

Machine Learning-Accelerated Development of Perovskite Optoelectronics Toward Efficient Energy Harvesting and Conversion

Baian Chen, Rui Chen,* and Bolong Huang*

For next-generation optoelectronic devices with efficient energy harvesting and conversion, designing advanced perovskite materials with exceptional optoelectrical properties is highly critical. However, the conventional trial-and-error approaches usually lead to long research periods, high costs, and low efficiency, which hinder the efficient development of optoelectronic devices for broad applications. The machine learning (ML) technique emerges as a powerful tool for materials designs, which supplies promising solutions to break the current bottlenecks in the developments of perovskite optoelectronics. Herein, the fundamental workflow of ML to interpret the working mechanisms step by step from a general perspective is first demonstrated. Then, the significant contributions of ML in designs and explorations of perovskite optoelectronics regarding novel materials discovery, the underlying mechanisms interpretation, and large-scale information process strategy are illustrated. Based on current research progress, the potential of ML techniques in cross-disciplinary directions to achieve the boost of material designs and optimizations toward perovskite materials is pointed out. In the end, the current advances of ML in perovskite optoelectronics are summarized and the future development directions are shown. This perspective supplies important insights into the developments of perovskite materials for the next generation of efficient and stable optoelectronic devices.


silicon-based solar cells as reported in 2019.^[10] Higher PCEs exceeding 25% for single-junction devices and 29% for tandem perovskite-Si cells have been produced as a result of recent advancements in perovskite-based solar cells.^[11,12] To date, perovskite materials contain a variety of subcategories, such as inorganic oxide perovskites, halide perovskites, and hydride perovskites, which generally exhibit a wide range of activities in the field of optoelectronics. Multiple attempts have been made to screen or design an ideal optoelectronics functional material from the enormous perovskite materials family. In nature, compared with conventional semiconductors, promising perovskite materials exhibit high photoluminescence quantum yields (PLQYs), strong light absorption, tunable emissions, low-cost processing, etc.^[13–16] These advantageous properties of perovskite materials support their great potential in the field of optoelectronics, providing unique opportunities to be competitive candidates for next-generation optoelectronic devices. However, in the research and commercialization of perovskite materials, several experimental challenges have slowed further advancement,

including lead toxicity, lattice instability, limited model data, and unclear underlying mechanisms. For the toxicity of lead in perovskite materials, the synthesis of lead-free double perovskites as an effective solution has been proposed in recent years. Structural optimization against the lattice instability of perovskite materials has also been proven to be feasible. Unfortunately,

1. Introduction

Perovskite materials have attracted intensive attention in recent decades because of their emerging outstanding performances in optoelectronic applications such as lasers, solar cells, and light-emitting diodes (LEDs).^[1–9] Notably, the power conversion efficiency (PCE) of perovskite solar cells is almost equal to that of

B. Chen, B. Huang
Department of Applied Biology and Chemical Technology
The Hong Kong Polytechnic University
Hung Hom, Kowloon, Hong Kong SAR 999077, China
E-mail: bhuang@polyu.edu.hk

 The ORCID identification number(s) for the author(s) of this article can be found under <https://doi.org/10.1002/aesr.202300157>.

© 2023 The Authors. Advanced Energy and Sustainability Research published by Wiley-VCH GmbH. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

DOI: 10.1002/aesr.202300157

B. Chen, R. Chen
Department of Electrical and Electronic Engineering
Southern University of Science and Technology
Shenzhen 518055, China
E-mail: chenr@sustech.edu.cn

B. Huang
Research Centre for Carbon-Strategic Catalysis
The Hong Kong Polytechnic University
Hung Hom, Kowloon, Hong Kong SAR 999077, China

although the understanding of perovskite materials is deepening and the fabrication process is constantly improving, the long development cycle and high expenses still hinder the large-scale commercial uses of perovskite materials in the field of optoelectronics. In detail, the current production process of an effective and efficient optoelectronic device based on perovskite materials should involve rational compositional design, careful synthesis, systemic fabrication, activity characterization, and sufficient stability and aging tests. For each step, the time-consuming, high repetition of data acquisition, and high complexity of data analysis in traditional experimental approaches will obviously prolong the research period to reach the target devices with satisfying performances. To solve these issues, the amalgamation of numeric simulation-guided computational applications and human intuition-guided traditional materials science has become one of the main trends in recent years to break the persisting bottleneck in the development of novel optoelectronic materials.

In fact, various computing technologies have been widely applied in materials sciences such as the density functional theory (DFT) calculations, molecular dynamics (MD) simulations, machine learning (ML), etc. For the DFT calculations, it is a type of computational method used to investigate the electronic structure of many-body systems based on the quantum mechanical model. For the MD simulations, they realize the computational simulation of the physical movements of atoms and molecules based on Newton's equations of motion. Unlike these two computing technologies, ML is a subfield of artificial intelligence and relies on data-driven functions rather than parsed physical laws. In recent years, ML has developed rapidly with an unprecedented ability to enable versatile techniques in a wide range of applications such as computer vision, speech recognition, and weather forecasting.^[17–22] ML belongs to the field of computational science that analyzes and interprets the patterns and structures in data in order to achieve the purpose of learning, reasoning, and decision-making without human interaction. In simple terms, ML enables users to feed large amounts of data to computer algorithms, and then the computer will analyze the data and make data-driven recommendations and decisions solely based on the input data. If the algorithm identifies any corrections, it will integrate the corrected information to improve future decisions. As an interdisciplinary technology, ML combines different domain knowledge of computer science, statistics engineering, materials science, etc. Here, domain knowledge refers to the comprehensive knowledge reserved in a specialized discipline, which is sufficient to deal with complicated scientific problems. With fast computation speed and strong generalization ability, ML technology effectively handles complex problems that are difficult to solve by traditional experimental and computational methods. In particular, the applications of ML in materials science researches have shown explosive growth in recent decades, especially in the synthetic designs of new materials, performance predictions, in-depth characterizations of material microstructures, and improvements of material computational simulation methods. Compared to other computing technologies, ML has the strong ability to classify and predict patterns within a dataset and discern unforeseen trends that are otherwise impossible for a human observer to identify. Depending on this advantage, ML shows significant potential in perovskite optoelectronics for accelerating the discovery of advanced material

candidates, decoding the underlying mechanisms of observed phenomena, and high-throughput processing of material information.^[23–28] The current challenging area of ML technology used in the field of materials science is how to acquire a significant number of experimental or theoretical data and develop a matching effective dataset. To overcome the shortcomings of ML techniques in data mining, considerable efforts have been made in recent years to combine the different advantages of multiple computing technologies.^[29–31] For example, Yang et al. presented a deep learning approach to investigate the influencing degree of ionic defects on the stability of CsPbI₃ ternary systems based on 7,730 DFT-calculated structures and large-scale MD simulations.^[32] Although the thoughtful implementation of ML in materials science is still at its preliminary stage, it is necessary for researchers to maintain a sustained concern in this promising field to mitigate current experimental impediments and achieve breakthroughs in current materials science.

In this perspective, we have highlighted and summarized the recent research progresses of applying ML techniques as a powerful tool to facilitate the development of advanced perovskite optoelectronics, which starts from a concise overview of the important ML-related concepts. In particular, we demonstrate the pivotal role of ML in boosting scientific innovation of perovskite optoelectronics from three main aspects including discovery of advanced materials, interpretation of underlying mechanisms, and high-throughput processing of large-scale information. Through the review of these recent works, we shed the light on the great potential of ML techniques for designing novel perovskite materials to benefit the fast development of advanced optoelectronics for energy harvesting and storage. In the end, we have also pointed out the remaining challenges, proposed the possible multidisciplinaries, and supplied insightful perspectives for further accelerating the performances of perovskite optoelectronics through ML techniques. All the ML-related concepts and abbreviations in this work are supplied in **Table 1** and **2**.

2. Fundamental Understanding of ML and Its Workflow

As an essential branch of computer science, ML generates specific models based on existing databases and algorithms to complete target tasks, which is an effective approach to achieve the artificial intelligence in solving practical problems. For applications in materials science, ML algorithms are frequently divided into two broad categories: unsupervised and supervised learnings.^[33] The main difference between these two categories of algorithms is feature engineering, which represents the converting process of raw data into features through statistical or ML approaches.^[34] Here, we emphasize the features as the measurable characteristics of a specific object such as a dataset or an image. Extracting informative and discriminating features is the key to perform effective feature engineering on the target objects. Unsupervised learning algorithms explore patterns from unlabeled data, while supervised learning algorithms only treat available datasets with labeled features. Such characteristics make these two kinds of algorithms suitable for different task scenarios. For the supervised learning, the algorithm builds

Table 1. The brief explanations for ML-related concepts in this perspective.

Concepts	Brief Explanations
Domain knowledge	The comprehensive knowledge reserved in a specialized discipline to sufficiently deal with complicated scientific problems.
Unsupervised learning	Explore patterns from unlabeled data to discover the inherent structure of unlabeled data.
Supervised learning	Treat available datasets with labeled features, which require human interventions in advance to label the data appropriately.
Features	The measurable characteristics of a specific dataset.
Feature engineering	The converting process of raw data into features through statistical or ML approaches.
Validation set	Reflect the true performance of the targeted trained ML model. It generally contains at least 10% contents of the original dataset, which is distributed across the entire range of the values.
Correlation coefficient	An effective measure to identify the association degree between the target properties and labeled variables.
Transfer learning	Improve the new predictions of ML based on the knowledge gained from previous learning processes.
Image recognition	Classify the category of image content through the statistics of pixel distributions, colors, textures, and other characteristics in images.
Monitoring	Evaluate the performance of ML models during training and real-time deployment.

Table 2. The definitions for ML-related abbreviations in this perspective.

Abbreviations	Definitions
NN	Neural Networks
ANN	Artificial Neural Network
CNN	Convolution Neural Network
DNN	Deep Neural Networks
BP-ANN	Back Propagation Artificial Neural Network
SVC	Support Vector Classification
SVR	Support Vector Regression
MAML	Model-Agnostic Meta-Learning
PLATIPUS	Probabilistic LATent model for Incorporating Priors and Uncertainty in few-Shot
SNOBFIT	Stable Noisy Optimization by Branch and FIT
RL	Reinforcement Learning
MAOSIC	Materials Acceleration Operating System In Cloud
PLS	Partial Least Squares
Magpie	Materials Agnostic Platform for Informatics and Exploration
XGBR	eXtreme Gradient Boosting Regression
SVMs	Support Vector Machines
MAE	Mean Absolute Error
RMSE	Root Mean Square Error

the ML model by iteratively generating predictions on the data and adjusting for the correct responses. While supervised learning models tend to be more accurate than unsupervised learning models, human interventions in advance are required to label the data appropriately. Therefore, supervised learning is able to process accurate data mining tasks, where the representative examples are classification and regression with limited sample data. In contrast, unsupervised learning models work individually to discover the inherent structure of unlabeled data, which are more applicable for tasks based on large sample data volume, including clustering, association, and dimensionality reduction tasks. However, the unsupervised learning models still have some drawbacks, where some unsupervised learning models still require human intervention to validate output variables. It is worth noting that although supervised and unsupervised learning is suitable for dealing with different types of tasks, the basic flow of application processes in materials science is broadly similar. Here, as shown in **Figure 1**, we have summarized a basic workflow for performing ML models in materials science to suggest the target issues solutions: 1) Identify the exact object to be addressed and determine the type of training data. At the start of training ML models, researchers should first make sure of the goals of the project and decide what type of data to collect. For example, the predictions of one property of a given perovskite optoelectronic device needs to collect the corresponding numerical results of this device. The automatic recognition of the morphology features of a perovskite optoelectronic device needs the corresponding morphology pictures as the training data. The identified object should be definable, quantifiable, and based on real physical principles; 2) Gather sufficient data to form a dataset for model training. As a data-driven technology, ML models possess powerful prediction and classification capabilities that heavily depend on the quantities as well as the qualities of the training data. The “sufficiency” of data is a relative concept that is determined by the accuracy of the final ML model. Common sources of available material data include experimental records, published literature, and online databases. In general, the experimental data are the most reliable sources to be used in training ML models due to the real environmental conditions. However, strict experiments make it difficult to efficiently produce large-scale useable data, which hinders the dataset construction for ML model training. In contrast, although DFT calculations have the advantages of automated data mining, these computational results are more diverse due to the model buildings, calculation settings, and simulation environments. To solve specific ML issues, researchers are suggested to accumulate the data based on the appropriate strategies in advance, where all data sources should be consistent to avoid systematic errors. For the limited datasets with strong correlations, the interpolation method can be applied to make an effective extension; 3) Determine the feature representations and the corresponding algorithm. Typically, the accuracy of the learned ML model strongly depends on its feature engineering process. The feature vector expanded by the input objectives should have enough information to describe the characteristics of training data for contributing to the accurate output. However, the number of features should not be too large because when the dimensionality of the feature vector increases, the volume of the space increases massively, leading to the sparseness of available data. The reasonable settings of

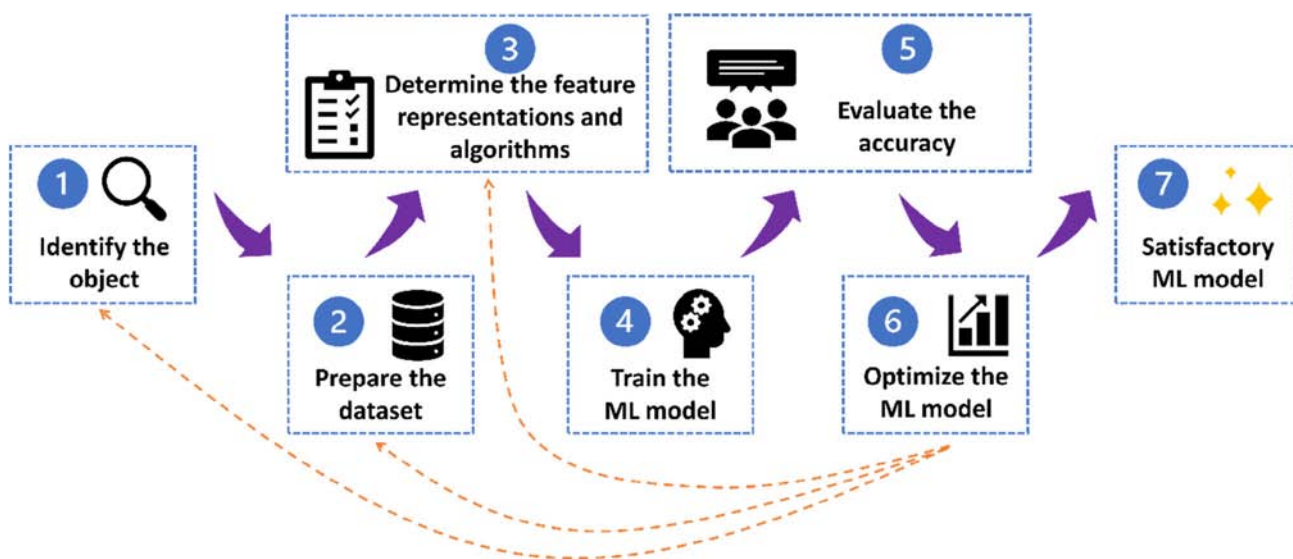


Figure 1. Schematic representations of basic workflow for performing ML models in materials science to solve target issues. Orange dashed lines indicate the parts that may need to be reconsidered during the ML model optimization process.

feature representations generally require an in-depth understanding of the target physical model and strict domain knowledge. The algorithms serve as the kernel of the ML models and are responsible for providing specific mathematical approaches to handle the target task. Multiple algorithms are selected for different projects, such as decision trees, support vector machines (SVMs), neural networks (NNs), etc.^[35–37] 4) Run the learning algorithm on the gathered training dataset to achieve an effective ML model. For each training process, all the data in the training dataset will be used to form an objective ML model. Although more training data seems to be more conducive to generate models, a high proportion of training data also easily leads to an overfitting result. It is necessary to retain a part of the original dataset as a validation set because an appropriate proportion of the validation set reflects the true performance of this targeted trained ML model. The meaningful validation set generally requires at least 10% contents of the original dataset and the validation data values need to be distributed across the entire range of the values. For the datasets with limited data amount, the proportion of the validation set should be correspondingly higher to avoid the inaccuracies caused by insufficient data during the validation process. Hyperparameters should be adjusted before the beginning of the training process to achieve the best predictive performance of the ML model; 5) Test and evaluate the accuracy of the trained ML model. In this step, the accuracy of the trained ML model is reflected by its performance on the validation set. Different algorithms usually have some unique evaluation criteria parameters. For example, common evaluation parameters for a linear regression ML model include mean absolute error (MAE), root mean square error (RMSE), coefficient of determination (R^2 score), and so on. Notably, due to the lack of sufficient evaluation, the conclusions drawn from a ML model appear to match the data can be flukes, leading researchers to misinterpret the actual ability of this ML model. Therefore, to obtain a reliable ML model, appropriate statistical model validation techniques

(e.g., cross-validation) should be used to test whether the corresponding ML model is still applicable to the data array; and 6) Optimize the ML model to satisfy the required accuracy for solving the target questions. For the trained ML models with poor performance, it is necessary to find out the possible intrinsic reasons for the results based on the data characteristics and then make the corresponding optimizations. To optimize the ML model, the first three steps should be reconsidered, as shown by the orange dashed lines in Figure 1. Unrealistic goals, insufficient data, and inappropriate feature representations as well as algorithms all lead to poor performance of ML models. The hyperparameters need to be redefined when the general settings change before the training process. To avoid invalid testing, the data used for testing should not overlap with the training set and validation set. Notably, although the general flow of materials science problems with ML techniques is similar, the exact ML model training and generation involved in various situations are still very different. Also, the impact of different steps on various problems is specific. The feature engineering step has a significant effect on those research objects possessing a strong intrinsic association with the descriptors. Meanwhile, for research objects that have a weak association-feature outcome, the influences of the feature engineering step are limited. Therefore, during the training process of ML applications, professional domain knowledge is highly required to guarantee efficient and accurate predictions.

3. Applying ML Techniques for Perovskite Developments

3.1. Accelerating the Discovery of New Advanced Perovskite Materials

As mentioned above, the expensive endeavor of traditional trial-and-error methods to explore new materials limits the fast

growth of perovskite optoelectronics. In general, designing an effective routine to predict whether an unknown compound meets the target conditions requires strict domain knowledge. This principle applies to both experimental and computational exploration of new materials. The first challenge is the determination of the target criteria, and with the emergence of new materials and the improvement of domain knowledge, mathematical criteria tend to become more complex and refined. Although contributing to the accurate identification of unknown materials, these developed criteria inevitably increase the amount of computation, which slows the verification process of target materials and hinders the discovery of new advanced materials.

The utilization of the strong predictive ability of ML techniques has proven to be efficient assistance in the discovery of new advanced materials. To explore more potential perovskite materials with formula ABX_3 , a mathematical criterion of “perovskites” should be proposed first before the analysis of all candidate materials, because not all ABX_3 stoichiometry are perovskite structures. In early studies, the Goldschmidt tolerance factor t usually plays the role in evaluating the formability and stability of perovskites ABX_3 .^[38] At present, more developed factors, such as the formation energy and the energy beyond the convex hull (E_{hull}), are used to evaluate the perovskite candidates more accurately.^[39] To avoid redundant computation and accelerate the evaluations of the thermodynamic stability for perovskite materials, Schmidt et al. developed a precise ML model for the predictions of E_{hull} of perovskite materials after training 20 000 samples.^[40] This work first created a dataset consisting of DFT calculations of about 250 000 cubic perovskite materials, which includes all potential perovskite and antiperovskite crystals produced with elements from hydrogen to bismuth, excluding rare gases and lanthanides. **Figure 2a** displays the histogram of the distribution of E_{hull} for all $\approx 250\,000$ cubic perovskite materials. Other 230 000 possible ABX_3 compounds were tested and finally, 641 structures were considered as thermodynamically stable candidates with low predicted $E_{\text{hull}} \leq 5 \text{ meV atom}^{-1}$. Liu et al. also performed ML techniques on the DFT database of 397 ABO_3 compounds and realized the classification of 891 possible structures based on the predicted E_{hull} values.^[41] Before the training process, this study first conducted effective feature engineering to develop the performance of linear ML models. The initial 9 features have been expanded to a maximum of 55 compound features and finally 25 features to obtain the most relevant results. **Figure 2b** visualizes the feature importance, which represents the relative significance of each feature in a dataset for a predictive model. It is found that the tolerance factor obviously dominates the classification of perovskites and nonperovskites. This work has successfully screened out 37 candidate perovskite materials with stable thermodynamics ($0 \text{ meV atom}^{-1} < E_{\text{hull}} < 36 \text{ meV atom}^{-1}$) for further synthesis and applications.

In addition to these representative works, many other studies have also applied ML techniques to the predictions of bandgap values (E_g) of perovskite materials, because E_g is a key index reflecting the photoelectric conversion ability for optoelectronic devices. Based on the support vector classification (SVC) and support vector regression (SVR) methods, Yang et al. proposed a two-step ML strategy to realize the rapid discovery of narrow-bandgap

oxide double perovskites.^[42] As shown in **Figure 2c**, the trained SVC classifiers are used to narrow down the enormous DFT datasets of samples, and SVR regression is applied to predict the E_g of candidate materials. As a result, 60 promising double perovskites for photovoltaic applications are screened out successfully from 6,529 samples, in which 19 structures show considerable photovoltaic potentials with excellent performance, which display a range of E_g from 1.25 to 1.45 eV.

ML-guided autonomous experimentation is another trend to enable the efficient exploration of advanced perovskite materials, in which the algorithms will automatically iterate new experimental contents based on the prior experimental databases with little human intervention.^[43–45] In one example of the ML-driven autonomous experimentations, Shekar et al. determined the optimal strategies to predict the growth of metal halide perovskite crystals by applying the model-agnostic meta-learning (MAML) model and the Probabilistic LATent model for Incorporating Priors and Uncertainty in few-Shot learning (PLATIPUS) approach.^[46] Incorporating the 1,870 historical reactions conducted with 19 different perovskite systems, the trained MAML models realized the predictions of reaction compositions in perovskite crystals and became more explainable for new chemical systems. In the evaluation stage, three active learning algorithms were tested by the prepared 20 experiments, and the PLATIPUS algorithm showed the best predictive capabilities for the new chemical systems. Li et al. also reported a solid approach to discover the optically active CsPbBr_3 nanocrystals through autonomous intelligent systems.^[47] They used the stable noisy optimization by branch and fit (SNOBFIT)-based reinforcement learning (RL) algorithm to optimize the corresponding parameter space consisting of the reaction temperature and the precursor concentrations. Supported by synthetic materials acceleration operating system in the cloud (MAOSIC) platform, their cloud experiments successfully detected the desired circular dichroism (CD) signal of the target nanocrystal from 250 iterations. They identified that the chiral origin of the synthesized nanoplates comes from their unique structures and screw dislocations. Another representative example is the work conducted by Abdel-Latif et al., in which an ML-guided autonomous approach for the synthesis of the lead halide perovskite (LHP) quantum dots (QDs) was introduced.^[48] In detail, the frequent operations between the different but connected modules realized the autonomous acquisition of large-scale spectroscopic data in real time. Using a pretraining stage with 200 experiments, the trained NN model is able to adjust the peak emission energies in further exploitation experiments. As a result, they combined the active learning strategies and modular QDs synthesizer to realize the autonomous and efficient exploration of desired LHP QDs at the objective peak emission energy with a minimized full-width-at-half-maximum (FWHM). With the advances in autonomous experimentation strategies, these studies demonstrate the ability of ML techniques to enhance the exploration efficiency among material researchers for the new perovskites.

The guidance provided by a precise ML model on the design of perovskite optoelectronics is reliable, which is often validated by actual experiments. Exemplified work includes a report from Bak et al. have fabricated a more efficient Sn-based perovskite solar cell following the ML suggestions than that of those

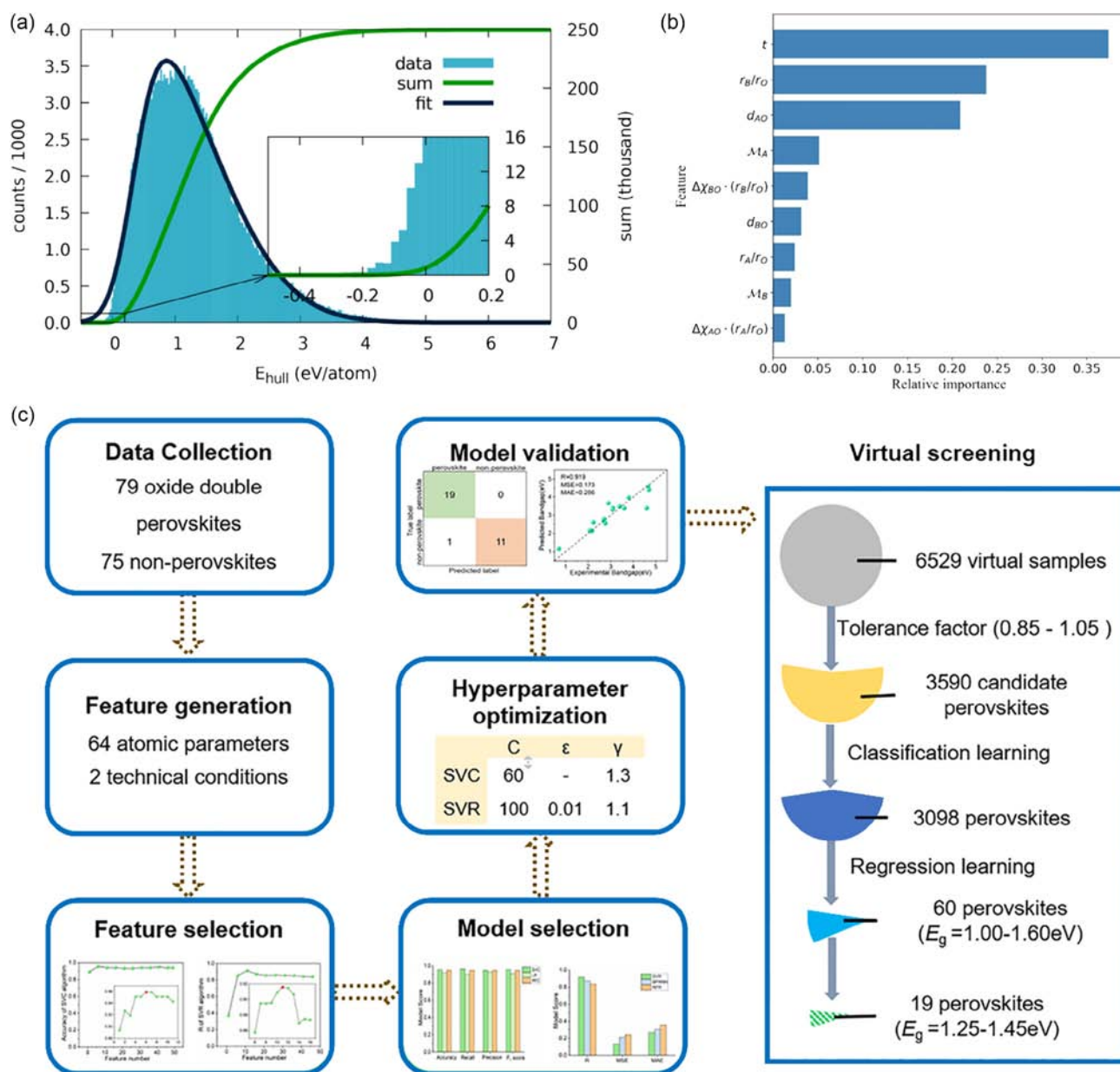


Figure 2. a) The histogram of the distribution for about 25 000 cubic perovskite structures. Reproduced with permission.^[40] Copyright 2017, American Chemical Society. b) The diagram represents feature importance. The rankings refer to the gradient boosting decision tree model. Reproduced with permission.^[41] Copyright 2020, Elsevier B.V. c) Brief workflow for material discovery based on CNN model. Reproduced with permission.^[42] Copyright 2021, Elsevier B.V.

perovskite materials prepared based on empirical experiences or trials and errors^[49] (Figure 3). Starting with the construction of deep neural networks (DNN), this work has successfully trained an accurate ML model with a limited amount of experimental data (Figure 3a–c). Recommendations given by the ML algorithm further facilitate the optimizations on designing perovskite-based solar cells. Accordingly, the PCE of the fabricated device based on ML suggestions reaches 5.57%, which is three times higher than the average PCE of unguided devices at 1.72% (Figure 3g). Above studies display the active assisting role of ML techniques in the accelerating discovery of new materials. In fact, as the most

extensive application scenarios of ML techniques in the current trend of materials science, this function of ML models has been widely reported to be effective in actual experiments. However, we notice that ML models used in perovskite optoelectronics studies are limited in a small range, where the training is accomplished based on some common parameters such as bandgap E_g , E_{hull} , and tolerance factor t , etc. The frequently repeated training of similar datasets is not conducive to the further development of ML techniques. For optoelectronics, optical properties, dielectric constants, and phonon frequencies are also available parameters, which are worthy of being learnt by ML models to explore specific

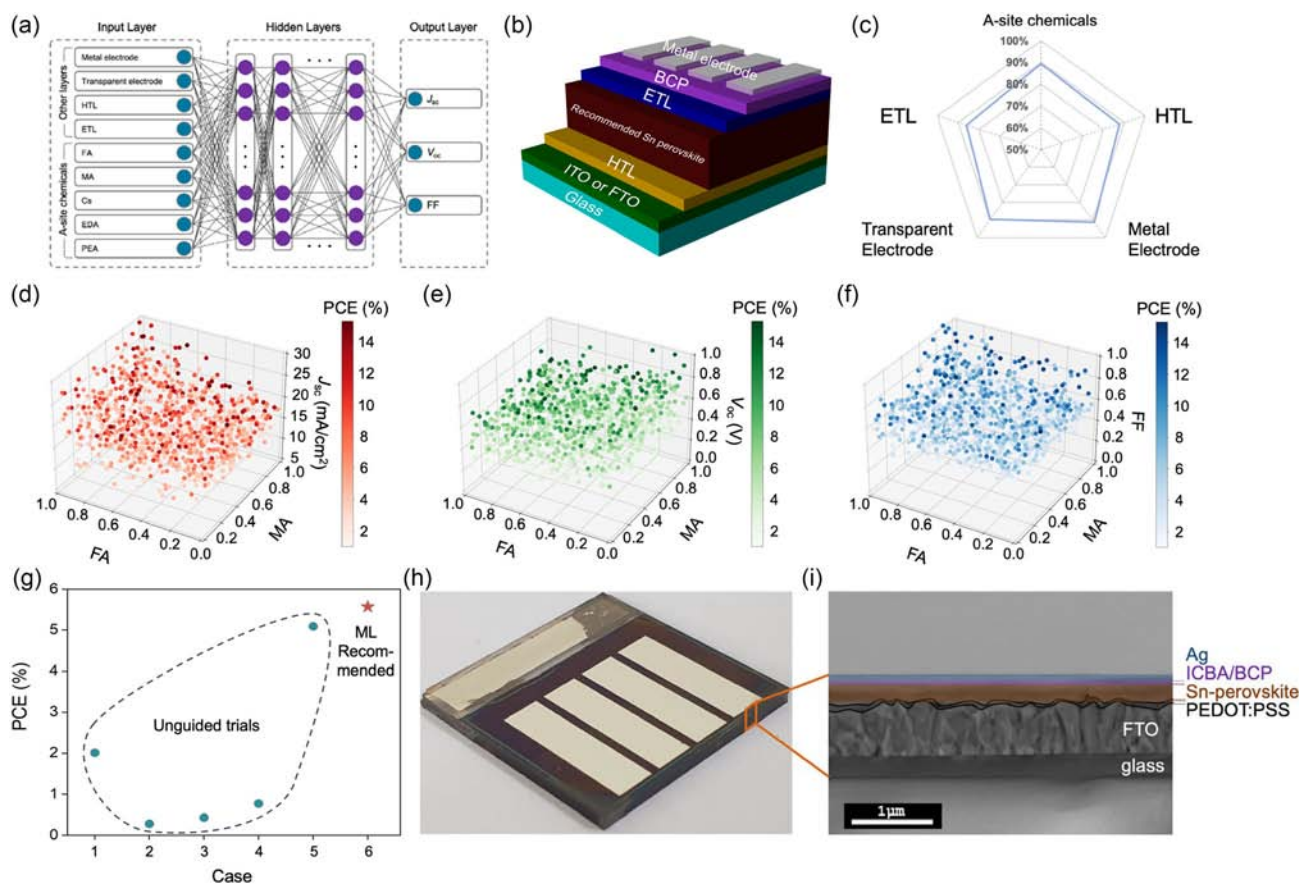


Figure 3. A brief demonstration of accelerated design of high-efficiency Sn-based perovskite solar cell via ML. a,b) Schematic of the a) trained DNN ML model and b) the suggested structure of the Sn-based perovskite solar cell. c) Radar graph representing the prediction accuracy of the DNN ML model for different parameters. d–f) Scatterplot of the recommendation results for each parameter d) Short-circuit current density (J_{sc}), e) open-circuit voltage (V_{oc}), and f) fill factor (FF). g) Comparison of PCE between the ML-guided and unguided trials. h) Photograph and i) cross-sectional SEM image of the fabricated perovskite-based solar cell based on the guidance of the ML model. Reproduced with permission.^[49] Copyright 2022, Springer Nature.

problems (e.g., the light–medium interactions). Until now, applying ML with these novel parameters in the designs of perovskite materials still needs further research efforts in the future. With these information, the long-term development of ML techniques in helping to discover new advanced perovskite optoelectronics is expected.

3.2. Improving the Interpretation of Underlying Mechanisms

During the investigations of perovskite materials, the underlying mechanism for the structure–property relationship enables an in-depth understanding of perovskite materials designs and optimizations. However, for some specific optoelectronic properties, the intrinsic mechanism for the performance is still unclear, which hinders the rapid progress of perovskite materials developments. Although many different variables and parameters are involved in ML, the correlation coefficients are an effective measure to identify the association degree between the target properties and labeled variables, which contains the essential information of the structure–property correlations. For instance, in the process of studying the reactivity of different amines on the organic–inorganic hybrid perovskite films, Yu et al. achieved

the exploration of hidden trends from the experimental results of correlation coefficients after extracting the feature importance based on ML algorithms.^[50] This work has pointed out the suitable types of amines that tend to have high compatibility with perovskite films, including amines with fewer hydrogen bond donors and acceptors, secondary and tertiary amines, and pyridine derivatives. Moreover, to understand the relationship between the electrical conductivity and the atomic properties in perovskite oxides, Liu et al. studied the correlation between the atomic parameters and the ionic conductivity of 117 perovskite oxide samples based on three ML algorithms: partial least squares (PLS), back propagation artificial neural network (BP-ANN), and SVR.^[51] Combining the experimental data and DFT calculation results, this research has found that the ratio of O–O charge population to the O–O band length, which was noted as P/L, has a quadratic curving relationship with the logarithm of oxide ion conductivity in some undoped perovskite-type oxides. Park et al. also addressed the impediment of lattice deformation for the accurate predictions of perovskite bandgap by investigating the correlation between the structural deformation and bandgap values.^[52] As shown in **Figure 4a–e**, this study analyzed the influence of octahedral structural deformation on

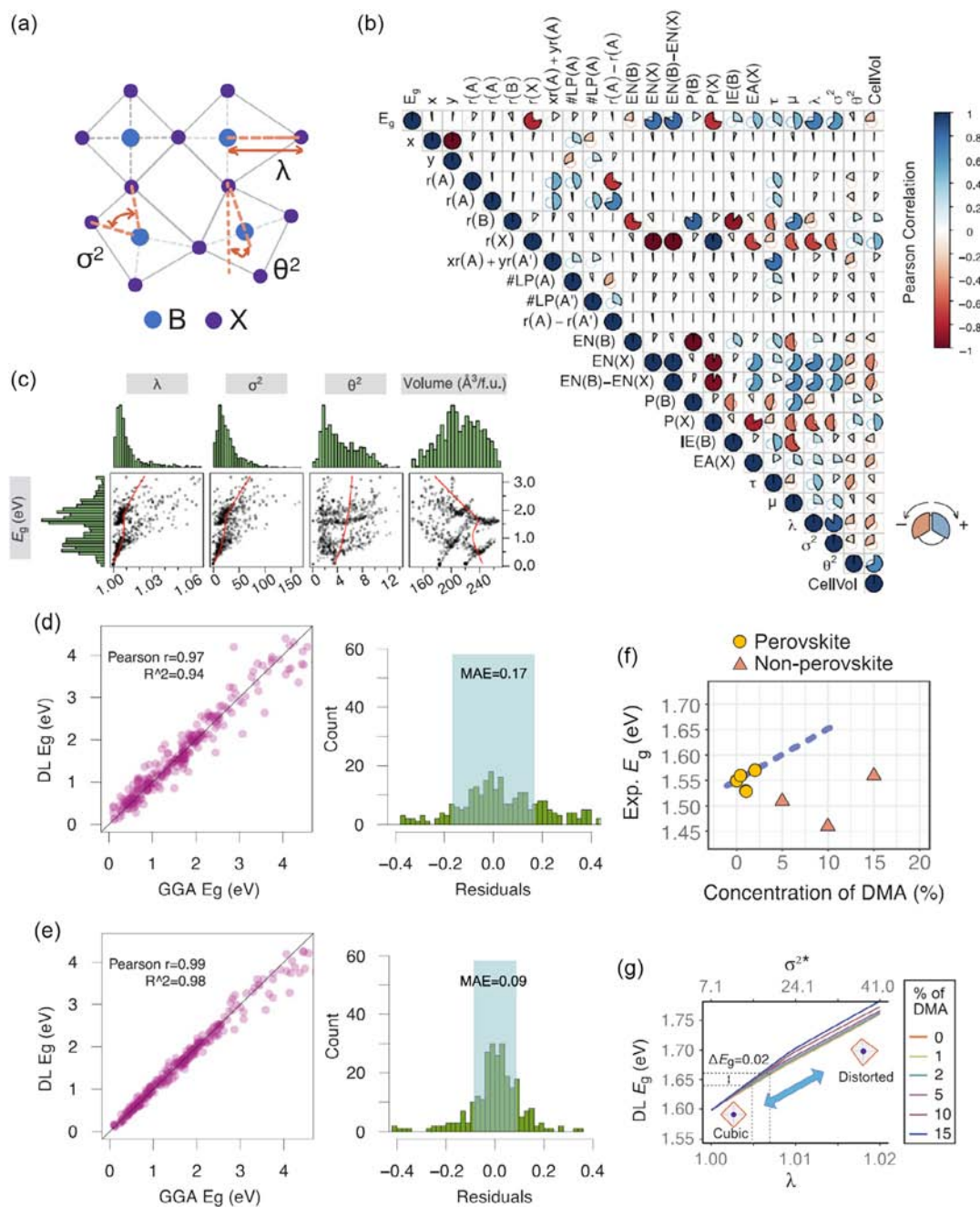


Figure 4. ML-guided correlation investigations between structural deformation and bandgap predictions in hybrid perovskites. a) Representation set for the octahedral deformation measurements of perovskites: λ , quadratic elongation of the octahedra; σ^2 , angle variance of the octahedra; and θ^2 , octahedral rotation. b) Calculated correlation coefficients between different descriptors. c) Overall data distribution and fitted correlation between bandgap values and structural deformation levels. d–e) Comparison of the predictive ability of the deep learning model trained e) with and d) without structural parameters. f) Experimentally observed and g) deep learning-calculated bandgaps of the mixed perovskites with the change of DMA concentration. Reproduced with permission.^[52] Copyright 2020, Elsevier B.V.

the bandgap predictions of hybrid perovskites and optimized an accurate predictor accordingly. The 0.02 eV bandgap difference observed in experimental samples is compatible with the change in λ predicted by the ML model (Figure 4f–g), which indicates that the bandgap increases upon dimethylammonium (DMA) doping is due to the local octahedra distortions

rather than volumetric effects. These works indicate that the visualized importance degrees of features not only guide feature engineering for ML model optimization but also supply innovative insights into the interpretation of the underlying mechanisms among variables. Although current ML techniques are still insufficient for the disclosure of associations between

variables, it still provides an opportunity for ML techniques to be widely applied in revealing photophysical processes in optoelectronic materials. More attempts and efforts are needed to further expand the compatibility of ML techniques in this field.

3.3. Facilitating the High-Throughput processing of Large-Scale information

Currently, the perovskite optoelectronic database tends to face the problem of large-scale information processing. This phenomenon is mainly attributed to two reasons: the variety of perovskite categories and the high-dimension space of perovskite parameters. For the first issue, in brief, both organic and inorganic perovskite materials have very diverse combinations of anions and ions, and the combination possibilities will be further increased if hybrid perovskites are considered. Accordingly, the complex element composition and diverse structures span the high-dimensional parameter space that defines perovskite materials. In addition, automated experiments and diverse databases enable rapid data acquisition in materials science. These reasons cause the screening of desired materials to be in a high-throughput fashion in the field of perovskite optoelectronics. However, existing routines for processing large-scale information require a considerable cost. The development of emerging ML techniques provides opportunities for the high-throughput processing of large-scale information that human researchers are otherwise unable to deal with.

Based on the transfer learning method and a hybrid descriptor set, Li et al. established a convolution neural network (CNN) ML model and realized the high-throughput screening of stable perovskite materials (Figure 5a).^[53] The unannotated perovskite materials obtained from DFT calculations datasets show high-precision structural information, which have been labeled first and used for the CNN model. The Materials Agnostic Platform for Informatics and Exploration (Magpie) descriptors have been applied in the training to get a generic screening model without requirements of structural information. The Magpie has included a large number of attributes to capture the physical/chemical properties of materials with any number of constituent elements. Transfer learning is able to improve the new predictions based on the knowledge gained from previous learning processes. Evidence proves that the trained CNN model shows the best performances in formation energy predictions of perovskites based on only composition information when compared to the ElemNet model and several other ML models. Finally, this work has successfully selected 625 potential candidates with low tolerance factors from the 21 316 perovskite samples, 98 of which were proved to be stable by DFT calculations. Similarly, Lu et al. developed a target-driven method combining ML techniques and DFT calculations to predict undiscovered hybrid organic–inorganic perovskites (HOIPs) for photovoltaics. During the feature engineering process, 30 initial features were analyzed, and the most 14 important ones are sorted out as an optimal set. Taking 212 samples as the training set, two HOIPs are screened out as potential photovoltaic materials with proper bandgaps and robust environmental stabilities from the preliminary 5158 unexplored samples (Figure 5b–d).^[54] Meanwhile, an

essential mapping of the close structure–property relationship in bandgaps of HOIPs was established.

In addition, to screen target materials from large-scale candidates, another utilization direction of the powerful high-throughput capabilities in ML techniques is image recognition in perovskite materials. For image recognition, its main task is to correctly classify the category of image content through the statistics of pixel distributions, colors, textures, and other characteristics in the image. In deep learning, the image recognition model not only performs its own task but also acts as a parameter extraction network for other tasks in computer vision.

Undoubtedly, the transplantation of image recognition concepts for research of perovskite optoelectronics will stimulate a lot of meaningful work to enable the optimizations of both structures and properties. To characterize the results of synthesized perovskite single crystals, Kirman et al. developed an automated experimental characterization system based on image recognition with the aid of CNNs.^[55] Using an optimized ML model to guide the sequence of ever-improved robotic synthetic trials, this work is able to perform high-throughput syntheses of perovskite single crystals with a protein crystallization robot. In addition, this work also characterizes the outcomes with the help of CNN-based image recognition. Following the predicted optimal conditions for the synthesis of a new perovskite single crystal by the trained ML model, the first synthesis of (3-PLA)₂PbCl₄ has been achieved. Taherimakhsoosi and colleagues also realized the identification and quantification of a variety of defects in thin films by applying ML.^[56] To illustrate the applicability of the proposed method in thin-film optimization, a specific CNN model trained by experimental dark-field images has been applied to resolve a 2D film morphology response surface in a set of experiments where both film composition and processing were varied. This approach automatically analyzes film morphologies in optical images and applies to multiple imaging conditions. The high-throughput ability of ML is also used in time-series forecasting as presented by Howard et al., who generated a 4 h time-series predictions of humidity-dependent photoluminescence (PL) intensity of perovskite thin film with <18% normalized RMSE (Figure 5e–g).^[57] These previous studies demonstrate the feasibility of applying ML solutions to highly complex materials science problems. Future ML techniques based on strong computing powers and efficient algorithms are expected to enable fully automated material screening, characterization of the synthesis process, and stability testing, which may induce a new revolution in materials engineering.

4. Future Opportunities and Challenges

4.1. Synergistic effect between ML and DFT calculations

Nowadays, DFT calculations have played an important role in the exploration of various functional optoelectronic materials including perovskites. Based on solid theory and reproducible settings, DFT calculations supply sufficient information about the object materials including the structural characteristics, electronic structures, optical properties, vibrational properties, etc. These valuable results contain the derivation of intrinsic nature and are able to offer evident recommendations as the extension of

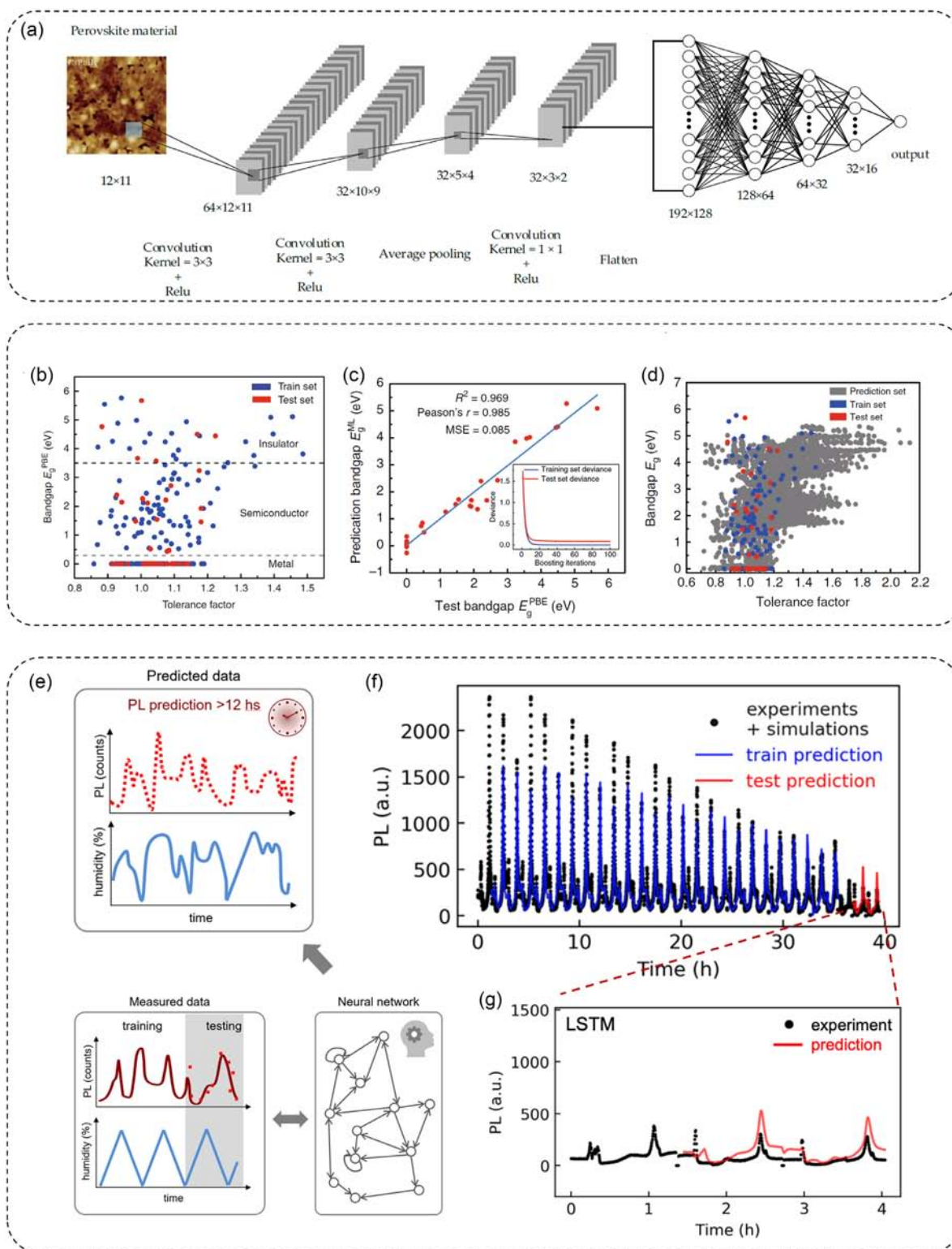


Figure 5. Examples of high-throughput processing via ML models in the field of perovskite optoelectronics. a) Schematic of CNN construction for the formation energy predictions of perovskites.^[53] Reproduced with permission.^[53] Copyright 2019, under the terms of CC-BY license, The Authors, published by MDPI. b–d) High-throughput screening of stable photovoltaic perovskites.^[54] b) Data visualization of training and test set. c) Predictive capability of the trained ML model and d) the predicted bandgaps against tolerance factors of all candidate perovskites. Reproduced with permission.^[54] Copyright 2018, under the terms of CC-BY license, The Authors, published by Springer Nature. e–g) Performance of ML models in time-series forecasting of perovskites.^[57] e) A brief demonstration of the ML model construction for time-series predictions. f–g) Performance of long short-term memory (LSTM) network on the test portion of the simulated degradation task. Reproduced with permission.^[57] Copyright 2022, American Chemical Society.

initial experiments. For the other aspect, as a data-driven technology, ML greatly improves the efficiency of data use. Given sufficient training data, ML is a powerful tool to achieve significant enhancements in the analyses, simulations, and predictions of optoelectronic properties for perovskite materials.^[58] Through the collection of experimental data in the previous literature, representative datasets can be effectively constructed to support the ML models. However, the generation of experimental data usually requires a long period, which induces a mismatch with the fast data processing ability of ML models. This has hindered the potential of ML techniques in boosting the research of perovskite optoelectronics. In contrast, DFT calculations significantly accelerate the data acquisition of optoelectronic properties for perovskite materials and make it possible to establish a broad-range dataset efficiently. The applications of ML models trained on DFT computational data become new and competitive solutions to optoelectronic problems in perovskite materials. Currently, the synergies between ML techniques and DFT calculations have been widely applied in the synthesis, screenings, and predictions of perovskite optoelectronics.^[59–61] The recent report from Gao et al. has proposed a novel search strategy combining ML and DFT calculations to screen 5,796 inorganic double perovskites.^[60] They carefully compared different ML models trained on various algorithms to get the best predictive power. As shown in **Figure 6a**, the eXtreme gradient boosting regression (XGBR) algorithm yielded the best accuracy in the predictions of bandgap when compared to the ANN and SVR algorithms. Two novel lead-free inorganic double perovskites, $\text{Na}_2\text{MgMnI}_6$ and K_2NaInI_6 , were finally obtained, where the predicted bandgap values and thermal stability are also confirmed by DFT calculations. We believe that in-depth combinations between ML techniques and DFT calculations will create seamless pipeline strategies for designing more advanced optoelectronic devices in the near future.

4.2. Identification of the Structure–Property correlation

The external optoelectronic behaviors of perovskite materials are greatly affected by the lattice structures and morphologies because the structural properties will determine the inner light–electron interactions. Obviously, the elucidation of a detailed correlation between structural and optoelectronic properties will benefit the design of structural characteristics to obtain better optoelectronic performances. Currently, several studies have proven the capability of ML algorithms in characterizing subtle structural information and constructing possible correlations through image recognition.^[62–64] Malkiel et al. empirically showed that a novel DNN trained with tens of thousands of simulation trials from COMSOL simulations is capable of solving the inverse problem as well as retrieving subwavelength dimensions only based on the far-field observations.^[62] As shown in **Figure 6b**, the trained DNN model based on synthetic experiments has realized the automated design of a gold plasmonic structure targeted to the dichloromethane with specific spectral polarization responses. The configurations and dimensions of the plasmonic structure are also found in this ML model. This work supplies an effective method to obtain significant information regarding optical elements with metasurfaces and

optimal nanostructures for designing the targeted chemicals and biomolecules, enabling broad applications in different fields. However, the robustness of these proposed methods still needs more evaluations and verifications in different fields. On the other hand, future optoelectronic materials may possess more complicated structures and performances, which require the facilitation of ML to enable accurate characterizations. Therefore, the high-precision, high-resolution, and high-throughput processing of ML technology is highly necessary for developing promising perovskite materials in future research.

4.3. Construction of Benchmark Database

Effective databases are vital fundamentals for training ML models to solve corresponding problems. Current existing large databases, such as Materials Project, contain integrated material information and can be utilized in diverse ML scenarios.^[65] However, there are still two aspects of the database construction that deserve the concerns of material researchers before the ML model training. The first issue is the general quality of the collected material data. Despite possessing impressive capabilities of data analysis, ML models trained on low-quality data always exhibit unsatisfied accuracy and are incapable of solving actual questions. Here, low-quality data refers to data with large errors and weak correlation with the target properties. The other aspect is the data specificity regarding the specific objectives. Although extracting useful data from integrated databases is necessary for solving specific questions, there are no strict criteria for data extraction, and material researchers can only rely on their self-judgments based on domain knowledge. Therefore, the construction of benchmark databases with an in-depth understanding and comprehensive domain knowledge for a small scope of materials will set an example for the data repository and contribute to the unifying guidelines of ML techniques in specific materials field. In addition, material researchers are encouraged to share their available databases in an open and reproducible manner for the further reuse of digital resources. Continuous collaborative actions are still required for the construction of benchmark databases with findable, accessible, interoperable, and reusable (FAIR) material data, which will enable the boosting of the future materials community with no doubt.^[66,67]

4.4. Establishment of Explainable Models for ML

The performances of the perovskite optoelectronic devices largely depend on the intrinsic properties of the perovskite materials, which is also the core part of explorations. However, the mechanisms of most current emerging ML models are unclear, especially deep learning models, which exhibit “black box” properties for the predictions of target materials.^[68] Such phenomena originated from the data-driven nature of the ML algorithms, which makes it more difficult for researchers to understand and explain the in-depth mechanisms or correlations behind the outputs proposed by the ML models. In this scenario, the efforts for investigating explainable ML models are valuable since it allows the researchers to inspect the underlying mechanisms of ML to achieve accurate predictions, which are particularly important for rational design and optimizations of perovskite materials.

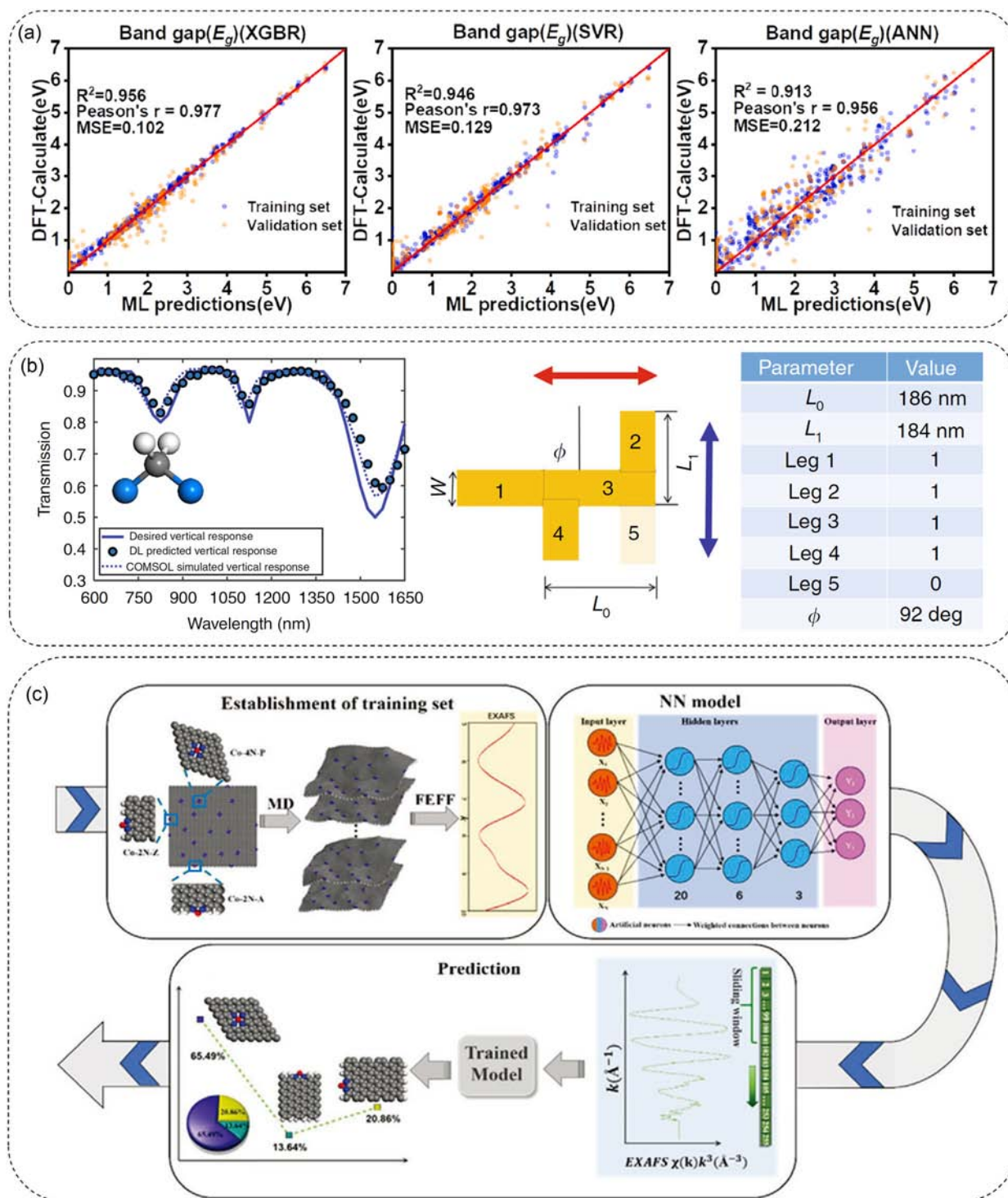


Figure 6. a) Comparison diagram of bandgap prediction models based on different ML algorithms: XGBR, SVR, and ANN, from left to right, respectively. Reproduced with permission.^[60] Copyright 2021, Elsevier B.V. b) Predicted geometry of the nanostructure based on trained DNN model. The left diagram displays the ML design of a gold plasmonic structure targeted to the organic molecule dichloromethane with specific spectral polarization responses. The right diagram represents the configuration and dimensions of the plasmonic structure found by the DNN model. Reproduced with permission.^[62] Copyright 2018, under the terms of CC-BY license, The Authors, published by Springer Nature. c) Brief scheme of ML strategy used to interpret the EXAFS. Reproduced with permission.^[69] Copyright 2021, Wiley-VCH GmbH.

In general, two main directions are considered to improve the interpretabilities of ML models: improve the feedback path of human experience during the training and increase the interactive presentation of the ML model contents during the decision-making process. Through the effective feedback path, professional domain knowledge is fed into the ML models. Such human-in-the-loop ML models combine the advantages of human and machine intelligence to better provide solid results with high interpretabilities.^[58] For the other aspect, ML models also should actively provide more interactive information during the training process to improve communications and feedback to scientists. Common interactive information refers to visual and readable descriptions, such as process diagrams, error warnings, parameter explanations, etc. For instance, Bayesian deep learning is a new ML model to accurately quantify the prediction uncertainties, which helps to avoid overfitting issues. These viable approaches supply important insights into the “black boxes” and improve the compatibility of ML models in diverse situations.

4.5. Applications of ML in Spectral Image Processing

At present, the photoelectric performances of perovskite devices are mostly based on experimental characterizations based on spectral images. These spectral images provide indispensable perspectives to analyze the light–matter interactions of functional perovskite materials at the microscopic scale. An ideal optical spectrometer should possess high spectral response, optical resolution, signal-to-noise ratio, stability, and operational convenience. However, achieving highly robust characterization performances with low costs in most optical spectrometers is still challenging, which requires more efforts in the future to improve the analysis of spectral images. Accordingly, image processing becomes a promising technique to overcome current challenges in spectrometers, which is one of the most fundamental functions in ML. It has been verified that ML algorithms show good performance in several specific fields of image processing such as resolution enhancement, feature recognition, image reconstruction, etc. For catalysts designs, Liu et al. successfully applied supervised learning techniques to interpret the measured synchrotron spectrum of Co single-atom catalysts (SACs) (Figure 6c).^[69] Based on the MD-extended X-ray absorption fine structure (MD-EXAFS) calculations of Co–N-doped graphene with different proportions of Co–4N–P, Co–2N–A, and Co–2N–Z, the training of the NN model is accomplished by including an input layer of the EXAFS spectrum. The highest prediction consistency of local structural proportion from the experimental EXAFS measurement is 63.94% for Co–4N–P. The accurate extraction of structural information by ML confirms that the improved hydrogen generation performances of the Co SACs are induced by the edge effect. The comprehensive characterization of the catalysts assisted by supervised learning offers a new approach to obtain precise structural information on materials. Inspired by this work, trained ML models are capable of upgrading current optical spectral analysis in accuracy, resolution, and processing rate based on powerful image processing capabilities. Although spectral image processing has not been widely used in perovskite materials, these contributions

further prove the assisting role of ML techniques in the field of perovskite optoelectronics.

4.6. Improvements in Accuracy by Monitoring ML Models

ML models are generally trained with historical data, most of which are obtained from multiple experiments with different settings, materials, and environments. These factors in experiments will induce inevitable regular differences in the data distribution. Therefore, a trained ML model is possible to become outdated in a new environment and lose its accuracy over time, which is attributed to the data drift effect.^[70] To address the drift effect and keep the accuracy of trained ML models, continuous monitoring and transfer learning are implemented to make ML models more robust to maintain accurate and reliable predictions in different scenarios. The monitoring process allows us to evaluate the performance of ML models during training and real-time deployment, which guarantees the elimination of poor generalization and changing parameters to realize the stability of predictions. Meanwhile, the transfer learning uses the pretrained models as the starting point, which optimizes the progress speed and the performance of ML for the new training. For example, the population stability index (PSI) is a model monitoring metric measure, which compares the distribution of a categorical variable in two different datasets. This method is an effective tool used to examine distributional shifts for all model-related attributes.^[71] In the field of perovskite optoelectronics, more learning paradigms should be investigated to develop a proper criterion, which helps to evaluate the characterization of data topology. Correspondingly, although current promising ML techniques are rapidly growing in materials science, we believe that more monitoring strategies are needed against the generalization errors induced by the degradation of ML models.

5. Conclusion

In conclusion, we highlight the current advances of ML techniques in the development of perovskite optoelectronics and summarize the future opportunities as well as challenges in this perspective. Starting with a brief introduction of ML-related concepts, we have presented a general workflow of performing ML including: 1) identifying the object; 2) preparing the dataset; 3) determining the feature representations and algorithms; 4) training the ML model; 5) evaluating the accuracy; and 6) optimizing the ML model. The applications of ML models in perovskite optoelectronics have achieved some progresses and received continuous research attention. Accordingly, we have discussed the examples of utilizing ML in developing novel perovskite optoelectronics regarding the discovery of new advanced perovskite materials, the interpretation of underlying mechanisms, and the high-throughput processing of large-scale information. Based on the strong predictive capabilities of ML, the screening, discovery, and predictions of new advanced perovskite materials have been realized effectively in recent years. With the accumulation of databases and the development of efficient algorithms, future ML techniques are expected to significantly accelerate the material evolutions in perovskite optoelectronics. Meanwhile, ML is also able to quantify the

relationship between different features of perovskites through the correlation coefficients during the training process, which allows ML models to discover and explains the underlying mechanisms of the structure–property relationship. In addition, the high-throughput process for large-scale data in ML is not only reflected in the screening of large databases but also in the recognition of images and the forecasting of time-series properties. Taking advantages of these capabilities, ML algorithms are able to process more complicated characterizations of perovskite optoelectronics regarding the surface morphologies, optical spectra, and spatial–temporal spectroscopies. In addition, we also discuss the future potentials of ML techniques in the field of perovskite optoelectronics explorations from six different aspects including: 1) synergistic effect between ML and DFT calculations; 2) identification of structure–property correlation; 3) construction of benchmark database; 4) establishment of explainable models for ML; 5) applications of ML in spectral image processing; and 6) improvements in accuracy by monitoring ML models.

6. Outlook

According to the current trends, there is no doubt that ML techniques will play an increasingly important role in the future of materials science. However, some concerns and challenges still need to be addressed properly to ensure the right direction of ML development. The first issue is the reliability of ML models, which includes not only the model accuracy but also the compatibility and robustness. The most efficient way to improve the reliability of ML models is to construct benchmark databases and corresponding monitoring systems. Benchmark databases fed with professional domain knowledge ensure that the trained ML model is consistent with the true model to ensure good compatibility. Monitoring systems further avoid data drift and improve the sustainability of original ML models. Continuous updating and correction of ML models is the basis for maintaining prediction reliability. Another issue is the interpretability of comprehensive ML models, where many models have poor interpretability for the output results. Although researchers easily obtain accurate output from ML models, the decision-making process is unclear. Such a “black box” effect will hinder the communications between human and machine intelligence, which is not conducive to the long-term development of ML technology. In this aspect, effective feedback paths and interactive presentations should be implemented in ML models to solve this challenge. In the big picture, ML technology still has significant development potential in crossdisciplinary fields of materials science, where the synergies with theoretical calculations, spectral image processing, and characterization of structural–optoelectronic correlation open up new directions for ML development in the future. Therefore, the applications of ML offer promising design and optimization strategies to accelerate the development of next-generation perovskite optoelectronics, which further brings current materials science into a new stage.

Acknowledgements

The authors gratefully acknowledge support from the National Key R&D Program of China (2021YFA1501101), National Natural Science

Foundation of China/Research Grant Council of Hong Kong Joint Research Scheme (N_PolyU502/21), National Natural Science Foundation of China/Research Grants Council of Hong Kong Collaborative Research Scheme (CRS_PolyU504_22), the funding for Projects of Strategic Importance of The Hong Kong Polytechnic University (project code: 1-ZE2V), Shenzhen Fundamental Research Scheme-General Program (JCY20220531090807017), Natural Science Foundation of Guangdong Province (2023A1515012219), and Departmental General Research Fund (project code: ZVUL) from The Hong Kong Polytechnic University. The authors also thank the support from Research Centre for Carbon-Strategic Catalysis (RC-CSC), Research Institute for Smart Energy (RISE), and Research Institute for Intelligent Wearable Systems (RI-IWEAR) of the Hong Kong Polytechnic University.

Conflict of Interest

The authors declare no conflict of interest.

Keywords

high-throughput, machine learning, material designs, optoelectronics, perovskites

Received: August 2, 2023
Revised: August 17, 2023
Published online: August 30, 2023

- [1] C. Huang, C. Zhang, S. Xiao, Y. Wang, Y. Fan, Y. Liu, N. Zhang, G. Qu, H. Ji, J. Han, L. Ge, Y. Kivshar, Q. Song, *Science* **2020**, 367, 1018.
- [2] Q. Shang, M. Li, L. Zhao, D. Chen, S. Zhang, S. Chen, P. Gao, C. Shen, J. Xing, G. Xing, B. Shen, X. Liu, Q. Zhang, *Nano Lett.* **2020**, 20, 6636.
- [3] Q. Zhang, Q. Shang, R. Su, T. T. H. Do, Q. Xiong, *Nano Lett.* **2021**, 21, 1903.
- [4] M. A. Green, E. D. Dunlop, J. Hohl-Ebinger, M. Yoshita, N. Kopidakis, A. W. Y. Ho-Baillie, *Prog. Photovoltaics Res. Appl.* **2020**, 28, 3.
- [5] J. Y. Kim, J.-W. Lee, H. S. Jung, H. Shin, N.-G. Park, *Chem. Rev.* **2020**, 120, 7867.
- [6] Y. Rong, Y. Hu, A. Mei, H. Tan, M. I. Saidaminov, S. I. Seok, M. D. McGehee, E. H. Sargent, H. Han, *Science* **2018**, 361, eaat8235.
- [7] A. Fakhruddin, M. K. Gangishetty, M. Abdi-Jalebi, S.-H. Chin, A. R. Bin Mohd Yusoff, D. N. Congreve, W. Tress, F. Deschler, M. Vasilopoulou, H. J. Bolink, *Nat. Electron.* **2022**, 5, 203.
- [8] D. Ma, K. Lin, Y. Dong, H. Choubisa, A. H. Proppe, D. Wu, Y.-K. Wang, B. Chen, P. Li, J. Z. Fan, F. Yuan, A. Johnston, Y. Liu, Y. Kang, Z.-H. Lu, Z. Wei, E. H. Sargent, *Nature* **2021**, 599, 594.
- [9] J. Chen, H. Xiang, J. Wang, R. Wang, Y. Li, Q. Shan, X. Xu, Y. Dong, C. Wei, H. Zeng, *ACS Nano* **2021**, 15, 17150.
- [10] *Nature Energy* **2019**, 4, 1.
- [11] M. V. Khenkin, E. A. Katz, A. Abate, G. Bardizza, J. J. Berry, C. Brabec, F. Brunetti, V. Bulović, Q. Burlingame, A. Di Carlo, R. Cheacharoen, Y.-B. Cheng, A. Colmann, S. Cros, K. Domanski, M. Duszka, C. J. Fell, S. R. Forrest, Y. Galagan, D. Di Girolamo, M. Grätzel, A. Hagfeldt, E. von Hauff, H. Hoppe, J. Kettle, H. Köbler, M. S. Leite, S. Liu, Y.-L. Loo, J. M. Luther, et al., *Nat. Energy* **2020**, 5, 35.
- [12] M. Green, E. Dunlop, J. Hohl-Ebinger, M. Yoshita, N. Kopidakis, X. Hao, *Prog. Photovoltaics Res. Appl.* **2021**, 29, 3.
- [13] L. Zhang, C. Sun, T. He, Y. Jiang, J. Wei, Y. Huang, M. Yuan, *Light Sci. Appl.* **2021**, 10, 61.
- [14] L. Yue, B. Yan, M. Attridge, Z. Wang, *Sol. Energy* **2016**, 124, 143.

- [15] J. Leng, T. Wang, Z.-K. Tan, Y.-J. Lee, C.-C. Chang, K. Tamada, *ACS Omega* **2022**, *7*, 565.
- [16] S. Sajid, A. M. Elseman, H. Huang, J. Ji, S. Dou, H. Jiang, X. Liu, D. Wei, P. Cui, M. Li, *Nano Energy* **2018**, *51*, 408.
- [17] T. Benbarrad, M. Salhaoui, S. B. Kenitar, M. Arioua, *J. Sens. Actuator Networks* **2021**, *10*, 7.
- [18] D. P. Penumuru, S. Muthuswamy, P. Karumbu, *J. Intell. Manuf.* **2020**, *31*, 1229.
- [19] L. Deng, X. Li, *IEEE Tran. Audio Speech Lang. Process.* **2013**, *21*, 1060.
- [20] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, K. Shaalan, *IEEE Access* **2019**, *7*, 19143.
- [21] S. Scher, G. Messori, *Q. J. R. Meteorolog. Soc.* **2018**, *144*, 2830.
- [22] N. Singh, S. Chaturvedi, S. Akhter, presented at 2019 Int. Conf. Signal Process. Commun. (ICSC), Noida, India March **2019**.
- [23] F. Mayr, M. Harth, I. Kouroudis, M. Rinderle, A. Gagliardi, *J. Phys. Chem. Lett.* **2022**, *13*, 1940.
- [24] X.-Y. Ma, J. P. Lewis, Q.-B. Yan, G. Su, *J. Phys. Chem. Lett.* **2019**, *10*, 6734.
- [25] C. L. Ritt, M. Liu, T. A. Pham, R. Epszstein, H. J. Kulik, M. Elimelech, *Sci. Adv.* **2022**, *8*, eabl5771.
- [26] S. P. Ong, *Comput. Mater. Sci.* **2019**, *161*, 143.
- [27] E. W. Huang, W.-J. Lee, S. S. Singh, P. Kumar, C.-Y. Lee, T.-N. Lam, H.-H. Chin, B.-H. Lin, P. K. Liaw, *Mater. Sci. Eng.: R: Rep.* **2022**, *147*, 100645.
- [28] M. Sun, T. Wu, A. W. Dougherty, M. Lam, B. Huang, Y. Li, C.-H. Yan, *Adv. Energy Mater.* **2021**, *11*, 2003796.
- [29] V. Botu, R. Ramprasad, *Int. J. Quantum Chem.* **2015**, *115*, 1074.
- [30] M. Gastegger, J. Behler, P. Marquetand, *Chem. Sci.* **2017**, *8*, 6924.
- [31] P. Pattnaik, S. Raghunathan, T. Kalluri, P. Bhimalapuram, C. V. Jawahar, U. D. Priyakumar, *J. Phys. Chem. A* **2020**, *124*, 6954.
- [32] W. Yang, J. Li, X. Chen, Y. Feng, C. Wu, I. D. Gates, Z. Gao, X. Ding, J. Yao, H. Li, *ChemPhysChem* **2022**, *23*, e202100841.
- [33] E. Alpaydin, *Introduction to Machine Learning*, The MIT Press, Cambridge, Massachusetts **2020**.
- [34] M. Alloghani, D. Al-Jumeily, J. Mustafina, A. Hussain, A. J. Aljaaf, in *Supervised and Unsupervised Learning for Data Science* (Eds: M. W. Berry, A. Mohamed, B. W. Yap), Springer International Publishing, Cham **2020**, pp. 3–21.
- [35] B. Charbuty, A. Abdulazeez, *J. Appl. Sci. Technol. Trends* **2021**, *2*, 20.
- [36] M. Somvanshi, P. Chavan, S. Tambade, S. V. Shinde, presented at 2016 Int. Conf. Comput. Commun. Control Autom. (ICCUBEA), Pune, India August **2016**.
- [37] Y. Hong, B. Hou, H. Jiang, J. Zhang, *WIREs Comput. Mol. Sci.* **2020**, *10*, e1450.
- [38] V. M. Goldschmidt, *Naturwissenschaften* **1926**, *14*, 477.
- [39] R. Armiento, B. Kozinsky, G. Hautier, M. Fornari, G. Ceder, *Phys. Rev. B* **2014**, *89*, 134103.
- [40] J. Schmidt, J. Shi, P. Borlido, L. Chen, S. Botti, M. A. L. Marques, *Chem. Mater.* **2017**, *29*, 5090.
- [41] H. Liu, J. Cheng, H. Dong, J. Feng, B. Pang, Z. Tian, S. Ma, F. Xia, C. Zhang, L. Dong, *Comput. Mater. Sci.* **2020**, *177*, 109614.
- [42] X. Yang, L. Li, Q. Tao, W. Lu, M. Li, *Comput. Mater. Sci.* **2021**, *196*, 110528.
- [43] A. E. Siemenn, Z. Ren, Q. Li, T. Buonassisi, *NPJ Comput. Mater.* **2023**, *9*, 79.
- [44] R. E. Kumar, A. Tiuhonen, S. Sun, D. P. Fenning, Z. Liu, T. Buonassisi, *Matter* **2022**, *5*, 1353.
- [45] J. A. Bennett, M. Abolhasani, *Curr. Opin. Chem. Eng.* **2022**, *36*, 100831.
- [46] V. Shekar, G. Nicholas, M. A. Najeeb, M. Zeile, V. Yu, X. Wang, D. Slack, Z. Li, P. W. Nega, E. M. Chan, A. J. Norquist, J. Schrier, S. A. Friedler, *J. Chem. Phys.* **2022**, *156*, 064108.
- [47] J. Li, J. Li, R. Liu, Y. Tu, Y. Li, J. Cheng, T. He, X. Zhu, *Nat. Commun.* **2020**, *11*, 2046.
- [48] K. Abdel-Latif, R. W. Epps, F. Bateni, S. Han, K. G. Reyes, M. Abolhasani, *Adv. Intell. Syst.* **2021**, *3*, 2000245.
- [49] T. Bak, K. Kim, E. Seo, J. Han, H. Sung, I. Jeon, I. D. Jung, *Int. J. Precis. Eng. Manuf. Green Technol.* **2023**, *10*, 109.
- [50] Y. Yu, X. Tan, S. Ning, Y. Wu, *ACS Energy Lett.* **2019**, *4*, 397.
- [51] L. Xu, L. Wencong, P. Chunrong, S. Qiang, G. Jin, *Comput. Mater. Sci.* **2009**, *46*, 860.
- [52] H. Park, R. Mall, A. Ali, S. Sanvito, H. Bensmail, F. El-Mellouhi, *Comput. Mater. Sci.* **2020**, *184*, 109858.
- [53] X. Li, Y. Dan, R. Dong, Z. Cao, C. Niu, Y. Song, S. Li, J. Hu, *Appl. Sci.* **2019**, *9*, 5510.
- [54] S. Lu, Q. Zhou, Y. Ouyang, Y. Guo, Q. Li, J. Wang, *Nat. Commun.* **2018**, *9*, 3405.
- [55] J. Kirman, A. Johnston, D. A. Kuntz, M. Askerka, Y. Gao, P. Todorović, D. Ma, G. G. Privé, E. H. Sargent, *Matter* **2020**, *2*, 938.
- [56] N. Taherimaksousi, B. P. MacLeod, F. G. L. Parlange, T. D. Morrissey, E. P. Booker, K. E. Dettelbach, C. P. Berlinguette, *NPJ Comput. Mater.* **2020**, *6*, 111.
- [57] J. M. Howard, Q. Wang, M. Srivastava, T. Gong, E. Lee, A. Abate, M. S. Leite, *J. Phys. Chem. Lett.* **2022**, *13*, 2254.
- [58] J. Zhou, B. Huang, Z. Yan, J.-C. G. Bünzli, *Light Sci. Appl.* **2019**, *8*, 84.
- [59] Z. Wang, M. Yang, X. Xie, C. Yu, Q. Jiang, M. Huang, H. Algadi, Z. Guo, H. Zhang, *Adv. Compos. Hybrid Mater.* **2022**, *5*, 2700.
- [60] Z. Gao, H. Zhang, G. Mao, J. Ren, Z. Chen, C. Wu, I. D. Gates, W. Yang, X. Ding, J. Yao, *Appl. Surf. Sci.* **2021**, *568*, 150916.
- [61] X. Zhai, F. Ding, Z. Zhao, A. Santomauro, F. Luo, J. Tong, *Commun. Mater.* **2022**, *3*, 42.
- [62] I. Malkiel, M. Mrejan, A. Nagler, U. Arieli, L. Wolf, H. Suchowski, *Light Sci. Appl.* **2018**, *7*, 60.
- [63] T. Zhang, J. Wang, Q. Liu, J. Zhou, J. Dai, X. Han, Y. Zhou, K. Xu, *Photonics Res.* **2019**, *7*, 368.
- [64] P. R. Wiecha, A. Lecestre, N. Mallet, G. Larrieu, *Nat. Nanotechnol.* **2019**, *14*, 237.
- [65] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K. A. Persson, *APL Mater.* **2013**, *1*, 011002.
- [66] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. 't Hoen, R. Hoof, T. Kuhn, R. Kok, J. Kok, et al., *Sci. Data* **2016**, *3*, 160018.
- [67] L. C. Brinson, L. M. Bartolo, B. Blaiszik, D. Elbert, I. Foster, A. Strachan, P. W. Voorhees, *MRS Bull.* **2023**, <https://doi.org/10.1557/s43577-023-00498-4>.
- [68] F. Bodria, F. Giannotti, R. Guidotti, F. Naretto, D. Pedreschi, S. Rinzivillo, *Data Min. Knowl. Discovery* **2023**, <https://doi.org/10.1007/s10618-023-00933-9>.
- [69] X. Liu, L. Zheng, C. Han, H. Zong, G. Yang, S. Lin, A. Kumar, A. R. Jadhav, N. Q. Tran, Y. Hwang, J. Lee, S. Vasimalla, Z. Chen, S.-G. Kim, H. Lee, *Adv. Funct. Mater.* **2021**, *31*, 2100547.
- [70] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama, G. Zhang, *IEEE Trans. Knowl. Data Eng.* **2019**, *31*, 2346.
- [71] R. Taplin, C. Hunt, *Risks* **2019**, *7*, 53.



Bolong Huang received his Ph.D. in 2012 from the University of Cambridge and his B.Sc. from Peking University in 2007. Following systematic training periods of postdoc at Peking University, and in Hong Kong, he started his independent research at the Hong Kong Polytechnic University in 2015. He is now the associate professor at Department of Applied Biology and Chemical Technology and director of the Research Centre for Carbon-Strategic Catalysis. His main research fields are electronic structures of nanomaterials, energy materials, solid functional materials, and rare earth materials, as well as their applications in multiscale energy conversion and supply systems.