npj | Computational Materials

www.nature.com/npjcompumats

REVIEW ARTICLE OPEN

Check for updates

# Machine learning for perovskite materials design and discovery

Qiuling Tao[1,3], Pengcheng Xu[2,3], Minjie Li[1] ✉ and Wencong Lu[1,2] ✉

The development of materials is one of the driving forces to accelerate modern scientific progress and technological innovation. Machine learning (ML) technology is rapidly developed in many fields and opening blueprints for the discovery and rational design of materials. In this review, we retrospected the latest applications of ML in assisting perovskites discovery. First, the development tendency of ML in perovskite materials publications in recent years was organized and analyzed. Second, the workflow of ML in perovskites discovery was introduced. Then the applications of ML in various properties of inorganic perovskites, hybrid organic–inorganic perovskites and double perovskites were briefly reviewed. In the end, we put forward suggestions on the future development prospects of ML in the field of perovskite materials.

*npj Computational Materials* (2021)7:23 ; https://doi.org/10.1038/s41524-021-00495-8

## INTRODUCTION

Perovskite materials have attracted much attention in many scientific fields for the composition diversity, easily available synthetic conditions and a variety of attractive properties[1,2]. For instance, hybrid organic–inorganic perovskite has widely applied in the fields of solar cells, light-emitting diodes, lasers, and photodetectors due to its longer charge diffusion lengths both for electrons and holes, higher carrier mobility and broad tunable bandgap ($E_g$)[3,4]. $ABO_3$-type perovskite oxide has gradually become a research hotspot in modern industrial catalysis and thermoelectricity for the controllable structure, outstanding stability and low cost[5,6]. Inorganic double perovskite has aroused an interest in solar cells and light-emitting diodes because of adjustable photoelectric properties[7,8]. The trends of published papers searched on the website ‘web of science’ from 1961 to December 2020 are shown in Fig. 1. The number of papers under the keyword of perovskite shows an alarming increase. Especially after 2013, since the perovskite solar cell was proposed, the related publications has increased exponentially, indicating that perovskite materials have always been a hotspot for scientists.

The traditional way to develop materials is usually based on trial and error, continuous synthesis and characterization keep trying until the properties of virtual materials meet the target. The method requires a long-time study on a limited quantity of materials and complicated experimental procedures, which can be a time-consuming and expensive endeavor. Under this limitation, important scientific progress often comes from the researchers’ experience and intuition or even was discovered by accident[9,10]. Besides, the discovery of high-performance materials needs a long cycle from experimental design to commercialization. With the development of synthesis and characterization techniques, the corresponding data become more and more complex. It is a great challenge to figure out the relationship between materials descriptors and properties by traditional experimental methods. To overcome this shortcoming, material simulated methods, including Density Functional Theory (DFT)[11], Monte Carlo simulation[12] and molecular dynamics[13] are employed to explore the relationship between the structural, compositional, and technological descriptors and performance of materials at different scales. In particular, DFT could be used to obtain some key properties of the material without the need for experimental synthesis. However, most computational methods only aim at a specific system, leading to an unbearable amount of computation for complex systems. Some theoretical methods still cannot meet the requirements of quantitative description of material properties. Moreover, computational simulation methods require high computational costs and professional skills.

In recent years, artificial intelligence (AI), known as the ‘fourth paradigm of science’, has attracted worldwide attentions[14]. Since the 1980s, machine learning (ML) has been the core of AI for the power of reorganizing existing knowledge structures and mining implicit relationships. ML can extract valuable information from existing data, even failed experimental data[15,16]. For material science, ML has been becoming a powerful tool to assist design and screen various materials. A series of achievements about ML have been made in superconductor, photovoltaic materials and high entropy alloys[17–19]. As shown in Fig. 1, ML has also been applied widely in perovskite materials. Considering that ML has conducted a lot of researches in this field, reviewing their progress and providing an outlook for future work will be helpful in the development of perovskite materials.

In this review, we briefly discuss the successful application of ML in properties prediction and stability assessment of perovskite material. In section 2, the basic workflow of ML in material science is outlined. In section 3, we introduce the different types of perovskites and applications of ML in various properties of perovskite materials. In section 4, some of the current challenges and opportunities encountering in ML applications to perovskite design and discovery are briefly discussed. Our work is to provide practical guidance for accelerating the design of perovskite materials.

## WORKFLOW OF MACHINE LEARNING

ML is an interdisciplinary subject that combines knowledge of computer science, statistics, mathematics and engineering to form an important branch of artificial intelligence[20,21]. The most common application of ML is to construct a statistical model used for data analysis and prediction. The main purpose of ML aims at evaluating or predicting the objects after training the model with historical data and specific conditions[22]. With the

[1]Department of Chemistry, College of Sciences, Shanghai University, Shanghai, China. [2]Materials Genome Institute, Shanghai University, Shanghai, China. [3]These authors contributed equally: Qiuling Tao, Pengcheng Xu. ✉email: minjieli@shu.edu.cn; wclu@shu.edu.cn

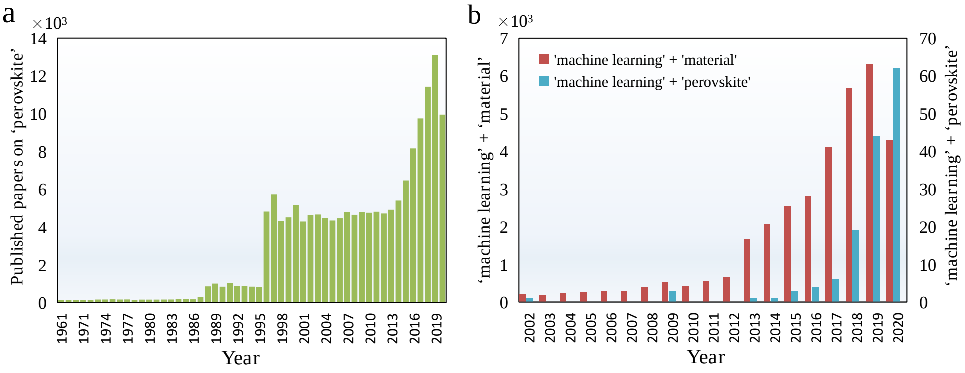

**Fig. 1** **Number of published papers.** **a** On keyword of 'perovskite' (from 1961 to December 2020). **b** On key words of 'machine learning and material' and 'machine learning and perovskite' (from 2002 to December 2020).

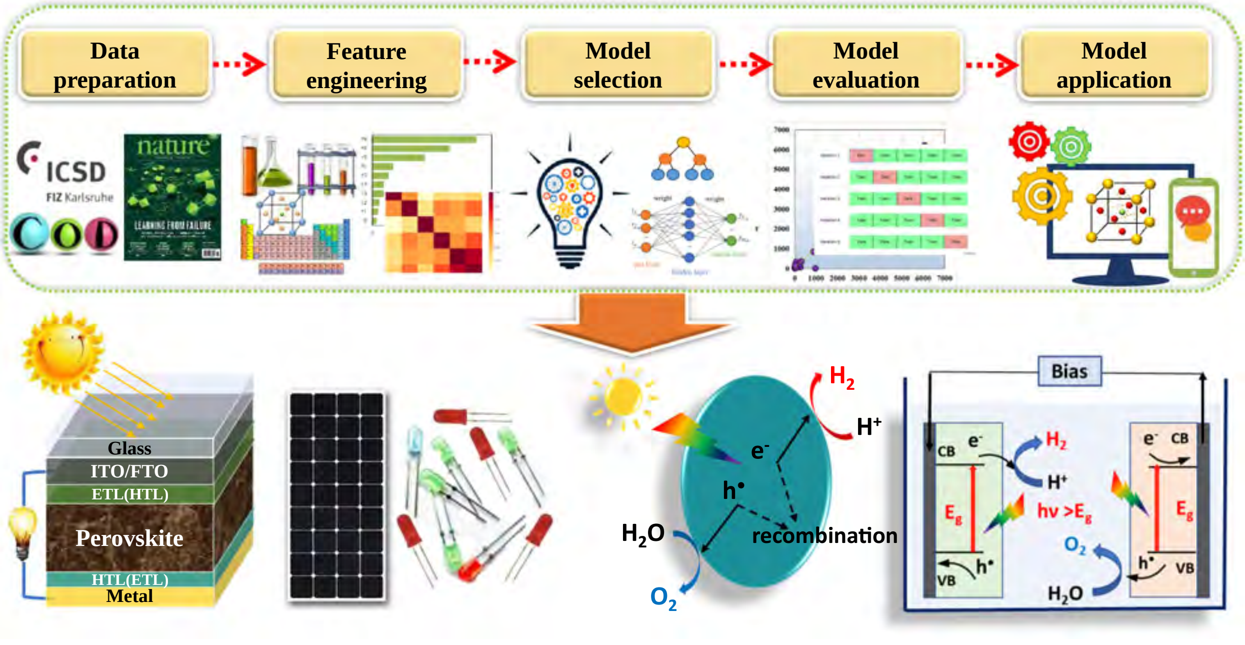

**Fig. 2** **The general workflow of ML in perovskite materials and related applications.** Workflow of ML in perovskite materials includes data preparation, feature engineering, model selection, model evaluation and model application.

development of materials genome initiative (MGI), ML keeps playing an important role in materials science with the ability of mapping the relationships and trends through the available data without the physical mechanisms. In addition, the constructed ML model could also be applied again for reverse discovery of high-performance materials. To screen virtual materials with desired properties, ML could be allowed to develop quantitative structure-property relationships to predict the properties of virtual materials. The general workflow (Fig. 2) of ML in material science includes data preparation, feature engineering, model selection, model evaluation, and model application.

## DATA PREPARATION

The dataset used for ML usually contains dependent and independent variables associated with the materials. Independent variables, also known as features or descriptors, refer to the representative information related to the structure and characteristics of materials, including the chemical composition, atomic or molecular parameters, structural parameters, as well as the technological conditions for synthesis process. The dependent variables refer to the target property of the materials affected by the independent variables, also known as the target variables[23,24]. the quantity and quality of data are key factors in the discovery of materials.

The amount of required samples depends on the ML model, but a general rule of thumb is that a reasonable ML model requires the number of data more than three times of descriptors at least, but some models such as neural networks and deep learning require large amounts of samples[25]. The quality of the data depends on the spatial coverage of the target properties and the uncertainties associated with the data. In general, data with a normal distribution is better for ML. Insufficient data of specific target or poor coverage of specific properties may not form an appropriate data distribution for ML. Also, data uncertainty, such as experimental error or calculation error, could affect the quality of data. The roughness of the modelling data directly determines the results of the constructed model. In general, the prediction error of the model is higher than the error of the training data. The methods adopted to reduce the roughness of the data include

**Table 1.** Publicly accessible databases of various materials.

| Database | Brief description | URL |
|---|---|---|
| Materials Project (MP) | Calculation data of properties of known and hypothetical materials | https://materialsproject.org |
| The Inorganic Crystal Structure Database (ICSD) | Experimental characterization data of inorganic crystal structure | https://icsd.fiz-karlsruhe.de/index.xhtml |
| Cambridge Structural Database (CSD) | The structure database of small molecules and metal-organic molecular crystals based on X-ray and neutron diffraction experiments collected by the Cambridge Crystallographic Data Centre | https://www.ccdc.cam.ac.uk/ |
| Aflow-Automatic-FLOW for Materials Discovery (AFLOW) | A data repository of structure and property of inorganic materials from high-throughput ab initio calculations | http://www.aflowlib.org |
| Crystallography Open Database (COD) | Structures data of organic, inorganic, and metal-organic compounds and minerals | http://cod.ensicaen.fr |
| Open Quantum Materials Database (OQMD) | Theoretical simulation calculation data of mostly hypothetical materials | http://www.oqmd.org/ |
| Springer Materials | The world's largest material data resource, a unique, high-quality numerical database | https://materials.springer.com |
| GDB | Hypothetical small organic molecule database | http://gdb.unibe.ch/downloads |
| ZINC | Commercially available organic molecules in two-dimensional and three-dimensional formats | https://zinc15.docking.org/ |
| Materiae | Topological material database | http://materiae.iphy.ac.cn/ |
| Materials Cloud | Structural calculation data of candidate two-dimensional materials | https://www.materialscloud.org/discover/2dstructures/dashboard/ptable |
| Materials Platform for Data Science (MPDS) | Peer-reviewed crystal structure, phase diagram, or physical property | https://mpds.io/#modal/menu |

deletion of missing values, completion of experimental conditions, data normalization and other methods.

Data can be collected from available databases (Table 1) or published papers. The database contains many types of data, which could be generated from experiments, simulations and ML. A large amount of data could be collected from database, while the reproducibility of the data is uncertain, which might not ensure the quality of data. To guarantee the quantity and quality, data should be collected from the authoritative databases if the data are available. Using autonomous workflows to generate data could be a very convenient and fast way, but the quality of data obtained in this way may be inferior to the data from the database. If the database or autonomous workflows does not obtain the needed data, the dataset could also be generated through lab-scale calculations. Lab-scale calculations could be performed in many existing open sources and various software platforms, such as Materials Studio (MS), Vienna Ab initio Simulation Package (VASP), Car-Parrinello Molecular Dynamics (CPMD). Lab-scale calculations could generate a large amount of required data with good reproducibility to guarantee the quality of data. Generally, ML models constructed by the calculated data have relatively good evaluation metrics. However, the calculations of complex materials may take up too many calculation resources and take a long time.

The origin data collected from computational simulations or experimental measurements are often presented with incompleteness, noise, and inconsistency[26]. Thus, data preprocessing should be performed to ensure consistency and integrity from origin data. Specifically, individual data need to be reformatted into a single tabular form, imputed missing values, eliminated erroneous or incomparable data points, and normalized and rescaled the data. Data standardization can improve model accuracy and convergence speed. The results of several ML algorithms can vary with whether any standardization or scaled. It is worth noting that both feature variables and target variables can be normalized or scaled[27].

## FEATURE ENGINEERING

The properties of each material depend on a specific set of features, also called descriptors. Before model construction, it is crucial to identify the key features closely related to the target properties[28]. Features are generally derived from known properties of the constituent elements, such as atomic radius and electronegativity. The quantity of features should be less than that of dataset samples for effectively training and avoiding overfitting. Therefore, feature selection should reduce the dimension of input space as much as possible without losing important information. In particular, redundant and high self-correlation features should be removed to guarantee the efficiency and accuracy of models[25,29]. Reasonable material features should meet the following three conditions of perfect representation of material properties, sensitive to target properties, and easy to obtain[30].

## MODEL SELECTION

ML algorithms could be briefly divided into two categories: supervised learning and unsupervised learning. Supervised learning is the process of using a set of samples with known labels to adjust the parameters of the models and achieve the required performance, which be further divided into regression and classification[31]. With the target property being a continuous value, the process is called regression. If the target is a discrete value, the process of searching the prediction function is called classification. Tables 2, 3 have summarized common ML algorithms in material design. Generally, the best model is obtained by comparing multiple algorithms. The criteria of algorithm selection are mainly based on the results of cross validation and independent test. The commonly used evaluation metrics include mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), determination coefficient ($R^2$), correlation coefficient (R) for regression; confusion matrix, precision, recall, receiver operating characteristic curve (ROC), and area under ROC curve (AUC) for classification.

**Table 2.** Common ML algorithms in material design.

| ML algorithms | Category | Brief description |
|---|---|---|
| Support Vector Machine (SVM) | Regression, classification | Support vector machine (SVM) includes support vector classification (SVC) and support vector regression (SVR). The main idea of SVC is to establish an optimal decision hyperplane to maximize the distance between the two kinds of samples closest to the plane on both sides of the plane. The basic idea of SVR is to map the data X into a higher-dimensional feature space F via a nonlinear mapping Φ and then to do linear regression in this space. SVM provides good generalization ability for classification and regression tasks. |
| Artificial Neural Networks (ANN) | Regression, classification | A neural network is composed of a large number of connected nodes (neurons). Samples are classified or regressed according to different connection modes and connection signals (weights) between nodes. |
| Random Forest (RF) | Regression, classification | An ensemble learning algorithm contains multiple decision trees. |
| Extra-Trees (ET)/Extremely Randomized Trees (ERT) | Regression, classification | An ensemble learning algorithm composed of multiple decision trees is similar to random forest. |
| Gradient Boosting Regression (GBR) | Regression | An ensemble learning algorithm contains multiple decision trees (regression trees). |
| Multiple Linear Regression | Regression | Solve the regression problem when the relationship between multiple independent variables and one dependent variable is linear. |
| Ridge Regression (RR) | Regression | The linear least-squares method with regularization. |
| Kernel Ridge Regression (KRR) | Regression | The algorithm combines ridge regression with kernel function. |
| Gaussian Process Regression (GPR) | Regression | A nonparametric model that uses Gaussian process priors to perform regression analysis on data. |
| Partial Least-Squares Regression | Regression | Least-squares fit of the output data to the input features |
| Decision Tree (DT) | Classification | Generate a tree based on the features and categories of the data and classify the unknown data with this tree. |
| k-Nearest Neighbors (KNN) | Classification | Classification is performed by measuring the distance between different feature values. |
| Naive Bayes | Classification | Solve the occurrence probability of each classification category, then divide samples into the category with the largest occurrence probability. |
| Logistic Regression | Classification | Explain the relationship between a dependent variable and one or more independent variables. |
| k-means clustering | Clustering | A typical partition clustering algorithm. It uses a clustering center to represent a cluster. The algorithm can only treat numeric-type data. |
| k-modes clustering | Clustering | Extension of k-means algorithm. Adopt the simple matching method to measure the similarity of classification type data. |
| k-prototypes clustering | Clustering | Combine two algorithms of k-means with k-modes, which can handle mixed data. |

It is important for researchers to construct optimal models that map existing data without overfitting or underfitting[23]. After the model selection, it is generally necessary to optimize the internal hyper-parameters of the model algorithm to balance overfitting and underfitting[32]. Overfitting refers to that the model focuses too much on each individual data point in the training set, the unknown data cannot be well predicted. While underfitting means that the model is too simple to capture the general information of the training set, resulting in a large deviation. Models insensitive to hyper-parameters usually do not require repeated adjustments to achieve satisfying results[33,34]. For the hyper parametric sensitive model, the performance of the model will be closely related to the selection. It is necessary to adjust different hyper-parameters to achieve the robust ML model.

## MODEL EVALUATION

The core of ML is to realize accurate prediction of unknown samples based on known information. There are inevitably some statistical errors in the calculation, which should rationally be checked and evaluated in the process model evaluation for the subsequent model application. There are three commonly used model evaluation methods: independent test, cross validation, and bootstrapping.

In general, the generalization error of the model can be evaluated by testing, but the goal of the model is the prediction of unknown samples. Therefore, a testing set is needed to test the generalization ability. The error obtained with the testing set can be taken as an approximation of the generalization error. The smaller error of independent test usually indicates the stronger generalization of the model available. It is worth noting that the independent testing set should be mutually exclusive to the training set[35].

Cross validation (CV) could be used to evaluate the reliability of the ML models. In the CV, the input data is divided into *k* mutually exclusive subsets of the similar size, each subset is generated by 'stratified samples'. Then, the union of *k-1* subsets is used as the training set with the remaining one used as the testing set. After *k* times of training and testing, all test results are averaged to represent the final ML performance. The stability and fidelity of the evaluation results of the CV method depend to a large extent on the value of *k*. Hence, the CV method is usually called *k*-fold cross validation. In the *k*-fold CV, *k* is a specified number, the commonly used values are 5, 10, and 20. When k is equal to the sample number of input data, this method is called leave-one-out cross validation (LOOCV). LOOCV is not affected by random sample partitioning and the results are often considered to be

**Table 3.** Overview of application cases of ML in perovskite materials.

| Ref. | Research object | Target properties | Algorithms | Prediction accuracy | Main achievements |
|---|---|---|---|---|---|
| 60 | $ABX_3$ perovskites | Stability | ERT | MAE = 121 meV/atom | Predict the thermodynamic stability of around 230,000 possible compounds and screen out the stable candidates. |
| 61 | $ABO_3$ perovskites | Thermodynamic phase stability | ET and KRR | ET (accuracy 93%), KRR(RMSE = 28.5 meV/atom) | Four stable perovskites were presented ($La_{0.5}Y_{0.5}Co_{0.5}Mn_{0.5}O_3$, $Y_{0.75}Sr_{0.25}VO_3$, $CeReO_3$ and $Dy_{0.75}Nd_{0.25}RuO_3$). |
| 62 | $ABO_3$ perovskites | Stable and metastable | GBDT | Accuracy 94.6% | 37 stable $ABO_3$ perovskites and 13 metastable perovskites were predicted according to $E_{hull}$ for further synthesis and application. |
| 63 | $ABX_3$ perovskite | Stability | GBR and convolutional neural network | GBR ($R^2 = 0.91$, RMSE = 0.28), CNN ($R^2 = 0.85$, RMSE = 0.34) | Promising new perovskite materials were selected from 21316 hypothetical perovskite structures by screening model. A novel ML method, transfer learning that can dispose of the problem of small datasets for large-scale screening of assumed perovskites proposed. |
| 64 | $ABX_3$ perovskites | Formability and stability | SVM | Training set (accuracy 93.8%); testing set (accuracy92.1%) | 40 potential $ABX_3$ perovskite halides with high perovskite crystal structure formation probability were determined. |
| 65 | $ABO_3$ perovskites | Formability and cubic structure stability | RF and Gradient boosting tree | The accuracy of both classification models is over 90% | 87 promising perovskite candidate materials were presented. Discovered new perovskites may be located in (a) A and B atoms being lanthanide or actinide elements; (b) Atom A being an alkali metal, alkali earth metal, or late transition metal atom; (c) B being a p-block element. |
| 66 | $ABX_3$ perovskites | Formability and interfacial properties | SVM | Training set (accuracy 94%); testing set (accuracy96%) | It is an important step towards a basic understanding of the interfacial properties of perovskite, facilitating further breakthroughs in photovoltaic technology. Proposed two promising stable candidate materials, $RbSnCl_3$ and $RbSnBr_3$, for future photovoltaic and related applications. |
| 67 | $ABX_3$ perovskites | Phase stability | XGBoost | $\lambda(R^2 = 0.87)$, $\sigma^2(R^2 = 0.88)$, $\Delta H_c$ ($R^2 = 0.99$) | The ML models were employed to predict the perovskite octahedral deformation and its related stability with high accuracy. |
| 69 | $ABX_3$ perovskites | Bandgap | RF | Mean score is 0.98, standard deviation is 0.002. | Propose 11 undiscovered Li (Na) based perovskite materials with ideal bandgap and formation energy ranges for solar cell applications. |
| 45 | $ABO_3$ perovskites | Formation energy and bandgap | GBR | Formation energy prediction precision ($R^2 = 0.964$); bandgap prediction precision ($R^2 = 0.855$) | Develop a novel progressive learning method with instrumental variables, which provides a new way to expand the feature set and reduce the calculation amount by avoiding expensive DFT calculations. |
| 68 | $ABX_3$ perovskite | Bandgap | Alternating conditional expectations | $R^2 = 0.824$, RMSE = 0.836 eV | Proposed an ML technology suitable for small datasets—alternating conditional expectations (ACE). |
| 75 | $ABO_3$ perovskites | Curie temperature | SVR | R = 0.8549, RMSE = 28.7 | Potential perovskite materials with higher *Tc* were discovered from virtual samples. |
| 76 | $ABO_3$ perovskites | Curie temperature | RR, SVM and ERT | The mean error is 13.9 K | Two lead-based perovskite ferroelectric solid solution materials with higher Curie temperature are proposed, and the predicted values are 481 °C and 466 °C respectively. |

**Table 3** continued

| Ref. | Research object | Target properties | Algorithms | Prediction accuracy | Main achievements |
|---|---|---|---|---|---|
| 79 | $ABO_3$ perovskites | Neel temperature | SVM | The MRE of training set and testing set were 8.0% and 16.4%, respectively. | An effective method for predicting the Neel temperature of $ABO_3$ perovskite using ML method is proposed. |
| 85 | $ABO_3$ perovskites | Maximum magnetic entropy change | GPR | The RMSE and R were 0.0121 and 99.997%, respectively. | Presented an ML method that can predict the maximum magnetic entropy change of $ABX_3$ perovskite with high accuracy. |
| 91 | $ABX_3$ perovskites | Dielectric breakdown strength | KRR, RF and least absolute shrinkage and selection operator (LASSO) | - | They proposed that boron-containing perovskites may be extremely tolerant to high electric fields, and two perovskites ($BSiO_2F$ and $SrBO_2F$) whose predicted results show a breakdown strength of nearly 2 GV/m. |
| 93 | $Ba(Ti_{1\text{-}x\%}Hf_{x\%})O_3$ | Dielectric permittivity | GPR | - | Predict the dielectric permittivity of $Ba(Ti_{1\text{-}x\%}Hf_{x\%})O_3$ based on the ML method, and successfully synthesize the perovskites material with the larger dielectric permittivity under the guidance of the model |
| 97 | $ABO_3$ perovskites | Ionic conductivity | SVM | RMSE = 1.08, MRE = 16.63% | The first use of ML method to predict the ionic conductivity of $ABO_3$ perovskite |
| 98 | $BaNbO_2N$ perovskites | The stability of anion ordering in $BaNbO_2N$ supercells. | RR | Accuracy 94% | Indicates that the most stable perovskite $BaNbO_2N$ supercells had each Nb atom coordinated with two N atoms, along with NbN chains in a cis conformation. |
| 37 | $ABO_3$ perovskites | Specific surface area (SSA) | SVM | R = 0.935 | Five new perovskites with larger SSA and photocatalytic potential are proposed. Develop the established model into an online predictive application for public use. |
| 100 | $ABX_3$ perovskites | Stability, bandgap and spontaneous polarization | GBC and GBR | The prediction accuracy of both energy difference and $E_g$ regressions exceeds 90% | They presented eight candidates of ferroelectric photovoltaic perovskites with excellent thermal stability, appropriate bandgap, and considerable spontaneous polarization. |
| 116 | HOIPs | Bandgap | GBR | The $R^2$ and MSE were 97.0% and 0.086 of testing set | Six Pb-free HOIPs with suitable bandgap and thermal stability at room temperature were selected |
| 118 | HOIPs | Stability | Deep learning (DL) | $R^2 = 0.98$, MAE = 9.52 meV/ion | A series of perovskite $Cs_xMA_{0.85\text{-}x}DMA_{0.15}PbI_3$ was synthesized under the guidance of the model, and it was demonstrated that the cubic perovskite structure could be restored by doping Cs in $MA_{0.85}DMA_{0.15}PbI_3$. |
| 119 | HOIPs: $MAPbI_3$ | Electronic transport properties | GBR | $R^2 = 0.977$ | A rapidly and effectively ML method is proposed to predict the electron transmission coefficients of MAPbI3 with different metal electrodes and tunnel barrier lengths. |
| 120 | HOIPs | Power conversion efficiency (PCE) | RF and association rule | Regular cells: RMSE = 1.7, inverted cells: RMSE = 1.51 | The ML model for predicting the PCE of perovskite solar cells was established and the factors affecting the PCE were mined through association rules. |
| 121 | HOIPs | Bandgap, open-circuit voltage, short-circuit current density, fill factor and PCE | ANN | Bandgap(R = 0.97), PCE(R = 0.8) | The ML method for predicting the performance of perovskite solar cells is proposed, and the perovskites whose experimental bandgap value is highly consistent with the predicted value is synthesized under the guidance of the model. |
| 127 | DP oxides | Bandgap | KRR | The $R^2$ of training set and testing set were 0.993 and 0.947, respectively. | An ML model was developed to predict $AA'BB'O_6$ double perovskite bandgap with high accuracy. |

**Table 3** continued

| Ref. | Research object | Target properties | Algorithms | Prediction accuracy | Main achievements |
|---|---|---|---|---|---|
| 128 | DP oxides | Formability | Medium Tree | Accuracy 95% | Using ML technique to identify the perovskite formation of $A_2BB'O_6$ compounds. |
| 113 | Chalcogenide DPs | Bandgap | RF | Accuracy 86.4% | Screening of five most promising perovskite photovoltaic absorbers based on ML model. |
| 129 | DP halides | Decomposition energy | KRR | $R^2 = 0.92$ | Predicted the decomposition energies of 14,190 possible $A_2B(I)B(III)X_6$ double perovskite halides |
| 130 | DP oxides | Stability | RF | The accuracy and F1 score were 84.87% and 90.60%, respectively. | Twenty-one double perovskite compounds were found via model prediction, including seven ferromagnetic half metals, six ferromagnetic insulators, five antiferromagnetic insulators, two ferromagnetic metals and one antiferromagnetic metal. |
| 131 | DP oxides | Oxygen evolution reaction (OER) activity | GPR | Training set: RMSE 0.2–0.22 eV, testing set RMSE 0.4–0.46 eV | Nine stable cubic perovskites with the best catalytic OER reaction performance are proposed (OER overpotential is about 0.5 V, tolerance factor > 0.90) |

more accurate. However, LOOCV may cost a long time and a lot of computational resources, which is not suitable for large dataset.

Bootstrap method is based on bootstrap sampling[36]. Given a dataset *D* containing m samples, the bootstrapping method would randomly copy a sample from *D* to the dataset *D′* at a time until *D′* contains m samples. In this process, some data may be sampled repeatedly while some data may never be sampled. Finally, *D* is designated as the training set and *D*∪*D′* is used as the testing set. The number of training samples obtained by the bootstrapping method is equal to the original dataset. The bootstrapping method is effective under the condition of a small dataset. Nevertheless, the dataset generated by the bootstrap method may change the distribution of the original dataset and introduce the deviation.

## MODEL APPLICATION

The purpose of ML is to generalize the hidden patterns between descriptors and material properties of existing data samples. The properties could be accurately predicted with the constructed model. Therefore, the developed ML model can be applied to high-throughput screening. First, many virtual samples could be designed, and then the properties could be predicted with ML model. Finally, the materials with desired properties would be selected from the hypothetical samples for the experiments.

To further develop the application of ML model, it has also become one of the hotspots to develop the online prediction model for sharing. The network model enables more users to predict target properties. For example, Shi et al.[37] developed the online server for predicting the specific surface area of $ABO_3$ perovskites. Furmanchuk et al.[38] developed an online application to predict the Seebeck coefficient of crystalline materials. The approach of developing models into online servers not only exposes sharing models, but also makes model applications easier and faster.

## APPLICATIONS OF MACHINE LEARNING IN PEROVSKITE MATERIALS

Perovskite, named after Russian geologist Perovski, originally referred to a specific compound, calcium titanate ($CaTiO_3$). Now it is used to stand for a group of compounds with the same crystal structure as $CaTiO_3$[39]. The structural formula of perovskite material is usually represented as $ABX_3$ or $AA'BB'X_6$, where A and B are cations, the ionic radius of A is larger than that of B, and X usually means halogen ions or oxygen ions with small radius[40]. $ABX_3$ perovskite is called simple perovskite, while $AA'BB'X_6$ perovskite is known as double perovskite (DP). According to whether A-site cations are organic small molecules or metal ions, the simple perovskite could be classified into inorganic perovskite and hybrid organic–inorganic perovskite (HOIP)[39,41]. As shown in Fig. 3a, the ideal structure of simple perovskite generally presents a cubic structure. The eight vertex angles of the cube are occupied with inorganic cations or small organic groups A, the body center position is occupied with cations B, and the six face center

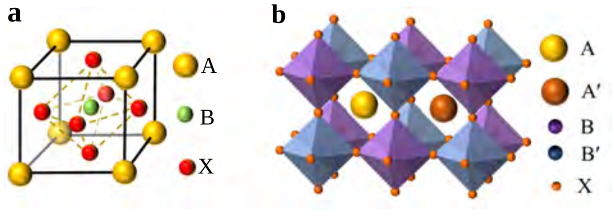

**Fig. 3 Structures of different perovskites. a** Simple perovskite cubic crystal structure and **b** Double perovskite crystal structure.

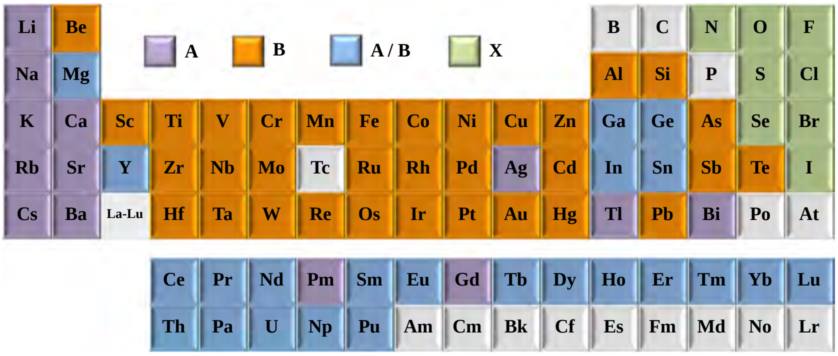
A
B
A / B
X
Li Be B C N O F
Na Mg Al Si P S Cl
K Ca Sc Ti V Cr Mn Fe Co Ni Cu Zn Ga Ge As Se Br
Rb Sr Y Zr Nb Mo Tc Ru Rh Pd Ag Cd In Sn Sb Te I
Cs Ba La-Lu Hf Ta W Re Os Ir Pt Au Hg Tl Pb Bi Po At
Ce Pr Nd Pm Sm Eu Gd Tb Dy Ho Er Tm Yb Lu
Th Pa U Np Pu Am Cm Bk Cf Es Fm Md No Lr

**Fig. 4 The elements of the periodic table to form perovskites.** Elements that have the probability of form the perovskite structure in the periodic table.

positions are occupied with anions X[42,43]. The $BX_6$ regular octahedron consists with six face-centered X anions and the body-centered B cations. Furthermore, the crystal structure of DP can be composed of the regular alternate arrangement of $BX_6$ and $B'X_6$ octahedrons (Fig. 3b). Normally, B and B′ are different transition metals, and A and A′ could be the same or different alkaline-earth or rare-earth metals[44]. Due to the flexibility of perovskite crystal structure, the ions at the Ca, Ti, or O positions of $CaTiO_3$ can be replaced by elements or groups with similar radius, making the types of perovskite rich and diverse. The number of potential perovskites could reach tens of thousands. Taking the element doping into consideration, the potential number of perovskites could easily exceed $10^7$ [41,45]. Up to now, there are about 1000 perovskites that have been developed through experiments[46]. There is still a huge space for stable perovskites to be excavated. It would be a time-consuming and inefficient project to find stable and high-performance perovskites simply by experiments or DFT calculations. Based on many existing experimental and computational data, ML technology has gradually played an important role in the perovskites discovery.

This review mainly introduces the application of ML in the discovery and rational design of $ABX_3$ inorganic perovskites, HOIPs, and DPs. In addition, two-dimensional layered perovskites are often classified as perovskites. However, the related works in layered perovskites are too seldom to discuss.

## $ABX_3$ INORGANIC PEROVSKITES

$ABX_3$ inorganic perovskite is one of the most active materials. The diversity and flexibility make them a wide variety, and also lead to many different material properties, such as ferroelectricity and piezoelectricity[47,48]. In many applications, these properties are unmatched by other known materials, which makes inorganic perovskites greatly important in various fields, such as magnetic refrigeration[49,50], solid oxide fuel cells[51,52], and photocatalysis[53,54].

Theoretically, most elements in the periodic table can replace the A or B of $ABX_3$ to form perovskites (Fig. 4). However, not all compounds with $ABX_3$ stoichiometry are perovskite structures. Therefore, finding an efficient way to determine whether a compound with the formula $ABX_3$ exhibits a perovskite structure has been the first challenge in perovskite discovery and design. In many researches, the Goldschmidt tolerance factor ($t$)[55] (Formula (1)) is usually used to judge the structure formability and phase stability of perovskite. However, the Goldschmidt tolerance factor is insufficient with an increasing variety of perovskites. Some researchers have proposed methods to determine the formability of perovskite structure. For instance, Sun et al.[56] proposed a descriptor based on the tolerance factor and the octahedral factor, which accuracy reached 90%. Bartel et al.[57] developed a tolerance factor (Formula (5)) that can be used to determine the formability of simple perovskite and double perovskite, defining when τ of the compounds less than 4.18 represents perovskite with 91% accuracy.

$$t = \frac{r_A + r_X}{\sqrt{2}(r_B + r_X)} \tag{1}$$

$$\mu = \frac{r_B}{r_X} \tag{2}$$

$$\eta = \frac{V_A + V_B + 3 \cdot V_X}{a^3} \tag{3}$$

$$(\mu + t)^\eta \tag{4}$$

$$\tau = \frac{r_X}{r_B} - n_A\left(n_A - \frac{\frac{r_A}{r_B}}{ln\left(\frac{r_A}{r_B}\right)}\right) \tag{5}$$

Where $r_A$, $r_B$, and $r_X$ are ionic radii of A, B, and X, respectively; $\mu$ is the octahedral factor; η is the atomic packing fraction; $V_A$, $V_B$, and $V_X$ are atomic volumes of A, B, and X, respectively, based on the rigid sphere model; $a$ is the lattice constant of cubic cell; $n_A$ is the oxidation state of A.

Although these descriptors can well evaluate the formability and stability of perovskite with high accuracy, researchers still try to find the factors and patterns controlling the formability of perovskite structure through ML to develop a method that can fully judge the formability and stability of perovskite in a faster and more accurate way. The energy beyond the convex hull ($E_{\mathrm{hull}}$) is a measure of the decomposition of the compound into a linear combination of the stable phases present on the phase diagram[58]. It is significant to evaluate the materials dynamic stability of. Normally, thermodynamically stable compounds have zero $E_{\mathrm{hull}}$, while more positive values of $E_{\mathrm{hull}}$ indicate decreasing stability[59]. $E_{\mathrm{hull}}$ can be calculated by the DFT, but the huge computational costs limit the power of DFT in materials with a large chemical search space. In 2017, Schmidt et al.[60] constructed a dataset containing 250,000 $ABX_3$ compounds, from which about 20,000 $ABX_3$ perovskite compounds were randomly extracted for model construction. An ML model with $E_{hull}$ as the target variable was built to predict the stability of the compound. The ML model was used to predict the thermodynamic stability of the remaining approximately 230,000 virtual samples, in which there were 641 formally candidates with $E_{hull}$ less than 5 meV/atom. Li et al.[61] developed a ML model to predict the thermodynamic phase stability of perovskite oxides using a dataset of more than 1900 $E_{hull}$ predicted by DFT. Two ML models were constructed respectively to classify and regress the $E_{hull}$, and then predict 15 perovskite candidate materials. Finally, four stable perovskites ($La_{0.5}Y_{0.5}$-$Co_{0.5}Mn_{0.5}O_3$、$Y_{0.75}Sr_{0.25}VO_3$、$CeReO_3$ and $Dy_{0.75}Nd_{0.25}RuO_3$) were

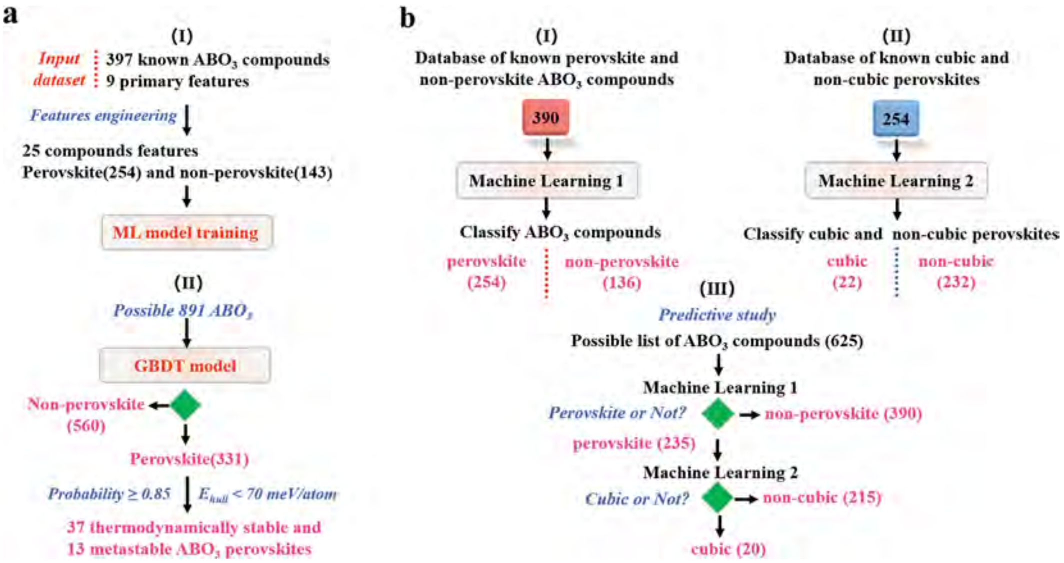

**Fig. 5 Workflows of ML in $ABO_3$ perovskites.** **a** Workflow for predicting stability and metastability of $ABO_3$ perovskite. Reproduced with permission from ref. [62]. Copyright Elsevier 2020 **b** Workflow for the $ABO_3$ cubic perovskite[65].

presented. In 2020, Liu et al.[62] screened stable and metastable $ABO_3$ perovskites using ML and the materials project based on the dataset of 397 $ABO_3$ compounds (Fig. 5a). The ML classification model was applied to divide 891 $ABO_3$ compounds into perovskite and non-perovskite compounds. The results showed that 331 compounds had perovskite structures, in which 174 had a formation probability of ≥85%. In addition, 37 thermodynamically stable $ABO_3$ perovskites (0 meV/atom < $E_{hull}$ < 36 meV/atom) and 13 metastable perovskites (36 meV/atom < $E_{hull}$ < 70 meV/atom) were screened through the ML regression model for further synthesis and application. These researches have proven that the ML model could provide effective guidance to determine the stability of various perovskite oxides.

In addition to $E_{hull}$, the formation energy of compounds could also be used to evaluate the formability and stability of perovskite. Li et al.[63] proposed a transfer learning strategy to evaluate the stability of the $ABX_3$ inorganic perovskites. First, an ML transfer learning model was constructed by taking the formation energies of 570 perovskites as the target variable and the physics-informed structural and elemental parameters of perovskites as descriptors. Then the transfer learning model was applied to predict the formation energies of 578 compounds with unknown target. With the combination of two datasets above, 1148 data were used to train a convolutional neural network model for high-throughput screening. Finally, 764 promising perovskite materials with the tolerance factor τ less than 4.8 were selected from 21316 assumed perovskites by the screening model, 98 of which have been validated to be stable by DFT calculation. In typical ML-based material discovery and large-scale screening of hypothetical perovskites, transfer learning is a recently developing ML method in dealing with the small data problems.

It is also an effective strategy to use the ML model constructed with experimental perovskite and non-perovskite to predict the formation probability of large quantities of unknown potential perovskites. In 2016, Pilania et al.[64] demonstrated the powerful function and practicality of ML via SVM based classifier, which used elemental parameters to evaluate the formability of $ABX_3$ halides in the perovskite crystal structure. After the exploration of vast descriptors, ionic radii, tolerance factor, and octahedral factor are identified as the most crucial related features for the model, indicating that steric and geometric packing effects have a great impact on the stability of these halides[64]. 40 $ABX_3$ with perovskite-type crystal structures were proposed through predicting the perovskite formability of 455 $ABX_3$ compounds with ML. Balachandran et al.[65] developed two decision tree classifiers to acquire many potential perovskite materials and cubic perovskites, as shown in Fig. 5b. Two models with accuracy more than 90% were trained to predict unknown 625 compounds, in which 235 were perovskite and 20 were cubic perovskites. Besides, 87 promising perovskite candidates were selected for further experimental guide. Analyzing the results, potential perovskites may locate at (a) A and B atoms are a lanthanide or actinide elements, (b) A atom is an alkali metal, alkali earth metal or late transition metal atom, or (c) B atom is a p-block element[65].

In 2019, Jain et al.[66] constructed an ML classification model based on SVM with 189 $ABX_3$ inorganic samples to predict the perovskite formability of 454 $ABX_3$ compositions, among which the formation probability of 45 compounds is equal to or higher than 0.8. After comparing the thermodynamic stability information of perovskite in MP, AFLOW, and OQMD, 18 compounds were subject to carry out the DFT-based bulk structural optimizations and electronic structure predictions. According to the overall DFT results, two promising stable photovoltaic candidates, $RbSnCl_3$ and $RbSnBr_3$, were represented for further study. This work is an important step towards a basic understanding of the interfacial properties of perovskites, facilitating further breakthroughs in photovoltaic technology.

Recently, Park et al.[67] proposed a method to identify the stability of perovskite. A series of ML models were developed for the target properties of the perovskite, namely, octahedral deformation parameters including the energy difference ($\Delta H_c$) between the relaxed and ideal cubic structures, quadratic elongation (λ), as well as octahedral angle variance ($\sigma^2$) (Fig. 6a). The possibility of a known cation embedded in the perovskite was systematically analyzed. The influence of A-site cation on the phase stability of the perovskite was evaluated by measuring the degree of octahedral deformation when a given cation embedded in $[BC_6]^{4-}$ [67]. This work shows that the combination of advanced electronic structure theory and ML analysis can provide an effective strategy that is superior to the conventional trial-and-error method in material design. More importantly, it provides a powerful guide for exploring a broad composition space of inorganic and mixed perovskites.

$E_g$ is a significant parameter in the applications of electrical conductivity, light-harvesting capability, photoelectric conversion, and other functions in perovskites, which is directly related with

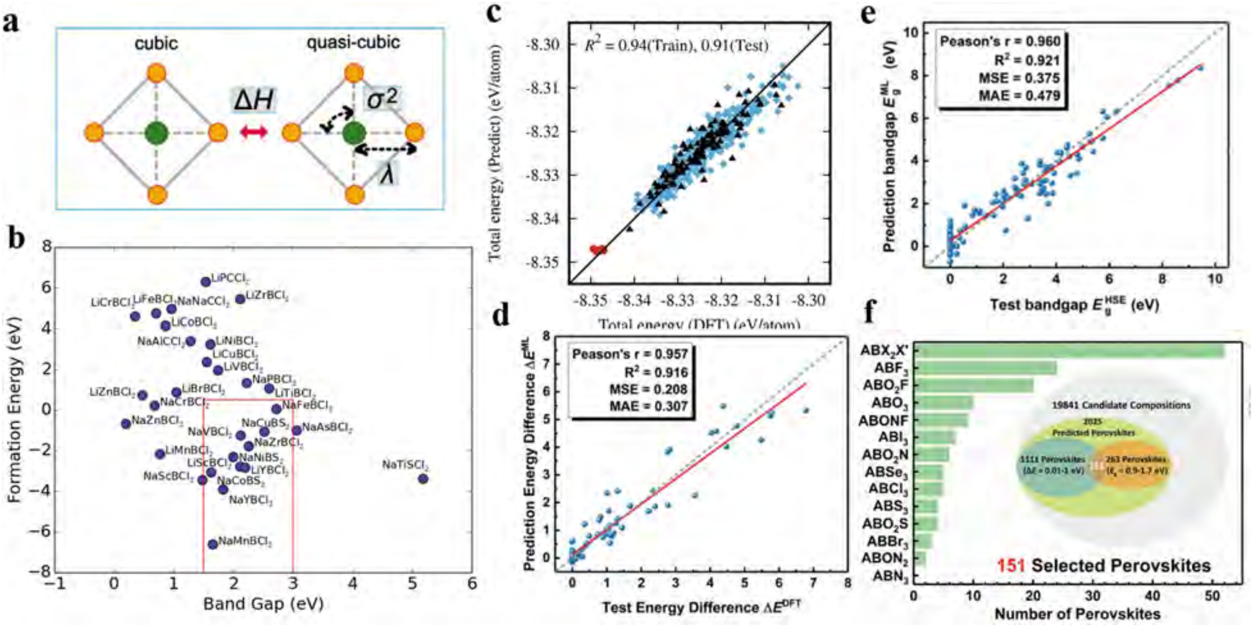

**Fig. 6 Applications of ML in inorganic perovskites.** **a** The parameters Δ*H*, *λ*, and $\sigma^2$ were used to quantify the distortion out of the ideal cubic perovskites[67]. **b** the formation energy and $E_g$ of predicted Li(Na)BX$_3$ perovskite with DFT. The perovskites in the red box have the ideal formation energy and $E_g$ values[69]. **c** Relationship between the total energy values for BaNbO$_2$N supercells in the training set (rhombuses) and the test set (triangles) predicted by RR and the data calculated by DFT. Reproduced with permission from ref. [98]. Copyright Elsevier 2019 **d** The predicted phase-transition energy difference Δ*E* versus DFT calculations. Reproduced with permission from ref. [100]. Copyright John Wiley and Sons, Inc. 2019 **e** Comparison between test bandgap $E_{HSE\ g}$and predicted bandgap $E_{ML\ g}$. Reproduced with permission from ref. [100]. Copyright John Wiley and Sons, Inc. 2019 **f** 151 promising perovskites with different types of X-site compositions. Reproduced with permission from ref. [100]. Copyright John Wiley and Sons, Inc. 2019.

the properties of various photovoltaic devices[68]. The theoretical models of $E_g$ could accelerate the discovery of perovskites, and help navigate the broad space of potential perovskite materials, and guide chemists to screen out candidates for experiments.

In 2018, Takahashi et al.[69] used the RF to predict the $E_g$ of ABX$_3$ perovskite to determine whether the $E_g$ values of the candidates meet the requirement of the applicable range of solar cells (1.7–3.0 eV). After model training with the $E_g$ data of 15,000 perovskite materials, 9328 potential perovskite materials with $E_g$ at the range of 1.7–3.0 eV were extracted from 414,736 candidates. Then the $E_g$ values of the selected candidates based on Li and Na were calculated and evaluated with DFT, where 11 undiscovered Li (Na) based perovskite materials fell into the ideal $E_g$ and formation energy ranges for solar cell applications (Fig. 6b). In addition to using the classification models to screen promising candidates with appropriate $E_g$, the ML regression models also have excellent predictive performance. Li et al.[45] constructed a ML model with ABO$_3$ perovskite formation energy ($E_f$) as the target. Then, the $E_f$ predicted by the model was used as the instrumental variable to build a progressive learning model to predict the $E_g$ of the perovskite materials. The results of the model indicated that the addition of predicted $E_f$ as an instrumental descriptor can promote the prediction accuracy of $E_g$ regression model ($R^2 =$ 0.855). This progressive learning strategy with instrumental descriptors provides an approach to widen the feature pool and reduce the computational effort instead of high-cost DFT calculations.

Even in the era of big data, limited samples are still the majority in material science. How to make full use of the limited samples for ML has also been a research potential in recent years. It is believed that the emergence of each method for small sample datasets would bring a bit of dawn to the development of ML in material science. Gladkikh et al.[68] presented an ML technology suitable for small datasets-alternating conditional expectations (ACE). ACE has an advantage in that it shows the results in a graphic form, which can help for model interpretation. The graphic form of the ACE transformations can view the impact of each descriptor on the target property. Furthermore, ACE does not suffer from the curse of dimensionality due to it is estimated by univariate functions. They used ACE to study nonlinear mappings between $E_g$ and descriptors of component elements and constructed a model to predict the $E_g$ of the perovskites. The $R^2$ and RMSE of the training set were 0.824 and 0.836 eV, respectively. Their study indicated that the $E_g$ values of ABO$_3$ perovskites mostly depend on the electronegativities, electron affinities, ionization energies, and atomic radii of the constituents.

The critical temperature at which ferroelectric materials convert from the ferroelectric to the paraelectric phase is called Curie temperature (*Tc*), also known as the Curie point[70]. *Tc* has been a key indicator in property measurement of ferroelectric materials. Most inorganic perovskites represented by ABO$_3$ structure have excellent ferroelectric properties and become one of the most promising materials for electronic and magnetic components such as multilayer capacitors and sensors[71,72]. *Tc* has a considerable influence on many applications of perovskite materials in the magnetic recording, sensor, actuators, and refrigeration[73,74]. Therefore, it is quite meaningful to predict *Tc* of perovskite materials quickly and effectively before experiments.

In 2018, Zhai et al.[75] developed a prediction model of *Tc* with physicochemical parameters based on ML. In the meanwhile, the potential perovskite material (La$_{0.66}$Sr$_{0.3}$Ba$_{0.04}$MnO$_3$) with high *Tc* of 390.35 °C were found from the virtual samples by the SVR model combined with the genetic algorithm search strategy. Similarly, Yang et al.[76] used RR, SVM, ERT and other ML methods to train the *Tc* of lead-based perovskite ferroelectrics. The ML model integrating with the above three algorithms was used to predict the *Tc* of more than 200,000 kinds of lead-based perovskite materials outside the dataset. Then two lead-based perovskite solid solution ferroelectric materials were screened with high *Tc* of 481 °C and 466 °C were selected for further experiments. In addition, the integrated ML model was also applied to analyze the *Tc* prediction results of $PbGa_{1/2}Nb_{1/2}$-$PbMn_{1/2}Nb_{1/2}O_3$-$PbTiO_3$ system.

Neel temperature ($T_N$) is the critical temperature at which antiferromagnetic material becomes paramagnetic[77]. It has been reported that $T_N$ is closely related to the applications of $ABO_3$ perovskite in the fields of magnetic refrigeration, colossal magnetoresistance, etc[78,79]. Therefore, accurate and rapid prediction of $T_N$ is a very significant work in the design and discovery of perovskite oxides. Xiao et al.[79] mapped the relationship between the main atomic parameters of Mn-based perovskite oxide and $T_N$ with SVM. It is worth noting that SVM is an algorithm especially appropriate for small sample datasets, which can build ML model with high generalization in a limited sample size. This work is helpful for the simple and rapid prediction of the $T_N$ of Mn-based perovskite.

Energy efficiency and sustainable development are the priority topics in modern society. However, refrigeration and air-conditioning consume a large amount of electric energy among various end-uses of energy in both commercial and residential areas[80]. Most refrigeration technologies rely on traditional conventional gas compression technologies, which have come under increasing criticism for their inefficiency and the use of air pollutant gases. The latest development of magnetic refrigeration technology based on the magnetocaloric effect of magnetic materials (especially near room temperature) has provided a promising alternative to vapor compression refrigeration[81,82]. In order to design a magnetic refrigerator with an operating temperature close to room temperature, much attention has been paid to the magnetocaloric material with a large maximum magnetic entropy change (MMEC) over a wide temperature range[83,84]. Zhang et al.[85] established a GPR model to elucidate the statistical relationship between the MMEC and lattice parameters of magnetocaloric lanthanum manganite perovskites. The model demonstrated a high accuracy and stability with RMSE, MAE and correlation coefficients being 0.0121, 0.0054, and 99.997%, respectively. In addition, the model could be used as part of ML to get a better understanding of magnetic phase transformations and magnetocaloric effects in various types of doped magnetocaloric lanthanum manganite.

The dielectric breakdown strength refers to the highest electric field strength that a material can withstand without being destroyed under the action of an electric field, which is the key property to assess the performance of electrical and electronic devices[86–88]. Dielectric materials with high dielectric breakdown strength are necessary for high energy density electric energy storage applications in combination with continued miniaturization of electronic devices[89,90]. It is not only determined by the intrinsic factors of the material (chemical constituents, nature of the chemical bonding, crystal structure, etc.) but also affected by the extrinsic factors (defects, morphology, impurities, degradation, interfaces, etc.)[91,92] Therefore, it is very challenging to accurately calculate the dielectric breakdown strength of complex materials entirely by DFT method and perform high-throughput screening from a large number of promising candidates. By contrast, ML may be a more potential approach to predict dielectric breakdown strength.

Kim et al.[91] applied the ML technology to train and validate on a limited amount of accurate data from DFT calculations, then to predict the dielectric breakdown strength of hundreds of $ABX_3$ compounds in a highly efficient manner. After making predictions on these compounds using the ML model, the dielectric breakdown strength of the most promising candidates was further validated by DFT calculations. The research results have shown that boron-containing perovskites may be extremely tolerant toward high electric fields. The prediction results of $BSiO_2F$ and $SrBO_2F$ showed a breakdown strength of almost 2 GV/m, which is worthy for further experimental studies. Gao et al.[93] studied the dielectric permittivity of perovskites based on ML. They employed the GPR algorithm to obtain the relationship between the composition of perovskites and the dielectric permittivity to find the maximum dielectric permittivity in Ba $(Ti_{1\text{-}x\%}Hf_{x\%})O_3$ ceramic material. According to ML prediction, the optimal composition is found to be x = 11 with the highest dielectric permittivity $\varepsilon_r = 4.5 \times 10^4$. The predicted materials are synthesized experimentally to further verify the accuracy of the model. This strategy combined with ML shows higher efficiency compared with the traditional experimental search.

$ABO_3$-type perovskite oxides have also been considered as the potential materials for solid electrolytes in solid oxide fuel cells (SOFCs). Conductivity is an essential parameter to describe the ease of charge flow in a material. In addition to external factors like oxygen pressures and operating temperature, the conductivity of perovskite oxides is also affected by its composition and structure[94–96]. To discover or design perovskite oxides with high ionic conductivity, it is necessary to figure out the relationships between the molecular composition parameters and the oxygen ionic conductivity of the perovskite oxides. Liu et al[97]. explored the correlation between atomic parameters and ionic conductivity properties of 117 perovskite oxide data via partial least squares, backpropagation artificial neural network and SVR, in which model constructed by SVR processed the best generalization. It was found that and the ratio of O–O charge population to the O–O band length (*P/L*) and logarithm of oxide ionic conductivity (*Lnσ*) have a quadratic curving relationship. The value of *P/L* is one of the important quantum chemical parameters to predict the ion conductivity of perovskite oxides. Based on the calculation of *P/L*, a semi-empirical formula can be used to predict the oxide ion conductivity of the doped $ABO_3$ perovskite.

Kaneko et al.[98] proposed a regression model built by ML based on the data with DFT calculations to predict the stability of anion ordering in perovskite-type $BaNbO_2N$ supercells. DFT was used to calculate the total energies of 560 small $BaNbO_2N$ supercells with random anion ordering. Using the total energy of 420 $BaNbO_2N$ supercells as the training set, an ML model was established with prediction accuracy reaching 94% (Fig. 6c). The conclusion indicates that the most stable perovskite $BaNbO_2N$ supercells had each Nb atom coordinated with two N atoms, along with NbN chains in a cis conformation[98]. This work has suggested an approach for the property predictions of complex-compositions materials at a reasonable computational cost and provided guidance for the design of stable perovskite oxynitrides.

The specific surface area (SSA) of the photocatalyst plays a significant role in the photocatalytic reaction. Generally, the larger the SSA of the photocatalyst is, the more reaction sites and the better the photocatalytic performance are. $ABO_3$ perovskite has been widely applied as the photocatalyst or photocatalytic active component in photocatalytic reactions. Shi Li et al.[37] used GA and SVM algorithms to explore the relationship between the SSA of perovskites and the composition as well as experimental conditions. After virtual screening with the developed model, five visual perovskites with larger SSA and photocatalytic potential were proposed. In addition, the author has also developed the established model into an online forecasting application, making the model more available to researchers to predict the required large SSA perovskites. This method should be extended to ML-aided design of other properties and other materials.

Zheng et al.[99] established a series of models by RF, RR and SVM with the electronegativity of atoms at A, B and the effective atomic radii of atoms at A, B, and X as descriptors for the predictions of four properties including density and formation energy, $E_g$ and crystal volume. The results showed that RF method could effectively predict the density and $E_g$ of perovskite materials; RR method could realize the prediction of density; SVM with linear kernel function method could achieve the prediction of formation energy. The research demonstrated that different ML algorithms have different sensitivity to the distribution of data samples. In the process of building ML models with different properties, different algorithms need to be evaluated and screened to optimize the evaluation function.

Lu et al.[100] combined DFT calculation and ML technology to propose a multistep screening scheme for all-inorganic perovskite with stability, high spontaneous polarization, and proper $E_g$. The phase-transition energy difference was adopted as the target property to directly judge whether the compound can be exposed spontaneous polarization. As shown in Fig. 6d, e, the ML prediction accuracy of both energy difference and $E_g$ regressions exceeds 90%, which is highly consistent with DFT calculations. After screening, 151 promising ferroelectric photovoltaic (FPV) perovskites were successfully extracted from 19,841 compositions (Fig. 6f). The accuracy of the ML predictions is further verified by DFT calculations, and 8 randomly selected FPV perovskites exhibited good thermal stability, appropriate $E_g$ (1.01–1.62 eV), and considerable spontaneous polarization (7.10–32.78 μC cm$^{-2}$). This scheme realized the ML for accelerating the material design of multi-property and the extension of materials database.

## HYBRID ORGANIC–INORGANIC PEROVSKITE

HOIPs have become a major hotspot in the field of optoelectronics in recent years due to its easy synthesis, low cost and excellent optoelectronic properties, such as tunable optical $E_g$ high optical absorption coefficient, high carrier mobility, and long load of diffusion length[101,102]. It has been widely applied in fields of solar cells[103,104], light-emitting diodes[105,106], and photodetectors[107,108], and its performance is comparable to traditional materials. In addition, the development of HOIPs is still in continuous improvement and breakthroughs.

In 2009, Kojima et al.[109] used perovskite-type organic–inorganic hybrid materials to prepare thin film solar cells and obtained a 3.8% power conversion efficiency (PCE). Since then, perovskite-type solar cells (PSCs) have attracted many interests of researchers for the huge development potential and the title of new hope in the field of photovoltaic[110]. The highest certified PCE of PSCs to date has reached 25.2%, according to the National Renewable Energy Laboratory[111,112]. It is reported that Pb is the key factor in the high performance of PSCs due to the strong antibonding coupling between the 6 s lone pairs of Pb and the 5p states of I, resulting in a small effectively masses and a direct $E_g$ with a p-p transition[113]. However, Pb based halide perovskites are easily degraded spontaneously under exposure to moisture, air, light, heat, and other environments, resulting in the degradation product of carcinogenic $PbI_2$[114,115]. These obvious shortcomings have hindered the industrial application of HOIPs solar cells, prompting researchers to seek high-performance perovskite materials with better chemical stability and environmentally friendly composition. ML method may accelerate the discovery of such materials.

The $E_g$ of HOIPs is an important parameter for evaluating high-efficiency photovoltaic perovskite materials. Electrons in the valence band could be excited to the conduction band only under the condition of enough energy. Therefore, a comprehensive understanding of $E_g$ and its relationship with HOIPs composition and structure would be very necessary before looking for photovoltaic materials with high light absorption coefficient. Lu et al.[116] developed a target-driven method based on ML and DFT calculations to discover stable Pb-free HOIPs. Taking the $E_g$ values of 212 HOIPs calculated by DFT as the training set, an ML model was built based on the GBR algorithm to predict the $E_g$ values of 5158 unexplored possible HOIPs. After further screening, six orthorhombic lead-free HOIPs with proper $E_g$ for solar cells and room temperature thermal stability were selected. And two of them have direct $E_g$ in the visible region, excellent thermal stability, and excellent environmental stability. As shown in Fig. 7a, the maximum error of $E_g$ obtained by ML prediction and DFT calculation is less than 0.1 eV, which shows that ML has a huge advantage in $E_g$ prediction, and its accuracy is comparable to DFT calculation. The workload of DFT calculation has greatly

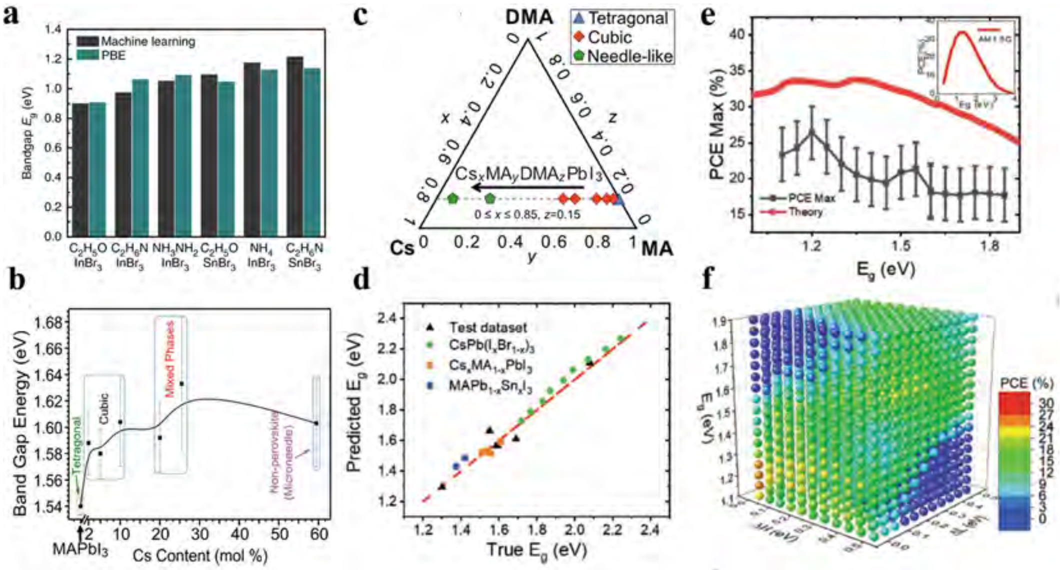

**Fig. 7 Applications of ML in HOIPs. a** A comparison of ML-predicted with DFT-calculated data of six HOIPs[116]. **b** The $E_g$ energy changes with Cs (mol %) in $Cs_xMA_{0.85\text{-}x}DMA_{0.15}PbI_3$ and the cubic structure begins to recover when x = 0.02[118]. **c** Ternary diagram denoting the different-stoichiometry crystal structures[118]. **d** The correlation of the ML $E_g$ data and the experimental $E_g$ data. Reproduced with permission from ref. [121]. Copyright John Wiley and Sons, Inc. 2019 **e** Blackline: the maximum PCE predicted by ML corresponding to each $E_g$ (1.2–1.3 eV); redline: the PCE of Shockley–Queisser limit. Reproduced with permission from ref. [121]. Copyright John Wiley and Sons, Inc. 2019 **f** 4D-plot of PCE with respect to $E_g$, $\Delta H$, and $\Delta L$, indicating the highest PCE values with the bandgap in the range of 1.2–1.3 eV. Reproduced with permission from ref. [121]. Copyright John Wiley and Sons, Inc. 2019.

reduced with the assist of ML, which is very important for large-scale screening of materials. Besides, HOIPs with small $E_g$ (less than 0.9 eV) can be used in infrared sensors, and large $E_g$ of HOIPs (larger than 3 eV) may serve as good insulating materials. Therefore, ML not only accelerates the prediction of $E_g$ in photovoltaic materials but also in other related fields.

In 2020, Saidi et al.[117] used DFT to calculate the $E_g$ and structural parameters of 862 HOIPs for modelling. Then a hierarchical convolutional neural network (CNN) was used to construct an ML model to predict the $E_g$ of HOIPs. The results show that the lattice constant and the octahedral till angle play the key role in the prediction of the $E_g$. When these two features are removed from the dataset, the RMSE increases from 0.07 to 0.16 eV. In addition, applying hierarchical CNN to alleviate problems related to the imbalanced target values is also the key to success. In material design, small samples are a common problem, which are usually unevenly distributed. And this well-designed hierarchical ML approach is expected to be used in the design of other materials with uneven data distribution.

In addition to $E_g$, stability is also a key parameter affecting the overall performance of PSCs. Recently, Ali et al.[118] constructed a dataset of A-site cation-doped HOIPs containing 852 data with the target of the energy difference ($\Delta H_C$) between the of cubic structure and the fully relaxed structure and 12 descriptors. These descriptors include the period and group numbers, the effective radius and the number of lone pairs to describe the A-site cations, the ionization energy and the electron affinity of the inorganic elements in B- and X-sites in combination with the tolerate factor and the octahedral factor[118]. The deep learning method was employed to train the model to predict the cubic phase stability, which was further applied to accelerate the search and discovery of HOIPs with stable cubic phase from the enormous material search space. A series of mixed-cation perovskites were synthesized under the guidance of the model, namely $Cs_xMA_{0.85-x}DMA_{0.15}PbI_3$, where x = 0, 0.02, 0.05, 0.1, 0.2, 0.255, 0.595, and 0.765[118]. The experimental characterization results indicate that the cubic structure began to be restored when x equals 2 mol% Cs in $Cs_xMA_{0.85-x}DMA_{0.15}PbI_3$ (Fig. 7b). This work also showed that the cubic structure could be recovered through converting the severely unstable double-cation perovskite ($MA_{0.85}DMA_{0.15}PbI_3$) in the cubic structure at room temperature into a triple-cation compound by the incorporation of Cs cation (Fig. 7c)[118]. The work shows that the ML perovskite structure stability prediction model has greatly sped up the experimental process of cubic perovskites and reduced experimental costs.

There are many factors that affect the applications of HOIPs in photovoltaics. Li et al.[119] using the ML approach and non-equilibrium Green's function together with DFT to explore the electronic transport properties of $MAPbI_3$. The band structure of $MAPbI_3$ calculated with DFT indicated that the ferroelectric and antiferroelectric dipole configurations have very little effect on the $E_g$[119]. They tested the tunnel junctions composed of $MAPbI_3$ and 48 different metal electrodes with the same fixed lattice constant as $MAPbI_3$ and found that the electron transmission coefficient of Mg electrodes is the highest, and the conductivity of the Pt electrodes is the least[119]. In addition, as the perovskite unit cell number increases, the electron transmission coefficients usually exponentially decrease. The ML algorithms were employed to explore the correlations of the transport properties of $MAPbI_3$ with different metal electrodes and tunnel barrier lengths[119]. This work could quickly and effectively predict the electron transmission coefficients of $MAPbI_3$ under different metal electrodes and different tunnel barrier lengths, thereby stimulating more experimental and theoretical interests in other tunnel junction systems and electron transport problems with the "DFT+ML" strategy[119].

PCE is a momentous indicator to evaluate the performance of solar cells. It would be strategically significant to study the PCE of reported PSCs with ML. In 2019, Odabaşı et al.[120] collected 1921 samples of HOIPs solar cell devices to propose an effective strategy to improve the PCE of PSCs. RF algorithm was used to build the ML model for predicting the PCE of PSCs. The RMSE for training set and testing set were 1.70 and 3.29 for regular cells, 1.51 and 2.91 for the inverted cells, respectively. In addition, the factors were explored with association rules to provide theoretical guidance for the design of PSCs with high PCE. The results revealed that the factors like mixed-cation perovskites, dimethylformamide and dimethyl sulfoxide as solvents, chlorobenzene as the antisolvent were crucial to obtain the PSCs with PCE higher than 18.0%. Li et al.[121] established a ML model for predicting $E_g$ of perovskite materials with the material composition as descriptors. Taking the perovskite $E_g$, the energy difference ($\Delta H$) between the HOMO of the hole transport layers and the HOMO of the perovskite material, and the energy difference ($\Delta L$) between the LUMO of the perovskite material and the LUMO of the electron transport layers as features, a series of ML models were established to predict the open-circuit voltage ($V_{oc}$), short-circuit current density ($J_{sc}$) and fill factor ($FF$) of PSCs. The performances of PSCs and the physical principles behind getting high-performance PSCs devices were fully studied based on the models. Moreover, perovskite materials were synthesized experimentally to verify the model. As shown in Fig. 7d, the $E_g$ of the synthesized perovskite materials was highly consistent with the result predicted by ML, which strongly proved the reliability of the ML model prediction. In addition, the PCE tendency of the PSCs predicted by the model was also consistent with that by the theory of the Shockley–Queisser limit (Fig. 7e). The relationship between $E_g$, $\Delta L$, $\Delta H$, and PCE was further analyzed to derive a strategy for developing high-performance PSCs with different $E_g$ (Fig. 7f). These findings indicate that ML has been very promising in terms of properties prediction and a deeper understanding of the physical phenomena associated with PSCs.

## DOUBLE PEROVSKITE

In order to solve the instability and toxicity of HOIPs and the wide b $E_g$ of $ABO_3$ perovskite, the researchers replaced A-site or B-site cations of perovskite with two cations, forming a type of stable perovskite called double perovskites (DPs)[122,123]. Theoretically, DPs could achieve both the excellent performance of HOIPs and the stability of $ABO_3$ inorganic perovskite, but the properties of DPs reported so far is not ideal. For example, $Cs_2AgBiBr_6$, which has been more popular in the DPs research direction recently, showed only 2.79% PCE in PSCs application, and the hydrogen production rate in photocatalytic water splitting was 48.9188 μmol $h^{-1}$ $g^{-1}$ [124,125]. The average oxygen production rate and hydrogen production rate of $Sr_2CoWO_6$ in the application of photocatalytic water splitting were 188 μmol $h^{-1}$ $g^{-1}$ and 30 μmolэh$^{-1}$ $g^{-1}$, respectively[126]. These properties are still relatively limited compared with more mature perovskite materials. Therefore, exploring high-performance DPs materials still has huge research and development prospects.

In 2016, Pilania et al.[127] proposed a robust ML model based on elemental descriptors, which effectively predicted the electronic $E_g$ values of $AA'BB'O_6$ double perovskite. The statistical learning model of KRR was used to train and test the dataset consisting of the $E_g$ values of ~1300 double perovskites calculated with Gritsenko, van Leeuwen, van Lenthe, and Baerends potential and further optimized for solids (GLLB-SC) functional. The most important chemical pattern derived from the adopted learning framework is that the $E_g$ is mainly controlled with the LUMO energy of the A-site 1 of the B-site. The $R^2$ of cross validation of the best model reached 0.993 and the RMSE was 0.132 eV; the $R^2$ of the test set was 0.947 and the RMSE was 0.36 eV. The results of the test set proved the strong generalization ability of the ML model and its high consistence with the DFT calculation results. In addition, this ML technique can be applied to any materials in a restricted chemical space with a given crystal structure to obtain the accurate prediction of $E_g$. In 2018, Xu et al.[128] developed a procedure to identify the perovskites formability of all $ABX_3$ and $AA'BB'X_6$ compounds stored in the Materials Projects database. This program could identify the perovskite-forming properties of $ABX_3$ and $A_2BB'X_6$ compounds with the crystal structure stored in the material project database. A variety of ML algorithms are employed to comprehensively analyze the correlation between atomic number, ionic radius, electronegativity, tolerance factor, and octahedral factor and perovskite formation to provide an intuitive view of these data. The prediction accuracy of best ML model reached more than 90%, which was used to identify suspicious data about the perovskite formation of $A_2BB'O_6$ compounds. Excluding those suspicious data, ML could achieve a prediction accuracy of up to 96.3%. In addition, the program also identified 11 $ABO_3$ compounds, which showed different formative properties compared with previous publications. This work has largely enriched the perovskite formability and corrected the possible errors in the previous data of the $ABO_3$ compounds.

In 2019, Agiorgousis et al.[113] used ML to explore chalcogenide DPs to identify photovoltaic absorbers that can replace $CH_3NH_3PbI_3$. After considering the thermodynamic stability, kinetic stability, and optical absorption, five promising perovskite photovoltaic absorbers ($Ba_2AlNbS_6$, $Ba_2GaNbS_6$, $Ca_2GaNbS_6$, $Sr_2InNbS_6$, and $Ba_2SnHfS_6$) were screened from more than 450 possible chalcogenide DPs candidates. Li et al.[129] proposed a strategy with the combination of ML and DFT to engineer stable halide DPs. By choosing 283 DFT-calculated perovskite decomposition energy ($\Delta H_D$) as the training set, the ML mapping between the stability of the perovskite and the compositional ionic radii was established. The ML model was applied to predict the $\Delta H_D$ of 14190 possible $A_2B(I)B(III)X_6$ type halide DPs, in which 2275 were stable ($\Delta H_D > 0$) and 11915 were unstable ($\Delta H_D < 0$). The ML method combined with DFT calculation could not only provide guidance for the experimental engineering of stable perovskites, but also offer enlightenment for the design and discovery of other materials without redundant experimental engineering and complex calculation simulation process.

Magnetism is the significant property of materials in many different applications. In 2019, Halder et al.[130] used a combination method of computational tools to predict virtual magnetic DPs: an ML technique for the screen of stable candidate DPs, an evolutionary algorithm for the determination of crystal structure, and DFT calculations for characterization of electronic and magnetic properties. ML technique was applied to screen the most likely B/B′ combination to predict a stable perovskite structure. Among the 412 screened candidates of $A_2BB'O_6$ composition with 3d, 4d or 5d transition metals at B and B′ sites, 33 compounds were found to form stable DP structures, 25 of which were further considered for characterization of their structure and properties. Twenty-one DPs with different magnetic and electronic properties are predicted, ranging from ferromagnetic half metals to ferromagnetic, from antiferromagnetic insulators to ferromagnetic metals, and then to a rare example of antiferromagnetic metals. This ML study is expected to help the discovery of magnetic DPs.

It is very challenging to solve the model overfitting caused by data scarcity. In 2020, Li et al.[131] developed an adaptive learning strategy to find high-performance $AA'B_2O_6$ cubic perovskites for catalyzing the oxygen evolution reaction (OER). Through mapping the correlations between a large amount of available informatics and the adsorption energies (i.e., *O and *OH), the probabilistic Gaussian processes quickly estimated the adsorption energies of reaction intermediates and the corresponding uncertainties of a rich material space. This adaptive learning strategy gradually improves the robustness of the model by verifying promising samples, albeit with large uncertainties. After iteratively validating/refining the candidates with theoretical overpotentials <0.5 V, an excellent ML model with RMSE less than 0.5 eV was attained. The model rapidly predicted nearly 4000 $AA'B_2O_6$ compounds and proposed nine stable cubic perovskite candidates with the optimal OER performance (OER overpotential is about 0.5 V, tolerance factor > 0.9): $KRbCo_2O_6$, $BaSrCo_2O_6$, $KBaCo_2O_6$, $KCaCo_2O_6$, $BaPbTi_2O_6$, $BaRbTi_2O_6$, $BaSnTi_2O_6$, $BaTnTi_2O_6$, $RbEuTi_2O_6$. Furthermore, they also revealed the potential relationship between the electronic structure descriptors and the OER activity of the perovskites, indicating that the orbital electronic structure characteristics of the B-site ion might be latent factors governing the OER activity. This work indicated that adaptive learning is a cost-effective strategy that can reduce the uncertainty of model predictions in high-dimensional feature spaces with the least computational cost.

## CONCLUSIONS AND OUTLOOK

This paper has briefly summarized the basic process of the ML method in material discovery and design and reviewed part of applications of ML in the large-scale screening and rational design of perovskite materials. The applications of ML in perovskites can be divided into the following four categories. The first type is using ML to explore the better evaluation indexes to describe the stability of perovskite materials. The second type aims to perform the high-throughput screen with the constructed ML model and many virtual samples to screen out the potential perovskite candidates with better properties for experimental guidance. The third type is to deeply dig out the relationship between the descriptors and perovskite properties to get a better understanding of the properties. The last type is the combination of ML and DFT calculation to deal with the problem of limited data. From this review, we have realized that ML has great potential and advantages in discovering materials and revealing the relationship between structural, compositional, and technological descriptors and performance based on known material information. In spite of some successful researches, the applications of ML in material research are still in its infancy and a lot of work needs to be further deepened in the future. Here, we propose some possible directions for ML in the field of perovskite materials:

(1) The combination of ML models and experiments/simulations: The ML development in the field of perovskites is still in its infancy. ML has focused on only a small part of the many excellent material properties of perovskite. Therefore, ML should be used to predict more material properties of perovskites and optimize the synthesis process of perovskites. Besides, ML could be effectively combined with DFT, molecular dynamics, Monte Carlo, and other theoretical simulation methods to accelerate the screening of large-scale perovskites and other materials. More importantly, experiments are the basis for the synthesis and characterization of materials. Therefore, it is necessary to strengthen the combination of ML and experiment to shorten experiment time, reduce experiment cost, and improve experiment efficiency. There are many recent studies on the combination of ML and perovskite experiments[132–135]. Sun et al.[132] used deep neural network methods to build ML models based on experimental X-ray diffraction data to assist in structural analysis. The data synthesized by the experiment could serve back to ML again to increase the amount of data and then improve the generalization ability of the ML model. For example, the perovskites with required specific surface area (SSA) values could be discovered by integrating ML with experiments. First, a ML model could be constructed with collected data to predict the perovskite SSA. The potential candidates would be screened out after visual screening for experiments. Then the experimental data could be added back to the dataset for model reconstruction. The loop would keep performing until the perovskites with targeted SSA are obtained.

(2) The establishment and sharing of perovskite databases: ML is the data-driven method that strongly depends on the quantity and quality of data. Compared with speech recognition, image processing, and other fields with millions of data, the amount of data in material science is extremely limited. The ML method is more prone to overfitting with limited data, leading to the reduction of the generalization ability of the ML model[126]. Although there has existed a database containing a large amount of materials data, more data in the published papers has not yet been entered into the database. It is necessary to establish a more comprehensive, more standard, and more general perovskite information database to speed up the realization of data sharing and reduce the barriers to data access. In the meanwhile, researchers could also obtain more theoretical data through high-throughput calculations, as well as develop methods for intelligently reading literature, access and obtain a large number of related experimental and theoretical data from publications and enter these data into the database.

(3) Development of ML algorithms for small samples: Many powerful ML algorithms have been developed to be successfully applied in various fields. However, these algorithms usually have their own limitations, such as not suitable for small sample data, difficult to adjust parameters, etc. Therefore, developing faster, more accurate, advanced, and intelligent learning algorithms to deal with the challenge of insufficient data would be very indispensable, especially when most data in publications about ML in perovskite materials belong to small samples. A common method to deal with small samples is meta-learning, that is, learning knowledge within or across a specific field[136,137]. The development of new technologies such as neural Turing machines[138] and imitation learning[139] could make it possible. It has recently been reported that the Bayesian program learning framework can reach the level of human experience through one-shot learning under limited data conditions[140]. This may have a huge boost in materials science with scarce and expensive data. It would greatly improve the applicability of the ML method in perovskites and improve the efficiency and generalization of the model.

(4) ML computation platform: The current ML work is more about using programming languages to call various ML algorithms for modeling. For many non-computer researchers, it would very inconvenient and difficult due to the lack of basic programming knowledge. Even if some ML platforms and toolkits have been developed, problems of higher fees, fewer algorithms, simple functions, and inaccessibility are particularly prominent. Therefore, it is urgent to develop computing platforms with free access, complete algorithms, powerful functions, and smart computing. In addition, associating computing platforms with various material databases is also an in-depth direction.

(5) Descriptor interpretation and construction: The predictions or decisions made by ML are mainly based on classical probability theory and mathematical statistics. The physical and chemical meanings of the model still need further research and explanation. Therefore, discovering physical descriptors and making the black box model of statistical ML interpretable is a promising direction for data-driven perovskites. It would not only help experimenters to quickly design and screen visual materials with desired targets, but also enable them to understand the underlying physical laws behind the characteristics for further perovskite design. Alternatively, the accurate and interpretable descriptors could be created with existing descriptors, domain knowledge and ML algorithms.

In summary, with the continuous improvement of high-tech requirements for materials and the rapid development of computer technology and computational methods, ML will be more widely applied in other materials. It is believed that ML will become an indispensable auxiliary tool for experiments and computations in the field of materials science in the future.

## DATA AVAILABILITY

All the data of the examples could be obtained from the corresponding references.

Received: 2 November 2020; Accepted: 6 January 2021;
Published online: 29 January 2021

## REFERENCES

1. Oró-Solé, J. et al. Synthesis, anion order and magnetic properties of RVO$_{3-x}$N$_x$ perovskites (R= La, Pr, Nd; 0 ≤ x ≤ 1). *J. Mater. Chem. C.* **2**, 2212–2220 (2014).
2. Shiogai, J. et al. Signature of band inversion in the perovskite thin-film alloys BaSn$_{1-x}$Pb$_x$O$_3$. *Phys. Rev. B* **101**, 125125 (2020).
3. Veldhuis, S. A. et al. Perovskite materials for light-emitting diodes and lasers. *Adv. Mater.* **28**, 6804–6834 (2016).
4. Wang, Y. et al. Mixed-dimensional self-assembly organic–inorganic perovskite microcrystals for stable and efficient photodetectors. *Mater. Chem. C.* **8**, 5399–5408 (2020).
5. Ekström, E. et al. The effects of microstructure, Nb content and secondary Ruddlesden–Popper phase on thermoelectric properties in perovskite CaMn$_{1-x}$Nb$_x$O$_3$ (x = 0-0.10) thin films. *RSC Adv.* **10**, 7918–7926 (2020).
6. Sydorchuk, V. et al. PrCo$_{1-x}$Fe$_x$O$_3$ perovskite powders for possible photocatalytic applications. *Res. Chem. Intermediat.* **46**, 1909–1930 (2020).
7. Li, L. et al. A novel double-perovskite LiLaMgTeO$_6$: Mn$^{4+}$ far-red phosphor for indoor plant cultivation white LEDs: Crystal and electronic structure, and photoluminescence properties. *J. Alloy. Compd.* **832**, 154905 (2020).
8. Zhao, D. et al. Facile deposition of high-quality Cs$_2$AgBiBr$_6$ films for efficient double perovskite solar cells. *Sci. China Mater.* **63**, 1518–1525 (2020).
9. Graser, J., Kauwe, S. K. & Sparks, T. D. Machine learning and energy minimization approaches for crystal structure predictions: a review and new horizons. *Chem. Mater.* **30**, 3601–3612 (2018).
10. Rajan, K. Materials informatics. *Mater. Today* **8**, 38–45 (2005).
11. Hohenberg, P. & Kohn, W. Inhomogeneous electron gas. *Phys. Rev.* **136**, B864–B871 (1964).
12. Hussain, A. et al. Monte Carlo simulation study of electron yields from compound semiconductor materials. *J. Appl. Phys.* **128**, 015305 (2020).
13. Alder, B. J. & Wainwright, T. E. Studies in molecular dynamics. I. General method. *J. Chem. Phy.* **31**, 459–466 (1959).
14. Agrawal, A. & Choudhary, A. Perspective: materials informatics and big data: realization of the "fourth paradigm" of science in materials science. *APL Mater.* **4**, 053208 (2016).
15. Raccuglia, P. et al. Machine-learning-assisted materials discovery using failed experiments. *Nature* **533**, 73–76 (2016).
16. Balachandran, P. V., Kowalski, B., Sehirlioglu, A. & Lookman, T. Experimental search for high-temperature ferroelectric perovskites guided by two-step machine learning. *Nat. Commun.* **9**, 1668 (2018).
17. Dai, D. et al. Using machine learning and feature engineering to characterize limited material datasets of high-entropy alloys. *Comput. Mater. Sci.* **175**, 109618 (2020).
18. Sun, W. et al. Machine learning-assisted molecular design and efficiency prediction for high-performance organic photovoltaic materials. *Sci. Adv.* **5**, eaay4275 (2019).
19. Stanev, V. et al. Machine learning modeling of superconducting critical temperature. *npj Comput. Mater.* **4**, 29 (2018).
20. Jordan, M. I. & Mitchell, T. M. Machine learning: trends, perspectives, and prospects. *Science* **349**, 255–260 (2015).
21. Rupp, M. Machine learning for quantum mechanics in a nutshell. *Int. J. Quantum Chem.* **115**, 1058–1073 (2015).
22. Goldsmith, B. R., Esterhuizen, J., Liu, J. X., Bartel, C. J. & Sutton, C. Machine learning for heterogeneous catalyst design and discovery. *AIChE J.* **64**, 2311–2323 (2018).
23. Lu, W., Xiao, R., Yang, J., Li, H. & Zhang, W. Data mining-aided materials discovery and optimization. *J. Materiomics* **3**, 191–201 (2017).
24. Wan, X. et al. Materials discovery and properties prediction in thermal transport via materials informatics: a mini review. *Nano Lett.* **19**, 3387–3395 (2019).
25. Chen, C. et al. A critical review of machine learning of energy materials. *Adv. Energy Mater.* **10**, 1903242 (2020).
26. Liu, Y., Zhao, T., Ju, W. & Shi, S. Materials discovery and design using machine learning. *J. Materiomics* **3**, 159–177 (2017).
27. Toyao, T. et al. Machine learning for catalysis informatics: recent applications and prospects. *ACS Catal.* **10**, 2260–2297 (2019).
28. Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A. & Kim, C. Machine learning in materials informatics: recent applications and prospects. *npj Comput. Mater.* **3**, 54 (2017).

29. Braham, E. J. et al. Machine learning-directed navigation of synthetic design space: a statistical learning approach to controlling the synthesis of perovskite halide nanoplatelets in the quantum-confined regime. *Chem. Mater.* **31**, 3281–3292 (2019).
30. Zhou, T., Song, Z. & Sundmacher, K. Big data creates new opportunities for materials research: a review on methods and applications of machine learning for materials design. *Engineering* **5**, 1017–1026 (2019).
31. 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.* **174**, 109474 (2020).
32. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).
33. 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.* **5**, 83 (2019).
34. Wang, H., Ji, Y. & Li, Y. Simulation and design of energy materials accelerated by machine learning. *WIREs Comput. Mol. Sci.* **10**, 1421 (2019).
35. Zhou, Z. *Machine Learning* (Tsinghua University Press, Bei Jing, 2016).
36. Efron, B. & Tibshirani, R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Stat. Sci.* **1**, 54–75 (1986).
37. Shi, L., Chang, D., Ji, X. & Lu, W. Using data mining to search for perovskite materials with higher specific surface area. *J. Chem. Inf. Model.* **58**, 2420–2427 (2018).
38. Furmanchuk, A. et al. Prediction of seebeck coefficient for compounds without restriction to fixed stoichiometry: a machine learning approach. *J. Comput. Chem.* **39**, 191–202 (2018).
39. Wang, Y. et al. Emerging perovskite materials for high density data storage and artificial synapses. *J. Mater. Chem. C.* **6**, 1600–1617 (2018).
40. Travis, W., Glover, E. N. K., Bronstein, H., Scanlon, D. O. & Palgrave, R. G. On the application of the tolerance factor to inorganic and hybrid halide perovskites: a revised system. *Chem. Sci.* **7**, 4548–4556 (2016).
41. Yin, W. J. et al. Oxide perovskites, double perovskites and derivatives for electrocatalysis, photocatalysis, and photovoltaics. *Energ. Environ. Sci.* **12**, 442–462 (2019).
42. Babu, R., Giribabu, L. & Singh, S. P. Recent advances in halide-based perovskite crystals and their optoelectronic applications. *Cryst. Growth Des.* **18**, 2645–2664 (2018).
43. Körbel, S., Marques, M. A. L. & Botti, S. Stability and electronic properties of new inorganic perovskites from high-throughput ab initio calculations. *J. Mater. Chem. C.* **4**, 3157–3167 (2016).
44. Saha-Dasgupta, T. Magnetism in double perovskites. *J. Supercond. Nov. Magn.* **26**, 1991–1995 (2012).
45. Li, C. et al. A progressive learning method for predicting the band gap of $ABO_3$ perovskites using an instrumental variable. *J. Mater. Chem. C.* **8**, 3127–3136 (2020).
46. Sun, Z. et al. Research progress and perspective of machine learning in material design (in Chinese). *Chin. Sci. B-Chin.* **64**, 3270–3275 (2019).
47. Bally, M. A. A. & Khan, F. A. Structural, dielectric and magnetic properties of La0.55Sr0.45MnO3 polycrystalline perovskite. *J. Magn. Magn. Mater.* **509**, 166897 (2020).
48. Shanker, J., Kumar, R. V., Rao, G. N. & Babu, D. S. Magnetic reversal in Fe substituted NdCrO3 perovskite nanoparticles. *Mater. Chem. Phys.* **251**, 123098 (2020).
49. AboZied, A. E. R. T., Ghani, A. A., Ali, A. I. & Salaheldin, T. A. Structure, magnetic and magnetocaloric properties of nano crystalline perovskite $La_{0.8}Ag_{0.2}MnO_3$. *J. Magn. Magn. Mater.* **479**, 260–267 (2019).
50. Mechi, N. et al. $La_{0.6}Ca_{0.2}Na_{0.2}MnO_3$ perovskite: structural, magnetic, critical, and magnetocaloric properties. *J. Supercond. Nov. Magn.* **33**, 1385–1393 (2019).
51. Li, L. et al. An efficient and durable perovskite electrocatalyst for oxygen reduction in solid oxide fuel cells. *Chem. Eng. J.* **396**, 125237 (2020).
52. Xia, W., Li, Q., Sun, L., Huo, L. & Zhao, H. Electrochemical performance of Sn-doped Bi0.5Sr0.5FeO3-δ perovskite as cathode electrocatalyst for solid oxide fuel cells. *J. Alloy. Compd.* **835**, 155406 (2020).
53. Carrasco-Jaim, O. A., Huerta-Flores, A. M., Torres-Martínez, L. M. & Moctezuma, E. Fast in-situ photodeposition of Ag and Cu nanoparticles onto $AgTaO_3$ perovskite for an enhanced photocatalytic hydrogen generation. *Int. J. Hydrog. Energ.* **45**, 9744–9757 (2020).
54. Zhou, J. et al. Photocatalytic degradation characteristics of tetracycline and structural transformation on bismuth silver oxide perovskite nano-catalysts. *Appl. Nanosci.* **10**, 2329–2338 (2020).
55. Goldschmidt, V. M. Die Gesetze der Krystallochemie. *Naturwissenschaften* **14**, 477–485 (1926).
56. Sun, Q. & Yin, W. J. Thermodynamic stability trend of cubic perovskites. *J. Am. Chem. Soc.* **139**, 14905–14908 (2017).
57. Bartel, C. J. et al. New tolerance factor to predict the stability of perovskite oxides and halides. *Sci. Adv.* **5**, eaav0693 (2019).
58. Armiento, R., Kozinsky, B., Hautier, G., Fornari, M. & Ceder, G. High-throughput screening of perovskite alloys for piezoelectric performance and thermodynamic stability. *Phys. Rev. B* **89**, 134103 (2014).
59. Liu, M. et al. Spinel compounds as multivalent battery cathodes: a systematic evaluation based on ab initio calculations. *Energy Environ. Sci.* **8**, 964–974 (2015).
60. Schmidt, J. et al. Predicting the thermodynamic stability of solids combining density functional theory and machine learning. *Chem. Mater.* **29**, 5090–5103 (2017).
61. Li, W., Jacobs, R. & Morgan, D. Predicting the thermodynamic stability of perovskite oxides using machine learning models. *Comput. Mater. Sci.* **150**, 454–463 (2018).
62. Liu, H. et al. Screening stable and metastable $ABO_3$ perovskites using machine learning and the materials project. *Comput. Mater. Sci.* **177**, 109614 (2020).
63. Li, X. et al. Computational screening of new perovskite materials using transfer learning and deep learning. *Appl. Sci.* **9**, 5510 (2019).
64. Pilania, G., Balachandran, P. V., Kim, C. & Lookman, T. Finding new perovskite halides via machine learning. *Front. Mater.* **3**, 19 (2016).
65. Balachandran, P. V. et al. Predictions of new $ABO_3$ perovskite compounds by combining machine learning and density functional theory. *Phys. Rev. Mater.* **2**, 043802 (2018).
66. Jain, D., Chaube, S., Khullar, P., Goverapet Srinivasan, S. & Rai, B. R. Bulk and surface DFT investigations of inorganic halide perovskites screened using machine learning and materials property databases. *Phys. Chem. Chem. Phys.* **21**, 19423–19436 (2019).
67. Park, H. et al. Learn-and-match molecular cations for perovskites. *J. Phys. Chem. A* **123**, 7323–7334 (2019).
68. Gladkikh, V. et al. Machine learning for predicting the band gaps of $ABX_3$ perovskites from elemental properties. *J. Phys. Chem. C.* **124**, 8905–8918 (2020).
69. Takahashi, K., Takahashi, L., Miyazato, I. & Tanaka, Y. Searching for hidden perovskite materials for photovoltaic systems by combining data science and first principle calculations. *ACS Photonics* **5**, 771–775 (2018).
70. Chen, Y. et al. Correlation of dielectric dispersion with distributed Curie temperature in relaxor ferroelectrics. *J. Appl. Phys.* **125**, 184104 (2019).
71. Luo, Z. et al. Growth and characterization of ternary $BiScO_3$-$Pb(Cd_{1/3}Nb_{2/3})O_3$-$PbTiO_3$ ferroelectric single crystals with high Curie temperature. *CrystEngComm* **22**, 4544–4551 (2020).
72. Shi, T., Li, G. & Zhu, J. Compositional design strategy for high performance ferroelectric oxides with perovskite structure. *Ceram. Int.* **43**, 2910–2917 (2017).
73. Jin, F., Zhang, H. & Chen, Q. Improved Curie temperature and temperature coefficient of resistance (TCR) in $La_{0.7}Ca_{0.3\text{-}x}SrMnO_3$: $Ag_{0.2}$ composites. *J. Alloy. Compd.* **747**, 1027–1032 (2018).
74. Pang, D. et al. Lead-reduced $Bi(Ni_{2/3}Ta_{1/3})O_3$-$PbTiO_3$ perovskite ceramics with high Curie temperature and performance. *J. Am. Ceram. Soc.* **102**, 1227–1239 (2019).
75. Zhai, X., Chen, M. & Lu, W. Accelerated search for perovskite materials with higher Curie temperature based on the machine learning methods. *Comp. Mater. Sci.* **151**, 41–48 (2018).
76. Yang, Z. X. et al. High critical transition temperature of lead-based perovskite ferroelectric crystals: a machine learning study. *Acta Phys. Sin.* **68**, 210502 (2019).
77. Amit, E., Keren, A., Lord, J. S. & King, P. A precise measurement of the oxygen isotope effect on the Néel temperature in cuprates. *Adv. Cond. Matter Phys.* **2011**, 1–5 (2011).
78. Chmaissem, O. et al. Relationship between structural parameters and the Néel temperature in $Sr_{1\text{-}x}Ca_xMnO_3(0 \le x \le 1)$ and $Sr_{1\text{-}y}Ba_yMnO_3(y \le 0.2)$. *Phys. Rev. B* **64**, 134412 (2001).
79. Xiao, L., Zhang, Q., Xu, X., Ji, X. & Lu, W. Support vector regression assisted predictions the néel temperature of perovskites manganites. *Comp. Appl. Chem.* **35**, 349–357 (2018).
80. Gutfleisch, O. et al. Magnetic materials and devices for the 21st century: stronger, lighter, and more energy efficient. *Adv. Mater.* **23**, 821–842 (2011).
81. Cao, G. et al. Enhanced magnetic entropy change and refrigeration capacity of $La(Fe,Ni)_{11.5}Si_{1.5}$ alloys through vacuum annealing treatment. *J. Alloy. Compd.* **800**, 363–371 (2019).
82. Phan, M. H. & Yu, S. C. Review of the magnetocaloric effect in manganite materials. *J. Magn. Magn. Mater.* **308**, 325–340 (2007).
83. Cao, F. et al. Effect of yttrium doping on magnetic properties and magnetic entropy change of bilayered perovskite $La_{1.3}Sr_{1.7}Mn_2O_7$. *J. Low. Temp. Phys.* **200**, 16–25 (2020).
84. Chen, F. et al. Large magnetic entropy change and refrigeration capacity around room temperature in quinary $Ni_{41}Co_{9\text{-}x}Fe_xMn_{40}Sn_{10}$ alloys ($x = 2.0$ and 2.5). *J. Alloy. Compd.* **825**, 154053 (2020).
85. Zhang, Y. & Xu, X. Machine learning the magnetocaloric effect in manganites from lattice parameters. *Appl. Phys. A* **126**, 341 (2020).

86. Xu, W. et al. Dielectric breakdown strength of alumina ceramics reinforced by fractal dendritic $Ca_9Al(PO_4)_7$ as the second crystalline phase. *J. Alloy. Compd.* **832**, 154811 (2020).

87. Yao, T. et al. Nano-BN encapsulated micro-AlN as fillers for epoxy composites with high thermal conductivity and sufficient dielectric breakdown strength. *IEEE Trans. Dielectr. Electr. Insul.* **27**, 528–534 (2020).

88. Zhang, T. et al. Effect of pores on dielectric breakdown strength of alumina ceramics via surface and volume effects. *J. Eur. Ceram. Soc.* **40**, 3019–3026 (2020).

89. Lu, Y. et al. Remarkable dielectric breakdown strength enhancement of a PVDF terpolymer using a 2D hybrid organic inorganic perovskite as a functional additive. *J. Mater. Chem. C.* **7**, 13390–13395 (2019).

90. Yang, L. et al. Perovskite lead-free dielectrics for energy storage applications. *Prog. Mater. Sci.* **102**, 72–108 (2019).

91. Kim, C., Pilania, G. & Ramprasad, R. Machine learning assisted predictions of intrinsic dielectric breakdown strength of $ABX_3$ perovskites. *J. Phys. Chem. C.* **120**, 14575–14580 (2016).

92. Zhang, C. et al. Achieving ultrahigh dielectric breakdown strength in MgO-based ceramics by composite structure design. *Mater. Chem. C.* **7**, 8120–8130 (2019).

93. Gao, J. et al. Designing high dielectric permittivity material in barium titanate. *J. Phys. Chem. C.* **121**, 13106–13113 (2017).

94. Bhattacharyya, R., Das, S. & Omar, S. High ionic conductivity of $Mg^{2+}$-doped non-stoichiometric sodium bismuth titanate. *Acta Mater.* **159**, 8–15 (2018).

95. Reis, S. L. & Muccillo, E. N. S. Influence of small amounts of gallium oxide addition on ionic conductivity of $La_{0.9}Sr_{0.1}Ga_{0.8}Mg_{0.2}O_{3-\delta}$ solid electrolyte. *Ceram. Int.* **44**, 115–119 (2018).

96. Verma, O. N., Jha, P. A., Singh, P., Jha, P. K. & Singh, P. Influence of iso-valent 'Sm' double substitution on the ionic conductivity of $La_{0.9}Sr_{0.1}Al_{0.9}Mg_{0.1}O_{3-\delta}$ ceramic system. *Mater. Chem. Phys.* **241**, 122345 (2020).

97. Liu, X., Lu, W., Peng, C., Sun, Q. & Guo, J. Two semi-empirical approaches for the prediction of oxide ionic conductivities in $ABO_3$ perovskites. *Comput. Mater. Sci.* **46**, 860–868 (2009).

98. Kaneko, M., Fujii, M., Hisatomi, T., Yamashita, K. & Domen, K. Regression model for stabilization energies associated with anion ordering in perovskite-type oxynitrides. *J. Energy Chem.* **36**, 7–14 (2019).

99. Zheng, W. D. et al. Performance prediction of perovskite materials based on different machine learning algorithms. *Chin. J. Nonfer. Met.* **29**, 803–807 (2019).

100. Lu, S., Zhou, Q., Ma, L., Guo, Y. & Wang, J. Rapid discovery of ferroelectric photovoltaic perovskites and material descriptors via machine learning. *Small Methods* **3**, 1900360 (2019).

101. Liao, K. et al. Aqueous solvent-regulated crystallization and interfacial modification in perovskite solar cells with enhanced stability and performance. *J. Power Sources* **471**, 228447 (2020).

102. Parrey, K. A., Ansari, S. G., Aziz, A. & Niazi, A. Enhancement in structural and optical properties of Cd doped hybrid organic-inorganic halide perovskite $CH_3NH_3Pb_{1-x}Cd_xI_3$ photo-absorber. *Mater. Chem. Phys.* **241**, 122387 (2020).

103. Wang, G. et al. An internally photoemitted hot carrier solar cell based on organic-inorganic perovskite. *Nano Energy* **68**, 104383 (2020).

104. Zhang, X., Wei, M. & Qin, W. Magneto-open-circuit voltage in organic-inorganic halide perovskite solar cells. *Appl. Phys. Lett.* **114**, 033302 (2019).

105. Kim, T., Kim, J. H. & Park, J. W. All-solution-processed hybrid organic-inorganic perovskite light-emitting diodes under Ambient Air. *Phys. Status Solidi A* **216**, 1900642 (2019).

106. Kim, T., Kim, J. H. & Park, J. W. Semi-transparent hybrid organic-inorganic perovskite light-emitting diodes fabricated under high relative humidity. *Solid State Electron.* **165**, 107749 (2020).

107. Xie, C., Liu, C. K., Loi, H. L. & Yan, F. Perovskite-based phototransistors and hybrid photodetectors. *Adv. Funct. Mater.* **30**, 1903907 (2019).

108. Xin, J. et al. Planar visible-near infrared photodetectors based on Hybrid organic-inorganic perovskite single crystal bulks. *J. Phys. D. Appl. Phys.* **53**, 414003 (2020).

109. Kojima, A., Teshima, K., Shirai, Y. & Miyasaka, T. Organometal halide perovskites as visible-light sensitizers for photovoltaic cells. *J. Am. Chem. Soc.* **131**, 6050–6051 (2009).

110. Xi, Z. Z., Wang, R. Q., Song, Z. C., Guo, Y. G. & Wu, X. Progressing on perovskite-based solar cells. *Mod. Chem. Ind.* **39**, 66–70 (2019).

111. Min, G., Yun, Y., Choi, H. J., Lee, S. & Joo, J. Hydrogen halide-free synthesis of organohalides for organometal trihalide perovskite solar cells. *J. Ind. Eng. Chem.* **89**, 375–382 (2020).

112. National Renewable Energy Laboratory. *NREL Efficiency Chart*. https://www.nrel.gov/pv/cell-efficiency.html/.

113. L. Agiorgousis, M. et al. Machine learning augmented discovery of chalcogenide double perovskites for photovoltaics. *Adv. Theor. Simul.* **2**, 1800173 (2019).

114. Ma, L. et al. Temperature-dependent thermal decomposition pathway of organic-inorganic halide perovskite materials. *Chem. Mater.* **31**, 8515–8522 (2019).

115. Zhang, Y. & Zhou, H. P. Intrinsic stability of Hybrid organic-inorganic perovskite. *Acta Phys. Sin.* **68**, 158804 (2019).

116. Lu, S. et al. Accelerated discovery of stable lead-free hybrid organic-inorganic perovskites via machine learning. *Nat. Commun.* **9**, 3405 (2018).

117. Saidi, W. A., Shadid, W. & Castelli, I. E. Machine-learning structural and electronic properties of metal halide perovskites using a hierarchical convolutional neural network. *npj Comput. Mater.* **6**, 36 (2020).

118. Ali, A. et al. Machine learning accelerated recovery of the cubic structure in mixed-cation perovskite thin films. *Chem. Mater.* **32**, 2998–3006 (2020).

119. Li, L. et al. Electronic transport of Hybrid organic-inorganic perovskites from first-principles and machine learning. *Appl. Phys. Lett.* **114**, 083102 (2019).

120. Odabaşı, Ç. & Yıldırım, R. Performance analysis of perovskite solar cells in 2013–2018 using machine-learning tools. *Nano Energy* **56**, 770–791 (2019).

121. Li, J., Pradhan, B., Gaur, S. & Thomas, J. Predictions and strategies learned from machine learning to develop high-performing perovskite solar cells. *Adv. Energy Mater.* **9**, 1901891 (2019).

122. Zhang, W., Hong, M. & Luo, J. Halide double perovskite ferroelectrics. *Angew. Chem. Int. Ed.* **59**, 9305–9308 (2020).

123. Zhao, X. G. et al. Rational design of halide double perovskites for optoelectronic applications. *Joule* **2**, 1662–1673 (2018).

124. Wang, B. et al. Photoactive Zn-chlorophyll hole transporter-sensitized lead-free $Cs_2AgBiBr_6$ perovskite solar cells. *Sol. RRL* **4**, 2000166 (2020).

125. Wang, T., Yue, D., Li, X. & Zhao, Y. Lead-free double perovskite $Cs_2AgBiBr_6$/RGO composite for efficient visible light photocatalytic $H_2$ evolution. *Appl. Catal. B Environ.* **268**, 118399 (2020).

126. Idris, A. M. et al. A novel double perovskite oxide semiconductor $Sr_2CoWO_6$ as bifunctional photocatalyst for photocatalytic oxygen and hydrogen evolution reactions from water under visible light irradiation. *Sol. RRL* **4**, 1900456 (2019).

127. Pilania, G. et al. Machine learning bandgaps of double perovskites. *Sci. Rep.* **6**, 19375 (2016).

128. Xu, Q., Li, Z., Liu, M. & Yin, W. J. Rationalizing perovskite data for machine learning and materials design. *J. Phys. Chem. Lett.* **9**, 6948–6954 (2018).

129. Li, Z., Xu, Q., Sun, Q., Hou, Z. & Yin, W. J. Thermodynamic stability landscape of halide double perovskites via high-throughput computing and machine learning. *Ad. Func. Mater.* **29**, 1807280 (2019).

130. Halder, A., Ghosh, A. & Dasgupta, T. S. Machine-learning-assisted prediction of magnetic double perovskites. *Phy. Rev. Mater.* **3**, 084418 (2019).

131. Li, Z., Achenie, L. E. K. & Xin, H. An adaptive machine learning strategy for accelerating discovery of perovskite electrocatalysts. *ACS Catal.* **10**, 4377–4384 (2020).

132. Sun, S. et al. Accelerated development of perovskite-inspired materials via high-throughput synthesis and machine-learning diagnosis. *Joule* **3**, 1437–1451 (2019).

133. Jiang, S. et al. Machine learning (ML)-assisted optimization doping of KI in MAPbI3 solar cells. *Rare Metals* (2020).

134. Weng, B. et al. Simple descriptor derived from symbolic regression accelerating the discovery of new perovskite catalysts. *Nat. Commun.* **11**, 3513 (2020).

135. Wu, W. & Sun, Q. Applying machine learning to accelerate new materials developmen t(in Chinese). *Sci. Sin. Phys. Mech. Astron.* **48**, 107001 (2018).

136. Su, X. et al. A wireless electrode-free QCM-D in a multi-resonance mode for volatile organic compounds discrimination. *IEEE T. Ind. Electron.* **305**, 111938 (2020).

137. Li, X., Li, H. & Dong, Y. Meta learning for task-driven video summarization. *Pattern Recogn. Lett.* **67**, 5778–5786 (2020).

138. Graves, A., Wayne, G. & Danihelka, I. Neural turing machines. Preprint at *ArXiv* https://arxiv.org/abs/1410.5401 (2014).

139. Duan, Y. et al. One-shot Imitation learning. *Adv. Neural Inf. Pro. Syst.* **30**, (2017).

140. Lake, B. M., Salakhutdinov, R. & Tenenbaum, J. B. Human-level concept learning through probabilistic program induction. *Science* **350**, 1332–1338 (2015).

## ACKNOWLEDGEMENTS

Financial support to this work from the National Key Research and Development Program of China (No. 2016YFB0700504) and the Science and Technology Commission of Shanghai Municipality (18520723500) is gratefully acknowledged.

## AUTHOR CONTRIBUTIONS

Q.T. is the first author who collected publications and completed the framework of the manuscript. P.X. is the co-first author and completed the initial manuscript. M.L. and W.L. revised the manuscript.

## COMPETING INTERESTS

The authors declare no competing interests.

## ADDITIONAL INFORMATION

**Correspondence** and requests for materials should be addressed to M.L. or W.L.

**Reprints and permission information** is available at http://www.nature.com/reprints

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.