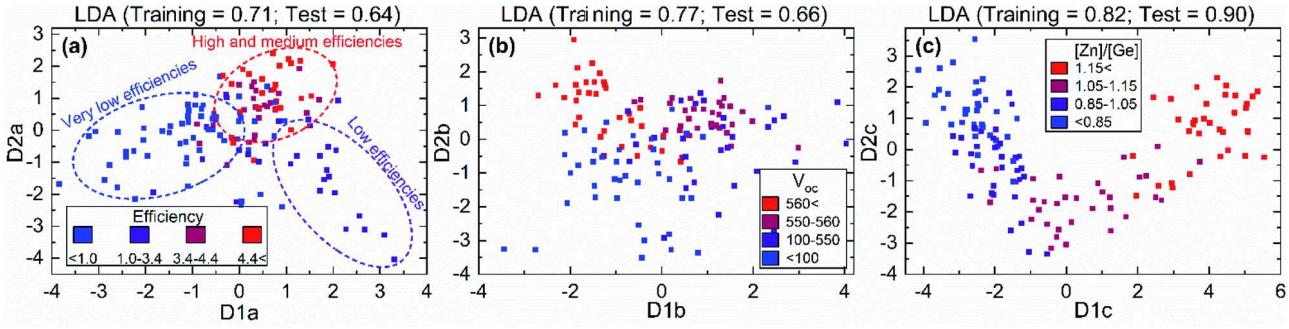


Fig. 8. Results of the ML analysis of the Raman spectroscopic data for different classification targets: (a) efficiency, (b) V_{oc}, and (c) [Zn]/[Ge] ratio.[123]

Quantum efficiency is an essential tool for assessing and optimizing the performance of photonic devices. DL provides strong support in designing highly quantum-efficient devices. Gómez-Bombarelli *et al.*[23] designed and synthesized a series of high-performance organic LED emitter materials. The researchers first conducted large-scale quantum chemistry calculations using DFT to systematically predict the optical properties of over one million emissive molecules. They then applied DL methods to prescreen the computational results and identified the most promising candidates. These molecules achieved high maximum external quantum efficiencies of up to 22% after synthesis and characterization.

Zhang *et al.*[124] developed an enhanced molecular information model based on DL to accurately predict the performance of perovskite LEDs using a small dataset of additive molecules. Using a predicted new additive, they achieved a high external quantum efficiency of 22.7% for perovskite LEDs, significantly higher than the previous best of 22%. This work demonstrated the power of DL for efficient molecular screening with limited data, providing new insights to boost the performance of emerging optoelectronic devices.

The data-driven approach based on real-world data, multidimensional data collection, and high-precision ML models can potentially support device optimization. Anshul *et al.*[125] tracked the maximum power of photovoltaic panel systems. The ensemble learning algorithm achieved the highest prediction accuracy of 98%. Wang *et al.*[126] significantly enhanced the performance of filter-free microspectrometers based on perovskite materials, providing important references for integrated microspectrometry systems.

Vandermause *et al.*[40] developed an interpretable GPR model and an efficient active learning method, enabling the rapid training of accurate machine-learned force fields with minimal first-principles computational data. They introduced a novel active learning approach named fast learning of atomistic rare events (FLARE), which successfully trained accurate potential energy fields for multiple materials. The obtained machine-learned potential energy fields exhibited a 1000-fold acceleration compared to first-principles calculations while maintaining accurate physical descriptions. In this study, the

authors demonstrated that the noise uncertainty of the GPR model accurately reflects baseline model errors, while the norm uncertainty assesses model extrapolation errors. The GPR model can actively select new training data points during molecular dynamics simulations using the FLARE method based on its own uncertainty estimates. This article lacks specific cases, but the adaptability and efficiency demonstrated by this method lead me to believe that it is highly suitable for application in the study of interfaces and surfaces of optoelectronic materials, providing a precise and efficient potential energy field description.

3.4. Material characterization

ML models can handle various types of data related to optoelectronic materials, including but not limited to images, spectral data, and electron microscopy imaging. This versatility in processing multimodal data empowers ML with remarkable flexibility and broad applicability in optoelectronic materials.

Using CNNs to recognize and classify material films and defects is a reasonable choice. Stern and Schellenberger[127] used training to test a fully convolutional network (FCN) for the defect classification of LED chips. The researchers demonstrated significant success using the DL model, particularly fully convolutional network (FCN), for LED chip defect classification tasks. Through techniques such as data augmentation, pretrained CNNs, improved connectivity mechanisms, and the application of weighted loss functions, researchers substantially enhanced the accuracy of defect classification. Furthermore, incorporating additional measurement values into the analysis further improved performance. These findings are significant for automated defect detection and quality control applications.

Imoto *et al.*[36] developed a method to automatically classify semiconductor process defect images using CNN and TL to reduce the expensive manual classification workload. The method adopts a two-stage TL strategy: firstly, a large amount of noisy labeled data is used for network pre-training, and then a tiny amount of accurately labeled data is used for model fine-tuning. Even with minimal training data, the method can achieve high classification accuracy. The results show that the method can accurately classify the main defect types and reduce the manual inspection workload by nearly 2/3, fully utilizing the advantages of TL in the case of insufficient training data and effectively reducing the labor cost of the semiconductor manufacturing process.

GAN can generate images useful for data scarcity in optoelectronic materials research. This approach is expected to improve the performance of DL models significantly and has potential for a wide range of applications. Vakharia *et al.*[128] employed ML techniques to analyze and predict surface properties of hybrid perovskite films. Initially, they utilized a GAN model known as SinGAN to generate multiple morphological images at varying scales from a single image. Subsequently, they extracted ten image quality assessment features from these generated images. To determine the most relevant feature subsets, they applied a heat transfer search algorithm for feature optimization. Finally, these selected feature subsets were input into four ML models: DT, SVM, KNN, and ANN for recognizing the morphology of different perovskite films. The findings demonstrated that combining optimized features with ANN yielded the highest prediction accuracy compared to other combinations. This study underscored the potential of using GAN-based data augmentation and ML models to recognize film morphology and defects in semiconductor manufacturing automatically.

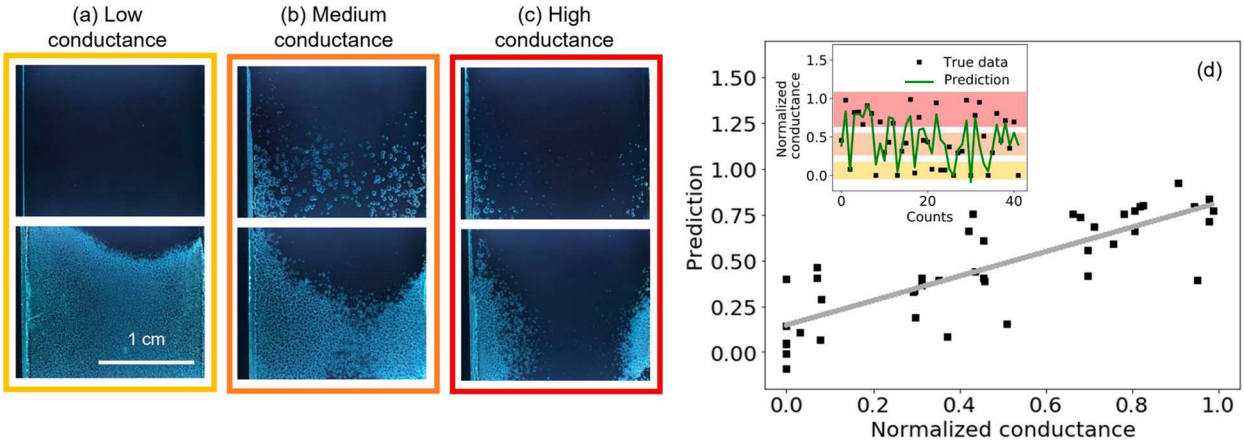


Fig. 9. Prediction of electrical conductivity using CNN. Panels (a), (b) and (c) show six DF example images representing low, medium, and high conductivity values. Panel (d) shows the predicted values compared with the true value.[131]

CNNs can also be applied to analyze properties, structures, and components. Chen *et al.*[129] developed a CNN model to establish correlations between the UV-vis absorption spectra and chemical compositions of gold nanoclusters, enabling the identification of nanocluster compositions from featureless absorption spectra. Wang *et al.*[130] demonstrated the

feasibility of CNN-based TEM image analysis for the structural control of 1T-phase $CrTe_2$. The approaches in this study could be extended to the structural control of other 1T or 2H phase 2D materials, providing vital support for emerging 2D material optoelectronic devices.

Srivastava *et al.*[131] applied fundamental ML models to address various problems, such as time-series photoluminescence intensity prediction, extrapolative device power output prediction, and extracting conductivity information from images. They proposed a holistic, integrated ML pipeline concept encompassing multiple stages of perovskite solar cell development, from component screening to material fabrication and device testing. This allowed the establishment of feedback loops between different stages, enabling end-to-end learning and optimization from material design to device enhancement.

We will use the prediction of conductivity as an example. They constructed a CNN model using dark field (DF) images. The researchers divided the DF image data of 210 spiro-OMeTAD films into training, test, and validation sets in the ratio of 16:5:4. Seven conductance measurements were acquired for each image at different locations using *I*–*V* data. Then, all measurements for each image were averaged and normalized. Data enhancement (rotation, scaling) was performed during the training process to obtain a model that predicts the conductance values of individual images. In Fig. 9, they compared the conductivity predicted using the CNN model with the actual value, yielding a correlation between the visible DF image and the conductivity.

3.5. Process optimization

Optoelectronic material processes typically involve tuning multiple parameters and variables with complex interrelationships.[132,133] ML can process high-dimensional data and find complex correlations between parameters to optimize the process.

Miyagawa *et al.*[134] optimized the process conditions of TiO_x/SiO_y/c-Si heterostructures in hydrogen plasma treatment (HPT) using the BO method to improve their surface interface passivation properties. The researchers first deposited 3 nm TiOx films on c-Si substrates, and some of the samples were oxidized by H_2O_2 to form a SiO_y interlayer before TiO_x deposition. Then, the team performed HPT on the samples, which involved six variable parameters: process temperature (T_{HPT}), process time (t_{HPT}), hydrogen pressure (p_{H2}), hydrogen flow rate (R_{H2}), radio frequency power (P_{RF}), and electrode distance (d). To optimize the process conditions for HPT, they first conducted ten sets of randomized experiments to obtain initial data, predicted the best combination of parameters based on a GPR model, and repeated the optimization cycle until the global optimum was found. The results showed that the six parameter combinations of HPT were optimized through only 15 rounds of optimization cycles, and the carrier lifetime was increased to more than three times than that of the untreated one. This study proved that BO is a practical method for efficient optimization of multidimensional parameters, which provides valuable inspiration for preparing high-efficiency silicon heterojunction solar cells using TiO_x.

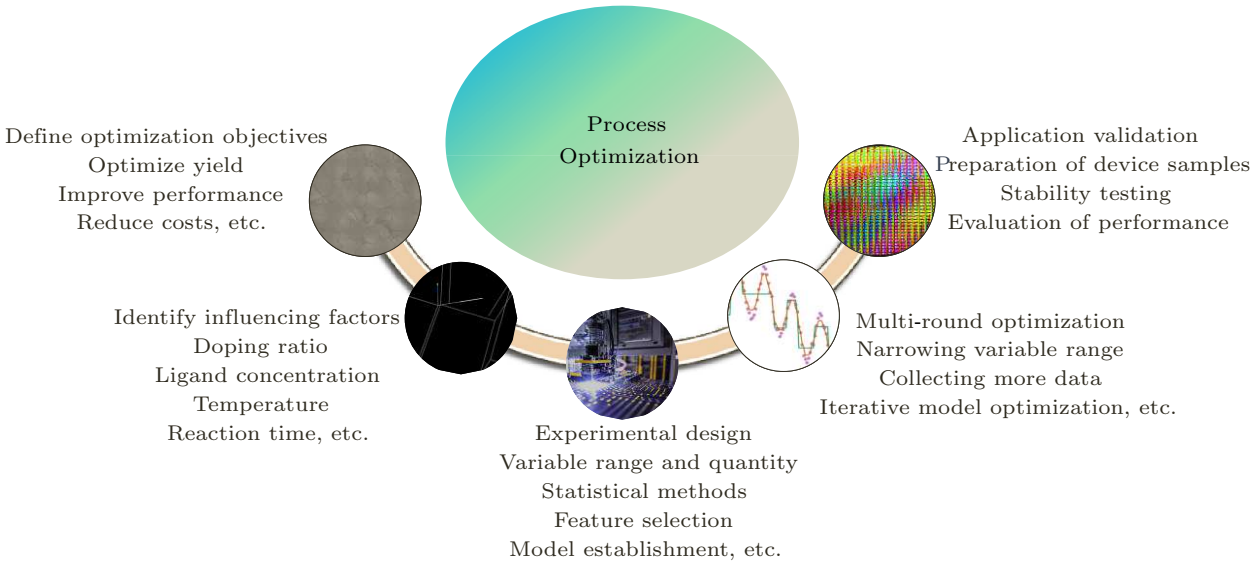


Fig. 10. Process optimization flowchart.

Wei *et al.*[135] investigated the use of ML to optimize the chemical deposition method for preparing p-type transparent conductive materials Cu–Zn–S thin films. The researchers strategically designed two rounds of orthogonal experiments: the first round examined the effects of four parameters, and the results of the first round guided the second round to optimize the parameters further. Based on the first round of experiments, the authors built an SVR model to obtain a predictive mapping map in the parameter space. The SVR model was built again for training and prediction to guide parame-

ter selection in the second round of experiments. The results show that the optimized Cu–Zn–S thin film can reach a conductivity of 430 S/cm with an average visible light transmittance of 80%, and the efficiency of the p–n heterojunction optoelectronic device it prepares is two times higher than that of the unoptimized film. This study demonstrates that ML optimization combined with the orthogonal design of experiments can effectively find the optimal parameters for preparing high-performance p-type transparent conductive CuZnS thin films and improve the efficiency of material and device development. This idea of multi-round optimization is worth learning and applying, and we have established the flow chart of process optimization in Fig. 10. The method can be extended to optimize other material synthesis processes.

ML can control parameters such as deposition rate, temperature, and gas flow rate by optimizing and adjusting the parameters of the fabrication process, and the optimization and adjustment of these parameters can improve the properties of the thin film layer and make the semiconductor device performance more reliable. Jin *et al.*[136] optimized a showerhead design for semiconductor manufacturing to improve airflow uniformity. The authors developed a CAD showerhead parameter model with 30 design variables, including the stem, back plate, porous baffle, and panel dimensions. The main design variables were the hole diameter distribution on the panel and the fillet radius of the stem. The authors collected data through CFD simulation and established the relationship between design variables and flow uniformity using linear regression and GPR models. Through sensitivity analysis, the authors found that the hole diameters in the inner regions and the second fillet radius significantly affected the airflow uniformity. The authors then used genetic algorithm (GA), stochastic gradient descent (SGD), and BO algorithms to optimize the panel design. Like conventional optimization algorithms, GA and SGD required a large amount of data and ran independently. Both algorithms found the optimal design, improving the airflow uniformity by 5.7% compared to the baseline. Based on Gaussian process modeling and acquisition functions, the BO algorithm utilized CFD feedback. It only needed one tenth of the baseline design data and improved the airflow uniformity by 9.6% compared to the baseline. The computational cost was reduced by 10 times.

However, when considering the fillet radius design variables, the performance of BO was inferior to GA and SGD algorithms. The main reasons were: first, the BO algorithm relies on the GPR model, and its fitting accuracy may not be as high as other models trained on large datasets. Second, the fillet radius variables had local optima, which BO was more prone to get trapped in. Finally, the large range of the fillet radius variables resulted in an overly large search space, which was also disadvantageous for BO. In comparison, GA and SGD were more capable of handling complex and high-dimensional design spaces through multi-round iterative search.

GA has shown potential for application in the optimal design of optoelectronic materials. For example, Zhu *et al.*[137] used a GA to optimize the design of III–V nitride light-emitting diodes and balanced luminous efficiency and operating voltage. Patra *et al.*[138] established a model based on a GA to predict and optimize the defect configurations in monolayer MoS_2, which improves the optoelectronic conversion efficiency. The BO algorithm excels at tackling highly noisy and non-convex complex problems, making it applicable to optimizing the structure and composition of optoelectronic materials. SGD, on the other hand, can be utilized for training large-scale DL models to predict the performance of optoelectronic materials. These algorithms can be employed individually or in combination to optimize optoelectronic materials cooperatively. They hold great promise in advancing the design and optimization of optoelectronic materials, thereby contributing to the further development of optoelectronic technologies.

We summarize representative ML methods in the last six years in Fig. 11, the application and development of which are expected to bring breakthroughs in optoelectronic materials. They improve the performance and efficiency of optoelectronic materials and expand the functions and applications of optoelectronic materials, providing strong support for developing a new generation of high-performance, low-cost, environmentally friendly, and sustainable optoelectronic technologies.

Figure 11(a) breaks through the computational barriers of DFT, making first-principles methods infinitely scalable. It represents a milestone in materials science. Figure 11(b) employs DL to predict complex spectra from composition and structure, avoiding the cumbersome first-principles calculations. This model is vital for advancing various functional materials, including optoelectronics. Figure 11(c), employing NLP techniques such as word embeddings, has been used for the first time to analyze scientific literature, extracting knowledge in materials science. This method has provided a fresh perspective for materials science research, and its framework can be extended to predict various functional materials and applied to paper in other scientific domains. Figure 11(d) has surpassed various limitations of empirical potentials and, relying on high precision, can be widely applied to crystalline, defective, liquid, and interfacial materials. They have opened up new frontiers in atomic simulation and provide robust support for theoretical exploration and optimization of optoelectronic materials. Zuo *et al.*[142] comprehensively evaluates the current major ML-IAPs, provides insights on the tradeoffs between accuracy, training data requirements, and computational cost, and provides benchmarks and guidance for future ML-IAPs development and applications. Figure 11(e) has

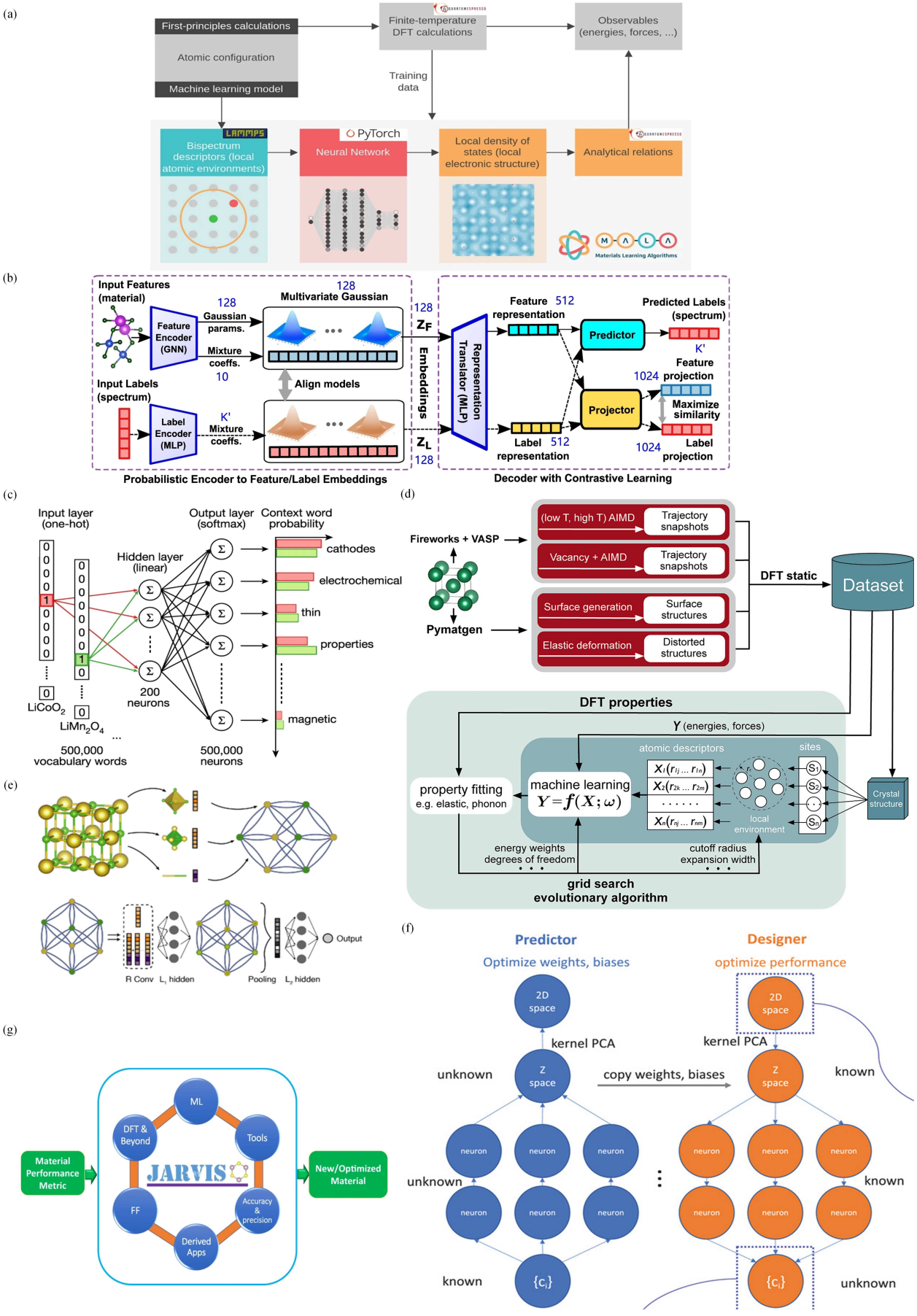


Fig. 11. Representative machine learning methods. (a) The materials learning algorithms (MALA) framework.[139] (b) The Mat2Spec model architecture.[140] (c) Corpus training process.[141] (d) ML-IAP development workflow.[142] (e) The CGCNN model.[92] (f) The inverse design network (IDN).[143] (g) An overview of the JARVIS infrastructure.[144]

pioneered a new field in crystal structure prediction and understanding using DL. It is a significant milestone in developing materials representation learning and GNNs. Figure 11(f) has introduced a new approach to materials reverse engineering. It generates structural parameters based on given spectral responses, representing an emerging research area in applying optoelectronic materials. Figure 11(g), utilizing DL for efficient first-principles calculations has significantly reduced the time required compared to traditional computational methods. It is one of the essential achievements in developing optoelectronic materials.

4. Conclusion and perspectives

We have outlined the general process of applying ML to optoelectronic materials. We have discussed standard methods for data collection and feature engineering in optoelectronic materials and have addressed the issue of overfitting. Our main focus has been on the applications of ML in optoelectronic materials. Through discussions of various research works, we have examined the impact of feature selection on research outcomes and summarized the applicability of different ML models in studying photonic materials and device properties. SVM and ensemble learning methods like GBR demonstrate exceptional performance in handling high-dimensional sparse data, making them invaluable for predicting material properties. The GPR model performs exceptionally well with small datasets, such as predicting band gaps after doping. RNNs are well-suited for managing time series data, such as assessing the stability of solar cells and estimating the lifetime of excited states. GNNs excel at modeling the intricate relationships between material structure and properties, encompassing electronic band structures, conductivity, and optical properties. CNNs are widely employed for defect detection and property prediction tasks. Reinforcement learning is effectively applied to optimize materials and device performance. TL proves its worth in scenarios where data is limited. Active learning offers an efficient means to reduce the search space and experimental costs through iterative feedback, making it a valuable tool for process optimization. ML-IAPs and MLFFs are pivotal in investigating molecular systems' properties, reactions, and behaviors.

ML has excellent potential in optoelectronic materials. However, the challenges in data requirements, transferability, and feature selection must be overcome to better apply them to real-world materials research and design.

(i) The main goal of optoelectronic databases is to collect, store, manage, and share various data, including experimental, computational, and literature data. Researchers can utilize NLP techniques to search for the required information. One of the goals of the materials and genomics initiative is to unify the materials innovation infrastructure (MII) by extending its scope to include properties, synthesis, processing, and fabrication methods of optoelectronic materials to build a comprehensive data infrastructure.

(ii) Property research is the focus of optoelectronic materials, and we take defect research as an example, where we first build a defect database containing structural and chemical information about defects and extract features based on this information. Hand-designed features can utilize domain knowledge and a priori information and apply them to semiconductor domains and tasks. However, automatically generated features can handle large-scale and complex datasets. Therefore, we can use automatic feature selection and dimensionality reduction techniques to simplify the feature space and discover more data patterns and relationships. Then, we can train different models, such as regression algorithms to predict defect properties and classification algorithms to predict defect energy level locations. According to the prediction results of the model, combined with the material's properties, we can screen out the eligible candidates.

(iii) According to the no free lunch theorem,[145] algorithm A, outperforming algorithm B on a specific dataset, must be accompanied by algorithm A's inferior performance to algorithm B in another dataset. This theorem reminds us to consider the particular material system and performance requirements when choosing an ML algorithm. Different algorithms may be better adapted to different types of materials and properties. Therefore, we must take a cautious and experimentally evaluated approach to determine which algorithm performs best in a given materials science problem. The ability of machines to actively collect and analyze data through active learning will be one of the future directions of ML for photovoltaic materials. Labeled data is often time-consuming and costly compared to unlabeled data, which is relatively easy to access, making semi-supervised learning potential for a wide range of applications in semiconductor materials science. Applying already trained models or knowledge to optoelectronic materials through TL across material systems can reduce the need for new data and increase the generalizability and adaptability of models.

With the advent of new supercomputers,[146] the development of more powerful computational capabilities and innovative algorithms will allow for the simulation of more realistic material models, which will be necessary for the discovery of tailored materials as well as for the prediction of material behaviors, and the concept of self-driving labs,[147] which combine ML, laboratory automation, and robotics to accelerate the discovery of new materials and molecules. We can expect more innovations and breakthroughs in optoelectronic materials from these emerging technologies.

Acknowledgments

Project supported by the National Natural Science Foundation of China (Grant No. 61601198) and the University of Jinan PhD Foundation (Grant No. XBS1714).

References

[1] Xing Y, Xu Y, Wu Q, Wang G and Zhu M 2021 *J. Mater. Chem. C* **9** 439
[2] Yan X, Zhen W L, Hu H J, Pi L, Zhang C J and Zhu W K 2021 *Chin. Phys. Lett.* **38** 068103
[3] Alexander Schmidt M, Argyros A and Sorin F 2016 *Adv. Opt. Mater.* **4** 13
[4] Wang D, Liu X, Fang S, Huang C, Kang Y, Yu H, Liu Z, Zhang H, Long R, Xiong Y, Lin Y, Yue Y, Ge B, Ng T K, Ooi B S, Mi Z, He J H and Sun H 2021 *Nano Lett.* **21** 120
[5] Zhang Y, Huang P, Guo J, Shi R, Huang W, Shi Z, Wu L, Zhang F, Gao L, Li C, Zhang X, Xu J and Zhang H 2020 *Adv. Mater.* **32** 2001082
[6] Tan Y J, Godaba H, Chen G, Tan S T M, Wan G, Li G, Lee P M, Cai Y, Li S, Shepherd R F, Ho J S and Tee B C K 2020 *Nat. Mater.* **19** 182
[7] Tan S, Huang T, Yavuz I, *et al.* 2022 *Nature* **605** 268
[8] Wang B H, Xing Y H, Dong S Y, Li J H, Han J, Tu H Y, Lei T, He W X, Zhang B S and Zeng Z M 2023 *Chin. Phys. B* **32** 098504
[9] Liu Y, Zhao T, Ju W and Shi S 2017 *J. Materiomics* **3** 159
[10] Hohenberg P and Kohn W 1964 *Phys. Rev.* **136** B864
[11] Alder B J and Wainwright T E 1959 *J. Chem. Phys.* **31** 459
[12] Bonate P L 2001 *Clin. Pharmacokinet* **40** 15
[13] Zhu L S and Zhao S J 2014 *Chin. Phys. B* **23** 063601
[14] Zafar M, Ahmed S, Shakil M, Choudhary M A and Mahmood K 2015 *Chin. Phys. B* **24** 076106
[15] Champagne A, Haber J B, Pokawanvit S, Qiu D Y, Biswas S, Atwater H A, Da Jornada F H and Neaton J B 2023 *Nano Lett.* **23** 4274
[16] Lugli P and Ferry D K 1986 *Phys. Rev. Lett.* **56** 1295
[17] Yang W, Li J, Huang Y and He B 2019 *Commun. Comput. Phys.* **25** 135
[18] Kalidindi S R, Brough D B, Li S, Cecen A, Blekh A L, Congo F Y P and Campbell C 2016 *MRS Bull.* **41** 596
[19] De Pablo J J, Jones B, Kovacs C L, Ozolins V and Ramirez A P 2014 *Curr. Opin. Solid State Mater. Sci.* **18** 99
[20] De Pablo J J, Jackson N E, Webb M A, *et al.* 2019 *NPJ Comput. Mater.* **5** 41
[21] Han G, Zhu W, Guo S, Zhou J, Liu Y and Deng Y 2023 *J. Mater. Sci. Technol.* **160** 18
[22] Li Y, Yang J, Zhao R, Zhang Y, Wang X, He X, Fu Y and Zhang L 2022 *J. Am. Chem. Soc.* **144** 16656
[23] Gómez-Bombarelli R, Aguilera-Iparraguirre J, Hirzel T D, *et al.* 2016 *Nat. Mater.* **15** 1120
[24] Takahashi K and Tanaka Y 2016 *Dalton Trans.* **45** 10497
[25] Ramprasad R, Batra R, Pilania G, Mannodi-Kanakkithodi A and Kim C 2017 *NPJ Comput. Mater.* **3** 54
[26] Mondal B, Westermayr J and Tonner-Zech R 2023 *J. Chem. Phys.* **159** 104702
[27] Saleh J, Haider S, Akhtar M S, Saqib M, Javed M, Elshahat S and Kamal G M 2023 *Molecules* **28** 1240
[28] Takahashi K, Takahashi L, Miyazato I and Tanaka Y 2018 *ACS Photonics* **5** 771
[29] Gao M, Cai B, Liu G, Xu L, Zhang S and Zeng H 2023 *Phys. Chem. Chem. Phys.* **25** 9123
[30] Pratik S, Liu P N, Ota J, Tu Y L, Lai G W, Ho Y W, Yang Z K, Rawat T S and Lin A S 2022 *ACS Omega* **7** 933
[31] Jeong M, Joung J F, Hwang J, Han M, Koh C W, Choi D H and Park S 2022 *NPJ Comput. Mater.* **8** 147
[32] Howard J M, Wang Q, Lee E, Howard R, Gong T, Srivastava M, Abate A and Leite M S 2020 arXiv: 2010.03702
[33] Takahashi K and Takahashi L 2019 *J. Phys. Chem. Lett.* **10** 283
[34] Ziletti A, Kumar D, Scheffler M and Ghiringhelli L M 2018 *Nat. Commun.* **9** 2775
[35] Zhang L, Zhuang Z, Fang Q and Wang X 2022 *Materials* **16** 334
[36] Imoto K, Nakai T, Ike T, Haruki K and Sato Y 2019 *IEEE Trans. Semicond. Manufact.* **32** 455
[37] Piprek J 2021 *Opt. Quantum Electron.* **53** 175
[38] Bassman Oftelie L, Rajak P, Kalia R K, Nakano A, Sha F, Sun J, Singh D J, Aykol M, Huck P, Persson K and Vashishta P 2018 *NPJ Comput. Mater.* **4** 74
[39] Flores R A, Paolucci C, Winther K T, Jain A, Torres J A G, Aykol M, Montoya J, Bajdich M and Bligaard T 2020 *Chem. Mater.* **32** 5854
[40] Vandermause J, Torrisi S B, Batzner S, Xie Y, Sun L, Kolpak A M and Kozinsky B 2020 *NPJ Comput. Mater.* **6** 20
[41] Hu M, Zhang J, Matkovic L, Liu T and Yang X 2023 *J. Appl. Clin. Med. Phys.* **24** e13898
[42] Epps R W, Volk A A, Reyes K G and Abolhasani M 2021 *Chem. Sci.* **12** 6025
[43] Deringer V L, Bernstein N, Csányi G, Ben Mahmoud C, Ceriotti M, Wilson M, Drabold D A and Elliott S R 2021 *Nature* **589** 59
[44] Sauceda H E, Gálvez-González L E, Chmiela S, Paz-Borbón L O, Müller K R and Tkatchenko A 2022 *Nat. Commun.* **13** 3733
[45] Song Y, Siriwardane E M D, Zhao Y and Hu J 2021 *ACS Appl. Mater. Interfaces* **13** 53303
[46] Zuo T, Qi F, Yam C and Meng L 2022 *Phys. Chem. Chem. Phys.* **24** 26948
[47] Chen J, Feng M, Zha C, Shao C, Zhang L and Wang L 2022 *Surf. Interfaces* **35** 102470
[48] Yılmaz B and Yıldırım R 2021 *Nano Energy* **80** 105546
[49] Masson J F, Biggins J S and Ringe E 2023 *Nat. Nanotechnol.* **18** 111
[50] Matsuo Y, LeCun Y, Sahani M, Precup D, Silver D, Sugiyama M, Uchibe E and Morimoto J 2022 *Neural Netw.* **152** 267
[51] Jordan M I and Mitchell T M 2015 *Science* **349** 255
[52] LeCun Y, Bengio Y and Hinton G 2015 *Nature* **521** 436
[53] Nadkarni P M, Ohno-Machado L and Chapman W W 2011 *J. Am. Med. Inform. Assoc.* **18** 544
[54] Zhang X, Cheng R, Feng L and Jin Y 2023 *IEEE Comput. Intell. Mag.* **18** 16
[55] Butler K T, Davies D W, Cartwright H, Isayev O and Walsh A 2018 *Nature* **559** 547
[56] Piccinotti D, MacDonald K F, Gregory S A, Youngs I and Zheludev N I 2021 *Rep. Prog. Phys.* **84** 012401
[57] Schleder G R, Padilha A C M, Acosta C M, Costa M and Fazzio A 2019 *J. Phys. Mater.* **2** 032001
[58] Choudhary K, DeCost B, Chen C, Jain A, Tavazza F, Cohn R, Park C W, Choudhary A, Agrawal A, Billinge S J L, Holm E, Ong S P and Wolverton C 2022 *NPJ Comput. Mater.* **8** 59
[59] Golmohammadi M and Aryanpour M 2023 *Mater. Today Commun.* **35** 105494
[60] Kadulkar S, Sherman Z M, Ganesan V and Truskett T M 2022 *Annu. Rev. Chem. Biomol.* **13** 235
[61] Liu D Y, Xu L M, Lin X M, Wei X, Yu W J, Wang Y and Wei Z M 2022 *Chip* **1** 100033
[62] Gražulis S, Daškevič A, Merkys A, Chateigner D, Lutterotti L, Quirós M, Serebryanaya N R, Moeck P, Downs R T and Le Bail A 2012 *Nucleic Acids Res.* **40** D420
[63] Zagorac D, Müller H, Ruehl S, Zagorac J and Rehme S 2019 *J. Appl. Crystallogr.* **52** 918
[64] Curtarolo S, Setyawan W, Wang S, Xue J, Yang K, Taylor R H, Nelson L J, Hart G L W, Sanvito S, Buongiorno-Nardelli M, Mingo N and Levy O 2012 *Comput. Mater. Sci.* **58** 227
[65] Jain A, Ong S P, Hautier G, Chen W, Richards W D, Dacek S, Cholia S, Gunter D, Skinner D, Ceder G and Persson K A 2013 *APL Mater.* **1** 011002
[66] Draxl C and Scheffler M 2019 *J. Phys. Mater.* **2** 036001
[67] Landis D D, Hummelshoj J S, Nestorov S, Greeley J, Dulak M, Bligaard T, Norskov J K and Jacobsen K W 2012 *Comput. Sci. Eng.* **14** 51
[68] Wilkinson M D, Dumontier M, Aalbersberg I J, *et al.* 2016 *Sci. Data* **3** 160018
[69] Lyngdoh G A, Zaki M, Krishnan N M A and Das S 2022 *Cem. Concr. Compos.* **128** 104414
[70] Magar R and Barati Farimani A 2023 *Comput. Mater. Sci.* **224** 112167
[71] Yoo J and Kang S 2023 *IEEE Access* **11** 26393
[72] Marukatat S 2023 *Artif. Intell. Rev.* **56** 5445
[73] Rausch J R and Kelley K 2009 *Behav. Res. Methods* **41** 85
[74] Potapov P and Lubk A 2019 *Adv. Struct. Chem. Imag.* **5** 4
[75] Wang J, Xu P, Ji X, Li M and Lu W 2023 *Materials* **16** 3134
[76] Bejani M M and Ghatee M 2021 *Artif. Intell. Rev.* **54** 6391

[77] Rupp M, Tkatchenko A, Müller K R and Von Lilienfeld O A 2012 *Phys. Rev. Lett.* **108** 058301
[78] Zhang K, Yin L and Liu G 2021 *Comput. Mater. Sci.* **186** 110071
[79] Parsaeifard B, Sankar De D, Christensen A S, Faber F A, Kocer E, De S, Behler J, Anatole Von Lilienfeld O and Goedecker S 2021 *Mach. Learn. Sci. Technol.* **2** 015018
[80] Bartók A P, Payne M C, Kondor R and Csányi G 2010 *Phys. Rev. Lett.* **104** 136403
[81] Zhang L, Han J, Wang H, Saidi W A and Car R 2018 *The 32nd Conference on Neural Information Processing Systems*, December 3–8, 2018, Montréal, Canada, p. 4441
[82] Hinton G E, Srivastava N, Krizhevsky A, Sutskever I and Salakhutdinov R R 2012 arXiv: 1207.0580v1
[83] Agiorgousis M L, Sun Y Y, Choe D, West D and Zhang S 2019 *Adv. Theory Simul.* **2** 1800173
[84] Wang A Y T, Murdock R J, Kauwe S K, Oliynyk A O, Gurlo A, Brgoch J, Persson K A and Sparks T D 2020 *Chem. Mater.* **32** 4954
[85] Fang J, Xie M, He X, Zhang J, Hu J, Chen Y, Yang Y and Jin Q 2022 *Mater. Today Commun.* **33** 104900
[86] Wei J, Chu X, Sun X, Xu K, Deng H, Chen J, Wei Z and Lei M 2019 *InfoMat* **1** 338
[87] Chen A, Zhang X and Zhou Z 2020 *InfoMat* **2** 553
[88] Xiong Z, Cui Y, Liu Z, Zhao Y, Hu M and Hu J 2020 *Comput. Mater. Sci.* **171** 109203
[89] Pant D, Pokharel S, Mandal S, Kc D B and Pati R 2023 *Sci. Rep.* **13** 3277
[90] Banko L, Maffettone P M, Naujoks D, Olds D and Ludwig A 2021 *NPJ Comput. Mater.* **7** 104
[91] Ren Z, Tian S I P, Noh J, Oviedo F, Xing G, Li J, Liang Q, Zhu R, Aberle A G, Sun S, Wang X, Liu Y, Li Q, Jayavelu S, Hippalgaonkar K, Jung Y and Buonassisi T 2022 *Matter* **5** 314
[92] Xie T and Grossman J C 2018 *Phys. Rev. Lett.* **120** 145301
[93] Chen C, Ye W, Zuo Y, Zheng C and Ong S P 2019 *Chem. Mater.* **31** 3564
[94] Karamad M, Magar R, Shi Y, Siahrostami S, Gates I D and Barati Farimani A 2020 *Phys. Rev. Mater.* **4** 093801
[95] Park C W and Wolverton C 2020 *Phys. Rev. Mater.* **4** 063801
[96] Choudhary K and DeCost B 2021 *NPJ Comput. Mater.* **7** 185
[97] Lee J and Asahi R 2021 *Comput. Mater. Sci.* **190** 110314
[98] Wang B, Fan Q and Yue Y 2022 *J. Phys. Condens. Matter* **34** 195901
[99] Liu Y H, Wang W, Xiao F, Xiong L B and Ming X 2021 *Chin. Phys. B* **30** 108102
[100] Zhu Z, Dong B, Guo H, Yang T and Zhang Z 2020 *Chin. Phys. B* **29** 046101
[101] Huang Y, Yu C, Chen W, Liu Y, Li C, Niu C, Wang F and Jia Y 2019 *J. Mater. Chem. C* **7** 3238
[102] Zhuo Y, Mansouri Tehrani A and Brgoch J 2018 *J. Phys. Chem. Lett.* **9** 1668
[103] Lu S, Zhou Q, Ouyang Y, Guo Y, Li Q and Wang J 2018 *Nat. Commun.* **9** 3405
[104] Im J, Lee S, Ko T W, Kim H W, Hyon Y and Chang H 2019 *NPJ Comput. Mater.* **5** 37
[105] Ma X Y, Lewis J P, Yan Q B and Su G 2019 *J. Phys. Chem. Lett.* **10** 6734
[106] Allam O, Holmes C, Greenberg Z, Kim K C and Jang S S 2018 *Chemphyschem* **19** 2559
[107] Choubisa H, Todorović P, Pina J M, Parmar D H, Li Z, Voznyy O, Tamblyn I and Sargent E H 2023 *NPJ Comput. Mater.* **9** 117
[108] Zhang Y and Xu X 2020 *Optik* **217** 164808
[109] Da Silva Macedo G, De Sousa Dias M R and Bezerra A T 2023 *Physica E* **146** 115513
[110] Boccard M and Holman Z C 2015 *J. Appl. Phys.* **118** 065704
[111] Heremans P, Tripathi A K, De Jamblinne De Meux A, Smits E C P, Hou B, Pourtois G and Gelinck G H 2016 *Adv. Mater.* **28** 4266
[112] Varley J B, Samanta A and Lordi V 2017 *J. Phys. Chem. Lett.* **8** 5059
[113] Mannodi-Kanakkithodi A, Toriyama M Y, Sen F G, Davis M J, Klie R F and Chan M K Y 2020 *NPJ Comput. Mater.* **6** 39
[114] Frey N C, Akinwande D, Jariwala D and Shenoy V B 2020 *ACS Nano* **14** 13406
[115] Wang B, Chu W, Wu Y, Casanova D, Saidi W A and Prezhdo O V 2022 *J. Phys. Chem. Lett.* **13** 5946
[116] Liu D, Wu Y, Vasenko A S and Prezhdo O V 2023 *Nanoscale* **15** 285
[117] Glasmann A, Kyrtsos A and Bellotti E 2021 *Mach. Learn. Sci. Technol.* **2** 025006
[118] Ramadhan R A A, Heatubun Y R J, Tan S F and Lee H J 2021 *Renew. Energ.* **178** 1006
[119] Wasmer S, Hübener K and Klöter B 2022 *Sol. RRL* **6** 2100477
[120] Jiang Q F, Lian J, Ying M J, Wei M Y, Wang C L and Zhang Y 2021 *Chin. Phys. B* **30** 097801
[121] Qian D D, Liu L, Xing Z X, Dong R, Wu L, Cai H K, Kong Y F, Zhang Y and Xu J J 2021 *Chin. Phys. Lett.* **38** 087801
[122] Li X, Hou Z, Gao S, Zeng Y, Ao J, Zhou Z, Da B, Liu W, Sun Y and Zhang Y 2018 *Sol. RRL* **2** 1800198
[123] Grau-Luque E, Anefnaf I, Benhaddou N, Fonoll-Rubio R, Becerril-Romero I, Aazou S, Saucedo E, Sekkat Z, Perez-Rodriguez A, Izquierdo-Roca V and Guc M 2021 *J. Mater. Chem. A* **9** 10466
[124] Zhang L, Li N, Liu D, Tao G, Xu W, Li M, Chu Y, Cao C, Lu F, Hao C, Zhang J, Cao Y, Gao F, Wang N, Zhu L, Huang W and Wang J 2022 *Angew. Chem. Int. Ed.* **61** e202209337
[125] Anshul A, Nitish K and Pawan D 2022 *Indian J. Pure Appl. Phys.* **60** 892
[126] Wang X L, Chen Y, Chu Y, Liu W J, Zhang D W, Ding S J and Wu X 2022 *ACS Appl. Mater. Interfaces* **14** 14455
[127] Stern M L and Schellenberger M 2021 *J. Intell. Manuf.* **32** 113
[128] Vakharia V, Shah M, Suthar V, Patel V K and Solanki A 2023 *Phys. Scr.* **98** 025203
[129] Chen T, Li J, Cai P, Yao Q, Ren Z, Zhu Y, Khan S, Xie J and Wang X 2023 *Nano Res.* **16** 4188
[130] Wang C, Zou Q, Cheng Z, Chen J, Luo C, Liang F, Cai C, Bi H, Lian X, Ji X, Zhang Q, Sun L and Wu X 2022 *Nanotechnology* **33** 085302
[131] Srivastava M, Howard J M, Gong T, Rebello Sousa Dias M and Leite M S 2021 *J. Phys. Chem. Lett.* **12** 7866
[132] Huang D M, Zhang J Y, Wang J H, Wei W Q, Wang Z H, Wang T and Zhang J J 2021 *Chin. Phys. Lett.* **38** 068101
[133] Xu Y, Wang J, Cao B and Xu K 2022 *Chin. Phys. B* **31** 117702
[134] Miyagawa S, Gotoh K, Kutsukake K, Kurokawa Y and Usami N 2021 *Appl. Phys. Express* **14** 025503
[135] Wei L, Xu X, Gurudayal, Bullock J and Ager J W 2019 *Chem. Mater.* **31** 7340
[136] Jin Z, Lim D D, Zhao X, Mamunuru M, Roham S and Gu G X 2023 *J. Intell. Manuf.* 1
[137] Zhu D, Schubert M F, Cho J, Schubert E F, Crawford M H, Koleske D D, Shim H and Sone C 2012 *Appl. Phys. Express* **5** 012102
[138] Patra T K, Zhang F, Schulman D S, Chan H, Cherukara M J, Terrones M, Das S, Narayanan B and Sankaranarayanan S K R S 2018 *ACS Nano* **12** 8006
[139] Fiedler L, Modine N A, Schmerler S, Vogel D J, Popoola G A, Thompson A P, Rajamanickam S and Cangi A 2023 *NPJ Comput. Mater.* **9** 115
[140] Kong S, Ricci F, Guevarra D, Neaton J B, Gomes C P and Gregoire J M 2022 *Nat. Commun.* **13** 949
[141] Tshitoyan V, Dagdelen J, Weston L, Dunn A, Rong Z, Kononova O, Persson K A, Ceder G and Jain A 2019 *Nature* **571** 95
[142] Zuo Y, Chen C, Li X, Deng Z, Chen Y, Behler J, Csányi G, Shapeev A V, Thompson A P, Wood M A and Ong S P 2020 *J. Phys. Chem. A* **124** 731
[143] Pei Z, Rozman K A, Doğan Ö N, Wen Y, Gao N, Holm E A, Hawk J A, Alman D E and Gao M C 2021 *Adv. Sci.* **8** 2101207
[144] Choudhary K, Garrity K F, Reid A C E, *et al.* 2020 *NPJ Comput. Mater.* **6** 173
[145] Yang X S 2012 *Int. J. Artif. Intell. Tools* **21** 1240010
[146] Chang C, Deringer V L, Katti K S, Van Speybroeck V and Wolverton C M 2023 *Nat. Rev. Mater.* **8** 309
[147] Abolhasani M and Kumacheva E 2023 *Nat. Synthesis* **2** 483