Computational Materials Science 226 (2023) 112215

Contents lists available at ScienceDirect

Computational Materials Science

journal homepage: www.elsevier.com/locate/commatsci

Full Length Article

Machine learning for perovskite solar cell design

Zhan Hui [a], Min Wang [b], Xiang Yin [a,*], Ya'nan Wang [a], Yunliang Yue [a,*]

[a] College of Information Engineering, Yangzhou University, Yangzhou 225127, People's Republic of China
[b] School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang 050018, People's Republic of China

ARTICLE INFO

Keywords:
Machine learning
Perovskites
Property predictions
Perovskite solar cells

ABSTRACT

As representatives of third-generation solar cells, perovskite solar cells (PSCs) have experienced rapid development. Suffering from inefficient traditional trial-and-error methods and huge search space, discovering superior performance of perovskite materials and high conversion efficiency and stability of PSCs is still a challenge. With the increased computational power and the establishment of large databases, data-driven machine learning (ML) is rapidly gaining momentum in the materials field. ML can predict the properties of potential perovskite materials as well as provide additional physical understanding to accelerate the advancement of PSCs. In this review, we first outline the basic steps and methods of ML. Then, we focus on recent advances in ML for perovskite property predictions and candidates screening, and research to find conditions for higher efficiency or stability in PSCs. We also analyzed the understanding provided by the ML approach and the relationship between the descriptors and the target properties. In addition, we summarize comments and opinions and discuss the current challenges and future opportunities in the field.

1. Introduction

Nowadays, clean energy discovery is still a research hotspot [1]. Solar energy is an important clean and renewable energy source, and maximizing its use can reduce dependence on chemical energy sources [2,3]. Photovoltaic cells that can generate electricity directly from sunlight are a major application of solar energy in recent decades [4,5]. Photovoltaic cells can be divided into three generations. The first and second-generation solar cells are respectively based on silicon wafers and thin films of CdTe and CuInGaSe, and the third-generation solar cells use organic, inorganic or hybrids structures [6]. Over the years, first and second-generation solar cells have limited their use due to high cost, elaborate fabrication process and environmental unfriendliness [7]. Therefore, researchers are looking for new cost-effective and non-polluting solar cell materials. Perovskites are generally recognized as a material with the same structure as $CaTiO_3$, a brand-new mineral material found in the Ural Mountains by the German mineralogist Gustav Rose [8]. Its general chemical formula is ABX_3, where the A-site is generally the larger cation, B-site is the smaller cation, and X-site is the anion coordinated with B-site cation. The B-site forms a BX_6 octahedron with the X-site, occupying the center of the octahedron. The BX_6 octahedron is angle-sharing, and the A-site cation fills the cavities. Perovskite solar cells (PSCs) are the novel photovoltaic cells that use perovskite materials as a light-absorbing layer. In PSCs, perovskites are sandwiched between the electron-transport layer (ETL) and the hole-transport layer (HTL), which creates a highly efficient energy conversion structure. Due to their remarkable light absorption efficiency [9], high carrier mobility [10], and low cost of preparation [11], PSCs have attracted extensive interest in solar energy harvesting [12,13]. Research on PSCs has focused on the light absorption layer material and the efficiency of PSCs [14,15]. Machine learning (ML) is a branch of artificial intelligence that strives to learn from vast data points and gain experience to apply in practice. Nowadays, abundant data is generated through theoretical computations and experimental simulations, which has led to the combination of machine learning and material science. By leveraging open-source databases, ML is beginning to play a role in the discovery of perovskite materials and PSCs [16–18].

PSCs have shown unprecedented growth rates compared to commercial silicon solar cells. Early in the research, much of the research focused on efficiency improvements. In 2009, Kojima [9] *et al.* achieved the first PSCs with a 3.8 % power conversion efficiency (PCE) using organic–inorganic halide perovskites as the light source. Since then, researchers have made significant strides in improving the PCE of PSCs, with Kim *et al.* [19] achieving a PCE of 9.7 % by employing $(CH_3NH_3)PbI_3$ perovskites as a light absorber. More recently, the performance of PSCs has improved to over 25.5 % [20]. The efficiency of PSCs reached a

* Corresponding authors.
E-mail addresses: yinxiang@yzu.edu.cn (X. Yin), yueyunliang@yzu.edu.cn (Y. Yue).

https://doi.org/10.1016/j.commatsci.2023.112215
Received 20 January 2023; Received in revised form 13 April 2023; Accepted 17 April 2023
Available online 30 April 2023

certain level, and stability issues began to raise new concerns and research subsequently [21–24]. In terms of composition and structural classification, most of the research on perovskites has focused on inorganic perovskites, hybrid organic–inorganic perovskites (HOIPs), and double perovskites [25–27]. Among them, the organic component in the structure of HOIPs adds extra functions and structural flexibility that cannot be achieved with pure inorganic perovskites. Additionally, their different structural and chemical variations provide ample chances to tune and modulate their physical properties with simple chemical modifications [28]. Double perovskites pave the way for more alternative options using different metal cations in A and B-sites, and it shows higher PCE and suitable bandgap [29]. Furthermore, inorganic perovskites typically perform with high thermal stability [30,31]. The combination of data-driven ML has shown impressive efficiency in screening candidates compared to traditional methods, as well as a high degree of accuracy in prediction through experimental validation [32–34].

However, with the gradual commercialization of PSCs, the issues of chemical composition screening of perovskites, PSCs performance, and toxicity gradually show up [35,36]. Toxicity problems are generally addressed by replacing toxic elements, such as Pb elements, which can solve the material toxicity problem [37]. Besides considering lead toxicity, the large amount of organic solvents used in the synthesis of perovskites is also a significant source of toxicity. For chemical composition screening, the composition space that could potentially compose the perovskite structure is huge. Up to now, only a few thousand perovskites have been identified [38,39]. In the past, high-throughput computational is a conventional method, enabling scientists to compute the properties of hundreds of compounds in a single investigation. In particular, Density Functional Theory (DFT) is currently a well-recognized method for determining the structural and behavioral properties of solids [40]. Although modern chemical simulation tools allow for predicting the properties of compounds before they are manufactured in the laboratory and can achieve the desired level of accuracy. Restricted by experimental conditions and theoretical foundations, traditional experimental and computational modelling usually consumes much time and resources. In addition, the performance testing of PSCs also requires a long time for experimental acquisition. Apparently, the traditional experiment or DFT methods no longer meet the current needs [41,42]. To address this, ML based on material informatics is becoming popular. For predicting single property of materials, using the DFT method may require hours of computation. However, for a trained ML model, it takes only a few seconds to predict the properties of thousands of materials. ML learns to recognize the relationship between given features and target attributes in massive data to obtain ML models for high-accuracy prediction [17]. At the same time, ML models can achieve an accuracy comparable to DFT [43–45]. In terms of PSCs testing, ML can also predict device performance by analyzing data from different experimental conditions of PSCs, which also greatly reduces the time required for experiments [46]. More importantly, the advantages of ML are not only in model prediction; ML can summarize physical rules that are difficult to observe by humans. With the understanding of ML, it is possible to apply this knowledge to material synthesis, thereby aiding experimental synthesis [47,48]. By leveraging the ML technology, researchers can develop predictive models that guide experimental synthesis efforts, leading to more efficient and targeted exploration of the vast chemical space of perovskite materials.

In this review, we first illustrate the general steps of applying ML. Then, we present the current specific ML applications for perovskite property predictions and composition screening, meanwhile analyzing the descriptors affecting the different property predictions. We also present the application of ML in the optimization of experimental conditions and performance prediction of PSCs. For example, ML techniques can speed up the identification of new perovskites with adequate bandgap and high stability [49,50]. Additionally, ML techniques can identify factors that are significantly correlated with device performance during experiments [51]. We have categorized research on ML technology in the field of PSCs, aimed at helping researchers gain insights into the trends of ML applications in studying PSCs. Moreover, we discuss the current challenges and future opportunities in this field.

2. Machine learning general steps

Big data and artificial intelligence have been dubbed the "fourth industrial revolution" [52], and there are increasing applications in the chemical and material fields. ML is one of the important branches of artificial intelligence. The core of ML applications is the statistical algorithm. At the same time, a growing number of ML tools (Scikit-learn [53], PyTorch [54], Tensorflow [55], etc.) are available to generate, test, and refine more powerful models. Complex problems with enormous combinatorial spaces or nonlinear processes that cannot be resolved by conventional methods can be tackled using ML. Raw data collection and pre-processing, feature engineering, model selection, model validation, and application make up the typical workflow of ML in material science (Fig. 1).

2.1. Raw data collection and pre-processing

The quality and quantity of the collected data significantly impact on the performance of the ML model. To underline the significance of huge data in the advancement of material research and promote the development of leading material databases, the Materials Genome Initiative (MGI) was launched in the United States back in 2011. For perovskite materials and PSCs, several data points were generated by high throughput calculations and experiments [56,57]. In addition, many large digital materials databases and application programming interfaces (APIs) are being built and improved to enable massive computational tasks and data processing. For example, Open Quantum Material Database [58], Computational Materials Repository [59], Inorganic Crystal Structure Database [60], Pymatgen [61], and AFLOWLIB [62]. There are normally some bad data points in the raw data, such as outliers, which are anomalous values that are far from other values [63]. If these data points are not eliminated, the trained model will have problems such as non-convergence or underfitting [64]. Therefore, in most cases, pre-processing of the data is particularly important. In the process of data pre-processing, redundant and low-quality data points can be addressed by removing these data points that may hinder model convergence [65]. For missing data points, after considering the size of the data, data deletion or padding is chosen [66]. For small data sets, data augmentation methods such as linear interpolation padding [67] are also performed on the data in some cases. There are automated tools for data pre-processing, for instance, AutoML, which is based on the scikit-learn package and includes pre-processing methods for new datasets [68].

2.2. Feature engineering

Feature, also known as a descriptor, is a set of values used to describe properties of materials. Ward *et al.* [69] classified inorganic material feature categories into four categories, including stoichiometric properties, elemental statistical properties, electronic structure properties, and ionic compound properties. Ghiringhelli *et al.* [70] state that a set of descriptors should satisfy the following four principles. First, descriptors uniquely characterize the material and the underlying processes associated with the properties. Second, different materials should be characterized by different descriptors. Third, descriptors must not involve computations as intensive as the predicted target. Fourth, the dimensionality of the descriptors should be as low as possible (under certain accuracy requirements). Therefore, redundant and highly autocorrelated features should be removed during feature selection. Table 1 shows some examples of the features and targets of ML induction. Nowadays, many packages for performing feature generation have appeared in the material field, such as Matminer [71], Pymatgen [61], the Smooth

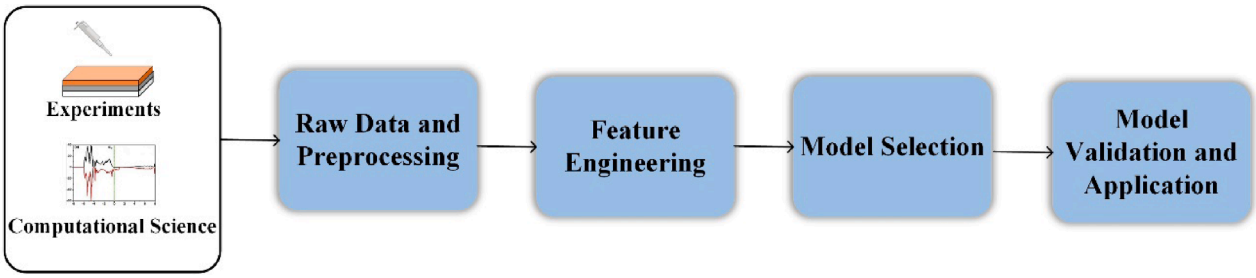

Fig. 1. The workflow of ML. First, the raw dataset including the correct molecules, structures and properties needs to be prepared. Second, feature engineering is performed based on molecular or structural information. Third, ML algorithms are selected to train the models. Finally, after tested on test sets, the well-performed model reveals the hidden relationships or predicts new candidates.

Table 1
Samples of features and related target attributes.

Properties	Features	References
Formability	Goldschmidt tolerance factor, octahedral factor, ion radius and other variants	[74–77]
Perovskites stability	Formation energy, tolerance factor, E_{hull}, electronegativity and polarizability	[78,79]
Bandgap	The lowest occupied energy level and electronegativity	[78,80]
PCE	HOMO, LUMO, ΔH, ΔL and bandgap	[47,81]
PSCs stability	Environmental factors, device composition, and manufacturing processes	[82]

Overlap of Atomic Positions (SOAP) [72], and Component-Based Feature Vector (CBFV) [73]. In this way, chemicals can be characterized by formulas or structures for feature encoding and it reduces the need for material knowledge. In addition, deep learning algorithms use automatic feature engineering to select a set of features relevant to the target output. This feature selection method minimizes the domain-specific knowledge used in the training model, which is friendly to non-material domain researchers and has good application prospects.

2.3. Model selection

Methods of ML applied to the property prediction of perovskite materials and PSCs may be roughly categorized into three categories: supervised learning, semi-supervised learning, and unsupervised learning. As one of the most commonly used methods, supervised learning uses a set of samples containing labels to train a model that can be used for regression or classification tasks [83]. Table 2 lists the commonly used supervised learning algorithms and a brief description.

Semi-supervised learning algorithms [84] are more specific. They are applied to situations where only a small amount of labelled data is available. A large amount of unlabeled data is used to improve the accuracy of the model, such as positive example unlabeled learning (PU learning) [85]. Usually, the parameter adjustment of the model is critical, and it is especially important to adjust the model parameters for different data sizes and distributions. The purpose of model parameter adjustment is to achieve a balance between the overfitting and underfitting of the model. Overfitting is when the trained model is highly fit to each data point in the training dataset, and therefore cannot get a good generalization capability. It shows better in the training set and badly in the test set. On the contrary, underfitting only fits a small portion of the training data. Grid Search [86] and Hyperopt [87] are the more commonly used methods for tuning parameters.

2.4. Model validation and application

A robust ML model must have strong generalization ability and high prediction performance for unknown data. To achieve this, the data is usually partitioned into a training set and a test set. The test data is generally not duplicated with the training data to ensure that the model can generalize well and perform accurately on unseen data. In general, there are five main evaluation metrics for models of regression tasks. Mean absolute error (MAE) is a commonly used method for error analysis, as it measures the error based on the absolute value of the difference. Mean squared error (MSE) and root mean squared error (RMSE) are more sensitive to outliers. The correlation coefficient (r) indicates the strength of linear correlation between variables. The coefficient of determination (R^2) is a statistical value that measures the degree of fitness of a model to the task [88]. In general, a smaller value of the first three metrics and a value close to 1 for the last two metrics indicate a stronger generalization ability and accuracy of the model.

Meanwhile, a cross-validation method is usually used to avoid coincidence in a set of data [89]. In that case, the dataset is separated into K equivalent parts, where one piece of data is used as the test set and the rest as the training set. The procedure is then cycled until each piece of the data is used as the test set once. Thus, the trained model can validly generalize to the entire dataset of all samples when cross-validation accuracy is high. In particular, when the number of samples is equal to K, indicating only one data in the test data, this method is called “leave-one-out cross-validation” (LOOCV). Although this method is highly accurate, it has a high computational overhead and is generally

Table 2
Machine learning algorithms.

ML Algorithm	Category	Description
Gaussian Process Regression (GPR)	Regression	A non-parametric model that uses the Gaussian Process for data regression analysis.
Bayesian Optimization (BO)	Regression	A global optimization strategy including a priori function and collection function core steps.
K-Nearest Neighbors (KNN)	Classification	Classification using the distance between points in a dataset.
Random Forest (RF)	Regression Classification	An algorithm that integrates multiple decision trees.
Extremely Randomized Trees (ERF)	Regression Classification	Similar to random forests, the selection of division points is more random.
Gradient Boost Regression Tree (GBRT)	Regression	An integrated learning algorithm consisting of multiple decision trees is used to fit each decision tree using gradient descent.
LightGBM	Regression Classification	A distributed gradient boosting framework based on decision trees algorithm.
Artificial Neural Networks (ANN)	Regression Classification	A neural network is an operational model made up of numerous nodes.
Deep Neural Networks (DNN)	Regression Classification	An algorithm for automatic feature learning of data using ANN (more than three hidden layers).
Graph Neural Networks (GNN)	Regression Classification	A branch of deep learning on the graph structure, which contains nodes and edges.

used when the amount of data is quite small. For small data sets, there is also the Bootstrapping method [90]. However, this approach may modify dataset distribution and introduce bias in estimation, so it is rarely used. For classification problems, the confusion matrix is an important method for model accuracy analysis. In the confusion matrix, model performance can be evaluated with the sum of diagonal elements divided by the sum of non-diagonal elements. The receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC) can also be used to estimate the accuracy of the classifier [91].

As mentioned before, the intention of ML is to summarize the hidden relationships between the descriptors of the data and the target property. Thus, it can support the experimental methods for screening perovskite materials and the performance evaluation of PSCs. Specifically, it can be divided into two steps. First, some unknown virtual samples can be mined to predict the target properties using ML models. Then, materials possessing the desired properties are selected among the hypothetical examples to further narrow the screening range for experimental validation.

3. Application of machine learning to perovskite material properties and candidate screening

The excellent photoelectric conversion efficiency of PSCs is attributed to the strong light absorption, the long diffusion length of carriers, minimal compounding and tolerance of defects in perovskites [92]. Perovskites are the focus of much of the research in the field of PSCs. The excellent properties (bandgap, formation energy, etc.) of the light-absorption layer (perovskites) play a critical role in the performance of PSCs.

3.1. Formability

The relationship between atomic and material properties was first exploited by Goldschmidt in 1926 [74], who proposed that the Goldschmidt tolerance factor (T_f) could be used to predict the probability of a pair of ions forming perovskite structures as shown in Eq. (1). In addition, the physical significance of the octahedral factor (O_f), as shown in Eq. (2), is the closeness among the anion X and the cation B, which reflects the structural relationship with respect to the anion X and the cation B. This also indicates the formability of the perovskite structure. Recently, Bartel *et al.* [75] developed a new tolerance factor (τ), which is more accurate than T_f for predicting the formability of perovskite structures of ABX_3 compounds with up to 92 % probability, as shown in Eq. (3).

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

$$O_f = r_B / r_X \tag{2}$$

$$\tau = \frac{r_X}{r_B} - n_A\left(n_A - \left(\frac{r_A}{r_B}/\ln\left(\frac{r_A}{r_B}\right)\right)\right) \tag{3}$$

Where r_A, r_B and r_X represents are A, B and X-site ionic radius, respectively; n_A denotes the oxidation state of the A-site.

In 2016, Pilania *et al.* [76] constructed an SVM-based classifier. The formability of a given composition in the crystal structure of ABX_3 halide perovskites was predicted by using only elemental features. Pilania learned by building the model from a dataset of 181 experimentally known ABX_3 compounds searched in the literature. In feature selection, Pilania determined that the ionic radius, T_f and O_f were the most crucial classification criterion, indicating that spatial and geometric stacking effects predominate in halide synthesis. Then, numerous unique ABX_3 compositions with perovskite crystal structures were effectively predicted by altering the elemental composition and training a high-accuracy model. Lu *et al.* [93] collected 539 samples of HOIPs and 24 samples of non-HOIPs, using 10 model algorithms and 10 sampling methods for unbalanced ML of formability and tested with an accuracy of more than 95 %. In addition, Shapley additive interpretation (SHAP) [94] was used to analyze the effect of different composition fragments on formability.

Deep learning is also a powerful tool for predicting material properties. Gu *et al.* [95] proposed a GNN model to evaluate the synthesizability of perovskites, and the model-building process is shown in Fig. 2a. Due to the insufficient number of samples, it is difficult to meet the demand of deep learning for the amount of data. Domain-specific semi-supervised and migration learning are applied to the GNN (Fig. 2b-d), and the model finally achieves a true positive rate of 0.957. Meanwhile, 179 virtual crystals have been synthesized out of 962 virtual crystals predicted to be synthesized by this ML model, reaffirming the reliability of the ML model for the prediction of perovskite synthesizability. Zhang *et al.* [77] combined ML with the SHAP method to accelerate the discovery of potential HOIPs. Based on the model predictions, 198 non-toxic candidates out of 18,560 virtual samples had a probability of formability greater than 99 %. SHAP analysis revealed that the radius and lattice constant of the B site were positively connected with formability. However, the ionic radius,T_f, and initial ionization energy of the A-site were negatively correlated.

3.2. Bandgap

The bandgap directly determines the response range of the solar spectrum and the conductivity of the PSCs [96]. Therefore, different methods have been used to calculate this electronic property. Certain traditional methods, such as Perdew-Burke-Ernzerhof (PBE) [97] and *meta*-GGA [98], have high accuracy for bandgap estimation. Nevertheless, these methods are usually resource-intensive and computationally expensive. ML provides an alternative approach to overcoming these problems. Pilania *et al.* [80] proposed a powerful ML model and a set of elemental descriptors for effective bandgap prediction of double perovskites and the workflow is shown in Fig. 3a. These optimal sets of descriptors based on elemental features are identified by searching the huge feature space. Based on a database containing 1300 bandgaps of double perovskites, the KRR-based ML model is trained and tested. They found that the lowest occupation energy level of the A-site element and the electronegativity of the B-site element determine the bandgap. In out-of-sample data, the trained model successfully predicts bandgap for several single perovskites and further demonstrates the generalization ability of the model. Vakharia *et al.* [99] used a dataset consisting of 240 organometallic halide perovskites to train two ML models, ElasticNet and a conservative regression model. ElasticNet and tenfold cross-validation results were calculated with a MAE of 0.09 eV for the Cs-based perovskites. In comparison, the perovskites based on methylammonium (MA) with conservative regression obtained a MAE of 0.34 eV. Omprakash *et al.* [100] trained the GNN model (MEGNet) [101] to predict all types of perovskites (organometallic, all inorganic, low-symmetry) bandgap and the MAE of the model was 0.28 eV. In addition, a pipeline was created, making the prediction of the bandgap for inorganic perovskites easier.

Besides the direct prediction of the perovskite bandgap via ML, the application of tunable bandgap is also effective. In the work of Park *et al.* [102], the bandgap variation due to organic cation mixing in heterogeneous organic–inorganic halide perovskites was investigated in depth using the GBRT model, and the relevance of the input features was also explored. Elemental and structural features were applied to the ML model with model errors and scatter plots shown in Fig. 3b, demonstrating that the bandgap increased by doping with dimethylamine (DMA) is related to the slight local distortion of octahedra. Li *et al.* [103] revealed the bandgap tuning strategy for cations and halide ions in lead halide perovskites is revealed by ML. The ANN predicts the bandgap with higher accuracy (RMSE of 0.05 eV) by considering the interaction of cations and halide ions. In addition, the ANN algorithm predicts the composition of mixed halide perovskites with an ideal bandgap and high

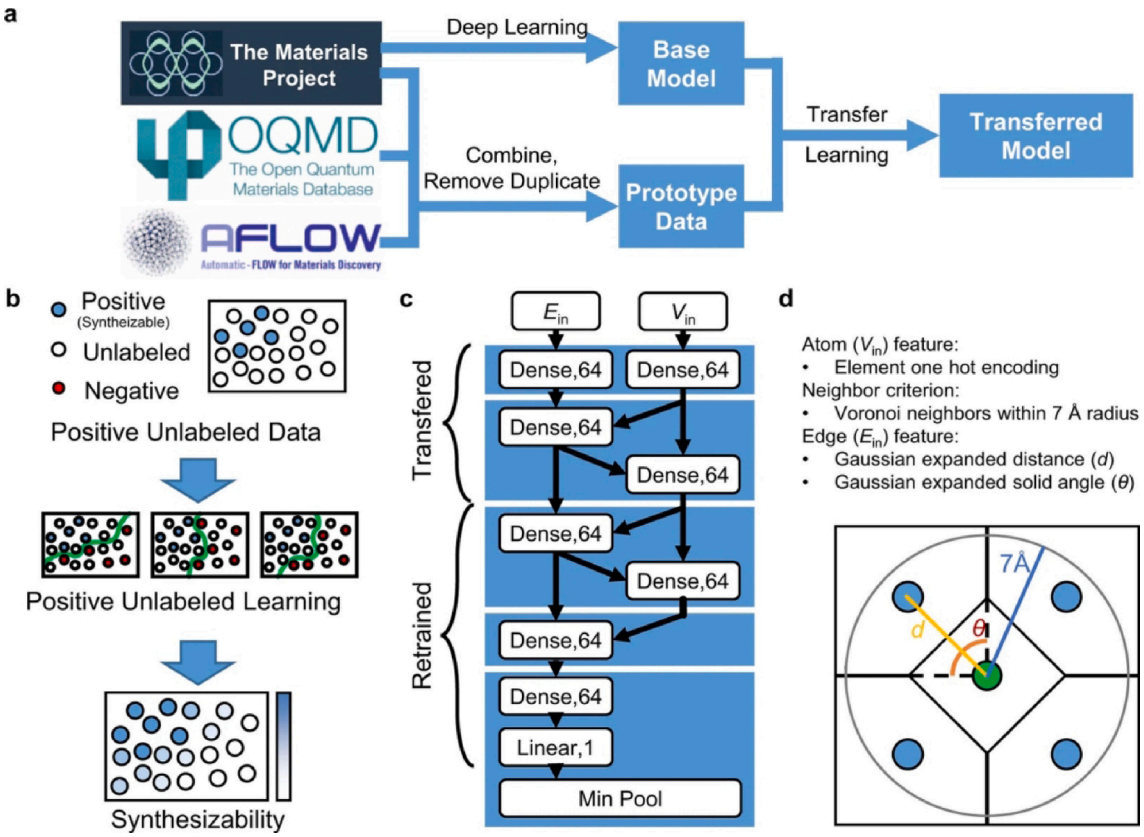

Fig. 2. (a) Domain-specific transfer learning workflow. (b) Semi-supervised learning strategies. (c) The architecture of deep learning networks. (d) Feature representations. Reproduced with permission [95].

iodide ratio and can suppress the suppressing halide segregation effect.

3.3. Stability

Researchers have explored the issue of the stability of perovskites for a long time [104,105]. The formation energy expresses the stability of a compound with respect to the composition of its elemental phases. Normally, the more negative value of the formation energy indicates the more stable the compound. Im *et al.* [78] used the GBRT algorithm to train the model in combination with the DFT calculated halide double perovskites dataset and the RMSE errors of formation energy and bandgap were respectively 0.024 (eV/atom) and 0.223 (eV). It needs to be noted that the formation energy error is comparable to the DFT calculation error value. The interpretable analysis of feature importance likewise reveals the importance of feature selection. Similarly, the GBRT algorithm was applied to the candidate halide double perovskites dataset to predict the formation energy and bandgap. Lu *et al.* [79] also used the GBRT model to predict 5158 unexplored HOIPs. With appropriate bandgap and thermal stability, six lead-free HOIPs were successfully screened. The workflow is shown in Fig. 4a. They noted that this outstanding performance is attributed to four characteristics: T_f, O_f, the electronegativity of metal ions, and the polarizability of organic molecules. This research provides a practical approach to the ideal HOIPs solar cells. Talapatra *et al.* [106] assembled a huge dataset of single and double perovskites by considering 68 elements in the periodic table. RF classifier was used to train the model with a mean accuracy of 0.9401 for formative classification and 0.9409 for stability classification. In particular, they pointed that the threshold range should be chosen carefully when determining the stability differentiation criteria.

The energy beyond the convex hull (E_{hull}) is important in assessing the dynamic synthesizability of the material. Schmidt *et al.* [108] calculated a dataset of 20,000 ABX_3-type perovskites containing E_{hull} by DFT. The extremely randomized trees model was used to predict the remaining 23,000 potential combination of perovskites. Finally, 641 potentially stable perovskite compounds were selected. Li *et al.* [107] used 1929 DFT-calculated ABO_3 compounds to train the ML model on a dataset of ABO_3 compounds. Furthermore, a set of 791 features was constructed using a combination of elemental attributes, and the first 70 features were selected using feature selection. Fig. 4b shows the model performance. The RMSE value was 28.5 (±7.5) meV/atom in predicting E_{hull}, demonstrating the availability for rapid screening of new candidate materials in a large combinatorial space by ML. Xie *et al.* [109] developed a crystal graph convolutional neural network (CGCNN) to use material structure and atomic information for material property predictions (Fig. 5a-b). The CGCNN was applied to train and predict E_{hull} for 18,928 ABX_3-type perovskites (Fig. 5c). The CGCNN model is highly accurate with a MAE value of 0.13 (Fig. 5d), while the B-position of groups 4–6 was found to be the most stable for the element. Similarly, the A-site is most stable when it is a large atom of groups 13–15. Subsequently, after limiting the search, 33 perovskites were found out of 378 compounds.

3.4. Other properties

As a relatively important parameter, the lattice constant affects the formation and stability of the structure and electronic structure, thus the material behavior. In addition, the crystal system identification of perovskites is of great interest. Zhang *et al.* [111] developed a GPR model to illustrate the relationship between the ionic radius and lattice constants of cubic perovskites. The result shown the model is highly accurate and stable and contributes to fast and robust lattice constant estimation. Behara *et al.* [110] used the LightGBM model to classify perovskites into different crystal systems (cubic, tetragonal, rhombohedral, or rhombic) using 13 elemental and structural features with a

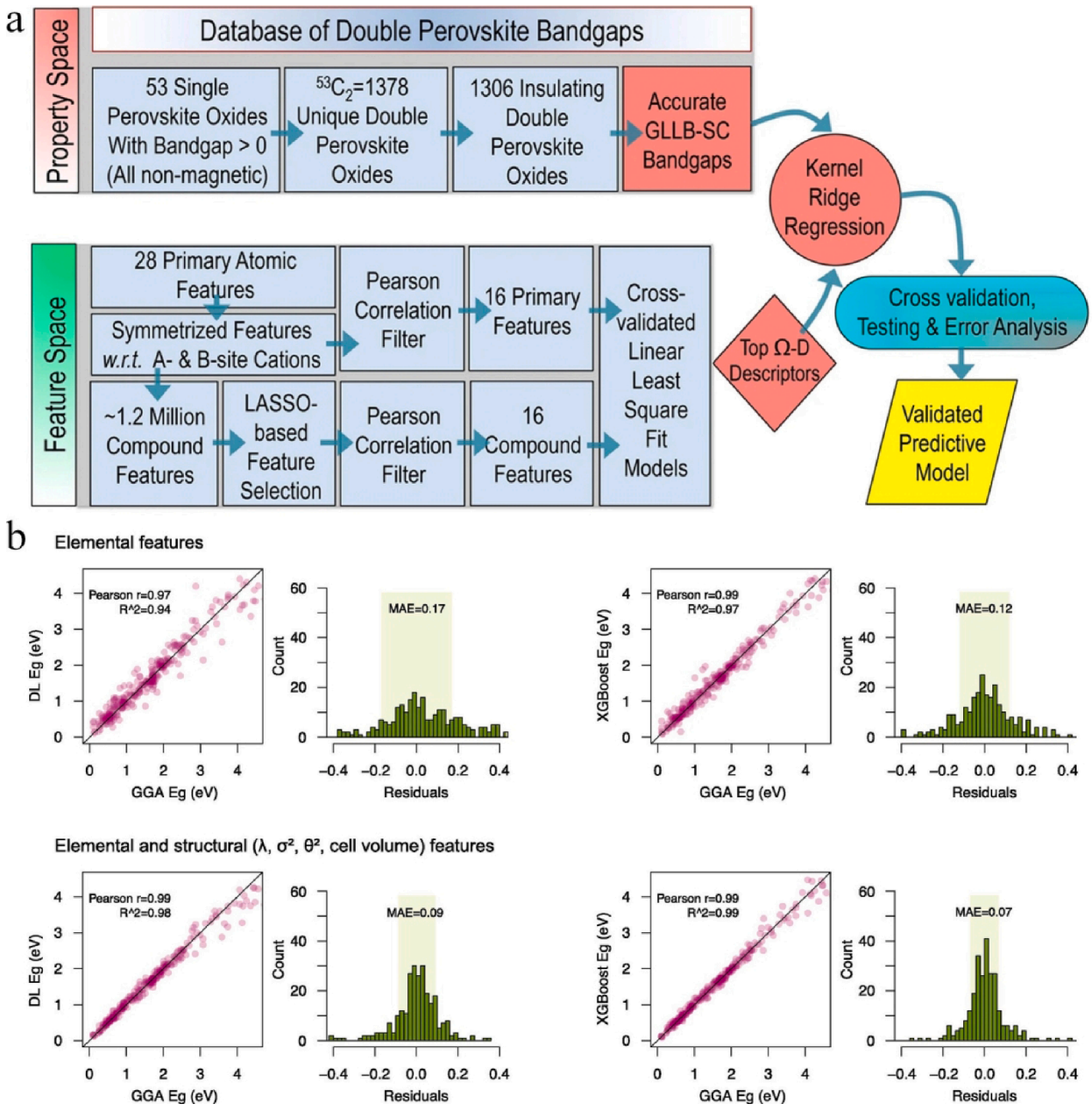

Fig. 3. (a) Flow diagram of ML predicted bandgap of double perovskites. (b) Error analysis diagram using elemental and structural features. Reproduced with permission [80,102].

prediction accuracy of 80.3 %. The mapping of features to crystal structures is illustrated by SHAP (Fig. 5e). Priya *et al.* [112] demonstrated a robust ML framework with 111 different features selected for efficient and accurate prediction of the total conductivity of perovskite materials. In addition, the total conductivity was used as an additional feature to classify the charge carriers according to their types at different temperatures and environmental conditions. Furthermore, the ML model screens out perovskites with high conductivity for different application situations.

Many applications have already emerged for ML coupled with perovskites data points obtained through theoretical calculations or experiments. Surprisingly, the accuracy of the prediction of multiple properties can be comparable to that of DFT calculations. It is believed that further combination of techniques such as ML and DFT can lead to more efficient exploitation of perovskite materials.

4. Application of machine learning in perovskite solar cells.

PSCs mainly consist of a cathode layer (metal), ETL, light-absorbing layer (perovskites), HTL, and transparent conducting oxide (TCO, anode) by functional layering (Fig. 6a). When sunlight is irradiated on the PSCs, the perovskites act as a light-absorbing layer to absorb light, generate excitons, and produce charge carriers (electrons and holes) as the exciton dissociates. Electrons and holes are respectively separated and injected into ETL and HTL. The working and counter electrode gather electrons and holes, then transport them to the external to generate current [113]. The working mechanism of PSCs is shown in Fig. 6b. The key research question in this field involves finding the optimal fabrication parameters to prevent material degradation. As a result, a large number of studies have explored the degradation patterns and mechanisms of perovskites under different environmental conditions [114–116]. However, much less research has applied high-

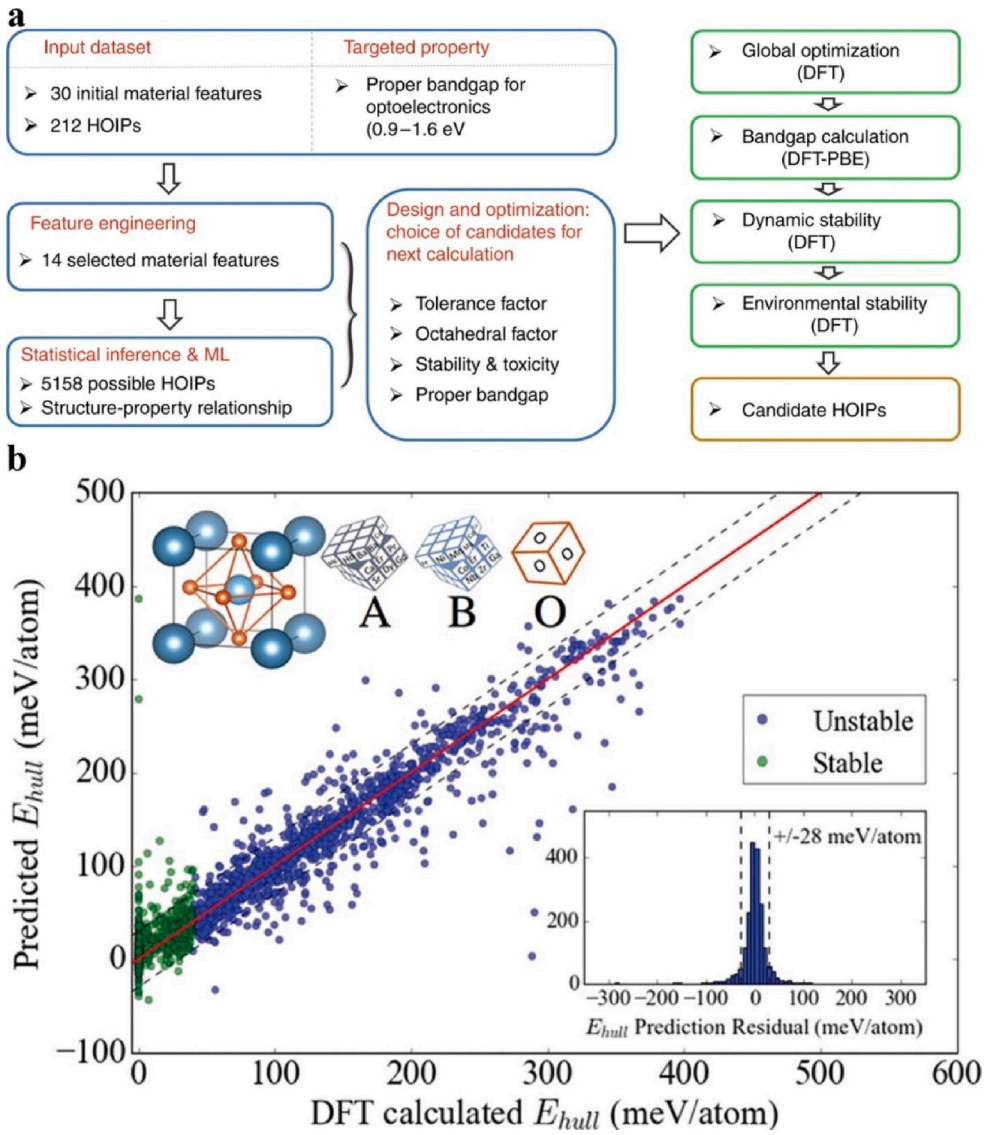

Fig. 4. (a) Screening workflow of HOIP. (b) Model performance and analysis. Reproduced with permission [79,107].

throughput experiments or ML models to design perovskite materials [117]. ML can reduce experiment time and provide avenues for future research. In addition, the analysis of features by the model can provide special physical insights.

4.1. *In the overall perovskite solar cell design*

ML can find potential insights from a large number of data points, which is instrumental in optimizing material composition and predicting the performance of PSCs. Li *et al.* [47] developed an efficient solar cell providing a two-step strategy. First, an ML model was developed to predict the bandgap of the perovskite materials. Then, a simple model was constructed to predict the performance of the PSCs considering the energy levels of the perovskites together with the HTL/ETL. Also, it is concluded from the ML analysis that for higher PCE, lower values of ΔH (HOMO energy difference between HTL and perovskites) and ΔL (LUMO energy difference between ETL and perovskites) are necessary for perovskites with lower bandgap. In contrast, higher values are required for perovskites with higher bandgaps. Liu *et al.* [51] further used ML for interpretable analysis of PSCs. They searched 814 data points from publications for ML model training, and the RMSE of the RF model with PCE reached 0.0158. By SHAP analysis, it is the composition of the light-absorbing layer that has the greatest effect on the PCE, and the increase of formamidine (FA) content has an obvious increase on the PCE. In addition, the reduction of energy level difference and the appropriate increase of electron mobility also improve the PCE. Gok *et al.* [118] apply a two-step ML approach. First, the bandgap of perovskites is predicted with ML and verified using experiments. Then, the performance of PSCs is evaluated using eight different perovskites. They found that the components with high electronegativity show a lower absorption onset point. In contrast, the lattice constants of A-site substitutions (except Rb) demonstrate a positive correlation with the absorption onset point. Bak *et al.* [119] developed a DNN-based strategy algorithm, and the DNN successfully determined the optimal device configuration. The proposed structure fabricated Sn-based PSCs exhibited a PCE of 5.57 %. The experimental dataset was collected from 49 selected publications on Sn-based PSCs, from which 122 different device configurations were obtained. The input characteristics were classified as perovskites component, metal electrode, transparent electrode, HTL, and ETL. Output properties consisted of photovoltaic parameters, i.e., short-circuit current density (J_{sc}), and open-circuit voltage (V_{oc}), and fill factor (*FF*).

The stability issue is one of the bottlenecks in the development of PSCs. It is important to explore the stability of PSCs and the factors

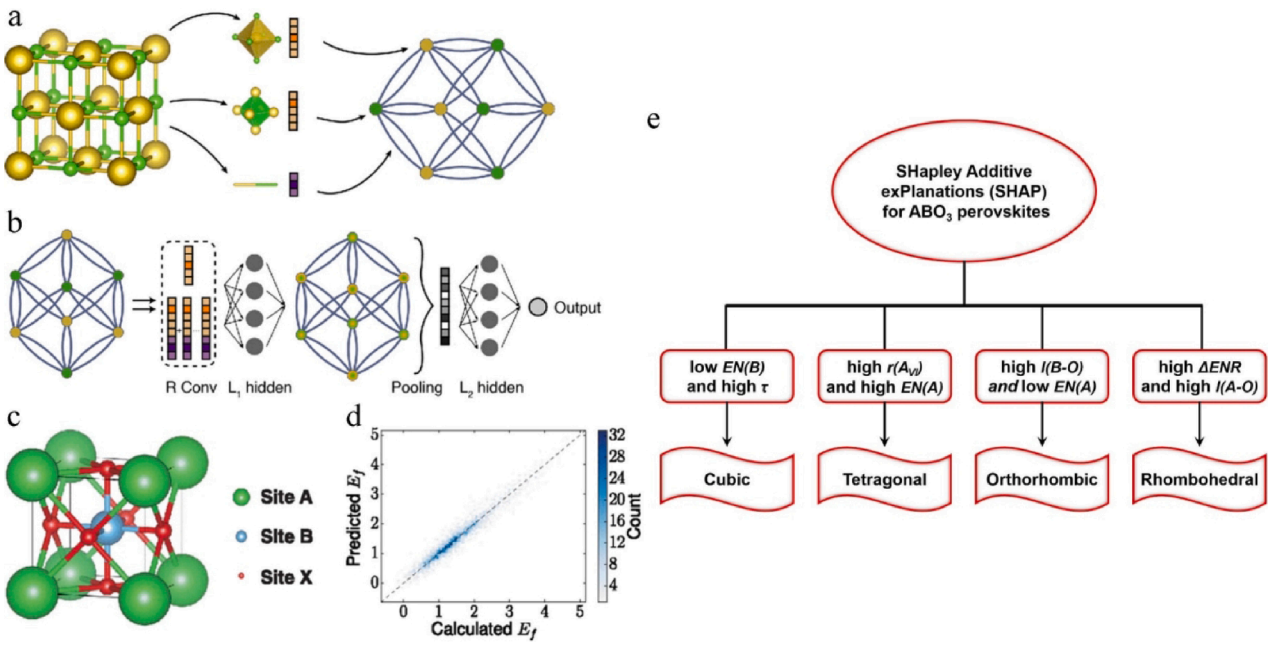

Fig. 5. (a) Crystal graph construction. (b) The architecture of convolutional neural networks. (c) The structure of ABX_3 Perovskites. (d) Scatter plot shows the model performance. (e) The understanding of perovskite structural classification by ML. Reproduced with permission [109,110].

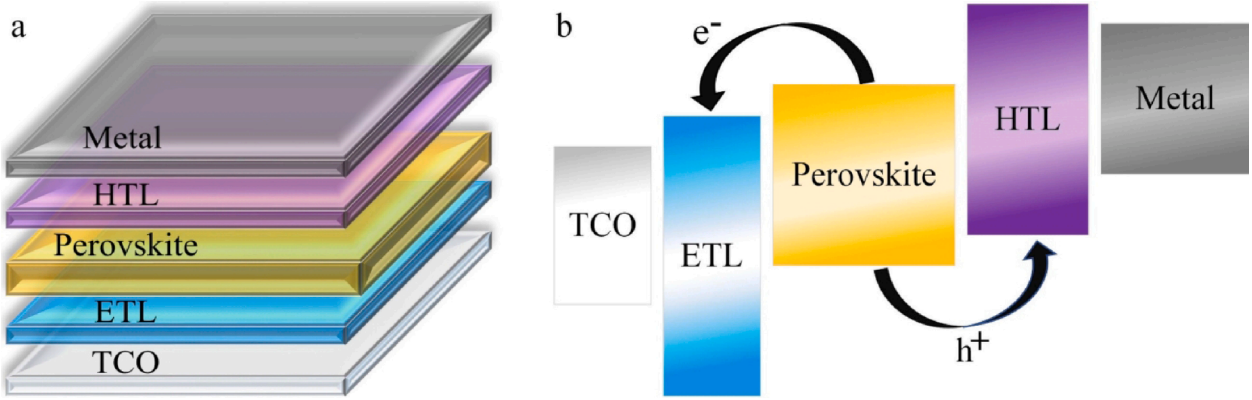

Fig. 6. (a) The structure of PSCs usually includes metal, ETL, light-absorbing layer (perovskites), HTL and TCO. (b) The working mechanism of PSCs. Sunlight absorbed by perovskite layer in PSCs generates excitons and charge carriers (electrons and holes) that are separated and injected into ETL and HTL, resulting in current generation at the working and counter electrodes.

affecting the stability. Odabaşı *et al.* [82] analyzed the effect of environmental conditions on the stability of PSCs. The effect of PSCs-related parameters (perovskites type, ETL, HTL, back contact, deposition details, etc.) was then investigated by association rule mining. Lastly, a decision tree was developed to identify the cell efficiency degradation on different days to deduce the rules and procedures that would aid in the creation of stable PSCs. Unfortunately, no standard has been established for testing the stability of PSCs. This limits the widespread use of ML technology because non-standard measurement and storage conditions, along with unclear and incomplete reporting, can increase the variability of the data. In predicting the performance of PSCs, HOMO and LUMO features play a key role [120]. However, recent research has shown that there is a need to investigate new descriptors in the PCE prediction of optimal organic photovoltaics (OPV), such as the near-degeneracy of frontier molecular orbitals in donor molecules [121,122]. Sahu *et al.* [81] trained models using multiple ML methods to predict the efficiency of OPV by collecting a dataset of 280 small-molecule OPV. Among them, the gradient boosting model (Pearson coefficient = 0.79) performed well, as shown in Fig. 7a–d. Besides, they found that the frontier molecular orbitals of donor molecules degenerate in almost all high-performance devices. Thus, orbitals close to HOMO/LUMO energetically are also involved in the exciton formation/dissociation and carrier transport, which is crucial for data-driven PCE studies, including PSCs.

Hole transport materials (HTMs) are important structures in PSCs that reduce charge complexation and improve the stability of the perovskite layer. Cueto *et al.* [124] developed an ML model to predict the PCE of PSCs. A series of features were used to describe HTM properties (fingerprints, structure and electron properties and additives), together with perovskites type and cell structure. The model could be applied to identify candidate HTMs more likely to present a larger PCE and demonstrate correlations between specific molecular fragments. She *et al.* [125] used several ML methods to investigate the effect of doped ETL on PCE. A dataset containing 2006 samples of PSCs was built, and ML models were trained based on this dataset. The result shows that having a high PCE is closely related to using SnO_2. The doping of ETL can modulate the conduction band minimum and fermi energy level, thus enhancing the PCE. In addition, two high-efficiency PSCs are

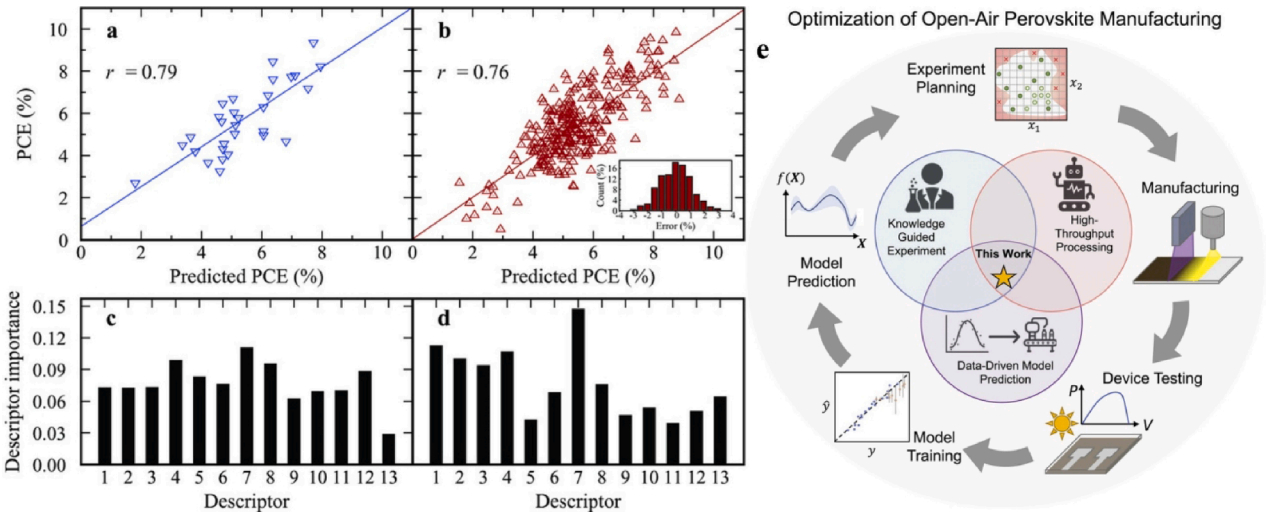

Fig. 7. (a) The LOOCV technique was used to validate the accuracy of the GBRT model. (b) For all data points of the GBRT model, the inset shows the probability density of the prediction error. The importance of descriptors for (c) the GBRT and (d) RF models, respectively. (e) The workflow for optimizing PSCs by RSPP and ML. Reproduced with permission [81,123].

predicted. Interface materials not only reduce surface defects at the interface but also modify the energy at the perovskites/HTL, facilitating carrier extraction. Liu *et al.* [126] used RF models to predict the PSCs efficiency of more than 100 perovskites/HTL. The study uses the features of the interface material, perovskites, and the PCE of the control device. The combination of ML and experiment shows that it performs best when the interface material is MA iodide. This study offers a workflow for ML-assisted PSCs screening of interface materials. Yan *et*

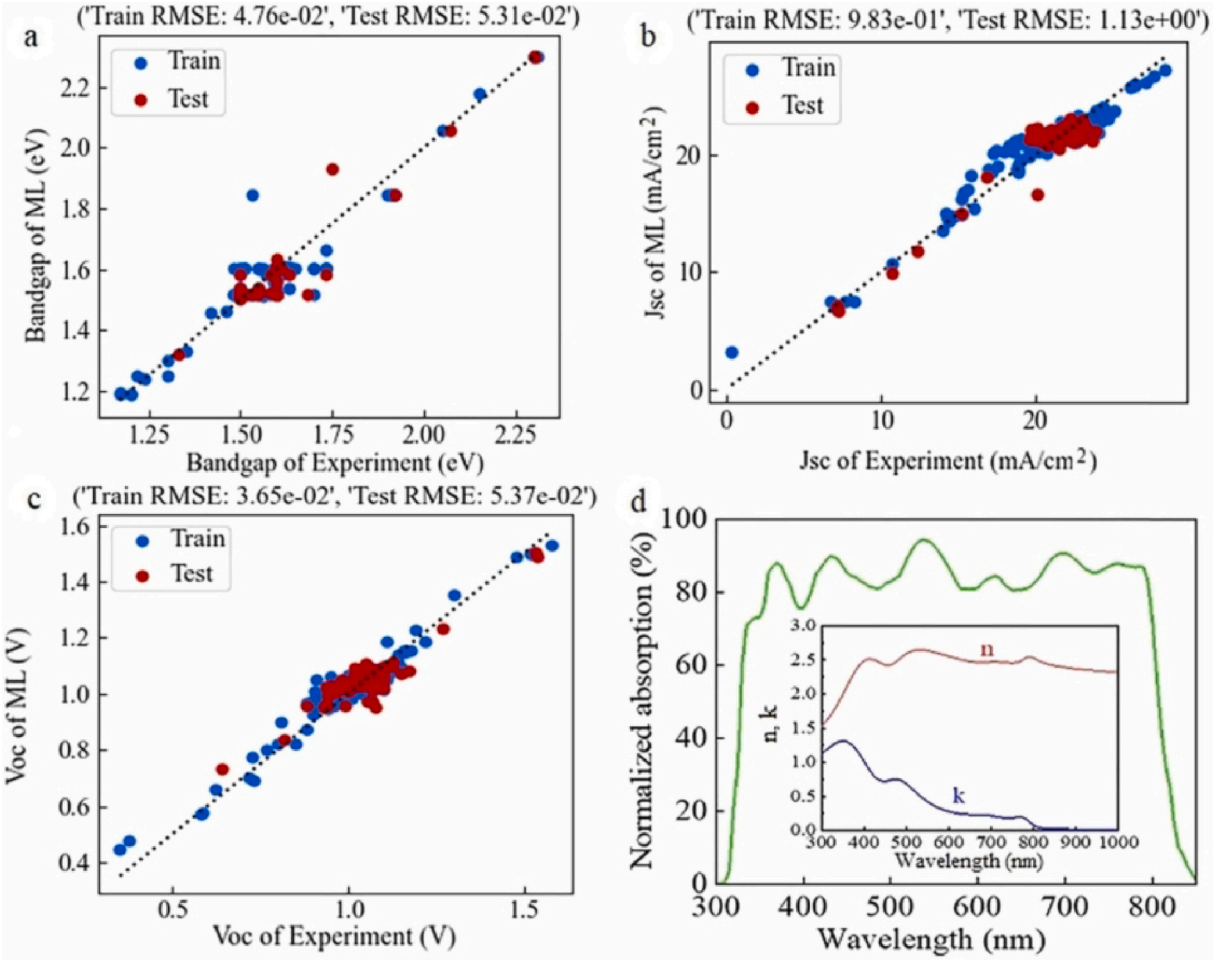

Fig. 8. Predicted versus experimental values: (a) bandgap, (b) J_{sc} and (c) V_{oc}. (d) Simulated optical absorption of PSCs in the wavelength range of 300–850 nm with curves for measuring the complex optical constants (n, k). Reproduced with permission [127].

al. [127] used ML to predict the bandgap, J_{sc}, and V_{oc} of five components composed of $(FAPbI_3)_x(MAPbBr_{2.8}Cl_{0.2})_{1-x}$ perovskites (Fig. 8). The results showed that the ML predictions were very close to the experimental values. It was found that the J_{sc} was improved by resisting the light management design using reflective PDMS nanocone arrays, resulting in a higher PCE.

4.2. In the optimization of processing and fabrication parameters for perovskite solar cells

In order to obtain stable and high PCE PSCs, process parameters for high-dimensional manufacturing technologies need to be optimized. Data-driven ML can accelerate this process and achieve promising accuracy. Liu *et al.* [123] proposed a BO framework that enables the integration of domain knowledge into the ML, as shown in Fig. 7e. With optimized PSCs by open-air rapid spray plasma processing techniques, the presented framework shows quicker optimization compared to other traditional experimental design approaches, achieving a device efficiency of 18.5 %. Yu *et al.* [128] applied ML to auxiliary material development. The suitability of $MAPbI_3$ perovskite films with different post-treatment amine types is discussed and analyzed. In addition, amine compositions with high suitability with perovskites films are identified. It is also pointed out that small primary amines with multiple hydrogen bonds tend to destroy perovskites during the post-treatment process and should be used with caution during latent liquid fabrication.

The photonic curing process allows rapid crystallization of perovskites and has tremendous commercial potential [129]. Xu *et al.* [130] used a BO algorithm for process optimization of the parameters affecting photonic curing. The BO yielded PCE values greater than ten per cent compared to conventional optimization methods. In addition, SHAP was used for interpretability analysis. Interfacial passivation can improve the PCE of PSCs due to the inhibition of nonradiative carrier complexation [131]. Zhi *et al.* [132] constructed a descriptor set containing molecular and chemical structures to predict the PCE of interfacial passivated PSCs under different ammonium salts using an ensemble learning approach. It is recognized that ML-assisted material screening and the interpretability of different features have a positive contribution to improving the performance and stability of PSCs.

Combination losses are essential in the process of PSCs photogeneration. Vincent *et al.* [134] used large-scale drift–diffusion simulations to better understand the light intensity dependence of V_{oc} and its correlation with the dominant combination process. An automatic identification tool using an ML approach was introduced to determine the dominant compounding losses using the light intensity-dependent performance as input, with >80 % classification accuracy. Kim *et al.* [133] investigated the interaction of charge transport kinetics and complex processes under environmental stimuli and illumination using the non-negative matrix factorization (NMF) method to understand the chemical and structural instability of HOIPs-based PSCs. As shown in Fig. 9, they further use a model-independent relaxation time distribution (DRT) method to analyze the impedance spectrum. This allows a better understanding of the interaction of charge transport under ambient and solar illumination.

ML technology has demonstrated successful applications in recognizing potential candidate perovskite materials for solar cell applications, including those with suitable bandgaps, lead-free compositions, and stability characteristics. Additionally, ML techniques have been employed in screening for high PCE and stable perovskite PSCs, as well as optimizing manufacturing environmental conditions. One of the key advantages of ML is its ability to process large amounts of data quickly and efficiently. By analyzing extensive datasets, ML algorithms can reveal underlying physical rules and correlations that may not be readily

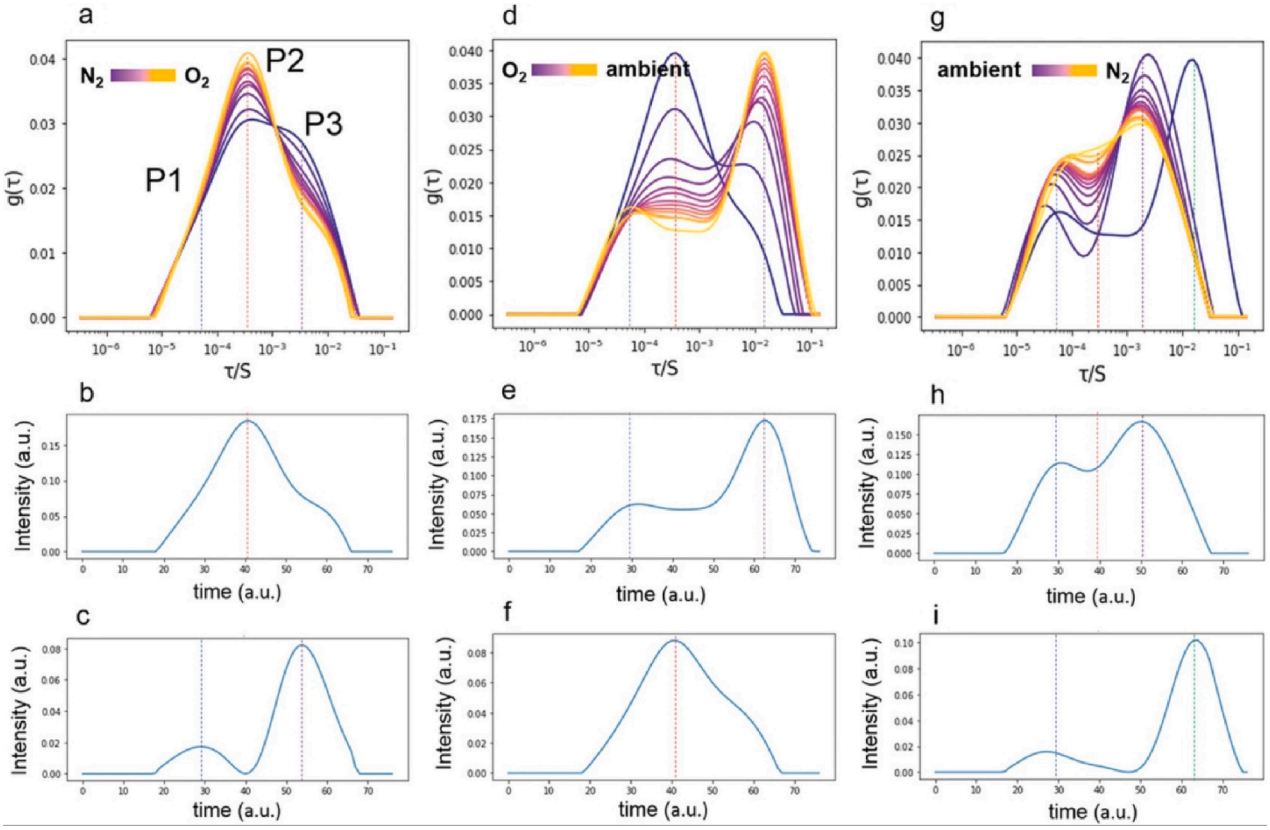

Fig. 9. Analysis of DRT and decomposition of components via ML. (a) DRT spectra with time and (b, c) DRT curves of two NMF components (from N_2 to O_2). (d) DRT spectra with time and (e, f) DRT curves of two NMF components (from O_2 to ambient condition). (g) DRT spectra with time and (h, i) DRT curves of two NMF components (from ambient conditions to N_2). Reproduced with permission [133].

apparent through traditional methods. This accelerates the discovery of novel PSCs with improved properties and opens up new avenues for exploration in this field.

5. Conclusions and perspectives

The combination of PSCs and ML has become a hot research topic in recent years due to the powerful data analysis capability of ML. Overall, this paper provides a comprehensive review of ML techniques, ranging from the screening of perovskite materials to their specific applications in PSCs. The application of ML techniques can not only reduce costs and save time, but also lead to improved properties and higher performance. Furthermore, when ML techniques are combined with other tools and fed with additional theoretical and experimental data, more accurate and reliable results can be obtained. With the ongoing development of open-source databases and computational tools, it is expected that this trend will continue to accelerate.

Although many studies have achieved good results in the application of ML to perovskite materials and PSCs research, there is still much work to be done to enhance its effectiveness further. First, data is the cornerstone of ML, and adequate and robust data is inseparable from its success. As mentioned above, many open-source databases have been established, but there are still limited databases like the one developed by Unger et al. [135] specifically for perovskite materials and PSCs designs. Therefore, a more comprehensive, standardized, and generalized database of perovskite information should be established to accelerate data sharing and reduce data access barriers. Secondly, on the feature engineering side, there is a need to propose more standardized feature or feature extraction methods, such as the feature generation method studied by Hirotomo *et al.* [136]. Thirdly, ML can be successful because of its ability to establish unique physical insights and interpretations of the selected features as well as the target attributes, which can aid in the discovery of underlying physical rules. A deeper study of algorithms specific to small data sets is needed at the ML algorithm level. In particular, for deep learning, the amount of data usually needs to be 10^4 [109], so deep learning is rarely applied to perovskite materials design and PSCs researches. Transfer learning is equivalent to the process of pre-training ML for small data sets, but its results often fall short of expectations [137]. Last but not least, ML should not be utilized in isolation and should be integrated more closely with other techniques. For example, experimental methods are accepted as an accurate way to discover and validate new materials at this stage, and ML should be combined with experimental methods and advanced circularly. The data obtained from experiments can be added to the input of ML to obtain more accurate predictions. It is also noteworthy that in the direction of PSCs, a large number of studies have explored the degradation patterns and mechanisms of perovskite materials under various environmental conditions. However, much fewer studies have applied high-throughput experiments or ML models to the design of perovskite materials, which is detrimental to the development of PSCs.

In summary, with the advancements of materials informatization and the rapid development of computational techniques and methods, ML will find wider applications in perovskite materials and PSCs. With increasingly robust open-source databases, ML will accelerate the discovery of perovskite materials, gain physical understanding and accelerate the progress of actual PSCs. Moreover, the convenience and efficiency of the ML approach surpass traditional methods in the development of perovskite materials and PSCs. It is believed that ML will become an indispensable tool for calculations and experiments in the field of materials science in the future.

CRediT authorship contribution statement

Zhan Hui: Investigation, Visualization, Writing – original draft, Writing – review & editing. **Min Wang:** Resources, Project administration, Validation. **Xiang Yin:** Conceptualization, Methodology, Formal analysis. **Ya'nan Wang:** Software, Data curation. **Yunliang Yue:** Conceptualization, Writing – review & editing, Supervision, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This research was supported by the Natural Science Foundation of Jiangsu Province, China (Grant No. BK20190878), and the Universities Natural Science Research Project of Jiangsu Province, China (Grant No. 19KJB510062).

References

[1] D.P. Tabor, L.M. Roch, S.K. Saikin, C. Kreisbeck, D. Sheberla, J.H. Montoya, S. Dwaraknath, M. Aykol, C. Ortiz, H. Tribukait, Accelerating the discovery of materials for clean energy in the era of smart automation, Nat. Rev. Mater. 3 (2018) 5–20.
[2] E. Kabir, P. Kumar, S. Kumar, A.A. Adelodun, K.-H. Kim, Solar energy: Potential and future prospects, Renew. Sustain. Energy Rev. 82 (2018) 894–900.
[3] I. Mathews, S.N. Kantareddy, T. Buonassisi, I.M. Peters, Technology and market perspective for indoor photovoltaic cells, Joule 3 (2019) 1415–1426.
[4] K. Ranabhat, L. Patrikeev, A. Antal'evna-Revina, K. Andrianov, V. Lapshinsky, E. Sofronova, An introduction to solar cell technology, J. Appl. Eng. Sci. 14 (2016) 481–491.
[5] P.K. Nayak, S. Mahesh, H.J. Snaith, D. Cahen, Photovoltaic solar cell technologies: analysing the state of the art, Nat. Rev. Mater. 4 (2019) 269–285.
[6] J. Yan, B.R. Saunders, Third-generation solar cells: a review and comparison of polymer: fullerene, hybrid polymer and perovskite solar cells, RSC Adv. 4 (2014) 43286–43314.
[7] N.S. Kumar, K.C.B. Naidu, A review on perovskite solar cells (PSCs), materials and applications, J. Materiomics. 7 (2021) 940–956.
[8] W. Li, Z. Wang, F. Deschler, S. Gao, R.H. Friend, A.K. Cheetham, Chemically diverse and multifunctional hybrid organic–inorganic perovskites, Nat. Rev. Mater. 2 (2017) 1–18.
[9] A. Kojima, K. Teshima, Y. Shirai, T. Miyasaka, Organometal halide perovskites as visible-light sensitizers for photovoltaic cells, J. Am. Chem. Soc. 131 (2009) 6050–6051.
[10] C. Wehrenfennig, G.E. Eperon, M.B. Johnston, H.J. Snaith, L.M. Herz, High charge carrier mobilities and lifetimes in organolead trihalide perovskites, Adv. Mater. 26 (2014) 1584–1589.
[11] H.J. Snaith, Perovskites: the emergence of a new era for low-cost, high-efficiency solar cells, J. Phys. Chem. Lett. 4 (2013) 3623–3630.
[12] M. Liu, M.B. Johnston, H.J. Snaith, Efficient planar heterojunction perovskite solar cells by vapour deposition, Nature 501 (2013) 395–398.
[13] M.M. Lee, J. Teuscher, T. Miyasaka, T.N. Murakami, H.J. Snaith, Efficient hybrid solar cells based on meso-superstructured organometal halide perovskites, Science 338 (2012) 643–647.
[14] X. Liu, G. Zhu, X. Wang, X. Yuan, T. Lin, F. Huang, Progress in black titania: a new material for advanced photocatalysis, Adv. Energy Mater. 6 (2016) 1600452.
[15] T. Torimoto, T. Tsuda, K.i. Okazaki, S. Kuwabata, New frontiers in materials science opened by ionic liquids, Adv. Mater. 22 (2010) 1196–1221.
[16] J. Wei, X. Chu, X.Y. Sun, K. Xu, H.X. Deng, J. Chen, Z. Wei, M. Lei, Machine learning in materials science, InfoMat. 1 (2019) 338–358.
[17] K.T. Butler, D.W. Davies, H. Cartwright, O. Isayev, A. Walsh, Machine learning for molecular and materials science, Nature 559 (2018) 547–555.
[18] S.M. Moosavi, K.M. Jablonka, B. Smit, The role of machine learning in the understanding and design of materials, J. Am. Chem. Soc. 142 (2020) 20273–20287.
[19] H.-S. Kim, C.-R. Lee, J.-H. Im, K.-B. Lee, T. Moehl, A. Marchioro, S.-J. Moon, R. Humphry-Baker, J.-H. Yum, J.E. Moser, Lead iodide perovskite sensitized all-solid-state submicron thin film mesoscopic solar cell with efficiency exceeding 9%, Sci. Rep. 2 (2012) 1–7.
[20] NREL, Best Research-Cell Efficiency chart. https://www.nrel.gov/pv/assets/pdfs/best-research-cell-efficiencies.20191104.pdf.
[21] A. Mingorance, H. Xie, H.S. Kim, Z. Wang, M. Balsells, A. Morales-Melgares, N. Domingo, W. Tress, J. Fraxedas, N. Vlachopoulos, Interfacial engineering of metal oxides for highly stable halide perovskite solar cells, Adv. Mater. Interfaces. 5 (2018) 1800367.

[22] N.-G. Park, M. Grätzel, T. Miyasaka, K. Zhu, K. Emery, Towards stable and commercially available perovskite solar cells, Nat. Energy. 1 (2016) 1–8.
[23] A. Ghaderian, N.H. Hemasiri, S. Ahmad, S. Kazim, Amplify the performance and stability of perovskite solar cells using fluorinated salt as the surface passivator, Energy Technol. (2023) 2200211.
[24] Q. Tu, I. Spanopoulos, E.S. Vasileiadou, X. Li, M.G. Kanatzidis, G.S. Shekhawat, V. P. Dravid, Exploring the factors affecting the mechanical properties of 2D hybrid organic–inorganic perovskites, ACS Appl. Mater. Interfaces. 12 (2020) 20440–20447.
[25] J. Liang, C. Wang, Y. Wang, Z. Xu, Z. Lu, Y. Ma, H. Zhu, Y. Hu, C. Xiao, X. Yi, All-inorganic perovskite solar cells, J. Am. Chem. Soc. 138 (2016) 15829–15832.
[26] H. Lei, D. Hardy, F. Gao, Lead-free double perovskite Cs2AgBiBr 6: fundamentals, applications, and perspectives, Adv. Funct. Mater. 31 (2021) 2105898.
[27] J. Zhao, Y. Deng, H. Wei, X. Zheng, Z. Yu, Y. Shao, J.E. Shield, J. Huang, Strained hybrid perovskite thin films and their impact on the intrinsic stability of perovskite solar cells, Sci. Adv. 3 (2017) eaao5616.
[28] J.M. Frost, K.T. Butler, F. Brivio, C.H. Hendon, M. Van Schilfgaarde, A. Walsh, Atomistic origins of high-performance in hybrid halide perovskite solar cells, Nano Lett. 14 (2014) 2584–2590.
[29] P.-K. Kung, M.-H. Li, P.-Y. Lin, J.-Y. Jhang, M. Pantaler, D.C. Lupascu, G. Grancini, P. Chen, Lead-free double perovskites for perovskite solar cells, Sol. RRL. 4 (2020) 1900306.
[30] J. Tian, Q. Xue, Q. Yao, N. Li, C.J. Brabec, H.L. Yip, Inorganic halide perovskite solar cells: progress and challenges, Adv. Energy Mater. 10 (2020) 2000183.
[31] W. Chen, X. Li, Y. Li, Y. Li, A review: crystal growth for high-performance all-inorganic perovskite solar cells, Energy Environ. Sci. 13 (2020) 1971–1996.
[32] T. Wu, J. Wang, Deep mining stable and nontoxic hybrid organic-inorganic perovskites for photovoltaics via progressive machine learning, ACS Appl. Mater. Interfaces. 12 (2020) 57821–57831.
[33] D. Jain, S. Chaube, P. Khullar, S.G. Srinivasan, B. Rai, Bulk and surface DFT investigations of inorganic halide perovskites screened using machine learning and materials property databases, PCCP. 21 (2019) 19423–19436.
[34] Z. Li, Q. Xu, Q. Sun, Z. Hou, W.J. Yin, Thermodynamic stability landscape of halide double perovskites via high-throughput computing and machine learning, Adv. Funct. Mater. 29 (2019) 1807280.
[35] J.-P. Correa-Baena, M. Saliba, T. Buonassisi, M. Grätzel, A. Abate, W. Tress, A. Hagfeldt, Promises and challenges of perovskite solar cells, Science 358 (2017) 739–744.
[36] P. Wang, Y. Wu, B. Cai, Q. Ma, X. Zheng, W.H. Zhang, Solution-Processable perovskite solar cells toward commercialization: Progress and challenges, Adv. Funct. Mater. 29 (2019) 1807661.
[37] W. Ke, M.G. Kanatzidis, Prospects for low-toxicity lead-free perovskite solar cells, Nat. Commun. 10 (2019) 1–4.
[38] A.A. Emery, C. Wolverton, High-throughput DFT calculations of formation energy, stability and oxygen vacancy formation energy of ABO3 perovskites, Sci. Data. 4 (2017) 1–10.
[39] C. Kim, T.D. Huan, S. Krishnan, R. Ramprasad, A hybrid organic-inorganic perovskite dataset, Sci. Data. 4 (2017) 1–11.
[40] K. Lejaeghere, G. Bihlmayer, T. Björkman, P. Blaha, S. Blügel, V. Blum, D. Caliste, I.E. Castelli, S.J. Clark, A. Dal Corso, Reproducibility in density functional theory calculations of solids, Science 351 (2016).
[41] P. Lopez-Varo, J.A. Jiménez-Tejada, M. García-Rosell, S. Ravishankar, G. Garcia-Belmonte, J. Bisquert, O. Almora, Device physics of hybrid perovskite solar cells: theory and experiment, Adv. Energy Mater. 8 (2018) 1702772.
[42] I.E. Castelli, T. Olsen, S. Datta, D.D. Landis, S. Dahl, K.S. Thygesen, K. W. Jacobsen, Computational screening of perovskite metal oxides for optimal solar light capture, Energy Environ. Sci. 5 (2012) 5814–5819.
[43] D. Dragoni, T.D. Daff, G. Csányi, N. Marzari, Achieving DFT accuracy with a machine-learning interatomic potential: Thermomechanics and defects in bcc ferromagnetic iron, Phys. Rev. Mater. 2 (2018), 013808.
[44] F.A. Faber, L. Hutchison, B. Huang, J. Gilmer, S.S. Schoenholz, G.E. Dahl, O. Vinyals, S. Kearnes, P.F. Riley, O.A. Von Lilienfeld, Prediction errors of molecular machine learning models lower than hybrid DFT error, J. Chem. Theory Comput. 13 (2017) 5255–5264.
[45] O.A. von Lilienfeld, K. Burke, Retrospective on a decade of machine learning for chemical discovery, Nat. Commun. 11 (2020) 4895.
[46] J.W. Hsu, W. Xu, Accelerate process optimization in perovskite solar cell manufacturing with machine learning, Matter. 5 (2022) 1334–1336.
[47] J. Li, B. Pradhan, S. Gaur, J. Thomas, Predictions and strategies learned from machine learning to develop high-performing perovskite solar cells, Adv. Energy Mater. 9 (2019) 1901891.
[48] W. Yan, Y. Liu, Y. Zang, J. Cheng, Y. Wang, L. Chu, X. Tan, L. Liu, P. Zhou, W. Li, Machine learning enabled development of unexplored perovskite solar cells with high efficiency, Nano Energy 99 (2022), 107394.
[49] Z. Gao, H. Zhang, G. Mao, J. Ren, Z. Chen, C. Wu, I.D. Gates, W. Yang, X. Ding, J. Yao, Screening for lead-free inorganic double perovskites with suitable band gaps and high stability using combined machine learning and DFT calculation, Appl. Surf. Sci. 568 (2021), 150916.
[50] Y. Wu, S. Lu, M.-G. Ju, Q. Zhou, J. Wang, Accelerated design of promising mixed lead-free double halide organic–inorganic perovskites for photovoltaics using machine learning, Nanoscale 13 (2021) 12250–12259.
[51] Y. Liu, W. Yan, S. Han, H. Zhu, Y. Tu, L. Guan, X. Tan, How machine learning predicts and explains the performance of perovskite solar cells, Sol. RRL. (2022) 2101100.
[52] A.D. Maynard, Navigating the fourth industrial revolution, Nat. Nanotechnol. 10 (2015) 1005–1006.
[53] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[54] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, Pytorch: An imperative style, high-performance deep learning library, Adv. Neural Inf. Process. Syst. 32 (2019).
[55] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G.S. Corrado, A. Davis, J. Dean, M. Devin, Tensorflow: Large-scale machine learning on heterogeneous distributed systems, *arXiv preprint arXiv:1603.04467,* 2016.
[56] X. Hu, C. Liu, Z. Zhang, X.F. Jiang, J. Garcia, C. Sheehan, L. Shui, S. Priya, G. Zhou, S. Zhang, 22% Efficiency inverted perovskite photovoltaic cell using cation-doped brookite TiO2 top buffer, Adv. Sci. 7 (2020) 2001285.
[57] T. Nakajima, K. Sawada, Discovery of Pb-free perovskite solar cells via high-throughput simulation on the K computer, J. Phys. Chem. Lett. 8 (2017) 4826–4831.
[58] J.E. Saal, S. Kirklin, M. Aykol, B. Meredig, C. Wolverton, Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD), JOM 65 (2013) 1501–1509.
[59] D.D. Landis, J.S. Hummelshøj, S. Nestorov, J. Greeley, M. Dułak, T. Bligaard, J. K. Nørskov, K.W. Jacobsen, The computational materials repository, Comput. Sci. Eng. 14 (2012) 51–57.
[60] A. Belsky, M. Hellenbrandt, V.L. Karen, P. Luksch, New developments in the inorganic crystal structure database (ICSD): accessibility in support of materials research and design, Acta Crystallogr Sect. B: Struct. Sci. 58 (2002) 364–369.
[61] S.P. Ong, W.D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K.A. Persson, G. Ceder, Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis, Comput. Mater. Sci. 68 (2013) 314–319.
[62] S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R.H. Taylor, L.J. Nelson, G. L. Hart, S. Sanvito, M. Buongiorno-Nardelli, AFLOWLIB. ORG: A distributed materials properties repository from high-throughput ab initio calculations, Comput. Mater. Sci. 58 (2012) 227–235.
[63] A. Smiti, A critical overview of outlier detection methods, Comput. Sci. Rev. 38 (2020), 100306.
[64] S.B. Kotsiantis, D. Kanellopoulos, P.E. Pintelas, Data preprocessing for supervised learning, Int. J. Comput. Sci. 1 (2006) 111–117.
[65] A.Y.-T. Wang, R.J. Murdock, S.K. Kauwe, A.O. Oliynyk, A. Gurlo, J. Brgoch, K. A. Persson, T.D. Sparks, Machine learning for materials scientists: an introductory guide toward best practices, Chem. Mater. 32 (2020) 4954–4965.
[66] G.E. Batista, M.C. Monard, An analysis of four missing data treatment methods for supervised learning, Appl. Artif. Intell. 17 (2003) 519–533.
[67] C. Oh, S. Han, J. Jeong, Time-series data augmentation based on interpolation, *Procedia Comput*, Sci. 175 (2020) 64–71.
[68] M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum, F. Hutter, Efficient and robust automated machine learning, Adv. Neural Inf. Process. Syst. 28 (2015).
[69] L. Ward, A. Agrawal, A. Choudhary, C. Wolverton, A general-purpose machine learning framework for predicting properties of inorganic materials, npj Comput Mater. 2 (2016) 1–7.
[70] L.M. Ghiringhelli, J. Vybiral, S.V. Levchenko, C. Draxl, M. Scheffler, Big data of materials science: critical role of the descriptor, Phys. Rev. Lett. 114 (2015), 105503.
[71] L. Ward, A. Dunn, A. Faghaninia, N.E. Zimmermann, S. Bajaj, Q. Wang, J. Montoya, J. Chen, K. Bystrom, M. Dylla, Matminer: An open source toolkit for materials data mining, Comput. Mater. Sci. 152 (2018) 60–69.
[72] A.P. Bartók, R. Kondor, G. Csányi, On representing chemical environments, Phys. Rev. B. 87 (2013), 184115.
[73] S.K. Kauwe, J. Graser, A. Vazquez, T.D. Sparks, Machine learning prediction of heat capacity for solid inorganics, Integr. Mater. Manuf. Innovation. 7 (2018) 43–51.
[74] V.M. Goldschmidt, Die gesetze der krystallochemie, Naturwissenschaften 14 (1926) 477–485.
[75] C.J. Bartel, C. Sutton, B.R. Goldsmith, R. Ouyang, C.B. Musgrave, L. M. Ghiringhelli, M. Scheffler, New tolerance factor to predict the stability of perovskite oxides and halides, Sci. Adv. 5 (2019).
[76] G. Pilania, P.V. Balachandran, C. Kim, T. Lookman, Finding new perovskite halides via machine learning, Front. Mater. 3 (2016) 19.
[77] S. Zhang, T. Lu, P. Xu, Q. Tao, M. Li, W. Lu, Predicting the formability of hybrid organic-inorganic perovskites via an interpretable machine learning strategy, J. Phys. Chem. Lett. 12 (2021) 7423–7430.
[78] J. Im, S. Lee, T.-W. Ko, H.W. Kim, Y. Hyon, H. Chang, Identifying Pb-free perovskites for solar cells by machine learning, npj Comput Mater. 5 (2019) 37.
[79] S. Lu, Q. Zhou, Y. Ouyang, Y. Guo, Q. Li, J. Wang, Accelerated discovery of stable lead-free hybrid organic-inorganic perovskites via machine learning, Nat. Commun. 9 (2018) 3405.
[80] G. Pilania, A. Mannodi-Kanakkithodi, B. Uberuaga, R. Ramprasad, J. Gubernatis, T. Lookman, Machine learning bandgaps of double perovskites, Sci. Rep. 6 (2016) 1–10.
[81] H. Sahu, W. Rao, A. Troisi, H. Ma, Toward predicting efficiency of organic solar cells via machine learning and improved descriptors, Adv. Energy Mater. 8 (2018) 1801032.
[82] Ç. Odabaşı, R. Yıldırım, S. Cells, Machine learning analysis on stability of perovskite solar cells, Sol. Energy Mater. Sol. Cells. 205 (2020), 110284.
[83] N. Burkart, M.F. Huber, A survey on the explainability of supervised machine learning, J. Artif. Intell. Res. 70 (2021) 245–317.

[84] Z.-H. Zhou, A brief introduction to weakly supervised learning, Natl, Sci. Rev. 5 (2018) 44–53.
[85] J. Bekker, J. Davis, Learning from positive and unlabeled data: A survey, Mach. Learn. 109 (2020) 719–760.
[86] P. Lerman, Fitting segmented regression models by grid search, Appl. Statist. 29 (1980) 77–84.
[87] B. Komer, J. Bergstra, C. Eliasmith, Hyperopt-sklearn, Automat. Machine Learning: Methods, Systems Challenges. 7 (2019) 97–111.
[88] S. Raschka, Model evaluation, model selection, and algorithm selection in machine learning, *arXiv preprint arXiv:1811.12808*, 2018.
[89] P. Refaeilzadeh, L. Tang, H. Liu, Cross-validation, Encyclopedia of database systems. 5 (2009) 532–538.
[90] S. Carey, Bootstrapping & the origin of concepts, Daedalus 133 (2004) 59–68.
[91] J. Huang, C.X. Ling, Using AUC and accuracy in evaluating learning algorithms, IEEE Trans. Knowl. Data Eng. 17 (2005) 299–310.
[92] L. Zhang, M. He, S. Shao, Machine learning for halide perovskite materials, Nano Energy 78 (2020), 105380.
[93] T. Lu, H. Li, M. Li, S. Wang, W. Lu, Predicting experimental formability of hybrid organic-inorganic perovskites via imbalanced learning, J. Phys. Chem. Lett. 13 (2022) 3032–3038.
[94] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst. 30 (2017).
[95] G.H. Gu, J. Jang, J. Noh, A. Walsh, Y. Jung, Perovskite synthesizability using graph neural networks, npj Comput Mater. 8 (2022) 1–8.
[96] N.J. Jeon, H. Na, E.H. Jung, T.-Y. Yang, Y.G. Lee, G. Kim, H.-W. Shin, S. Il Seok, J. Lee, J. Seo, A fluorene-terminated hole-transporting material for highly efficient and stable perovskite solar cells, Nat. Energy. 3 (2018) 682–689.
[97] J.P. Perdew, K. Burke, M. Ernzerhof, Generalized gradient approximation made simple, Phys. Rev. Lett. 77 (1996) 3865.
[98] J. Sun, A. Ruzsinszky, J.P. Perdew, Strongly constrained and appropriately normed semilocal density functional, Phys. Rev. Lett. 115 (2015), 036402.
[99] V. Vakharia, I.E. Castelli, K. Bhavsar, A. Solanki, Bandgap prediction of metal halide perovskites using regression machine learning models, Phys. Lett. A. 422 (2022), 127800.
[100] P. Omprakash, B. Manikandan, A. Sandeep, R. Shrivastava, P. Viswesh, D. B. Panemangalore, Graph representational learning for bandgap prediction in varied perovskite crystals, Comput. Mater. Sci. 196 (2021).
[101] C. Chen, W. Ye, Y. Zuo, C. Zheng, S.P. Ong, Graph networks as a universal machine learning framework for molecules and crystals, Chem. Mater. 31 (2019) 3564–3572.
[102] H. Park, R. Mall, A. Ali, S. Sanvito, H. Bensmail, F. El-Mellouhi, Importance of structural deformation features in the prediction of hybrid perovskite bandgaps, Comput. Mater. Sci. 184 (2020), 109858.
[103] Y. Li, Y. Lu, X. Huo, D. Wei, J. Meng, J. Dong, B. Qiao, S. Zhao, Z. Xu, D. Song, Bandgap tuning strategy by cations and halide ions of lead halide perovskites learned from machine learning, RSC Adv. 11 (2021) 15688–15694.
[104] M.-G. Ju, M. Chen, Y. Zhou, J. Dai, L. Ma, N.P. Padture, X.C. Zeng, Toward eco-friendly and stable perovskite materials for photovoltaics, Joule. 2 (2018) 1231–1241.
[105] J. Hieulle, X. Wang, C. Stecker, D.-Y. Son, L. Qiu, R. Ohmann, L.K. Ono, A. Mugarza, Y. Yan, Y. Qi, Unraveling the impact of halide mixing on perovskite stability, J. Am. Chem. Soc. 141 (2019) 3515–3523.
[106] A. Talapatra, B.P. Uberuaga, C.R. Stanek, G. Pilania, A machine learning approach for the prediction of formability and thermodynamic stability of single and double perovskite oxides, Chem. Mater. 33 (2021) 845–858.
[107] W. Li, R. Jacobs, D. Morgan, Predicting the thermodynamic stability of perovskite oxides using machine learning models, Comput. Mater. Sci. 150 (2018) 454–463.
[108] J. Schmidt, J. Shi, P. Borlido, L. Chen, S. Botti, M.A. Marques, Predicting the thermodynamic stability of solids combining density functional theory and machine learning, Chem. Mater. 29 (2017) 5090–5103.
[109] T. Xie, J.C. Grossman, Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties, Phys. Rev. Lett. 120 (2018), 145301.
[110] S. Behara, T. Poonawala, T. Thomas, Crystal structure classification in ABO3 perovskites via machine learning, Comput. Mater. Sci. 188 (2021), 110191.
[111] Y. Zhang, X. Xu, Machine learning lattice constants for cubic perovskite compounds, ChemistrySelect 5 (2020) 9999–10009.
[112] P. Priya, N. Aluru, Accelerated design and discovery of perovskites with high conductivity for energy applications through machine learning, npj Comput Mater. 7 (2021) 1–12.
[113] J. Chen, N.-G. Park, Inorganic hole transporting materials for stable and high efficiency perovskite solar cells, J. Phys. Chem. C. 122 (2018) 14039–14063.
[114] E. Bi, H. Chen, F. Xie, Y. Wu, W. Chen, Y. Su, A. Islam, M. Grätzel, X. Yang, L. Han, Diffusion engineering of ions and charge carriers for stable efficient perovskite solar cells, Nat. Commun. 8 (2017) 1–7.
[115] F. Zhang, K. Zhu, Additive engineering for efficient and stable perovskite solar cells, Adv. Energy Mater. 10 (2020) 1902579.
[116] C. Ma, M. Grätzel, N.-G. Park, Facet engineering for stable efficient perovskite solar cells, ACS Energy Lett. 7 (2022) 3120–3128.
[117] M. Srivastava, J.M. Howard, T. Gong, M. Rebello Sousa Dias, M.S. Leite, Machine learning roadmap for perovskite photovoltaics, J. Phys. Chem. Lett. 12 (2021) 7866–7877.
[118] E.C. Gok, M.O. Yildirim, M.P. Haris, E. Eren, M. Pegu, N.H. Hemasiri, P. Huang, S. Kazim, A. Uygun Oksuz, S. Ahmad, Predicting perovskite bandgap and solar cell performance with machine learning, Sol. RRL. 6 (2022) 2100927.
[119] T. Bak, K. Kim, E. Seo, J. Han, H. Sung, I. Jeon, I.D. Jung, Accelerated design of high-efficiency lead-free tin perovskite solar cells via machine learning, Int. J. Pr. Eng. Man-Gt. (2022) 1–13.
[120] M.C. Scharber, D. Mühlbacher, M. Koppe, P. Denk, C. Waldauf, A.J. Heeger, C. J. Brabec, Design rules for donors in bulk-heterojunction solar cells—Towards 10% energy-conversion efficiency, Adv. Mater. 18 (2006) 789–794.
[121] A. Kuzmich, D. Padula, H. Ma, A. Troisi, Trends in the electronic and geometric structure of non-fullerene based acceptors for organic solar cells, Energy Environ. Sci. 10 (2017) 395–401.
[122] H. Ma, A. Troisi, Modulating the exciton dissociation rate by up to more than two orders of magnitude by controlling the alignment of LUMO+ 1 in organic photovoltaics, J. Phys. Chem. C. 118 (2014) 27272–27280.
[123] Z. Liu, N. Rolston, A.C. Flick, T.W. Colburn, Z. Ren, R.H. Dauskardt, T. Buonassisi, Machine learning with knowledge constraints for process optimization of open-air perovskite solar cell manufacturing, Joule 6 (2022) 834–849.
[124] M. Del Cueto, C. Rawski-Furman, J. Arago, E. Orti, A. Troisi, Data-driven analysis of hole-transporting materials for perovskite solar cells performance, J. Phys. Chem. C. 126 (2022) 13053–13061.
[125] C. She, Q. Huang, C. Chen, Y. Jiang, Z. Fan, J. Gao, Machine learning-guided search for high-efficiency perovskite solar cells with doped electron transport layers, J. Mater. Chem. A. 9 (2021) 25168–25177.
[126] W. Liu, Y. Lu, D. Wei, X. Huo, X. Huang, Y. Li, J. Meng, S. Zhao, B. Qiao, Z. Liang, Screening interface passivation materials intelligently through machine learning for highly efficient perovskite solar cells, J. Mater. Chem. A. 10 (2022) 17782–17789.
[127] W. Yan, Y. Liu, Y. Zang, J. Cheng, Y. Wang, L. Chu, X. Tan, L. Liu, P. Zhou, W. Li, Machine learning enabled development of unexplored perovskite solar cells with high efficiency, Nano Energy 107394 (2022).
[128] Y. Yu, X. Tan, S. Ning, Y. Wu, Machine learning for understanding compatibility of organic–inorganic hybrid perovskites with post-treatment amines, ACS Energy Lett. 4 (2019) 397–404.
[129] M. Saliba, J.-P. Correa-Baena, C.M. Wolff, M. Stolterfoht, N. Phung, S. Albrecht, D. Neher, A. Abate, How to make over 20% Efficient perovskite solar cells in regular (n–i–p) and inverted (p–i–n) architectures, Chem. Mater. 30 (2018) 4193–4201.
[130] W. Xu, Z. Liu, R.T. Piper, J.W. Hsu, Bayesian optimization of photonic curing process for flexible perovskite photovoltaic devices, Sol. Energy Mater. Sol. Cells. 249 (2023), 112055.
[131] H. Min, D.Y. Lee, J. Kim, G. Kim, K.S. Lee, J. Kim, M.J. Paik, Y.K. Kim, K.S. Kim, M.G. Kim, Perovskite solar cells with atomically coherent interlayers on SnO2 electrodes, Nature 598 (2021) 444–450.
[132] C. Zhi, S. Wang, S. Sun, C. Li, Z. Li, Z. Wan, H. Wang, Z. Li, Z. Liu, Machine-learning-assisted screening of interface passivation materials for perovskite solar cells, ACS Energy Lett. 8 (2023) 1424–1433.
[133] D. Kim, E.S. Muckley, N. Creange, T.H. Wan, M.H. Ann, E. Quattrocchi, R. K. Vasudevan, J.H. Kim, F. Ciucci, I.N. Ivanov, Exploring transport behavior in hybrid perovskites solar cells via machine learning analysis of environmental-dependent impedance spectroscopy, Adv. Sci. 8 (2021) 2002510.
[134] V.M. Le Corre, T.S. Sherkar, M. Koopmans, L.J.A. Koster, Identification of the dominant recombination process for perovskite solar cells based on machine learning, Cell Rep. Phys. Sci. 2 (2021), 100346.
[135] E. Unger, T.J. Jacobsson, The perovskite database project: A perspective on collective data sharing, ACS Energy Lett. 7 (2022) 1240–1245.
[136] H. Moriwaki, Y.-S. Tian, N. Kawashita, T. Takagi, Mordred: A molecular descriptor calculator, J. Cheminf. 10 (2018) 1–14.
[137] K. Weiss, T.M. Khoshgoftaar, D. Wang, A survey of transfer learning, J. Big Data. 3 (2016) 1–40.