

Review

Open Access



Machine learning-enabled optoelectronic material discovery: a comprehensive review

Yu Shu, Naihua Miao*, Rize Li, Yucheng Lin, Siyu Han, Jian Zhou, Zhimei Sun*

School of Materials Science and Engineering, Beihang University, Beijing 100191, China.

*Correspondence to: Prof. Naihua Miao, Prof. Zhimei Sun, School of Materials Science and Engineering, Beihang University, No. 37 Xueyuan Road, Haidian District, Beijing 100191, China. E-mail: nhmiao@buaa.edu.cn; zmsun@buaa.edu.cn

How to cite this article: Shu, Y.; Miao, N.; Li, R.; Lin, Y.; Han, S.; Zhou, J.; Sun, Z. Machine learning-enabled optoelectronic material discovery: a comprehensive review. *J. Mater. Inf.* **2025**, *5*, 36. <https://dx.doi.org/10.20517/jmi.2025.13>

Received: 14 Mar 2025 **First Decision:** 11 Apr 2025 **Revised:** 23 Apr 2025 **Accepted:** 12 May 2025 **Published:** 29 May 2025

Academic Editors: Baisheng Sa, Sergei Manzhos **Copy Editor:** Pei-Yun Wang **Production Editor:** Pei-Yun Wang

Abstract

The development of advanced optoelectronic materials constitutes a pivotal frontier in modern energy and communication technologies, facilitating critical energy-photon-electron interconversion processes that underpin sustainable energy infrastructures and high-performance electronic devices. However, the discovery and optimization of novel optoelectronic materials face substantial hurdles arising from complicated structure-property interdependencies, prohibitive development costs, and protracted innovation cycles. Conventional empirical approaches and computational simulations usually exhibit limited efficacy in addressing the escalating demands for materials with superior stability, economic viability, and customizable electronic properties. The integration of machine learning (ML) with high-throughput screening has emerged as a transformative strategy to address these challenges. By rapidly processing large multidimensional datasets and predicting critical material properties such as electronic structure, thermodynamic stability, and charge transport behaviors, ML offers unprecedented capabilities in the efficient and rational design of high-performance optoelectronic materials. This review provides a comprehensive overview of cutting-edge ML-driven methodologies in efficient optoelectronic materials discovery with emphasis on critical workflows, data integration strategies, and model frameworks. We also discuss the challenges and prospects for ML applications, particularly in data standardization, model interpretability and closed-loop experimental validation. We further propose the potential of artificial intelligence and autonomous laboratories to build a powerful discovery pipeline to advance the development of high-performance optoelectronic materials.

Keywords: Optoelectronic materials, machine learning, high-throughput calculation, materials design



© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, sharing, adaptation, distribution and reproduction in any medium or format, for any purpose, even commercially, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.



INTRODUCTION

With the advancement of technology, optoelectronic materials have become increasingly important across various fields due to their unique and efficient energy-photon-electron conversion properties. These materials are fundamental to innovations in energy conversion, display technologies, environmental remediation, and information processing. Their diverse applications drive the development of efficient, eco-friendly, and intelligent systems^[1-4]. As shown in [Figure 1](#), optoelectronic materials can efficiently interconvert energy, photons, and electrons, playing a crucial role in multiple domains. First, these materials have extensive applications in solar cells, facilitating solar energy conversion into clean electricity^[4,5]. For example, perovskite solar cells achieve high optoelectronic conversion efficiency and offer low cost and simple fabrication advantages, revealing tremendous commercial potential^[6,7]. Additionally, optoelectronic materials are equally crucial in light-emitting devices, such as light emitting diodes (LEDs) and organic light emitting diodes (OLEDs)^[8,9]. These devices provide essential support for modern display and lighting technologies by efficiently converting photon-electron interconversion^[10]. In photocatalysis, optoelectronic materials drive redox reactions by absorbing light energy, enabling applications in environmental purification and clean energy synthesis^[11]. For instance, photocatalytic materials can decompose water and remove organic pollutants under light irradiation, thus offering novel solutions for environmental remediation and energy storage^[12,13]. Moreover, photodetectors represent another critical application of optoelectronic materials. With high sensitivity to light signals, they play an indispensable role in communications and medical imaging^[14]. By swiftly responding to light signals and converting them into electrical signals, photodetectors provide robust technological support for information transmission and processing^[15,16]. Therefore, optoelectronic materials exhibit vast application potential in clean energy conversion, display technologies, environmental purification, and information technology. Their diverse functionalities make them indispensable materials in advancing modern technology, meeting the demand for efficient, green, and intelligent solutions across a broad range of fields.

In enhancing the performance of optoelectronic materials, precise control over their electronic structures and achieving a balance between stability and cost-effectiveness are crucial. However, optoelectronic devices commonly face multifaceted challenges, including structural and environmental instability, inefficiencies in interfacial engineering, and performance bottleneck^[17-20]. Addressing these multifaceted challenges demands advanced materials design, precise interface engineering, and integrated ML-driven workflows to accelerate the identification, optimization, and experimental validation of next-generation optoelectronic materials^[21,22]. The evolution of scientific research paradigms has progressed from empirical science and theoretical science to computational science and data-driven science, as shown in [Figure 2A](#). Traditional methods for discovering new materials, such as empirical, theoretical, and density functional theory (DFT)-based methods, have long held a significant position in materials science^[23,24]. However, these methods have limitations, including long development cycles, low efficiency, and high costs, making it challenging to meet the demands of modern, rapidly advancing material science^[25,26]. With significant advances in computational power and numerical algorithms, data-driven science has emerged as a new research paradigm for rationalizing novel optoelectronic materials^[27]. This paradigm emphasizes discovering scientific laws by analyzing large datasets and fully utilizing technologies such as machine learning (ML) to process complex data sets. Unlike traditional methods, data-driven science integrates knowledge from multiple disciplines, including computer science, statistics, mathematics, and engineering, and is an important branch of artificial intelligence (AI)^[28-30]. Through ML techniques, computer algorithms can automatically improve and adapt based on experience, enabling computers to make decisions and learn. ML can leverage vast amounts of data and complex algorithms to accurately predict the properties and behaviors of materials^[31-33]. Recent studies highlight how the integration of ML and AI is transforming materials discovery by enabling faster, more efficient, and autonomous prediction, design, and synthesis of

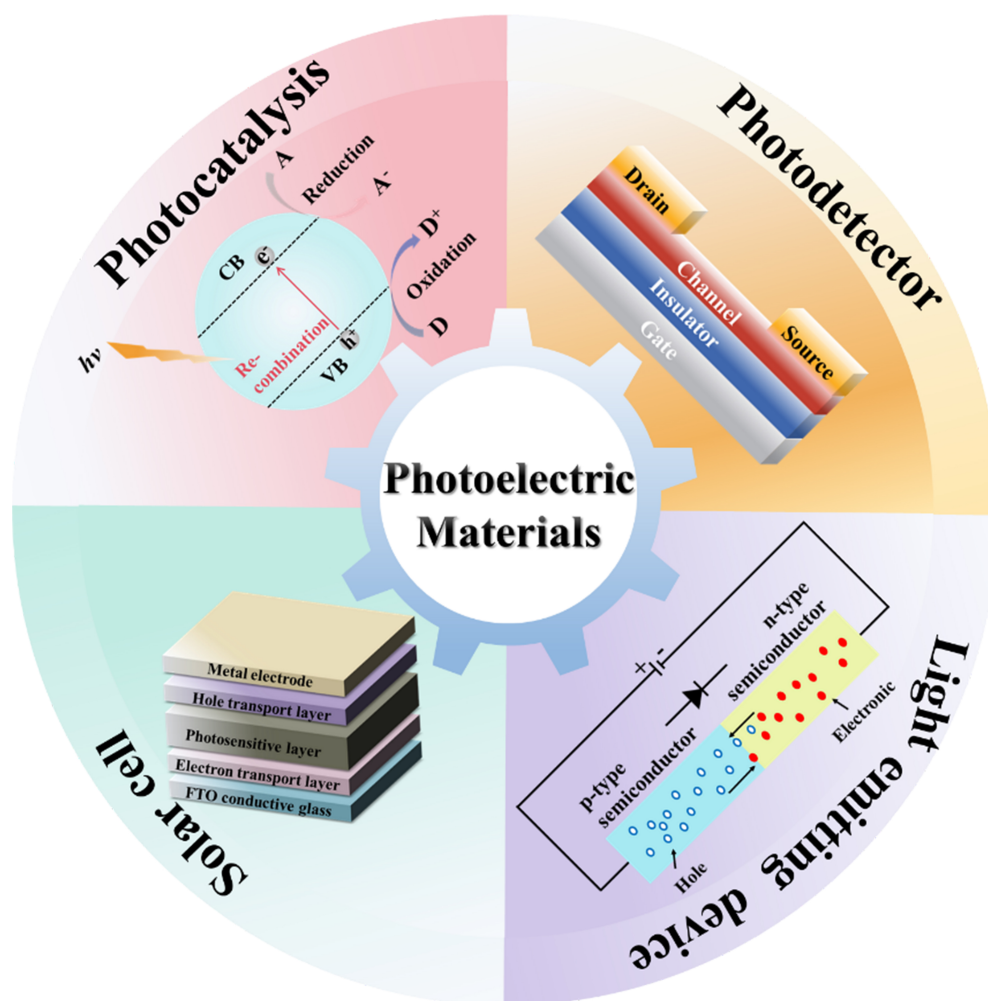


Figure 1. Application of optoelectronic materials.

advanced functional materials^[34]. Currently, we are in the fourth paradigm, dominated by data science, but the arrival of the fifth paradigm is inevitable. The fifth paradigm will fully integrate AI technologies, ushering in a new era of intelligent research. This paradigm will not only rely on big data and computer simulations but will also deeply integrate ML and AI to promote the automation and intelligence of scientific research, addressing challenges faced by traditional methods, especially in problems involving high computational complexity and uncertainty.

In this review, we briefly discuss the successful applications of high-throughput (HT) screening and ML in optoelectronic materials. Section “Overview of HT and ML” primarily introduces the basic workflow of ML in materials science and some fundamental ML algorithms. Section “HT and ML of optoelectronic materials” covers the application of HT methods in optoelectronic materials and the practical use of ML techniques in exploring electronic properties, stability, optimization of optoelectronic performance, and the design of novel optoelectronic semiconductors. Section “Summary and outlook” addresses the current challenges and opportunities in applying ML to the design and discovery of optoelectronic materials and provides some insights into the future development of this field.

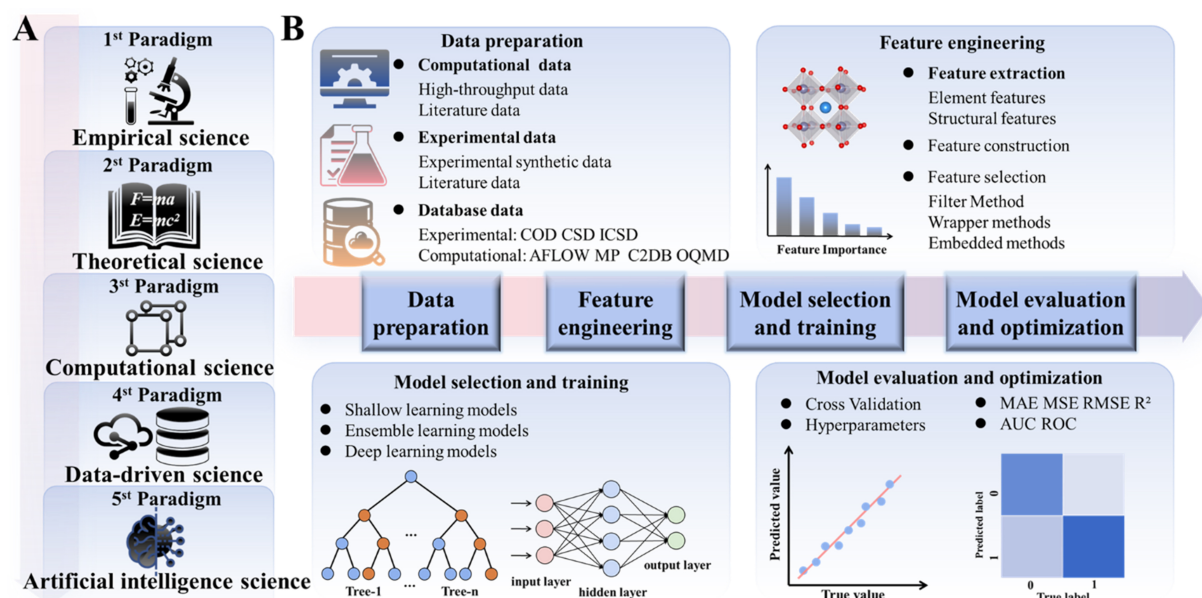


Figure 2. (A) The development of scientific research paradigms; (B) General workflow of ML in material design. From data preparation to feature engineering to model selection and training, and finally to model evaluation and optimization. ML: Machine learning.

OVERVIEW OF HT AND ML

High throughput computation

HT technology is an advanced method capable of analyzing many samples in a short period. Its core lies in simultaneously processing a vast number of samples, thereby significantly improving the efficiency of data collection and processing^[35–37]. In recent years, with the rapid development of computational capabilities, HT computational materials design has become an effective approach for discovering new functional materials. This method is widely used in various materials fields, including optoelectronic materials^[38,39], thermoelectric materials^[40,41], topological insulators^[42,43], and magnetic materials^[44,45]. HT computation utilizes first-principles calculations to establish large-scale databases from which potential candidate materials are selected under predefined constraints. This selection process relies on material descriptors accurately capturing the properties required for specific target applications. The construction of these descriptors directly influences the reliability of the screening outcomes, as they must establish a precise quantitative relationship between the intrinsic material properties and their macroscopic performance. This data-driven research and development approach, particularly in the high-precision prediction of critical parameters in optoelectronic materials, provides a breakthrough technological pathway for developing next-generation high-performance optoelectronic devices.

ML workflow

With the advancement of the Materials Genome Initiative, the importance of ML in materials science is steadily increasing^[46–48]. ML has become a powerful tool for designing and screening high-performance materials. It can reveal quantitative structure-property relationships between the physical and chemical properties of new materials and their atomic parameters, chemical compositions, process parameters, and other factors. Once a relational model between the data is successfully constructed, the relationship model can be used to predict the performance of the designed materials. These predictions can then be validated through synthesis and characterization, thereby screening for high-performance materials^[49–51]. The general process of using ML for materials design is illustrated in Figure 2B, where data collection, feature engineering, model selection and training, model validation and optimization are the most important steps.

Data preparation

Data preparation is a crucial step in the ML process, as the quality and relevance of the data directly impact the model performance and accuracy^[52]. Data for ML models can be obtained through various methods, including literature mining^[53], computational data^[54], experimental data^[55], and database data^[56,57]. While accurate data from literature or DFT calculations is highly valuable, it is often limited in quantity. Conversely, databases can quickly provide large volumes of data, though specific materials may not always be included. Experimental databases such as the Cambridge Structural Database (CSD)^[58], the Inorganic Crystal Structure Database (ICSD)^[59], PubChem, and the Crystallography Open Database (COD)^[60] offer essential structural data for ML studies. For example, the ICSD contains over 240,000 crystal structures of inorganic compounds. Key computational databases include AFLOW^[61], the Materials Project (MP)^[62], the Computational 2D Materials Database (C2DB)^[63], and the Open Quantum Materials Database (OQMD)^[64], which provide computed properties such as optimized structures, electronic band structures, and densities of states, supporting efficient ML predictions. MP alone includes ~144,000 inorganic compounds and ~63,000 molecules. These resources are invaluable for advancing material predictions through ML. After acquiring the data, the initial step is to assess its quality to determine its suitability for building a ML model. This assessment involves checking the representativeness of the samples, identifying outliers or erroneous samples, ensuring consistent parameters for sample labels, and verifying that the distribution of label values is balanced and close to a normal distribution. Following the quality assessment, data cleaning is essential to remove noise, errors, and inconsistencies, ensuring the quality and reliability of data. To handle missing values, one can either delete samples or features with missing values or use interpolation methods (such as mean, median, or regression) to fill in the gaps. Normalizing and standardizing the data is crucial to eliminate scale differences between features, ensuring comparability and consistency across the dataset. These preprocessing steps are fundamental to prepare the data for effective ML, leading to more robust and accurate models.

Feature engineering

Feature engineering is a preprocessing step in ML, referring to the process of transforming raw data into features that better represent the underlying problem, thereby improving the predictive performance of the model. It aids in representing the underlying problem to the predictive model more effectively, thus enhancing the accuracy of the model on unseen data. Predictive models consist of predictor variables and outcome variables, and the feature engineering process selects the most useful predictor variables for the model. Feature engineering in ML primarily includes feature extraction, construction, and selection. Feature extraction is the process of extracting useful information from raw data. The methods for feature extraction vary depending on the type of data. For instance, extracting information about crystal elements, atomic positions, atomic interactions, and local structures can help the model understand material properties^[57,65,66]. Feature construction involves creating new features through linear combinations of the original features to provide richer information and improve the model performance. Effective feature construction requires a deep understanding of the data and applying domain knowledge to innovate^[67]. In practical ML tasks, it is often necessary to repeatedly train ML models to evaluate the effectiveness of the current set of features and iterate through the three stages of feature engineering. Feature selection is selecting the most relevant subset of features from the original feature set by removing redundant, irrelevant, or noisy features. The goal is to choose the features with the highest predictive power from the original set^[68,69]. Common feature selection methods include filter methods^[70], wrapper methods^[71], and embedded methods^[72].

Model selection and training

Before building a ML model, it is essential to clearly define the type of addressed task and select the most suitable model from multiple candidates for the specific problem. Different types of ML tasks, such as regression, classification, and clustering, may correspond to different model selection strategies. There is

usually no one universal best ML algorithm for materials research. The selection of an appropriate model based on the available data and the prior knowledge and assumptions related to the specific research problem. Each algorithm has its advantages and is suitable for different types of problems. It is critical to select the appropriate model based on the specific characteristics of the dataset and the problem. This requires a comprehensive understanding of material properties and their relationship to system variables and parameters^[73-75]. ML can be divided into supervised, unsupervised, semi-supervised, and reinforcement learning^[76,77]. Among these, supervised learning is the most widely applied method in ML. Its primary goal is to predict new input data by learning the mapping relationship between input features and known target variables. Supervised learning tasks can be divided into two main categories: regression and classification^[78]. Additionally, ML models can be divided into three major categories based on their complexity and structure: shallow learning models, ensemble learning models, and deep learning models.

Shallow learning models

Shallow learning models are relatively simple ML algorithms typically used when the data is simple, and the relationships between features and target variables are linear or uncomplicated. Compared to more complex models such as deep learning networks, shallow learning models are computationally less expensive and easier to interpret, making them particularly useful when dealing with small datasets or when model interpretability is crucial. Linear regression is a common shallow learning model that predicts continuous values based on the assumption of a linear relationship between input features and the target variable. The mathematical representation of a linear regression model is given by:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \varepsilon \quad (1)$$

where y is the predicted value, β_0 is the intercept, $\beta_1, \dots, \beta_n$ are the coefficients, $x_1, \dots, x_n$ are the input features, and ε is the error term. While it performs well when there is a simple linear relationship between features and the target, its prediction performance may be limited when the data contains nonlinear patterns or complex relationships. Logistic regression^[79] is commonly used for binary classification tasks. The logistic regression model is defined as:

$$P(y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}} \quad (2)$$

where $P(y = 1|X)$ is the probability that the dependent variable y equals 1 given the input features x . The coefficients β_i are estimated to maximize the likelihood of the observed data. Logistic regression model “ P ” provide probabilities for classification, making it simple, efficient, and easy to interpret. However, it assumes that the decision boundary between classes is linear, which may not work well for complex nonlinear data. Support vector machines (SVM)^[80] are suitable for classification and regression tasks, particularly with high-dimensional data. For a binary classification problem, the decision boundary is determined by:

$$f(x) = \text{sign}(w \cdot x + b) \quad (3)$$

among them, $f(x)$ is the predicted category, w is the weight vector, b is the input feature vector, x is the bias (or intercept), and $\cdot$ represents the dot product of the vectors. SVM tries to find an optimal hyperplane to separate different classes, and with the use of kernel tricks, it can handle nonlinear decision boundaries. Although SVM is powerful in many scenarios, it is sensitive to the selection of hyperparameters and can be computationally expensive, especially for large datasets. K-nearest neighbors (KNN)^[81] is a non-parametric algorithm that makes predictions based on the proximity of data points. It is used for both classification and regression tasks. It is intuitive and easy to implement, especially when data has clear clusters. However,

KNN can be computationally expensive during prediction as it requires storing all training data, and it is sensitive to the scale of the data and the choice of distance metric. Naive Bayes^[82] is widely used in text classification and spam detection tasks. It is based on Bayes' theorem and assumes conditional independence among features. While simple and efficient, it may not perform well if the assumption of feature independence is not realistic in many real-world datasets. Overall, shallow learning models are highly effective when the data is simple, clean, or when a simple and interpretable model is required. However, they often struggle with complex, high-dimensional, or nonlinear data, making more powerful methods such as ensemble learning or deep learning models necessary.

Ensemble learning models

Ensemble learning models combine multiple base models (also known as weak learners) to create a more powerful and accurate predictive model. These methods are based on the idea that combining the predictions of several models can outperform any individual model. By aggregating the predictions from various base models, ensemble methods reduce the risk of overfitting and improve the model's overall accuracy^[83]. There are two main types of ensemble learning techniques: bagging and boosting. Bagging, which stands for bootstrap aggregating, includes models such as random forests and bagging trees^[84,85]. These models are used for classification, regression, and other predictive tasks, particularly concerning high model variance. Bagging works by training multiple models on different subsets of the data and combining their predictions, which helps to reduce overfitting and increases the robustness of the model. However, if the base model is weak, the method can still suffer from high bias, and training multiple models can be computationally expensive. On the other hand, boosting methods such as gradient boosting machines (GBM)^[86], AdaBoost^[87], and XGBoost^[88] are widely used for solving complex classification and regression problems, especially in competitive ML settings. Boosting works by sequentially correcting the errors made by previous models, often leading to high performance even in challenging tasks. It effectively addresses both bias and variance, but it can become prone to overfitting if the number of iterations is too large or if the model is not properly regularized. Another ensemble technique, Stacking^[89], involves combining predictions from multiple different models, where each model contributes its unique strengths to the overall prediction task. This method leverages the diversity of multiple models to create a more powerful predictive model. While stacking can provide excellent results, it is computationally expensive and requires careful selection and combination of the models. Ensemble methods are highly powerful because they harness the diversity of different base models to reduce errors and bias. These techniques are particularly effective when dealing with noisy, imbalanced, or complex data and are widely used in both research and industry for their ability to deliver top-tier predictive performance.

Deep learning models

Deep learning models are a powerful class of ML algorithms composed of multiple layers that transform input features into output predictions. These models are particularly well-suited for capturing complex, hierarchical data representations, enabling them to model intricate relationships that simpler algorithms may overlook^[90-92]. However, deep learning models typically require large amounts of labeled data and substantial computational resources, making them ideal for applications in fields that deal with big data. In the context of optoelectronic materials, deep learning can be leveraged to model the complex relationships between material properties and performance, which is critical for accelerating material discovery and design. One of the most common types of deep learning models is artificial neural networks (ANN)^[93]. ANN is highly flexible and powerful, capable of modeling complex, nonlinear relationships in data. However, ANN requires large labeled datasets for training, and the computational cost can be significant. Additionally, without proper regularization, ANN can be prone to overfitting, particularly when working with limited data or insufficient diversity in material properties. Another type of deep learning model is the

convolutional neural network (CNN)^[94], which is commonly used for image processing, object detection, and spatial data analysis. CNN excels at automatically learning hierarchical feature representations from raw data, making them particularly effective for tasks in materials science that involve complex spatial structures, such as analyzing material microstructures or predicting material properties from images. While CNN requires large labeled datasets and substantial computational power, its ability to reduce the need for manual feature extraction is invaluable in material science applications. Recurrent neural network (RNN)^[95] is designed to handle sequential data and is ideal for time-dependent tasks such as time series forecasting. In optoelectronic materials, RNN could be used to model the temporal evolution of material properties under varying conditions, such as light exposure or electrical stress. However, training RNN can be challenging, especially when dealing with long sequences, due to issues such as vanishing gradients. Generative adversarial networks (GAN)^[96] have gained prominence for their ability to generate realistic synthetic data, which is useful for augmenting datasets in material science. GAN can help create new material designs by generating plausible material configurations based on learned data distributions. However, GAN is challenging to train and can suffer from instability or mode collapse, which may hinder their effectiveness in practical applications. Recently, Transformer-based architectures have revolutionized deep learning, particularly in natural language processing (NLP) and sequence modeling^[97-99]. These architectures excel in handling long-range dependencies, which can be advantageous in materials science for predicting complex relationships between different material properties or performance over time. Transformers are highly scalable and flexible, making them an exciting avenue for developing advanced models in optoelectronic material design. They hold the potential to improve predictive modeling, accelerate the discovery of new materials, and optimize material performance. Despite their computational cost, Transformer-based models are increasingly becoming central to many cutting-edge AI applications, and their potential in materials science is beginning to be realized.

Before a model can be used for prediction, it should be able to adjust its parameters automatically to improve performance. During model training, an independent dataset is often used to validate the training results and detect potential overfitting. The validation set, which usually accounts for about 10% of the available data, monitors the training process, helps select hyperparameters, and prevents overfitting. After training the model, an external dataset (test set) is used to evaluate the generalization ability of the model. This test set, comprising about 10% of the available data, is critical for assessing the performance of the model on unseen data, thereby providing an unbiased evaluation of its predictive ability^[100]. The training, validation, and testing process in practical applications often repeats multiple times to optimize the model performance. This process is known as model iteration or model optimization. Through repeated iterations, the best model parameters and structure can be identified, thereby enhancing the predictive and generalization capabilities of the model.

Model evaluation and optimization

Upon completion of model training, evaluating the model is a crucial step to ensure its practical effectiveness. The purpose of evaluation is to measure the predictive performance of the model, including its accuracy and error^[101,102]. The following are common methods for model evaluation and optimization.

Cross-validation (CV)^[103]: CV is a widely used model evaluation technique to prevent overfitting. Common CV methods include k-fold and leave-one-out CV. In k-fold CV, the dataset is divided into k subsets; the model is trained on k-1 subsets and validated on the remaining one. This process is repeated k times, with each subset used once as the validation set. The final evaluation result is the average of the k-validation results. Leave-one-out CV uses one sample as the validation set, and the remaining as the training set each time. Although this method is computationally expensive, it is particularly effective for small datasets. Bootstrap sampling^[104]: Bootstrap sampling is a statistical resampling method that involves repeatedly

sampling with replacement from the dataset. The bootstrap method assesses statistical properties such as variance, confidence intervals, and bias. It is especially useful for small and imbalanced datasets and can be combined with ML algorithms to provide more accurate predictions. Holdout method^[26]: The Holdout method randomly divides the dataset into training, validation, and test sets. The model training on the training set, hyperparameters are tuned on the validation set, and its performance is finally evaluated on the test set. The holdout method is simple and intuitive but requires a sufficiently large dataset to avoid random evaluation results.

The performance metrics used in different ML models (e.g., classification, regression) vary. For classification models, performance metrics include accuracy, precision, recall, F1 score, receiver operating characteristic (ROC) curve, and area under the curve (AUC)^[105,106]. The confusion matrix is also commonly used to evaluate classification models, comparing actual and predicted classes^[107]. For regression models, performance metrics include root mean square error (RMSE), mean absolute error (MAE), mean squared error (MSE), and the coefficient of determination (R^2)^[33]. These metrics measure the difference between predicted and actual values, assessing the model accuracy and reliability. Ultimately, model performance should be evaluated using the test set. The test set must be entirely independent of the training and validation sets to ensure genuine model performance on unseen data. The test set selection should avoid overlapping with training data to prevent the model from encountering learned samples during testing, which could result in falsely high performance. Additionally, the test set should be sufficiently large and randomly selected to ensure the evaluation results are representative and reliable^[108,109].

In addition to testing on an independent dataset, model performance can be further enhanced through hyperparameter tuning. Hyperparameters (such as the learning rate, regularization strength, number of hidden layers in neural networks, and the number of trees in ensemble methods)^[110], unlike model parameters, are not learned during training but rather set prior to the learning process, significantly impacting the model accuracy and generalization. Hyperparameter tuning involves systematically searching for the optimal combination of hyperparameters to maximize model performance. Common methods for hyperparameter tuning include Grid Search^[111], an exhaustive approach that evaluates every possible combination within a predefined set of hyperparameter values. Although computationally intensive, grid search can be effective for smaller search spaces. Random Search^[79], rather than evaluating all combinations, randomly selects combinations, making it more efficient for large search spaces and often yielding satisfactory results with fewer trials. An enhancement to this approach is Hyperband^[112], which dynamically allocates resources to promising configurations while employing early stopping for less promising ones. This combination effectively allows Hyperband to explore large search spaces, often achieving competitive results with fewer evaluations than traditional random search methods. Bayesian Optimization (BO)^[113] uses probabilistic models to predict promising hyperparameter combinations, iteratively refining the search based on previous evaluations. BO is efficient and often yields better results than exhaustive methods for complex models.

By fine-tuning hyperparameters, models can balance bias and variance optimally, minimizing overfitting or underfitting and improving training and test data performance. The final model, selected after hyperparameter tuning, is evaluated on the independent test set to confirm its predictive power and generalizability.

HT AND ML OF OPTOELECTRONIC MATERIALS

HT techniques of optoelectronic materials

HT computational screening has revolutionized the search for and optimization of materials in optoelectronic applications, offering a rapid, cost-effective alternative to traditional experimental methods. HT methods provide critical insights into the complex relationships between structural characteristics and electronic properties by enabling the simulation and evaluation of large datasets. In recent years, HT screening has shown great promise across various material classes, including two-dimensional (2D) materials, perovskite materials, and traditional III-V semiconductors^[114,115]. These materials exhibit excellent optoelectronic properties and are widely applied in solar cells, light-emitting devices, photocatalytic, and photodetectors, supporting advancements in renewable energy technologies^[116-119]. With the integration of ML and data-driven approaches, HT screening provides a powerful pathway for exploring these material systems, uncovering novel compositions, and accelerating the rational design of next-generation optoelectronic devices.

Exploring the potential of 2D materials for photocatalytic applications, Gao *et al.* applied HT computational screening to identify 2D polar materials capable of efficient water splitting^[120]. Figure 3A shows a workflow for HT inverse design of photocatalysts, with two main steps: Step 1 (Inverse Design) and Step 2 (HT DFT Calculations). Step 1: Inverse design for 2D polar water-splitting photocatalyst candidates using C2DB database materials. Key criteria include high stability, non-magnetism, out-of-plane polarity ($dipz > 0.001$ e-Å/unit cell), and a DFT-calculated band gap (0.1 to 2.5 eV). After filtering and manual selection, they identified the candidate materials. Step 2: HT DFT calculations involve screening band alignment ($E_{VBM} < -5.67$ eV, $E_{CBM} > -4.44$ eV) and reaction free energy ($\Delta G_{HER}, \Delta G_{OER} < 0$). The final calculation confirmed that the 2D polar materials are suitable for photocatalytic water splitting, and the schematic diagram of photocatalytic water splitting is presented in Figure 3B. Under light irradiation, photogenerated electrons and holes move in opposite directions under the action of the internal electric field and accumulate on different surfaces. Energetically, electrons jump from the valence band (VB) to the conduction band (CB) after photoexcitation, and the electrons and holes in the CB and VB catalyze hydrogen evolution reaction (HER) and oxygen evolution reaction (OER), respectively. The blue and red bars represent the CB and VB, while the dotted lines represent the CBM and VBM. The researchers finally identified specific materials with optimal electronic properties for water splitting, where the inherent polar properties of the materials promote charge separation, thereby improving photocatalytic efficiency. This study demonstrates the ability of HT screening to identify promising candidate materials by evaluating electronic structure and energy band arrangement, thereby advancing the field of photocatalytic materials for sustainable energy production. Focusing on photovoltaic (PV) potential in materials typically overlooked for thin-film PV, Kangsabanik *et al.* introduced a HT method to evaluate indirect bandgap semiconductors^[121]. Their study developed a computationally efficient approach to model phonon-assisted absorption, a critical factor for indirect bandgap materials. A schematic diagram of the method is illustrated in Figure 3C, including the process of phonon-assisted optical absorption and the computational workflow for evaluating direct and indirect bandgap semiconductor PV performance parameters. By simplifying the modeling process with Γ -point phonons, they successfully screened 127 binary compounds, revealing 28 promising candidates, of which 20 are indirect bandgap materials. These findings highlight previously untapped PV potential in materials such as TiS_3 , CdP_4 , and BaP_3 , which exhibit favorable properties, including low recombination rates and high carrier mobility, essential for efficient charge separation and transport in PV devices. By including phonon-assisted absorption calculations, Kangsabanik *et al.* broadened the scope of HT screening to encompass more complex material classes, offering insight into the latent optoelectronic advantages of indirect bandgap materials^[121]. Their work suggests that such materials, often overlooked, could significantly expand the options for efficient, stable materials in thin-film PV, marking a new direction for sustainable energy technology.

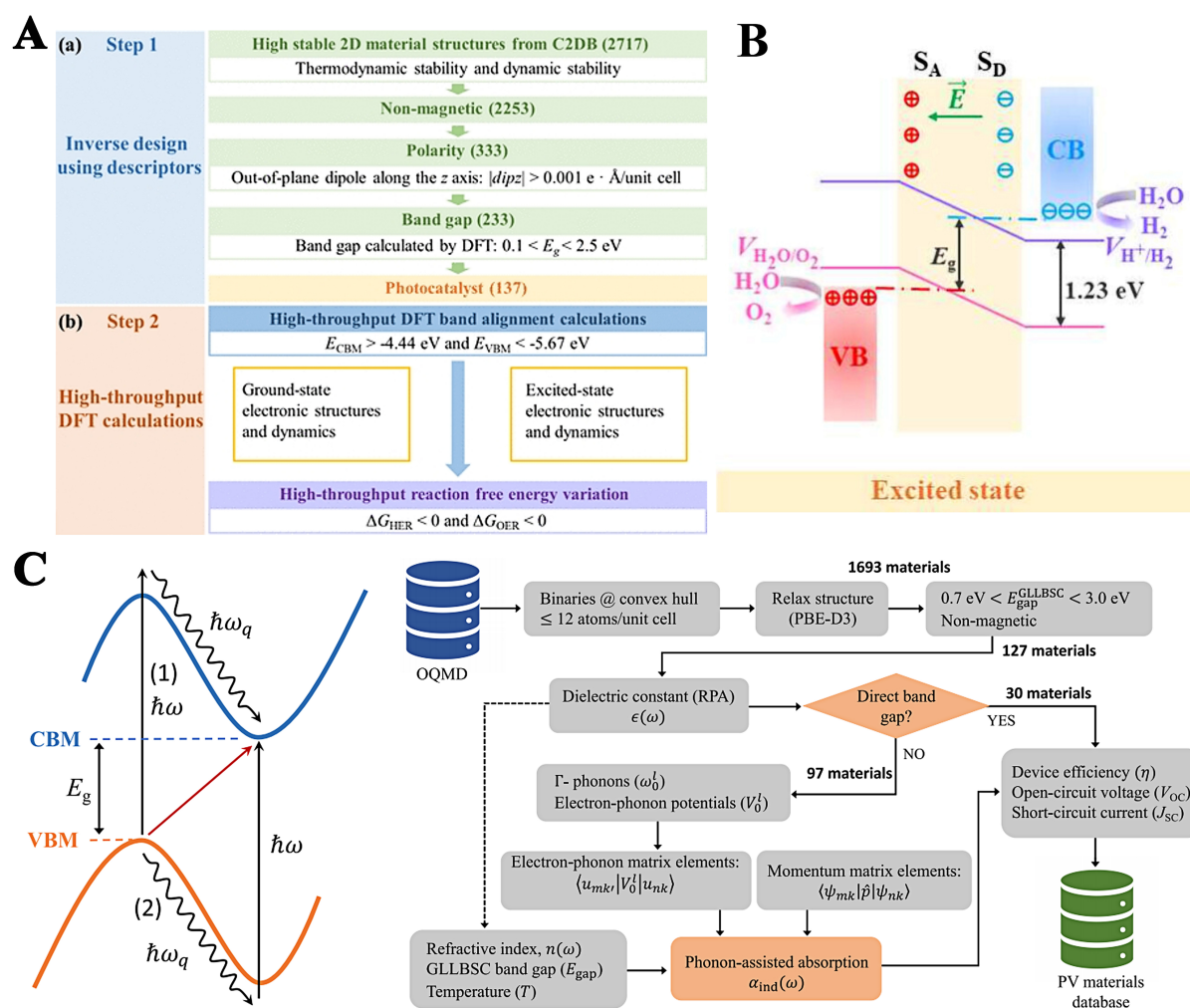
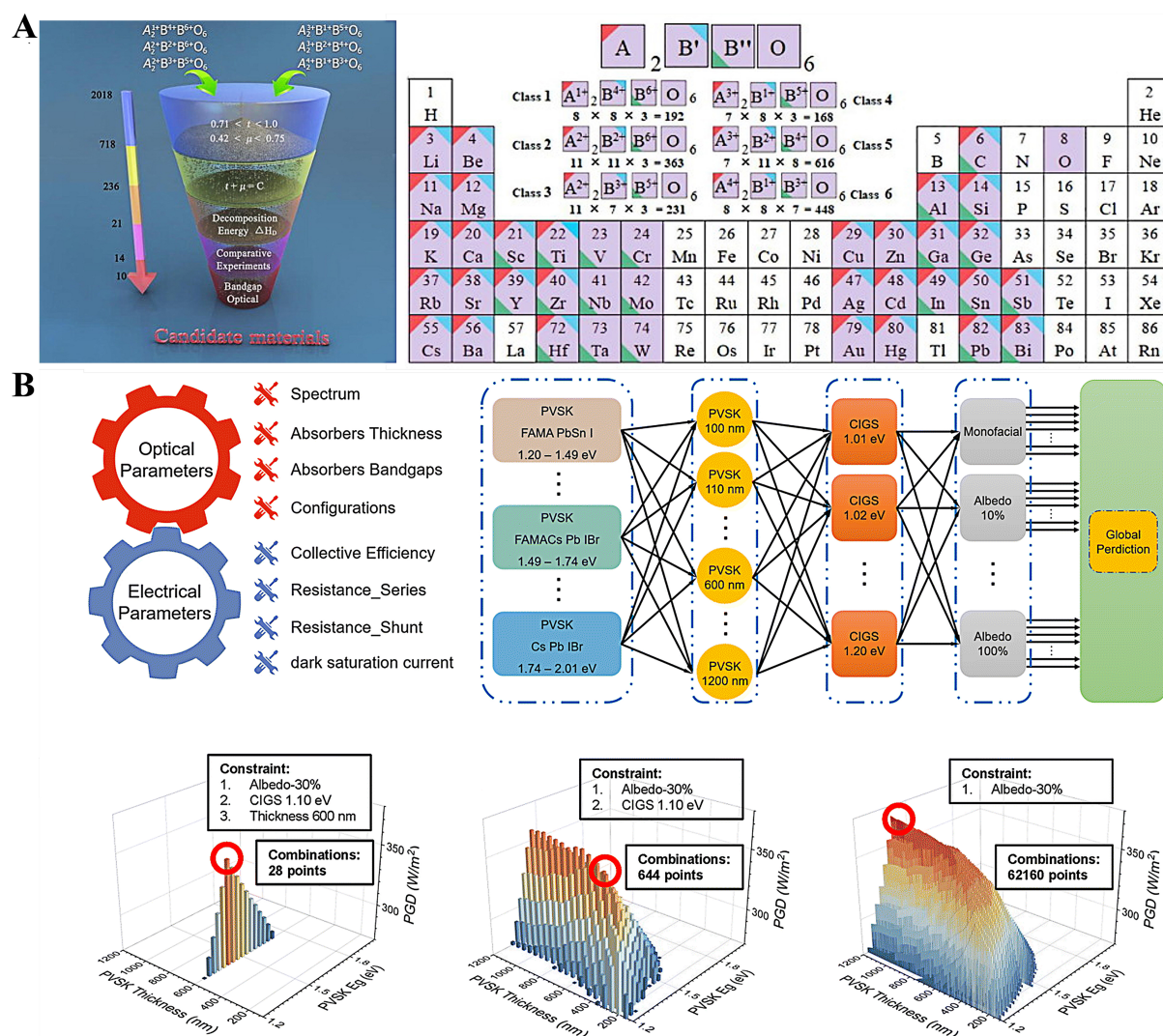


Figure 3. Flow process schematics of the HT calculation. (A) Workflow of HT inverse design for the photocatalyst, including step 1 (inverse design) and step 2 (HT DFT calculations); (B) Schematic illustration of photocatalytic water splitting on a 2D polar material. Copyright 2024, American Chemical Society, Reproduced with permission^[120]; (C) Schematic diagram of phonon-assisted optical absorption and computational workflow used to evaluate the PV performance parameters for direct and indirect band gap semiconductors. Copyright 2022, American Chemical Society, Reproduced with permission^[121]. HT: High-throughput; DFT: density functional theory; 2D: two-dimensional; PV: photovoltaic.

To further explore structural diversity in optoelectronics, Jiang *et al.* conducted HT computational screening to explore oxide double perovskites ($\text{A}_2\text{B}'\text{B}''\text{O}_6$) for applications in optoelectronics and photocatalysis, emphasizing these materials' structural flexibility and compositional diversity^[122]. As shown in Figure 4A, using first-principles DFT calculations, the team evaluated 2018 potential candidates based on stability and electronic properties. Through stability screening and phase diagram analysis, they identified 138 perovskites across three common crystallographic phases, with 21 being stable and 14 previously unreported. The selected candidates exhibited quasi-direct bandgaps, balanced electrons, hole effective masses, and strong optical absorptions, all favorable for optoelectronic and photocatalytic applications. Their work expands the library of perovskite materials and provides a methodical approach to screening large, chemically complex datasets for novel material discovery in energy applications. Additionally, expanding the application of HT modeling to device design, Tang *et al.* showcased the application of HT optoelectrical modeling to optimize the power generation density (PGD) of bifacial tandem solar cells^[123]. In Figure 4B, a schematic representation of this HT optoelectronic modeling and screening workflow



illustrates the process for identifying optimal tandem cell configurations, providing a visual summary of how to maximize PGD across a wide parameter space. Tang and colleagues developed a HT modeling approach capable of simulating hundreds of thousands of combinations of device variables, including thicknesses and bandgaps of the active layers, under various lighting conditions. Their findings demonstrated that a bifacial perovskite/Cu(In, Ga)Se₂ tandem cell with optimized parameters could achieve a PGD exceeding 495 W/m² under high albedo (reflectivity) conditions, outperforming traditional monofacial configurations. The four-terminal configuration, in particular, achieved remarkable improvements in power output by leveraging the bifacial design to capture light from both front and rear sides, thus maximizing energy conversion efficiency. This study demonstrates the economic and practical value of combining tandem and bifacial strategies, especially in high albedo environments such as snowy areas, where the increased reflected light significantly boosts power output.

Together, these studies underscore the transformative impact of HT screening in advancing optoelectronic materials research. By refining and optimizing materials at multiple stages of the design process, HT methods enable researchers to systematically explore electronic and structural parameters, thus providing a pathway for accelerated innovation in PV, photocatalytic, and solar cell applications. HT screening enriches the material landscape for optoelectronics and supports the development of scalable, efficient solutions for sustainable energy technologies.

Predicting optoelectronic properties

Accurate prediction of optoelectronic properties is crucial for advancing materials used in solar cells, light-emitting devices, photocatalytic materials, and photodetectors. Leveraging HT computational screening and ML has expanded the ability of researchers to evaluate materials across vast chemical and structural landscapes^[124,125]. This section explores recent advancements in ML-enabled predictive models focusing on crucial properties such as band gap tuning, band alignment, and charge carrier dynamics.

To enhance the understanding of impurity effects on conductivity, Mannodi-Kanakkithodi *et al.* developed an ML approach to predict impurity energy levels in Cd-based chalcogenides, a crucial factor for controlling conductivity in optoelectronic applications^[126]. As shown in Figure 5A, they used a DFT dataset to train regression models to predict impurity formation enthalpies and charge transition levels across different Cd-based compounds. This approach allowed for high-accuracy predictions of impurity effects in CdS, CdSe, and CdTe, enabling quick screening of impurity atoms that affect the Fermi level and conductivity type. The model successfully generalized various mixed anion compositions, demonstrating its power in guiding material design for tailored optoelectronic properties. Building on this concept of material property optimization, Wang *et al.* tackled the challenge of consistent and high-accuracy band gap prediction in perovskites, a class with extensive compositional diversity^[127]. They developed a robust band gap predictor through a ML model trained on a rigorously compiled dataset of band gaps verified by quasi-particle self-consistent GW calculations. Figure 5B provides a detailed depiction of the accuracy of the ML model in predicting the band gaps of perovskites. The figure compares the model predicted band gap with the reference DSH band gap values, evaluating the accuracy and robustness of the model in both the training and testing datasets. The model demonstrates low deviation in both datasets, characterized by MSE and an R^2 value exceeding 98%, indicating high predictive accuracy and generalizability across various perovskite compositions. Furthermore, the model effectively predicts band gap in single perovskite structures and shows excellent accuracy in double perovskite structures. This model demonstrated high transferability across 15,659 single and double perovskite compositions, identifying 14 lead-free perovskites with band gaps in the ideal range for PV applications, including MASnBr_3 and $\text{FA}_2\text{InBiBr}_6$. These materials exhibit direct band gaps, low effective masses, and minimal exciton binding energies, improving charge separation and light absorption efficiencies. Thus, the model broadens the perovskite material space and provides a reliable tool for discovering efficient, non-toxic PV materials.

Further extending the predictive capabilities of ML, Kim *et al.* explored the use of B-site alloying in metal halide perovskites (MHPs), focusing on their stability and band gap properties^[128]; the schematic workflow of the study is shown in Figure 5C, given the structural and chemical instability of lead-based perovskites, employed CNNs to analyze 41,400 B-site-alloyed MHP configurations, simulating each with DFT and estimating stability based on decomposition energy and band gap type. The validation results of the trained crystal graph convolutional neural networks (CGCNN) model are also shown in Figure 5C. The parity plots show the relationship between the decomposition energy (ΔH_{decomp}) and bandgap predictions predicted by CGCNN and the DFT calculated values, with very high MAE and R^2 values, indicating very high prediction accuracy. Also shown are the confusion matrix and key classification metrics (accuracy = 0.96, precision = 0.84, recall = 0.90, and F1 score = 0.87), validating the model performance in distinguishing between indirect

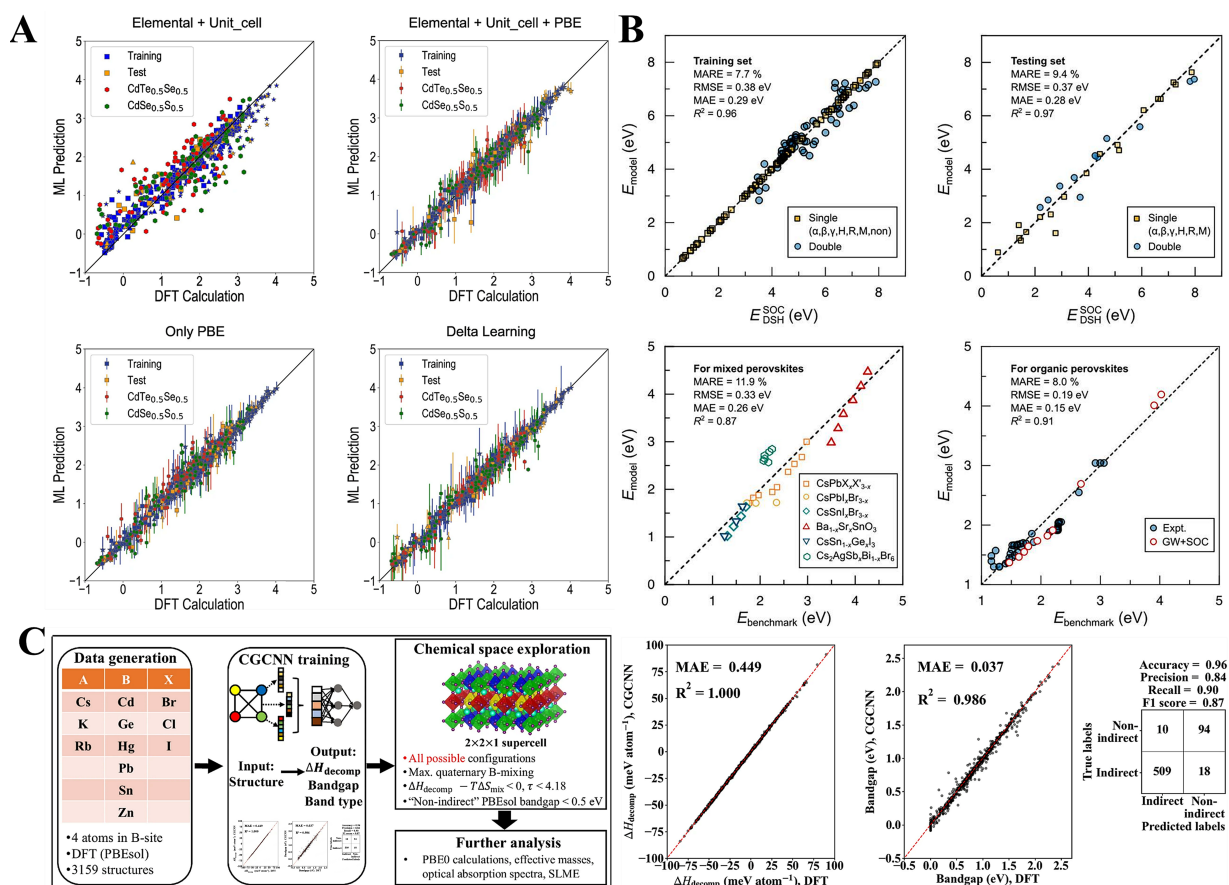


Figure 5. (A) Parity plots are shown for predictive models trained using 90% of the CdTe + CdSe + CdS dataset as the training set, with performances shown for the training, test and out-of-sample points using the elemental and unit cell defect descriptors. Copyright 2020, Springer Nature, Reproduced with permission^[126]; (B) Performance of the developed model for band gap estimation. Copyright 2024, American Chemical Society, Reproduced with permission^[127]; (C) Schematic workflow of the present study and parity plot between CGCNN-predicted and DFT (PBEsol)-calculated. Copyright 2024, Springer Nature, Reproduced with permission^[128]. CGCNN: Crystal graph convolutional neural network; DFT: density functional theory.

and non-indirect bandgap materials. By examining the band structure, the study identified $\text{CsGe}_{0.3125}\text{Sn}_{0.6875}\text{I}_3$ and $\text{CsGe}_{0.0625}\text{Pb}_{0.3125}\text{Sn}_{0.625}\text{Br}_3$ as leading candidates for single-junction and tandem solar cells, respectively. This validation confirms that the CGCNN model can accurately predict decomposition energy and bandgap, making it a valuable tool for HT screening of MHP compositions for stability and electronic properties.

In another approach, Mahal *et al.* focused on predicting band alignment types in 2D hybrid perovskites, known for their environmental stability and unique quantum-well-like structures [Figure 6A]^[129]. Using ML classifiers trained on molecular and elemental descriptors, they categorized perovskites into specific band alignment types: I_a, I_b, II_a, and II_b, directly influencing device suitability. For instance, type I alignments favor devices with localized exciton transitions. In contrast, type II structures, which separate carrier populations across the organic and inorganic layers, are optimal for PV applications due to extended carrier lifetimes. The work of Nayak *et al.* provides a systematic classification framework for 2D hybrid perovskites, enabling the rapid identification of materials with the necessary band alignments for specific optoelectronic functions^[130]. This classification is particularly valuable for selecting perovskites with optimized charge separation and carrier recombination characteristics. Complementing these ML-driven predictions, Nayak

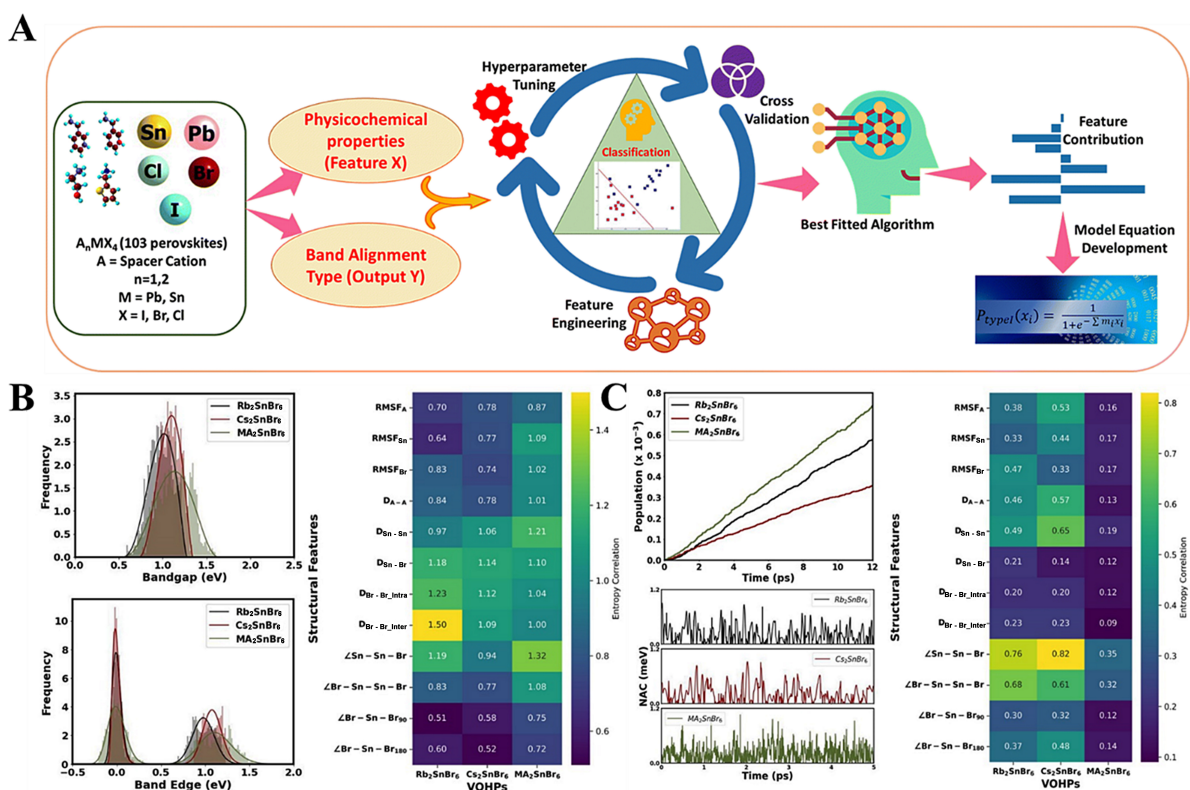


Figure 6. (A) ML workflow for predicting band alignment types in 2D hybrid perovskites. Feature contributions of the finally selected nine features towards output Y. Confusion matrix for the type I and II classification of considered 2D perovskite. Copyright 2023, Royal Society of Chemistry, Reproduced with permission^[129]; (B) Electronic structures of VOHPs fluctuate over a 5 ps (5,000 snapshots) period at 300 K, with histograms showing band gaps, VBM, and CBM energies along AIMD trajectories, and mutual information between band gap and critical structural features; (C) Displays excited-state charge carrier dynamics under ambient conditions, tracking nonradiative electron-hole recombination over time, absolute NAC values between VBM and CBM over 5 ps, and mutual information of NAC with structural features. Copyright 2024, American Chemical Society, Reproduced with permission^[130]. ML: Machine learning; 2D: two-dimensional; VOHPs: vacancy-ordered halide perovskites; VBM: valence band maximum; CBM: conduction band minimum; AIMD: ab initio molecular dynamics; NAC: nonadiabatic coupling.

et al. explored the role of A-site cations in influencing charge carrier dynamics within vacancy-ordered halide perovskites (VOHPs)^[130]. Using non-adiabatic molecular dynamics and ML to analyze electron-phonon coupling, they examined VOHPs with different A-site cations [e.g., Cs, Rb, and methylammonium (MA)] to observe how these choices impact carrier lifetimes. In particular, **Figure 6B** provides a detailed view of the effect of these cations on the lattice dynamics and the resulting impact on nonradiative recombination rates, while **Figure 6C** showcases the influence on carrier lifetimes in VOHPs with different cation compositions. Their findings indicate that inorganic cations such as Cs and Rb suppress lattice dynamics, reducing nonradiative recombination and, consequently, longer carrier lifetimes. Conversely, organic cations such as MA amplify lattice fluctuations, increasing recombination rates and decreasing carrier lifetimes. This study underscores the importance of structural dynamics in controlling optoelectronic performance and suggests pathways for designing VOHPs with extended carrier lifetimes for more efficient, sustainable devices.

Together, these studies demonstrate the transformative role of HT screening and ML in accelerating the discovery and optimization of optoelectronic materials. By improving the prediction of key properties such as band gap, band alignment, and charge carrier dynamics, these models are paving the way for the development of high-performance, sustainable materials for next-generation optoelectronic devices.

Assessing the stability of optoelectronic materials

The stability of optoelectronic materials is a critical factor for their long-term performance and durability, particularly in applications such as PV devices and catalysis, where material degradation over time can significantly affect efficiency and sustainability. Accurately predicting the stability of materials is essential for designing durable devices and ensuring their practical viability^[131,132]. Recent developments in ML have provided powerful tools for predicting the stability of optoelectronic materials^[133,134]. This section reviews recent advancements in ML-based stability prediction models, highlighting their ability to enhance material screening and guide the development of stable, high-performance materials. In recent studies, Bartel *et al.* proposed an improved tolerance factor τ to more accurately predict the stability of perovskite structures, addressing the limitations of the traditional Goldschmidt tolerance factor^[135]. Based on simple geometric relationships, the Goldschmidt factor is somewhat effective for predicting the stability of oxide perovskites, but performs inadequately in screening complex materials, especially halide-containing perovskites. To overcome these limitations, Bartel and colleagues applied the sure independence screening and sparsifying operator (SISSO) method to develop a new one-dimensional descriptor τ , incorporating information about ionic radii and oxidation states. This significantly enhances prediction accuracy, achieving a 92% success rate for the oxide, fluoride, and chloride perovskites dataset, demonstrating superior performance in structural screening compared to the Goldschmidt factor. **Figure 7A** assesses the performance of τ in predicting perovskite stability, illustrating the model accuracy across different chemical combinations and its applicability to halide perovskites. Notably, the improved τ factor classifies perovskite stability and provides a continuous stability probability estimate $P(\tau)$. By applying Platt scaling, Bartel *et al.* transformed the model output into continuous probability values, allowing greater adaptability across various perovskite types^[135]. **Figure 7B** shows a stability map for predicted double perovskites, with the lower triangle depicting stability probabilities for $\text{Cs}_2\text{BB}'\text{Cl}_6$ compounds and the upper triangle showing $\text{La}_2\text{BB}'\text{O}_6$ compounds. These probability maps, generated from Platt-scaled $P(\tau)$ values, use color gradients to indicate the likelihood of forming stable perovskite structures across different ion combinations, highlighting potential stability. This probability estimate provides researchers with an efficient screening tool for exploring potential perovskite materials across broader chemical spaces. Additionally, extensive experimental validation enabled the model to predict the stability of 23,314 potential double perovskites, providing a prioritized material list for further investigation. This predictive approach lays a solid theoretical foundation for the functional design and application of perovskite materials, with significant implications for their development in fields such as PVs and electrocatalysis.

Building on the advancements in stability prediction by Bartel *et al.*, Gu *et al.* introduced a novel approach focused on the synthesizability of perovskite materials based on a graph convolutional neural network (GCNN) combined with positive-unlabeled (PU) learning, addressing the limitations of traditional methods in predicting the practical synthesizability of materials^[136]. Traditional stability prediction models, such as convex hull energy calculations, primarily assess the thermodynamic stability of materials but are less effective at predicting synthesizability under experimental conditions. To overcome this limitation, Gu and colleagues pre-trained the GCNN model on actual synthesis data from the MP database and then applied it to a perovskite-specific dataset^[136]. Using transfer learning, the model achieved significantly improved synthesizability prediction accuracy across different perovskite structures, achieving a true positive rate of 95.7%. **Figure 7C** illustrates the model prediction accuracy and data distribution, highlighting the enhanced performance of the GCNN and PU learning combination in identifying experimentally synthesized materials as positive samples. Furthermore, the model demonstrated its effectiveness in predicting virtual perovskites, identifying 179 out of 11,964 virtual perovskite candidates as having potential synthesizability based on the model-generated crystal-likeness (CL). **Figure 7D** compares the GCNN model with screening results based on the Goldschmidt rule. The lower left triangle in this figure presents synthesizability scores for ABO_3 perovskite compounds, with a green gradient indicating synthesizability likelihood (deeper green

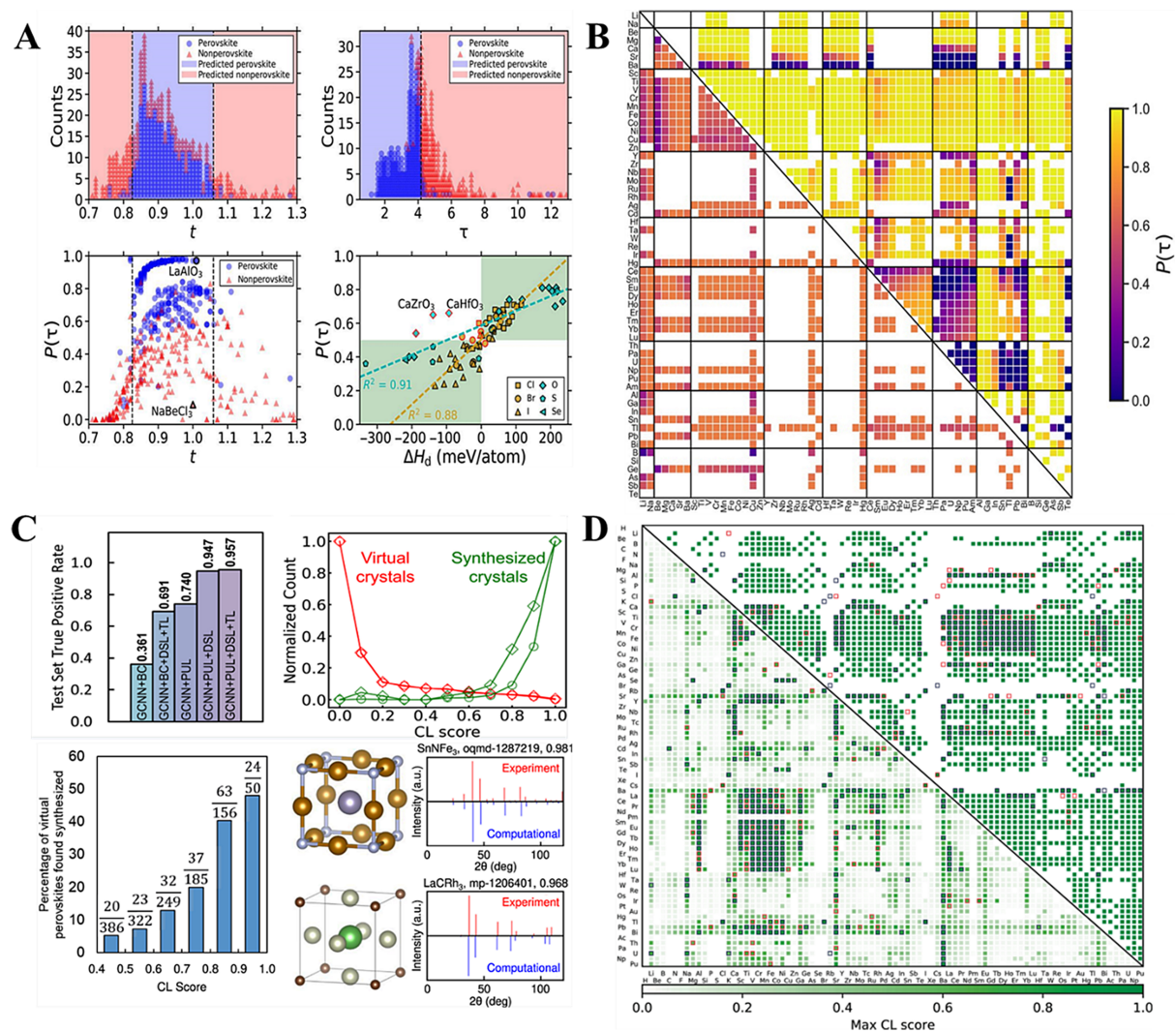


Figure 7. (A) Assessing the performance of the improved tolerance factor τ and the map of predicted double perovskite oxides and halides; (B) Map of predicted double perovskite oxides and halides. Lower triangle: Probability of forming a stable perovskite with the formula $\text{Cs}_2\text{BB}'\text{Cl}_6$ as predicted by τ . Upper triangle: Probability of forming a stable perovskite with the formula $\text{La}_2\text{BB}'\text{O}_6$ as predicted by τ . Copyright 2019, American Association for the Advancement of Science, Reproduced with permission^[135]; (C) Model accuracy and data distribution, model validation; (D) Synthesizability of ABO_3 perovskite compounds for the model (lower left triangle) and Goldschmidt-rule-based screening (upper right triangle). Copyright 2022, Springer Nature, Reproduced with permission^[136].

reflects higher synthesizability probability). The upper right triangle shows the Goldschmidt-rule-based screening results, where green marks indicate compounds that pass the screening criteria. This comparison reveals that the GCNN model can identify a broader range of perovskite structures with synthesizability potential, especially for perovskites with covalent bonds, halides, and anti-perovskites, highlighting a scope and flexibility beyond the limits of the Goldschmidt rule. The model exhibits considerable potential for screening perovskites with diverse bonding characteristics and structural types, particularly for applications in PV and solid electrolytes. This model provides a valuable tool for exploring novel materials with potential experimental realizability.

These studies collectively contribute to the growing field of stability and synthesizability prediction by demonstrating the effectiveness of ML models in broadening optoelectronic materials screening capabilities,

potentially enhancing the discovery of stable and high-performing materials.

Optimizing optoelectronic performance

The optimization of optoelectronic performance in semiconductor materials is essential for enhancing the efficiency and applicability of these materials in optoelectronic and related applications. The optimization of optoelectronic performance can directly influence the energy conversion efficiency, response speed, and stability of devices, thereby determining the overall performance of optoelectronic components^[137,138]. To address the challenge of pinpointing high-efficiency optoelectronic materials, Cai *et al.* employed a ML model combined with HT screening to identify high-performance 2D PV candidates from a database of 187,093 inorganic crystal structures^[139]. Through classification-based ML algorithms, their model filtered 2DPV candidates based on features correlated with high PV conversion efficiency, such as the packing factor (P_f), which emerged as a primary indicator of PV potential. **Figure 8A** categorizes the selected 2DPV candidates according to structural prototypes, illustrating that those materials with specific space groups (e.g., $P3m1$) tend to exhibit favorable PV properties. Moreover, three materials - Sb_2Se_2Te , Sb_2Te_3 , and Bi_2Se_3 - were highlighted for their high conversion efficiencies, making them promising candidates for PV applications. This approach demonstrates the power of ML-assisted HT screening in identifying structurally promising candidates for further development, showcasing how structure-related features can serve as indicators for enhanced optoelectronic performance. Building upon this ML optimization approach, Liu *et al.* introduced a BO framework with knowledge constraints to optimize the fabrication conditions for perovskite solar cells using rapid spray plasma processing (RSPP)^[140]. The BO framework incorporated observations as probabilistic constraints to focus the parameter search on high-quality films. Their method achieved a 5-round, 18.5% power conversion efficiency with fewer than 100 process conditions through this iterative optimization approach. **Figure 8B** outlines this sequential optimization process, mapping the relationship between process parameters and efficiency improvements across successive rounds. Additionally, **Figure 8C** provides a detailed visualization of process conditions and predicted efficiency correlations, highlighting critical variable interdependencies that inform experimental adjustments, such as the alignment of spray flow rates and substrate speed. This framework highlights the advantage of integrating sequential learning with empirical knowledge, allowing researchers to efficiently navigate complex parameter spaces and optimize fabrication conditions for maximum efficiency gains.

Designing novel optoelectronic materials

The design of novel optoelectronic materials is crucial for the continued advancement of optoelectronic applications. As the demand for materials with enhanced performance, stability, and environmental sustainability increases, researchers are turning to innovative approaches to identify and develop novel materials with optimized properties^[4,141]. Recent studies demonstrate that ML and HT screening are invaluable tools for accelerating material discovery, enabling the efficient exploration of vast chemical spaces to uncover materials with desirable characteristics for next-generation applications^[142,143]. To initiate the exploration of new optoelectronic materials, Ma *et al.* employed a HT ML framework to screen for potential 2D PV materials within the family of octahedral oxyhalides (OOHs)^[144]. By training ML algorithms on structural and electronic properties from DFT data, the model evaluated a dataset of 5,000 OOH compounds, ultimately identifying six candidate materials with optimal band gaps and high electron mobilities suitable for PV applications. **Figure 9A** visualizes the model workflow, highlighting the multi-stage screening process and the structural attributes of top-performing candidates, such as $Bi_2Se_2Br_2$. Expanding the scope of material discovery, Jin *et al.* applied ensemble ML techniques to explore the compositional space of all-inorganic, lead-free perovskites, which address concerns over toxicity and stability associated with traditional lead-based perovskites^[145]. By developing a physics-inspired multicomponent neural network, they screened nearly 12 million $AA'BB'X_3X'$ compositions. **Figure 9B** illustrates the band gap prediction accuracy achieved through this ensemble model, showing the close

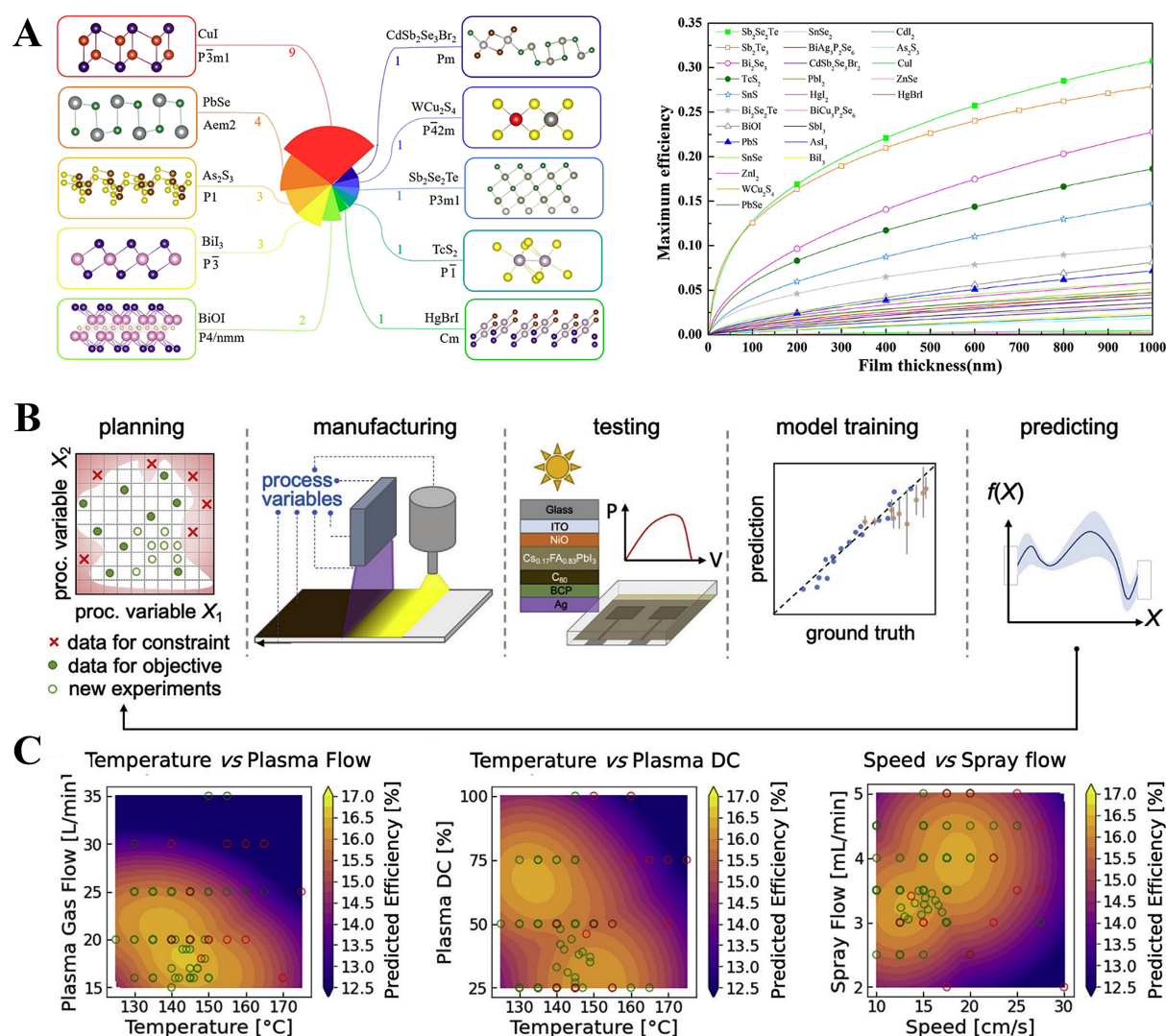


Figure 8. (A) 2D PV structural prototypes based on space group and calculated maximum efficiencies of 26 2D PV candidates as a function of absorber thickness. Copyright 2020, American Chemical Society, Reproduced with permission^[139]; (B) Schematic of sequential learning optimization of perovskite solar cells with probabilistic constraints; (C) Visualization of the process-efficiency relation based on the trained regression models. Copyright 2022, Elsevier, Reproduced with permission^[140]. 2D: Two-dimensional; PV: photovoltaic.

correlation between predicted and DFT calculated values. The figure also displays a detailed screening workflow, ultimately narrowing down thousands of candidates to a select group with optimal stability and electronic properties.

In another targeted exploration of material classes, Wang *et al.* implemented a ML framework designed to expedite the discovery of stable spinel materials with direct band gaps, essential for advanced optoelectronic applications^[146]. Figure 10A illustrates the detailed scheme of the target-driven approach. This includes the initial selection of A-, B-, and X-site elements, followed by data generation based on the combinations of these elements and a preliminary filter using the tolerance factor. Next, the process involves feature engineering and the application of ML techniques. Finally, crystal structure calculations, electronic structure analysis, and thermodynamic stability evaluations of the selected candidates are performed using DFT. Ultimately, the researchers used the XGBoost algorithm to screen 3,880 potential spinel compositions,

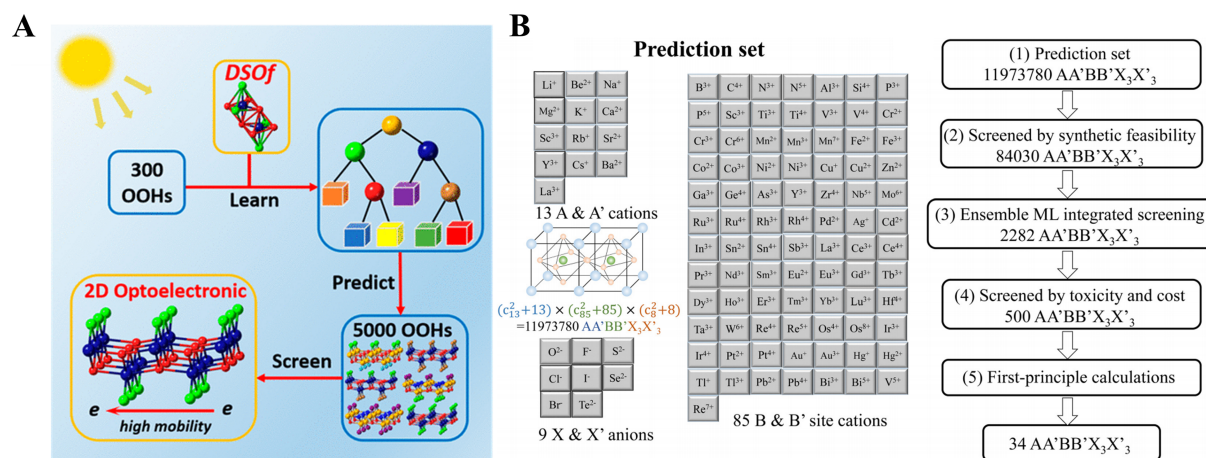


Figure 9. (A) Schematic structure of OOHs and the procedure of prediction and screening. Copyright 2019, American Chemical Society, Reproduced with permission^[144]; (B) The composition and structure of AA'BB'X₃X'₃ perovskites in the prediction set and the multi-step screening process of discovering novel AA'BB'X₃X'₃ perovskites according to the combination of ML and DFT calculation for PV application. Copyright 2023, Royal Society of Chemistry, Reproduced with permission^[145]. OOHs: Octahedral oxyhalides; ML: machine learning; DFT: density functional theory; PV: photovoltaic.

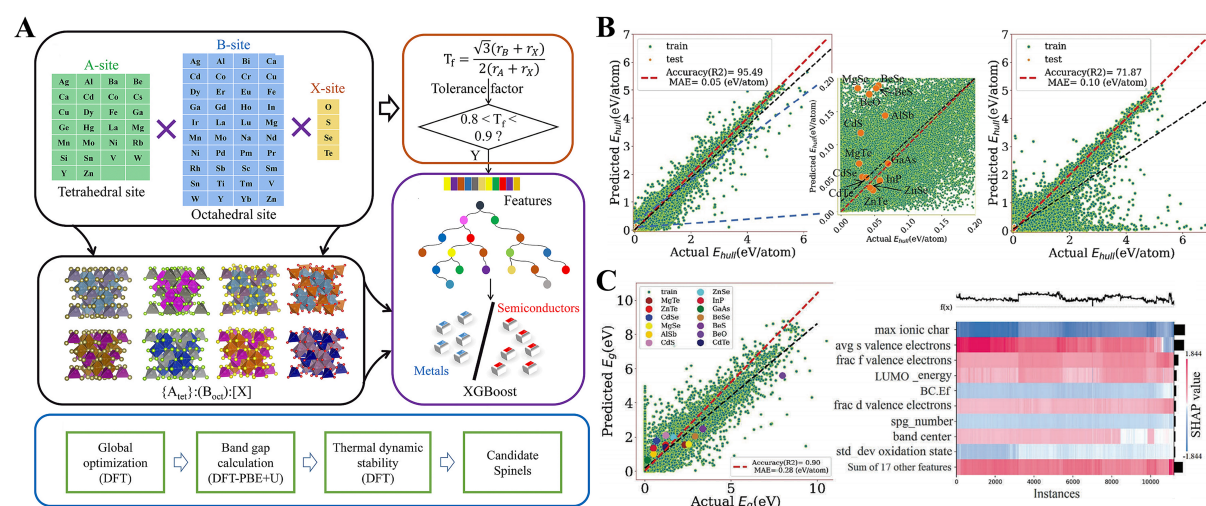


Figure 10. (A) Scheme of the proposed target-driven method. Selection of A-, B- and X-site elements, data generation based on the combination of selected elements, use of tolerance factor. Schematic diagram of feature engineering and ML technology. Calculations of crystal structures, electronic structures and thermodynamic stabilities of final candidates by DFT. Copyright 2021, Elsevier, Reproduced with permission^[146]; (B) The impact of formation energy on the prediction results of energy above the hull is depicted; (C) The prediction results of the bandgap and the heatmap of features based on the SHAP values. Copyright 2024, Wiley, Reproduced with permission^[147]. ML: Machine learning; DFT: density functional theory; SHAP: Shapley Additive exPlanations.

narrowing down to eight promising candidates demonstrating optimal direct band gaps and thermal stability under ambient conditions. This selection process was specifically targeted toward identifying materials with applicability in energy systems, such as CaAl₂O₄ and CaGa₂S₄. Another work focuses on Haeckelite structures, characterized by their unique square-octagonal frameworks and potential applications in various technological domains. Alibaghari *et al.* introduced an ML-based approach for identifying synthesizable Haeckelite structures, a distinctive class known for their square-octagonal framework^[147]. This study evaluated 1,083 candidate Haeckelite structures by analyzing formation energy, phase stability, and electronic bandgap properties to identify structures with optimal stability and optoelectronic characteristics.

Figure 10B details the effect of formation energy on stability predictions, showing the prediction results with formation energy as a descriptor, demonstrating how it improves the model accuracy in predicting the stability of various Haeckelite configurations. And the distribution of the 13 Haeckelite compounds is identified as stable within the output space, while providing a comparison with regression results when formation energy is excluded as a descriptor. Figure 10C presents the predicted bandgap values for the screened Haeckelite compounds, highlighting those within an ideal range for optoelectronic applications. Additionally, it displays a heatmap generated using Shapley Additive exPlanations (SHAP) values, which assesses the relative impact of each feature on the bandgap prediction. Key descriptors, including d-electron fraction and atomic radius, proved influential, confirming the robustness of the RFR model in selecting Haeckelite structures with desirable electronic properties. As a result, the study identified 13 Haeckelite structures with excellent stability and suitable bandgaps, which are highly compatible with optoelectronic applications. These structures exhibit phase stability and promising electron mobilities and absorption coefficients, making them strong candidates for optoelectronic devices where efficient light absorption and charge transport are crucial.

Further expanding on structural diversity, Li *et al.* focused on hybrid heterostructure semiconductors (HHSs) by combining ML and HT screening to identify hybrid organic-inorganic semiconductor superlattices with desirable optoelectronic properties^[148]. Targeting organic-inorganic semiconductor superlattices, the model screened over 200 structural variants to analyze their thermodynamic stability, electronic structures, and optoelectronic characteristics. Through this approach, 96 HHS candidates were identified with stable band gaps and efficient carrier mobility, attributes ideal for PV applications. As depicted in Figure 11A and B, hybrid organic-inorganic semiconductors are structurally classified into Type I (hybrid ion-substituting semiconductors) and Type II (hybrid heterostructured semiconductors). The model predictions, comparing the formation energies (E_{form}) of these structures, demonstrated high prediction accuracy, providing a strong framework for material discovery and optimization in advanced optoelectronic applications. Meanwhile, Chen *et al.* examined double hybrid organic-inorganic perovskites (DHOIPs) by integrating ML and HT screening to identify candidates for solar energy applications^[149]. By integrating ML with HT screening, the study assessed a vast pool of 78,400 DHOIP candidates based on critical criteria: charge neutrality, stability, non-toxicity, and bandgap suitability. This ML-driven framework is depicted in Figure 11C, which outlines the sequential screening and refinement process used to filter down the candidate pool, enabling the systematic identification of perovskites with desirable bandgap ranges tailored for solar cells. The ML model employed in this study incorporated specific structural features, notably the anisotropic shapes of organic cations, which significantly enhanced the prediction accuracy for optoelectronic properties. As a result, the model was capable of accurately screening perovskite structures that balanced electronic properties with physical stability. Ultimately, the approach narrowed the initial candidates to a promising list of 19 DHOIPs. These shortlisted compounds were further validated through DFT, which confirmed their stability and suitability as PV materials.

To develop more efficient, stable, and sustainable next-generation optoelectronic materials, Chen *et al.* developed a comprehensive data-driven platform to support the discovery and exploration of 2D HOIPs^[150]. Figure 12A illustrates the detailed ML workflow, encompassing stages from data collection and feature engineering to model selection, training, and final prediction. This structured ML pipeline enables accurate assessments of structure-property relationships by incorporating relevant descriptors that expand beyond conventional classifications. Figure 12B showcases the platform's extensive database, structured around geometric descriptors and an ML model, offering a powerful 2D HOIPs exploration tool. This platform addresses limitations in traditional classification methods for structure-property relationships by integrating structural descriptors, known as structure factors (SFs), into the ML framework. This platform integrates

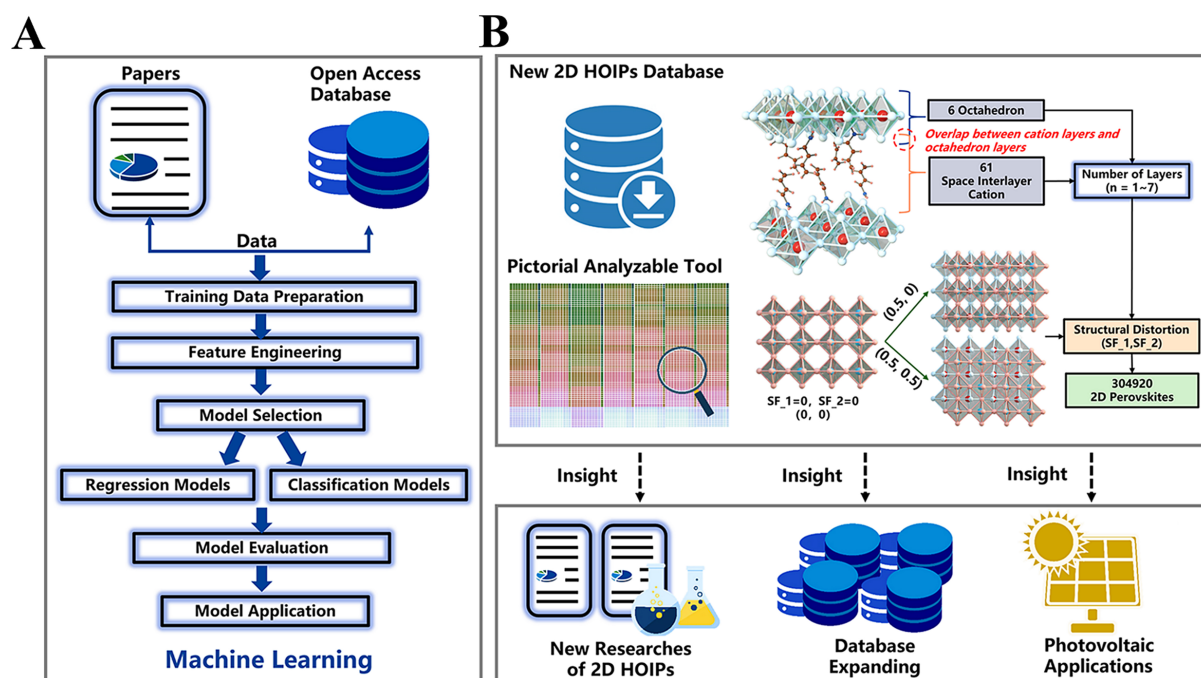


Figure 12. (A) An illustration of the ML process includes data collection, feature engineering, model selection and training, and model prediction; (B) The established database is based on geometric descriptors and ML model and a 2D HOIPs exploration platform integrating searching, download, analysis, and online prediction, which provide a useful tool for the further research of 2D HOIPs and other fields in materials, energy, and engineering, especially PV field. Copyright 2023, American Chemical Society, Reproduced with permission^[150]. ML: Machine learning; 2D: two-dimensional; HOIPs: hybrid organic-inorganic perovskites; PV: photovoltaic.

accuracy in areas such as band gap tuning, thermal stability, charge carrier dynamics, and synthesizability, offering significant insights into complex materials with targeted optoelectronic performance.

Firstly, a key challenge in applying ML to optoelectronic materials lies in the dataset limitations, as the data for materials are often sparse, heterogeneous, and lacking in standardization. As noted in previous research, inconsistent data quality can hamper ML model training, leading to unreliable predictions. To overcome this challenge, future efforts should focus on establishing standardized databases, which are crucial for property prediction. Improving data quality and quantity will lead to more accurate predictions, creating a positive feedback loop where high-quality data further refine ML model performance.

Secondly, current ML methods face limitations regarding interpretability and generalization. Although effective for prediction, many ML models operate as “black boxes”, making it difficult to understand the influence of individual features on material properties. Enhancing model transparency through feature importance analyses and interpretive tools, such as SHAP values, can help bridge this gap, providing insights into how model predictions relate to physical properties. Furthermore, incorporating domain expertise in materials science can further improve the accuracy and interpretability of ML models. For instance, understanding the fundamental impact of key properties, such as bandgap, carrier mobility, exciton binding energy, dielectric constant, and light absorption coefficient enables the selection of more physically meaningful descriptors, thereby guiding models toward more reliable and scientific predictions.

Thirdly, predicting the optoelectronic performance of materials, such as band gaps, charge transport, and stability, is inherently complex due to many influencing factors, including composition, crystal structure, and defect states. ML models have shown promise in handling these multi-parameter challenges by

Table 1. A year-by-year comparison of ML studies, highlighting the types of materials investigated, the ML models employed, and the key results

Year	Material type	ML models	Key results	Ref.
2019	Perovskite oxides/halides	SISSO	τ factor 92% stability prediction accuracy	[135]
2019	2D OOHs	GBR	Several candidates with excellent bandgap/mobility have been identified	[144]
2020	Cd-based chalcogenides	RFR	Accurately predict impurity formation enthalpy and charge transition levels	[139]
2020	2D inorganic crystal	GBC	Developed HT and ML methods to identify 2DPV candidate materials	[126]
2021	Spinel	XGBoost	Successfully screened out 8 spinels	[146]
2022	Perovskite	GCNN	A model is proposed to evaluate the synthesizability of perovskites	[136]
2022	Hybrid heterostructured semiconductors	GBRT	96 stable HHSs were found	[148]
2022	DHOIPs	Δ -GBR	19 promising DHOIPs were screened	[149]
2022	Perovskite	BO	Achieved 18.5% power conversion efficiency	[140]
2023	2D hybrid Organic-inorganic lead-halide perovskites	XGBR	Develop a 2D HOIPs exploration platform	[150]
2023	All-inorganic lead-free perovskites	Multi-component neural network	34 lead-free AABBX ₃ X ₃ identified	[145]
2023	2D hybrid perovskites	Bagging classifier	Accurate classification of type I/II band alignment	[129]
2024	Multi-element MHPs	CGCNN	Achieved high prediction accuracy in stability and band gap	[128]
2024	Perovskites	SISSO	Developing a universal ML descriptor	[127]
2024	Haeckelite structures	RFR and CNN	Identified 13 stable Haeckelite configurations with ideal band gaps	[147]
2024	Lead-free VOHPs	Unsupervised ML	Revealed the influence of A-site cations on carrier lifetimes	[130]

ML: Machine learning; SISSO: sure independence screening and sparsifying operator; 2D: two-dimensional; OOHs: octahedral oxyhalides; GBR: gradient boosting regressor; RFR: random forest regressor; GBC: gradient boosting classifier; HT: high-throughput; PV: photovoltaic; GCNN: graph convolutional neural network; GBRT: gradient boosting regression trees; HHSs: hybrid heterostructure semiconductors; DHOIPs: double hybrid organic-inorganic perovskites; BO: Bayesian Optimization; XGBR: XGBoost regressor; HOIPs: hybrid organic-inorganic perovskites; MHPs: metal halide perovskites; CGCNN: crystal graph convolutional neural network; CNN: convolutional neural network; VOHPs: vacancy-ordered halide perovskites.

efficiently exploring large compositional spaces and optimizing for specific properties. These features are particularly advantageous for identifying high-performance candidate materials with ideal energy band alignment and environmental stability. Future advances may include combining ML predictions with experimental feedback to enable more precise tuning to optimize material properties.

Finally, while ML models excel at identifying statistical correlations within data, their lack of direct physical interpretation can limit their ability to predict mechanisms accurately. Unlike traditional physics-based models, ML models are not inherently bound by the laws of physics, which restricts their applicability in certain contexts. Combining ML techniques with first-principles methods or physics-driven models could enhance the predictive accuracy and physical validity of ML outcomes, fostering a balanced approach that leverages the strengths of both paradigms.

Looking ahead, the integration of AI and ML techniques has revolutionized materials research. The shift from traditional experimental approaches to HT screening and ML-driven prediction has allowed researchers to rapidly assess vast chemical and structural spaces, enabling the systematic exploration of key parameters that influence optoelectronic properties. AI-driven ML methods have proven particularly effective in predicting material properties^[151]. For instance, DeepMind of Google developed GNoME, a

GNN-based model trained on data from the MP, which predicted over 380,000 new stable materials, significantly expanding the known crystalline materials landscape^[152]. Furthermore, autonomous A-Lab of Lawrence Berkeley National Laboratory synthesized 41 out of 58 AI-predicted materials within 17 days, demonstrating the potential of combining AI-driven predictions with automated synthesis^[153]. Notably, generative models such as MatterGen have emerged as powerful tools in materials design. MatterGen employs a diffusion-based generative process to create stable and diverse inorganic crystals across the periodic table. It can be fine-tuned to target specific property constraints, enabling the generation of materials with desired chemistry, symmetry, and electronic properties^[154]. These developments underscore the transformative impact of ML in accelerating the discovery and optimization of optoelectronic materials. The convergence of AI with autonomous laboratory platforms, capable of performing automated material synthesis and characterization, will enable the rapid development of materials with superior efficiency, stability, and sustainability^[155-157]. This integrated approach will accelerate the design of next-generation optoelectronic devices, helping to meet the growing demand for energy-efficient technologies. Advances in dataset standardization, model interpretability, and AI frameworks will play a pivotal role in shaping the future of materials innovation.

In summary, ML transforms optoelectronic material design by offering rapid, high-accuracy predictions across vast chemical spaces. Continued advancements in dataset standardization, model interpretability, property-specific tuning, and hybrid ML-physics frameworks will foster a new generation of optoelectronic materials optimized for efficiency, stability, and sustainability. The integration of ML with experimental and AI holds promise for accelerating material innovation, ultimately contributing to developing next-generation optoelectronic devices that meet the increasing demands for energy-efficient technologies.

DECLARATIONS

Authors' contributions

Data analysis, interpretation and manuscript draft: Shu, Y.

Performed data acquisition and collected references: Li, R.; Lin, Y.; Han, S.; Zhou, J.

Provided revision, acquired funding and supervision: Miao, N.; Sun, Z.

Availability of data and materials

Not applicable.

Financial support and sponsorship

This work was supported by the National Natural Science Foundation of China (52222101).

Conflicts of interest

All authors declared that there are no conflicts of interest.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Copyright

© The Author(s) 2025.

REFERENCES

1. Zhang, L.; Mei, L.; Wang, K.; et al. Advances in the application of perovskite materials. *Nanomicro. Lett.* **2023**, *15*, 177. DOI PubMed PMC
2. Liu, J.; Zhang, S.; Wang, W.; Zhang, H. Photoelectrocatalytic principles for meaningfully studying photocatalyst properties and photocatalysis processes: from fundamental theory to environmental applications. *J. Energy. Chem.* **2023**, *86*, 84-117. DOI
3. Tan, S.; Huang, T.; Yavuz, I.; et al. Stability-limiting heterointerfaces of perovskite photovoltaics. *Nature* **2022**, *605*, 268-73. DOI PubMed
4. Lu, T.; Li, M.; Lu, W.; Zhang, T. Recent progress in the data-driven discovery of novel photovoltaic materials. *J. Mater. Inf.* **2022**, *2*, 7. DOI
5. Wen, J.; Rong, K.; Jiang, L.; et al. Copper-based perovskites and perovskite-like halides: a review from the perspective of molecular level. *Nano. Energy.* **2024**, *128*, 109802. DOI
6. Aydin, E.; Ugur, E.; Yildirim, B. K.; et al. Enhanced optoelectronic coupling for perovskite/silicon tandem solar cells. *Nature* **2023**, *623*, 732-8. DOI PubMed
7. Isikgor, F. H.; Zhumagali, S.; Merino, L. V. T.; De Bastiani, M.; McCulloch, I.; De Wolf, S. Molecular engineering of contact interfaces for high-performance perovskite solar cells. *Nat. Rev. Mater.* **2023**, *8*, 89-108. DOI
8. Zhang, Z. Light-emitting materials for wearable electronics. *Nat. Rev. Mater.* **2022**, *7*, 839-40. DOI
9. Jang, E.; Jang, H. Review: quantum dot light-emitting diodes. *Chem. Rev.* **2023**, *123*, 4663-92. DOI PubMed
10. Han, T.; Jang, K. Y.; Dong, Y.; Friend, R. H.; Sargent, E. H.; Lee, T. A roadmap for the commercialization of perovskite light emitters. *Nat. Rev. Mater.* **2022**, *7*, 757-77. DOI
11. Xiong, R.; Zhang, L.; Wen, C.; Anpo, M.; Ang, Y. S.; Sa, B. Ferroelectric switching driven photocatalytic overall water splitting in the As/In₂Se₃ heterostructure. *J. Mater. Chem. A.* **2025**, *13*, 4563-75. DOI
12. Chen, Z.; Yao, D.; Chu, C.; Mao, S. Photocatalytic H₂O₂ production systems: design strategies and environmental applications. *Chem. Eng. J.* **2023**, *451*, 138489. DOI
13. Zhou, P.; Navid, I. A.; Ma, Y.; et al. Solar-to-hydrogen efficiency of more than 9% in photocatalytic water splitting. *Nature* **2023**, *613*, 66-70. DOI PubMed
14. Fan, Y.; Huang, W.; Zhu, F.; et al. Dispersion-assisted high-dimensional photodetector. *Nature* **2024**, *630*, 77-83. DOI PubMed
15. Li, Z.; Yan, T.; Fang, X. Low-dimensional wide-bandgap semiconductors for UV photodetectors. *Nat. Rev. Mater.* **2023**, *8*, 587-603. DOI
16. Wang, H.; Li, Z.; Li, D.; et al. Van der Waals integration based on two-dimensional materials for high-performance infrared photodetectors. *Adv. Funct. Mater.* **2021**, *31*, 2103106. DOI
17. Saliba, M.; Matsui, T.; Seo, J. Y.; et al. Cesium-containing triple cation perovskite solar cells: improved stability, reproducibility and high efficiency. *Energy. Environ. Sci.* **2016**, *9*, 1989-97. DOI PubMed PMC
18. Zhang, Y.; Chen, Y.; Liu, G.; et al. Nonalloyed α -phase formamidinium lead triiodide solar cells through iodine intercalation. *Science* **2025**, *387*, 284-90. DOI
19. Tan, T.; Jiang, X.; Wang, C.; Yao, B.; Zhang, H. 2D material optoelectronics for information functional device applications: status and challenges. *Adv. Sci.* **2020**, *7*, 2000058. DOI PubMed PMC
20. Dastgeer, G.; Afzal, A. M.; Nazir, G.; Sarwar, N. p-GeSe/n-ReS₂ heterojunction rectifier exhibiting a fast photoresponse with ultra-high frequency-switching applications. *Adv. Mater. Inter.* **2021**, *8*, 2100705. DOI
21. Liu, Z.; Na, G.; Tian, F.; Yu, L.; Li, J.; Zhang, L. Computational functionality-driven design of semiconductors for optoelectronic applications. *InfoMat* **2020**, *2*, 879-904. DOI
22. Khan, D.; Qu, G.; Muhammad, I.; Tang, Z.; Xu, Z. Overcoming two key challenges in monolithic perovskite-silicon tandem solar cell development: wide bandgap and textured substrate - a comprehensive review. *Adv. Energy. Mater.* **2023**, *13*, 2302124. DOI
23. Zhang, C. Z.; Fu, X. Q. Applications and potentials of machine learning in optoelectronic materials research: an overview and perspectives. *Chinese. Phys. B.* **2023**, *32*, 126103. DOI
24. Hu, Y.; Chen, J.; Wei, Z.; He, Q.; Zhao, Y. Recent advances and applications of machine learning in electrocatalysis. *J. Mater. Inf.* **2023**, *3*, 18. DOI
25. Yuan, J.; Li, Z.; Yang, Y.; et al. Applications of machine learning method in high-performance materials design: a review. *J. Mater. Inf.* **2024**, *4*, 14. DOI
26. Liu, Y.; Zhao, T.; Ju, W.; Shi, S. Materials discovery and design using machine learning. *J. Materiomics.* **2017**, *3*, 159-77. DOI
27. Yang, X.; Zhou, K.; He, X.; Zhang, L. Methods and applications of machine learning in computational design of optoelectronic semiconductors. *Sci. China. Mater.* **2024**, *67*, 1042-81. DOI
28. Himanen, L.; Geurts, A.; Foster, A. S.; Rinke, P. Data-driven materials science: status, challenges, and perspectives. *Adv. Sci.* **2019**, *6*, 1900808. DOI PubMed PMC
29. Brunton, S. L.; Kutz, J. N. Data-driven science and engineering: machine learning, dynamical systems, and control. Cambridge University Press: 2019. DOI
30. Chen, Z.; Yang, Y. Data-driven design of eutectic high entropy alloys. *J. Mater. Inf.* **2023**, *3*, 10. DOI
31. Chen, M.; Yin, Z.; Shan, Z.; et al. Application of machine learning in perovskite materials and devices: a review. *J. Energy. Chem.* **2024**, *94*, 254-72. DOI

32. He, H.; Wang, Y.; Qi, Y.; Xu, Z.; Li, Y.; Wang, Y. From prediction to design: recent advances in machine learning for the study of 2D materials. *Nano. Energy*. **2023**, *118*, 108965. DOI
33. Chen, J.; Feng, M.; Zha, C.; Shao, C.; Zhang, L.; Wang, L. Machine learning-driven design of promising perovskites for photovoltaic applications: a review. *Surf. Interfaces*. **2022**, *35*, 102470. DOI
34. Nematov, D.; Hojamberdiev, M. Machine learning - driven materials discovery: unlocking next-generation functional materials - a minireview. *arXiv* **2025**, arXiv:2503.18975. <https://doi.org/10.48550/arXiv.2503.18975>. (accessed 27 May 2025)
35. Li, Y.; Yang, K. High-throughput computational design of halide perovskites and beyond for optoelectronics. *WIREs. Comput. Mol. Sci.* **2021**, *11*, e1500. DOI
36. Shen, L.; Zhou, J.; Yang, T.; Yang, M.; Feng, Y. P. High-throughput computational discovery and intelligent design of two-dimensional functional materials for various applications. *Acc. Mater. Res.* **2022**, *3*, 572-83. DOI
37. Xu, D.; Zhang, Q.; Huo, X.; Wang, Y.; Yang, M. Advances in data-assisted high-throughput computations for material design. *MGE. Adv.* **2023**, *1*, e11. DOI
38. Gan, Y.; Miao, N.; Lan, P.; Zhou, J.; Elliott, S. R.; Sun, Z. Robust design of high-performance optoelectronic chalcogenide crystals from high-throughput computation. *J. Am. Chem. Soc.* **2022**, *144*, 5878-86. DOI PubMed
39. Lan, P.; Miao, N.; Gan, Y.; et al. High-throughput computational design of 2D ternary chalcogenides for sustainable energy. *J. Phys. Chem. Lett.* **2023**, *14*, 10489-98. DOI
40. Bai, S.; Zhang, X.; Zhao, L. D. Rethinking SnSe thermoelectrics from computational materials science. *Acc. Chem. Res.* **2023**, *56*, 3065-75. DOI
41. Deng, T.; Qiu, P.; Yin, T.; et al. High-throughput strategies in the discovery of thermoelectric materials. *Adv. Mater.* **2024**, *36*, e2311278. DOI PubMed
42. Xu, Y.; Elcoro, L.; Song, Z. D.; et al. High-throughput calculations of magnetic topological materials. *Nature* **2020**, *586*, 702-7. DOI PubMed
43. Cao, G.; Ouyang, R.; Ghiringhelli, L. M.; et al. Artificial intelligence for high-throughput discovery of topological insulators: the example of alloyed tetradymites. *Phys. Rev. Mater.* **2020**, *4*, 034204. DOI
44. Zhang, X.; Meng, W.; Liu, Y.; Dai, X.; Liu, G.; Kou, L. Magnetic electrodes: high-throughput material screening, intriguing properties, and applications. *J. Am. Chem. Soc.* **2023**, *145*, 5523-35. DOI
45. Miao, N.; Sun, Z. Computational design of two-dimensional magnetic materials. *WIREs. Comput. Mol. Sci.* **2022**, *12*, e1545. DOI
46. de Pablo, J. J.; Jackson, N. E.; Webb, M. A.; et al. New frontiers for the materials genome initiative. *npj. Comput. Mater.* **2019**, *5*, 173. DOI
47. de Pablo, J. J.; Jones, B.; Kovacs, C. L.; Ozolins, V.; Ramirez, A. P. The Materials Genome Initiative, the interplay of experiment, theory and computation. *Curr. Opin. Solid. State. Mater. Sci.* **2014**, *18*, 99-117. DOI
48. Yu, Q.; Ma, N.; Leung, C.; Liu, H.; Ren, Y.; Wei, Z. AI in single-atom catalysts: a review of design and applications. *J. Mater. Inf.* **2025**, *5*, 9. DOI
49. Jordan, M. I.; Mitchell, T. M. Machine learning: trends, perspectives, and prospects. *Science* **2015**, *349*, 255-60. DOI PubMed
50. Xu, P.; Ji, X.; Li, M.; Lu, W. Small data machine learning in materials science. *npj. Comput. Mater.* **2023**, *9*, 1000. DOI
51. Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559*, 547-55. DOI PubMed
52. Schleder, G. R.; Padilha, A. C. M.; Acosta, C. M.; Costa, M.; Fazzio, A. From DFT to machine learning: recent approaches to materials science - a review. *J. Phys. Mater.* **2019**, *2*, 032001. DOI
53. Jacobsson, T. J.; Hultqvist, A.; García-Fernández, A.; et al. An open-access database and analysis tool for perovskite solar cells based on the FAIR data principles. *Nat. Energy*. **2022**, *7*, 107-15. DOI
54. Mannodi-Kanakithodi, A.; Chan, M. K. Y. Data-driven design of novel halide perovskite alloys. *Energy. Environ. Sci.* **2022**, *15*, 1930-49. DOI
55. Ma, B.; Wu, X.; Zhao, C.; et al. An interpretable machine learning strategy for pursuing high piezoelectric coefficients in (K_{0.5}Na_{0.5})NbO₃-based ceramics. *npj. Comput. Mater.* **2023**, *9*, 1187. DOI
56. Cheng, G.; Gong, X. G.; Yin, W. J. Crystal structure prediction by combining graph network and optimization algorithm. *Nat. Commun.* **2022**, *13*, 1492. DOI PubMed PMC
57. Chen, C.; Zuo, Y.; Ye, W.; Li, X.; Deng, Z.; Ong, S. P. A critical review of machine learning of energy materials. *Adv. Energy. Mater.* **2020**, *10*, 1903242. DOI
58. Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. The Cambridge Structural Database. *Acta. Crystallogr. B. Struct. Sci. Cryst. Eng. Mater.* **2016**, *72*, 171-9. DOI PubMed PMC
59. Bergerhoff, G.; Hundt, R.; Sievers, R.; Brown, I. D. The inorganic crystal structure data base. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 66-9. DOI
60. Gražulis, S.; Daškevič, A.; Merkys, A.; et al. Crystallography Open Database (COD): an open-access collection of crystal structures and platform for world-wide collaboration. *Nucleic. Acids. Res.* **2012**, *40*, D420-7. DOI PubMed PMC
61. Curtarolo, S.; Setyawan, W.; Wang, S.; et al. AFLOWLIB.ORG: a distributed materials properties repository from high-throughput ab initio calculations. *Comput. Mater. Sci.* **2012**, *58*, 227-35. DOI
62. Jain, A.; Ong, S. P.; Hautier, G.; et al. Commentary: The Materials Project: a materials genome approach to accelerating materials innovation. *APL. Mater.* **2013**, *1*, 011002. DOI

63. Gjerding, M. N.; Taghizadeh, A.; Rasmussen, A.; et al. Recent progress of the Computational 2D Materials Database (C2DB). *2D. Mater.* **2021**, *8*, 044002. DOI
64. Kirklin, S.; Saal, J. E.; Meredig, B.; et al. The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. *npj. Comput. Mater.* **2015**, *1*, BFnjcompumats201510. DOI
65. Damewood, J.; Karaguesian, J.; Lunger, J. R.; et al. Representations of materials for machine learning. *Annu. Rev. Mater. Res.* **2023**, *53*, 399-426. DOI
66. Li, S.; Liu, Y.; Chen, D.; Jiang, Y.; Nie, Z.; Pan, F. Encoding the atomic structure for machine learning in materials science. *WIREs. Comput. Mol. Sci.* **2022**, *12*, e1558. DOI
67. Oh, S. H. V.; Hwang, W.; Kim, K.; Lee, J. H.; Soon, A. Using feature-assisted machine learning algorithms to boost polarity in lead-free multicomponent niobate alloys for high-performance ferroelectrics. *Adv. Sci.* **2022**, *9*, e2104569. DOI PubMed PMC
68. Schmidt, J.; Marques, M. R. G.; Botti, S.; Marques, M. A. L. Recent advances and applications of machine learning in solid-state materials science. *npj. Comput. Mater.* **2019**, *5*, 221. DOI
69. Li, J.; Cheng, K.; Wang, S.; et al. Feature selection: a data perspective. *ACM. Comput. Surv.* **2018**, *50*, 1-45. DOI
70. Hsu, H.; Hsieh, C.; Lu, M. Hybrid feature selection by combining filters and wrappers. *Expert. Syst. Appl.* **2011**, *38*, 8144-50. DOI
71. Rodriguez-Galiano, V. F.; Luque-Espinar, J. A.; Chica-Olmo, M.; Mendes, M. P. Feature selection approaches for predictive modelling of groundwater nitrate pollution: an evaluation of filters, embedded and wrapper methods. *Sci. Total. Environ.* **2018**, *624*, 661-72. DOI PubMed
72. Liu, H.; Zhou, M.; Liu, Q. An embedded feature selection method for imbalanced data classification. *IEEE/CAA. J. Autom. Sinica.* **2019**, *6*, 703-15. DOI
73. Zhang, Z.; Liu, S.; Xiong, Q.; Liu, Y. Strategic integration of machine learning in the design of excellent hybrid perovskite solar cells. *J. Phys. Chem. Lett.* **2025**, *16*, 738-46. DOI
74. Gladkikh, V.; Kim, D. Y.; Hajibabaei, A.; Jana, A.; Myung, C. W.; Kim, K. S. Machine learning for predicting the band gaps of ABX₃ perovskites from elemental properties. *J. Phys. Chem. C.* **2020**, *124*, 8905-18. DOI
75. Gou, F.; Ma, Z.; Yang, Q.; et al. Machine learning-assisted prediction and control of bandgap for organic-inorganic metal halide perovskites. *ACS. Appl. Mater. Interfaces.* **2025**, *17*, 18383-93. DOI PubMed
76. Wang, A. Y.; Murdock, R. J.; Kauwe, S. K.; et al. Machine learning for materials scientists: an introductory guide toward best practices. *Chem. Mater.* **2020**, *32*, 4954-65. DOI
77. Wei, J.; Chu, X.; Sun, X.; et al. Machine learning in materials science. *InfoMat* **2019**, *1*, 338-58. DOI
78. Orupattur, N. V.; Mushrif, S. H.; Prasad, V. Catalytic materials and chemistry development using a synergistic combination of machine learning and ab initio methods. *Comput. Mater. Sci.* **2020**, *174*, 109474. DOI
79. Ali, Y.; Awwad, E.; Al-Razgan, M.; Maarouf, A. Hyperparameter search for machine learning algorithms for optimizing the computational complexity. *Processes* **2023**, *11*, 349. DOI
80. Cherkassky, V.; Ma, Y. Practical selection of SVM parameters and noise estimation for SVM regression. *Neural. Netw.* **2004**, *17*, 113-26. DOI PubMed
81. Cover, T.; Hart, P. Nearest neighbor pattern classification. *IEEE. Trans. Inform. Theory.* **1967**, *13*, 21-7. DOI
82. Yang, F. J. An implementation of Naive Bayes Classifier. In *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*, Las Vegas, USA. Dec 12-14, 2018. IEEE; 2018; pp. 301-6. DOI
83. Liu, Y.; Zhou, Q.; Cui, G. Machine learning boosting the development of advanced lithium batteries. *Small. Methods.* **2021**, *5*, e2100442. DOI
84. Dong, X.; Yu, Z.; Cao, W.; Shi, Y.; Ma, Q. A survey on ensemble learning. *Front. Comput. Sci.* **2020**, *14*, 241-58. DOI
85. Sagi, O.; Rokach, L. Ensemble learning: a survey. *WIREs. Data. Min. Knowl.* **2018**, *8*, e1249. DOI
86. Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189-232. <http://www.jstor.org/stable/2699986>. (accessed 27 May 2025)
87. Freund, Y.; Schapire, R. E. Experiments with a new boosting algorithm. In *Proceedings of the Thirteenth International Conference*, 1996. pp. 148-56. <http://www.jstor.org/stable/2699986>. (accessed 27 May 2025)
88. Chen, T.; Guestrin, C. XGBoost: a scalable tree boosting system. In *Proceedings of the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, USA. 2016; pp. 785-94. DOI
89. Pavlyshenko, B. Using stacking approaches for machine learning models. In *2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP)*, Lviv, Ukraine, Aug 21-25, 2018. IEEE; 2018. pp. 255-8. DOI
90. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436-44. DOI PubMed
91. Liu, J.; Liang, L.; Su, B.; et al. Transformativ strategies in photocatalyst design: merging computational methods and deep learning. *J. Mater. Inf.* **2024**, *4*, 33. DOI
92. Choudhary, K.; Decost, B.; Chen, C.; et al. Recent advances and applications of deep learning methods in materials science. *npj. Comput. Mater.* **2022**, *8*, 734. DOI
93. Jain, A. K.; Mao, J.; Mohiuddin, K. M. Artificial neural networks: a tutorial. *Computer* **1996**, *29*, 31-44. DOI
94. Shelhamer, E.; Long, J.; Darrell, T. Fully convolutional networks for semantic segmentation. *IEEE. Trans. Pattern. Anal. Mach. Intell.* **2017**, *39*, 640-51. DOI PubMed
95. Zhang, H.; Wang, Z.; Liu, D. A comprehensive review of stability analysis of continuous-time recurrent neural networks. *IEEE. Trans. Neural. Netw. Learning. Syst.* **2014**, *25*, 1229-62. DOI

96. Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; et al. Generative adversarial networks. *arXiv* **2014**, arXiv:1406.2661. <https://doi.org/10.48550/arXiv.1406.2661>. (accessed 27 May 2025)
97. Zheng, Z.; Zhang, O.; Borgs, C.; Chayes, J. T.; Yaghi, O. M. ChatGPT chemistry assistant for text mining and the prediction of MOF synthesis. *J. Am. Chem. Soc.* **2023**, *145*, 18048–62. DOI PubMed PMC
98. OpenAI: Optimizing language models for dialogue. 2023. <https://openai.com/blog/chatgpt/>. (accessed 27 May 2025).
99. Chowdhary, K. R. Natural language processing. In: Fundamentals of artificial intelligence. New Delhi: Springer India; 2020. pp. 603–49. DOI
100. Zhu, J. J.; Yang, M.; Ren, Z. J. Machine learning in environmental research: common pitfalls and best practices. *Environ. Sci. Technol.* **2023**, *57*, 17671–89. DOI PubMed
101. Li, Z.; Yoon, J.; Zhang, R.; et al. Machine learning in concrete science: applications, challenges, and best practices. *npj. Comput. Mater.* **2022**, *8*, 810. DOI
102. Artrith, N.; Butler, K. T.; Coudert, F. X.; et al. Best practices in machine learning for chemistry. *Nat. Chem.* **2021**, *13*, 505–8. DOI
103. Wong, T. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern. Recognit.* **2015**, *48*, 2839–46. DOI
104. Efron, B.; Tibshirani, R. J. An introduction to the bootstrap. Chapman and Hall/CRC: 1994. <https://www.hms.harvard.edu/bss/neuro/bornlab/nb204/statistics/bootstrap.pdf>. (accessed 27 May 2025).
105. Palanivinaayagam, A.; El-Bayeh, C. Z.; Damaševičius, R. Twenty years of machine-learning-based text classification: a systematic review. *Algorithms* **2023**, *16*, 236. DOI
106. Sebastiani, F. Machine learning in automated text categorization. *ACM. Comput. Surv.* **2002**, *34*, 1–47. DOI
107. Chicco, D.; Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC. Genomics.* **2020**, *21*, 6. DOI PubMed PMC
108. Ho, S. Y.; Phua, K.; Wong, L.; Bin, G. W. W. Extensions of the external validation for checking learned model interpretability and generalizability. *Patterns* **2020**, *1*, 100129. DOI PubMed PMC
109. Xiong, Z.; Cui, Y.; Liu, Z.; Zhao, Y.; Hu, M.; Hu, J. Evaluating explorative prediction power of machine learning algorithms for materials discovery using k-fold forward cross-validation. *Comput. Mater. Sci.* **2020**, *171*, 109203. DOI
110. Probst, P.; Bischl, B.; Boulesteix, A. L. Tunability: importance of hyperparameters of machine learning algorithms. *arXiv* **2018**, arXiv:1802.09596. <https://doi.org/10.48550/arXiv.1802.09596>. (accessed 27 May 2025)
111. Bischl, B.; Binder, M.; Lang, M.; et al. Hyperparameter optimization: foundations, algorithms, best practices, and open challenges. *WIREs. Data. Min. Knowl.* **2023**, *13*, e1484. DOI
112. Li, L.; Jamieson, K.; DeSalvo, G.; Rostamizadeh, A.; Talwalkar, A. Hyperband: a novel bandit-based approach to hyperparameter optimization. *arXiv* **2016**, arXiv:1603.06560. <https://doi.org/10.48550/arXiv.1603.06560>. (accessed 27 May 2025)
113. Victoria, A. H.; Maragatham, G. Automatic tuning of hyperparameters using Bayesian optimization. *Evol. Syst.* **2021**, *12*, 217–23. DOI
114. Sa, B.; Hu, R.; Zheng, Z.; et al. High-throughput computational screening and machine learning modeling of Janus 2D III–VI van der Waals heterostructures for solar energy applications. *Chem. Mater.* **2022**, *34*, 6687–701. DOI
115. Mooraj, S.; Chen, W. A review on high-throughput development of high-entropy alloys by combinatorial methods. *J. Mater. Inf.* **2023**, *3*, 4. DOI
116. Sa, Z.; Liu, F.; Zhuang, X.; et al. Toward high bias-stress stability P-type GaSb nanowire field-effect-transistor for gate-controlled near-infrared photodetection and photocommunication. *Adv. Funct. Mater.* **2023**, *33*, 2304064. DOI
117. Kang, Y.; Hou, X.; Zhang, Z.; et al. Ultrahigh-performance and broadband photodetector from visible to shortwave infrared band based on GaAsSb nanowires. *Chem. Eng. J.* **2024**, *501*, 157392. DOI
118. Kang, Y.; Hou, X.; Zhang, Z.; et al. Enhanced visible-NIR dual-band performance of GaAs nanowire photodetectors through phase manipulation. *Adv. Opt. Mater.* **2025**, 2500289. DOI
119. Li, D.; Lan, C.; Manikandan, A.; et al. Ultra-fast photodetectors based on high-mobility indium gallium antimonide nanowires. *Nat. Commun.* **2019**, *10*, 1664. DOI PubMed PMC
120. Gao, Y.; Zhang, Q.; Hu, W.; Yang, J. First-principles computational screening of two-dimensional polar materials for photocatalytic water splitting. *ACS. Nano.* **2024**, *18*, 19381–90. DOI
121. Kangsabanik, J.; Svendsen, M. K.; Taghizadeh, A.; Crovetto, A.; Thygesen, K. S. Indirect band gap semiconductors for thin-film photovoltaics: high-throughput calculation of phonon-assisted absorption. *J. Am. Chem. Soc.* **2022**, *144*, 19872–83. DOI PubMed
122. Jiang, X.; Yin, W. High-throughput computational screening of oxide double perovskites for optoelectronic and photocatalysis applications. *J. Energy. Chem.* **2021**, *57*, 351–8. DOI
123. Tang, J.; Xue, J.; Xu, H.; et al. Power generation density boost of bifacial tandem solar cells revealed by high throughput optoelectrical modelling. *Energy. Environ. Sci.* **2024**, *17*, 6068–78. DOI
124. Xie, T.; Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* **2018**, *120*, 145301. DOI PubMed
125. Liang, C.; Rouzhahong, Y.; Ye, C.; Li, C.; Wang, B.; Li, H. Material symmetry recognition and property prediction accomplished by crystal capsule representation. *Nat. Commun.* **2023**, *14*, 5198. DOI PubMed PMC
126. Mannodi-Kanakithodi, A.; Toriyama, M. Y.; Sen, F. G.; Davis, M. J.; Klie, R. F.; Chan, M. K. Y. Machine-learned impurity level prediction for semiconductors: the example of Cd-based chalcogenides. *npj. Comput. Mater.* **2020**, *6*, 296. DOI

127. Wang, H.; Ouyang, R.; Chen, W.; Pasquarello, A. High-quality data enabling universality of band gap descriptor and discovery of photovoltaic perovskites. *J. Am. Chem. Soc.* **2024**, *146*, 17636–45. [DOI](#) [PubMed](#)
128. Kim, J.; Noh, J.; Im, J. Machine learning-enabled chemical space exploration of all-inorganic perovskites for photovoltaics. *npj. Comput. Mater.* **2024**, *10*, 1270. [DOI](#)
129. Mahal, E.; Roy, D.; Manna, S. S.; Pathak, B. Machine learning-driven prediction of band-alignment types in 2D hybrid perovskites. *J. Mater. Chem. A.* **2023**, *11*, 23547–55. [DOI](#)
130. Nayak, P. K.; Mora Perez, C.; Liu, D.; Prezhd, O. V.; Ghosh, D. A-cation-dependent excited state charge carrier dynamics in vacancy-ordered halide perovskites: insights from computational and machine learning models. *Chem. Mater.* **2024**, *36*, 3875–85. [DOI](#)
131. Wang, S.; Yousefi Amin, A. A.; Wu, L.; Cao, M.; Zhang, Q.; Ameri, T. Perovskite nanocrystals: synthesis, stability, and optoelectronic applications. *Small. Struct.* **2021**, *2*, 2000124. [DOI](#)
132. Liu, J.; Yang, Z.; Ye, B.; et al. A review of stability-enhanced luminescent materials: fabrication and optoelectronic applications. *J. Mater. Chem. C.* **2019**, *7*, 4934–55. [DOI](#)
133. Liu, H.; Cheng, J.; Dong, H.; et al. Screening stable and metastable ABO₃ perovskites using machine learning and the materials project. *Comput. Mater. Sci.* **2020**, *177*, 109614. [DOI](#)
134. Burlingame, Q.; Ball, M.; Loo, Y. It's time to focus on organic solar cell stability. *Nat. Energy.* **2020**, *5*, 947–9. [DOI](#)
135. Bartel, C. J.; Sutton, C.; Goldsmith, B. R.; et al. New tolerance factor to predict the stability of perovskite oxides and halides. *Sci. Adv.* **2019**, *5*, eaav0693. [DOI](#) [PubMed](#) [PMC](#)
136. Gu, G. H.; Jang, J.; Noh, J.; Walsh, A.; Jung, Y. Perovskite synthesizability using graph neural networks. *npj. Comput. Mater.* **2022**, *8*, 757. [DOI](#)
137. Fu, Y.; Zhu, H.; Chen, J.; Hautzinger, M. P.; Zhu, X.; Jin, S. Metal halide perovskite nanostructures for optoelectronic applications and the study of physical properties. *Nat. Rev. Mater.* **2019**, *4*, 169–88. [DOI](#)
138. Li, J.; Duan, J.; Yang, X.; Duan, Y.; Yang, P.; Tang, Q. Review on recent progress of lead-free halide perovskites in optoelectronic applications. *Nano. Energy.* **2021**, *80*, 105526. [DOI](#)
139. Cai, X.; Li, Y.; Liu, J.; Zhang, H.; Pan, J.; Zhan, Y. Discovery of all-inorganic lead-free perovskites with high photovoltaic performance via ensemble machine learning. *Mater. Horiz.* **2023**, *10*, 5288–97. [DOI](#)
140. Liu, Z.; Rolston, N.; Flick, A. C.; et al. Machine learning with knowledge constraints for process optimization of open-air perovskite solar cell manufacturing. *Joule* **2022**, *6*, 834–49. [DOI](#)
141. Chen, T.; Pang, Z.; He, S.; et al. Machine intelligence-accelerated discovery of all-natural plastic substitutes. *Nat. Nanotechnol.* **2024**, *19*, 782–91. [DOI](#) [PubMed](#) [PMC](#)
142. Mai, H.; Le, T. C.; Chen, D.; Winkler, D. A.; Caruso, R. A. Machine learning for electrocatalyst and photocatalyst design and discovery. *Chem. Rev.* **2022**, *122*, 13478–515. [DOI](#)
143. Osman, A. I.; Nasr, M.; Eltaweil, A. S.; et al. Advances in hydrogen storage materials: harnessing innovative technology, from machine learning to computational chemistry, for energy storage solutions. *Int. J. Hydrogen. Energy.* **2024**, *67*, 1270–94. [DOI](#)
144. Ma, X. Y.; Lewis, J. P.; Yan, Q. B.; Su, G. Accelerated discovery of two-dimensional optoelectronic octahedral oxyhalides via high-throughput ab initio calculations and machine learning. *J. Phys. Chem. Lett.* **2019**, *10*, 6734–40. [DOI](#) [PubMed](#)
145. Jin, H.; Zhang, H.; Li, J.; et al. Discovery of novel two-dimensional photovoltaic materials accelerated by machine learning. *J. Phys. Chem. Lett.* **2020**, *11*, 3075–81. [DOI](#)
146. Wang, Z.; Zhang, H.; Li, J. Accelerated discovery of stable spinels in energy systems via machine learning. *Nano. Energy.* **2021**, *81*, 105665. [DOI](#)
147. Alibagheri, E.; Ranjbar, A.; Khazaei, M.; Kühne, T. D.; Vaez Allaei, S. M. Remarkable optoelectronic characteristics of synthesizable square-octagon haeckelite structures: machine learning materials discovery. *Adv. Funct. Mater.* **2024**, *34*, 2402390. [DOI](#)
148. Li, Y.; Yang, J.; Zhao, R.; et al. Design of organic-inorganic hybrid heterostructured semiconductors via high-throughput materials screening for optoelectronic applications. *J. Am. Chem. Soc.* **2022**, *144*, 16656–66. [DOI](#) [PubMed](#)
149. Chen, J.; Xu, W.; Zhang, R. Δ -Machine learning-driven discovery of double hybrid organic-inorganic perovskites. *J. Mater. Chem. A.* **2022**, *10*, 1402–13. [DOI](#)
150. Chen, A.; Wang, Z.; Gao, J.; et al. A data-driven platform for two-dimensional hybrid lead-halide perovskites. *ACS. Nano.* **2023**, *17*, 13348–57. [DOI](#)
151. Liu, Y.; Madanchi, A.; Anker, A. S.; Simine, L.; Deringer, V. L. The amorphous state as a frontier in computational materials design. *Nat. Rev. Mater.* **2025**, *10*, 228–41. [DOI](#)
152. Merchant, A.; Batzner, S.; Schoenholz, S. S.; Aykol, M.; Cheon, G.; Cubuk, E. D. Scaling deep learning for materials discovery. *Nature* **2023**, *624*, 80–5. [DOI](#) [PubMed](#) [PMC](#)
153. Szymanski, N. J.; Rendy, B.; Fei, Y.; et al. An autonomous laboratory for the accelerated synthesis of novel materials. *Nature* **2023**, *624*, 86–91. [DOI](#) [PubMed](#) [PMC](#)
154. Zeni, C.; Pinsler, R.; Zügner, D.; et al. A generative model for inorganic materials design. *Nature* **2025**, *639*, 624–32. [DOI](#) [PubMed](#) [PMC](#)
155. Wu, J.; Torresi, L.; Hu, M.; et al. Inverse design workflow discovers hole-transport materials tailored for perovskite solar cells. *Science* **2024**, *386*, 1256–64. [DOI](#) [PubMed](#)

156. Lu, J. M.; Wang, H. F.; Guo, Q. H.; et al. Roboticized AI-assisted microfluidic photocatalytic synthesis and screening up to 10,000 reactions per day. *Nat. Commun.* **2024**, *15*, 8826. DOI PubMed PMC
157. Zhang, J.; Hauch, J. A.; Brabec, C. J. Toward self-driven autonomous material and device acceleration platforms (AMADAP) for emerging photovoltaics technologies. *Acc. Chem. Res.* **2024**, *57*, 1434-45. DOI PubMed PMC