IOP Publishing Journal of Physics: Condensed Matter
J. Phys.: Condens. Matter **36** (2024) 355901 (12pp) https://doi.org/10.1088/1361-648X/ad512d

Predicting photovoltaic parameters of perovskite solar cells using machine learning

Zhan Hui[1], Min Wang[2,3], Jialu Chen[1], Xiang Yin[1,*], Yunliang Yue[1,*] and Jing Lu[3,*]

[1] College of Information Engineering, Yangzhou University, Yangzhou 225127, People's Republic of China
[2] School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang 050018, People's Republic of China
[3] State Key Laboratory of Artificial Microstructure and Mesoscopic Physics, School of Physics, Peking University, Beijing 100871, People's Republic of China

E-mail: yinxiang@yzu.edu.cn, yueyunliang@yzu.edu.cn and jinglu@pku.edu.cn

Received 15 January 2024, revised 20 May 2024
Accepted for publication 28 May 2024
Published 7 June 2024

Abstract

Perovskite solar cells (PSCs) have garnered significant attention owing to their highly power conversion efficiency (PCE) and cost-effectiveness. Traditionally, screening for PSCs with superior photovoltaic parameters relies on resource-intensive trial-and-error experiments. Nowadays, time-saving machine learning (ML) techniques serve as an artificial intelligence approach to expedite the prediction of photovoltaic parameters using accumulated research datasets. In this study, we employ seven supervised ML methods to forecast key photovoltaic parameters for PSCs such as PCE, short-circuit current density (J_{sc}), open-circuit voltage (V_{oc}), and fill factor (FF). Particularly, we design an artificial neural network (ANN) architecture that incorporates residual connectivity and layer normalization after the linear layers to enhance the scope and adaptability of the network. For PCE and J_{sc}, ANN demonstrates superior prediction accuracy, yielding root mean square errors of 2.632% and 2.244 mA cm^{-2}, respectively. The Random Forest (RF) model exhibits exceptional prediction performance for V_{oc} and FF. Additionally, an interpretability analysis of the model is conducted to elucidate the impact of features on PCE prediction, offering a novel approach for accurate and interpretable ML methods in the context of PSCs.

Supplementary material for this article is available online

Keywords: perovskite solar cells, machine learning, performance prediction, artificial neural networks, model interpretability

1. Introduction

Perovskite solar cells (PSCs) have attracted considerable attention as a promising photovoltaic technology due to their high energy conversion efficiency and relatively low fabrication costs [1]. Since the first demonstration of PSCs achieving a power conversion efficiency (PCE) of 3.8% in 2009 by Kojima *et al* [2], there has been remarkable progress, with recent PSCs reaching efficiencies exceeding 26% [3]. This rapid development over just a decade underscores the potential of PSCs. Predicting and optimizing the performance of PSCs remains a daunting challenge [4, 5]. The performance is influenced by various factors, including the physicochemical properties of perovskite materials, crystal structures, and fabrication process parameters [6, 7]. Traditional methods of

* Authors to whom any correspondence should be addressed.

obtaining performance rely heavily on experiment methods, involving specialized steps such as one-step/two-step preparation methods, solution deposition, or evaporation techniques [8], which are labor-intensive, time-consuming, and subject to the influence of environmental factors like temperature and humidity [9].

To address these challenges and improve the accuracy and efficiency of performance prediction, machine learning (ML) techniques have emerged as a promising research avenue [10–12]. PSCs can be generalized to certain structures and thus can be combined with ML methods. In general, the structure of PSCs can be categorized into two main types: the conventional structure (n-i-p) and the inverted structure (p-i-n) [13]. Despite differences in layout between these two structures, they both follow a similar energy conversion mechanism [14]. ML leverages its ability to automatically uncover patterns within data, offering novel methodologies for predicting PSC performance [15]. By analyzing extensive experimental data and simulation results, ML models can provide a deeper understanding of PSCs and offer insights into performance prediction [16–18]. Additionally, tools like SHapley Additive exPlanations (SHAP) aid in interpreting model predictions [19].

Recently, ML has gained traction in PSCs research [20]. ML techniques are increasingly used to optimize device structures and analyze the properties of PSCs. The optimization of device structure includes optimizing the perovskite layer, ETL, HTL, etc for better performance [21]. By determining the structure of the device, ML can quickly make fast and accurate predictions on the properties of PSCs. In a study by Gok *et al* [22], a two-step ML approach was developed to investigate PSC performance. Initially, ML was employed to predict the bandgap of perovskites, validated through experimentation. Subsequently, the performance of PSCs was evaluated using eight different perovskite compositions. The study revealed that highly electronegative compositions exhibited lower absorption onsets, while a positive correlation was observed between lattice parameters and absorption onsets. Additionally, Del Cueto *et al* [23] devised an ML model to predict PCE in PSCs, leveraging features describing hole transport materials (HTMs), perovskite types, and cell structures. The model identified HTMs candidates more likely to achieve higher PCE and elucidated correlations among specific molecular segments. Mishra *et al* [24] predicted the photovoltaic properties of indoor PSCs using the ML method and performed an interpretable analysis using the correlation matrix and SHAP. Yan *et al* [25] used ML to screen out candidate PSCs, which were experimentally validated and optimized to achieve stable PSCs with PCE of 23.6%. Overall, ML algorithms enable the discovery of physical rules and relationships that traditional methods may struggle to uncover, expediting the development of novel PSCs with enhanced performance and paving the way for further exploration in this domain [26, 27].

In our investigation, ML techniques were applied to assess the performance of PSCs. Initially, we meticulously curated data points related to PSCs, eliminating redundancy. Subsequently, seven supervised ML methods, including Linear Regression (LR), *K*-Nearest Neighbors (KNN), Support Vector Regression (SVR), Random Forest (RF), Gradient Boosting Regression Trees (GBRT), Extreme Gradient Boosting Regression Trees (XGBRT), and Artificial Neural Network (ANN), were employed to analyze performance indicators such as Power Conversion Efficiency (PCE), short-circuit current density (J_{sc}), open-circuit voltage (V_{oc}), and fill factor (FF) for PSCs. Additionally, a simple ANN was designed, incorporating layer normalization (LayerNorm) after the linear layer and incorporating residual connections to enhance feature extraction and model fitting. Our findings revealed superior performance of the designed ANN model in PCE and J_{sc} prediction, while the RF model demonstrated optimal performance in V_{oc} and FF. Finally, the SHAP method was utilized for interpretability analysis, facilitating an exploration of the relationships between features and target properties using ML. This study accelerates the prediction of photovoltaic properties of PSCs using ML, demonstrating accurate and fast prediction capabilities and revealing the intricate relationship between various features and target properties.

2. Methods

The flowchart for predicting the performance of PSCs using ML is shown in figure 1. First, data points with information about PSCs needs to be extracted from literature or databases. The data points should contain the target performance and the type of data that can be transformed into features for feature extraction. Then to choose a suitable ML model for training and evaluation. Model tuning and optimization are essential for prediction performance improvement. Meanwhile, insights on the model will be instructive for performance prediction. Finally, models with high prediction accuracy can be used for unknown PSC performance prediction and applications.

2.1. Data collection

The data points used in this study were obtained from Li *et al* [28] and Liu *et al* [29], who collected 333 and 463 PSCs datasets, respectively. Data on PSCs dating back 3–5 years before 2019 were collected by Li *et al.* Whereas Liu *et al* collected data up to 2022, and all the parameters of the PSCs were recorded at AM1.5G sunlight and 100 mW cm^2. Both of them ensured that the bandgap in the data were experimentally acquired, which also ensured the quality of the data. Meanwhile, due to partial overlap in time, there were overlapping data points between these sources. During the integration of the datasets, we removed data points with 11 features that were completely duplicated and retain data points with only component duplicates to ensure the comprehensiveness of the dataset. The final dataset consists of 615 data points, as shown in table S1 in the supplemental material. Each data point in the dataset consists of three different types of data. These data include the elemental composition of the perovskite materials, the energy gap between the perovskite materials and the hole/electron transport layer (HTL/ETL),

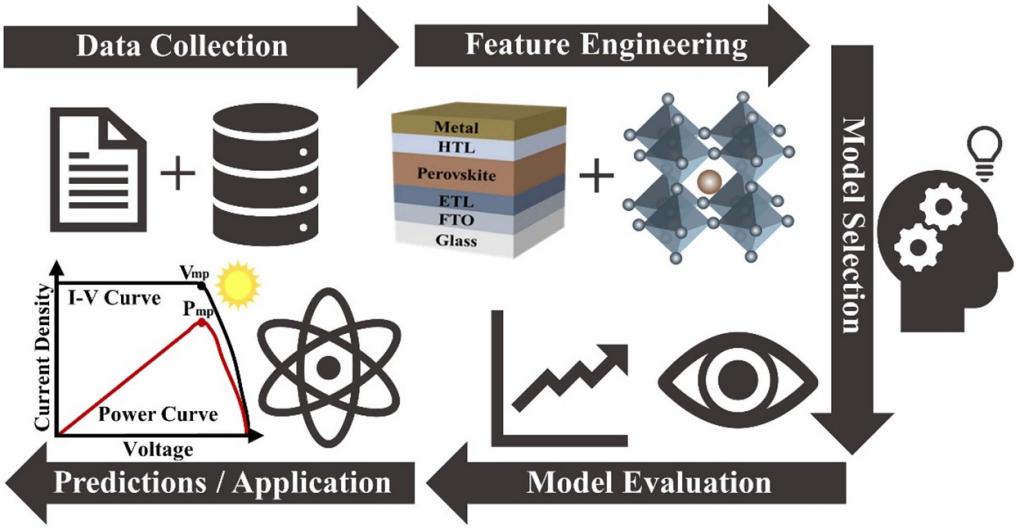

Figure 1. Flowchart for ML to predict the performance of PSCs.

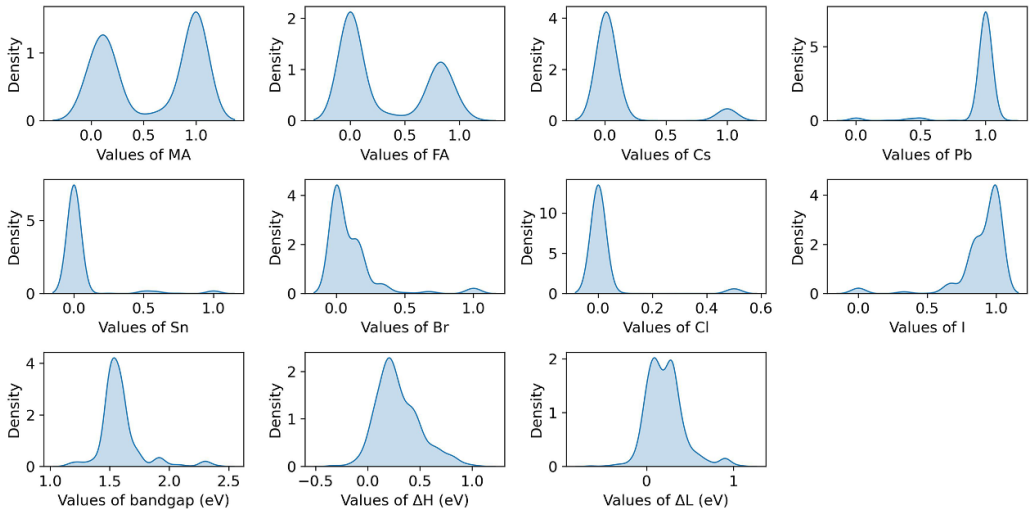

Figure 2. Feature density plot. The X-axis represents the distribution range of the feature values and the Y-axis represents the density of the features.

and the photovoltaic parameters of the PSCs. The elemental composition of the perovskite followed an ABX_3 structure, where the A-site elements included Cs, methylammonium (MA) and formamidinium (FA), the B-site elements involved Pb and Sn, and the X-site elements encompassed Cl, Br, and I. The photovoltaic parameters of the PSCs encompassed four aspects: open-circuit voltage (V_{oc}), short-circuit current (J_{sc}), fill factor (FF), and power conversion efficiency (PCE).

2.2. Feature engineering

The features selected in the experiments are mainly derived from the common features of the source data. We selected the composition of the perovskite material, bandgap and the energy gap between the HTL (ΔH) and ETL (ΔL) as features. The light absorbing layer is crucial for the performance of the PSCs. Therefore, the elemental composition of the perovskite layer is processed through one-hot coding to encode the ratio of the elements involved. Meanwhile, the bandgap of the perovskite material directly affects the spectral absorption ability of the PSCs. Regarding the macroscopic features, the HOMO/LUMO and properties of ETL and HTL also affect the performance of PSCs. Usually the values of ΔH (ΔL) are greater than or equal to zero because a negative ΔH (ΔL) introduces a potential barrier between the perovskites and the transport layer, while too large ΔH (ΔL) induces a significant energy loss at the interface and creates a transport barrier between the transport layer and the electrode [30]. Figure 2 displays density plots of different features. This step is to ensure that our dataset is of good quality and that there are no obvious bad data points. The distribution of the elemental composition of perovskites lies between 0 and 1, representing different proportions of elements contained. The bandgap of the perovskite layer are almost distributed between

1.5 and 1.9 eV, which aligns with physical laws observed in highly efficient PSCs reported to date [31]. As the data was extracted from experimental papers, most experiments selected bandgaps of the light-absorbing layer within the range between 1.5 and 1.9 eV, resulting in high-efficiency photovoltaic cells due to this specific range. Similarly, ΔH and ΔL values range between 0.2 and 0.6 eV. This attribute to that the ΔH and ΔL relevant to carrier transport, while too small energy differences affect conductivity [32]. Thus, achieving values between 0.2–0.6 eV corresponds to higher PCE. In addition, we apply Z-score standardization to eliminate the scale difference between features, which helps to accelerate the convergence process of the model. Unlike max-min normalization, which adjusts data based on their maximum and minimum values, Z-score standardization rescales the data to a standard normal distribution with a mean of 0 and a standard deviation of 1. Moreover, its reliance on mean and standard deviation makes it robust against outliers, aligning well with the characteristics of our dataset.

2.3. Model selection

We selected seven commonly used supervised ML models to predict the performance of PSCs, including LR, KNN, SVR, RF, GBRT, XGBRT, and ANN algorithms. LR is a fundamental statistical method that establishes the relationship between independent and dependent variables by fitting a linear model. However, its major limitation lies in assuming a linear relationship between inputs and outputs, which might not hold for complex, non-linear data patterns. KNN is an instance-based learning method that performs classification or regression based on neighboring points in the feature space. SVR is a regression analysis technique that utilizes Support Vector Machine theory to handle linear and nonlinear data. RF is an ensemble learning method consisting of multiple decision trees that make predictions through voting or averaging. It offers insights into feature importance but might struggle with noisy data. GBRT is an iterative ensemble learning algorithm that enhances prediction performance by sequentially optimizing decision trees. XGBRT is an optimized gradient-boosting algorithm that combines regularization and parallel processing to improve accuracy and efficiency. ANN is a model inspired by the structure and functionality of the human neural system, capable of handling complex nonlinear relationships and large-scale data, exhibiting strong fitting and generalization capabilities. ANN has been widely used in property prediction and has achieved satisfactory results [33]. It offers high flexibility but demands a large amount of data for training and poses challenges in model interpretability due to its black-box nature [34].

In these models, numerous hyperparameters require optimization to achieve better fitting performance. The parameters in the models are multi-dimensional, and attempting them one by one results in an enormous search space, which is time-consuming and computationally wasteful. Several automated parameter tuning methods have emerged, such as random search, grid search, and Bayesian optimization techniques [35]. Random search explores a wide range of parameter space due to its randomness but lacks directionality, potentially spending extended periods in less critical areas. Grid search exhaustively explores parameters at fixed intervals and is suitable for smaller parameter spaces and lower dimensions. However, its computational cost increases exponentially as the number of parameters increases, leading to lower efficiency. The Bayesian optimization method can alleviate this dilemma by exploring the parameter space autonomously by using information from previous sampling points to guide the next sampling. It estimates the performance of a parameter by constructing a probabilistic model and then selects the next most promising sampling point based on that model. As a result, it is able to find the optimal solution in a relatively small number of sampling points and iterations. Therefore, we choose the Bayesian optimization method as the final optimization method.

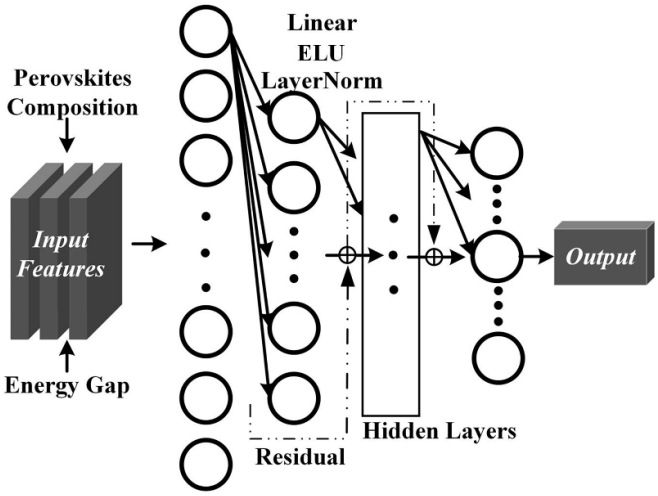

Figure 3. The architecture of the ANN.

In addition, we designed a simple ANN model with the network structure shown in figure 3. The ANN network overall is based on the TensorFlow framework [36], referencing code from the work of Cherukara and Mannodi-Kanakkithodi [37]. First, LayerNorm normalization is added after the linear layer to further improve the generalization of the model. LayerNorm helps to stabilize the internal distribution of the network during the training process, which accelerates the convergence speed of the network and improves the training effect of the model [38]. Compared to the batch normalization method, LayerNorm is more suitable for processing smaller batches of data while reducing the sensitivity to the learning rate. By keeping the inputs of each layer relatively consistently distributed, LayerNorm can effectively reduce the gradient vanishing and explosion problems, thus making the network easier to train. The introduction of residual connections then better addresses the gradient vanishing problem in deeper networks, facilitating the flow of information and the transfer of features. Residual connections allow the network to directly pass some of the input signals across multiple layers, allowing the gradient

to propagate more freely. In addition, residual connectivity can prompt the model to model features at a deeper level, which further enhances the expressive power of the model [39].

2.4. Model evaluation

The evaluation metrics are root mean square error (RMSE) and pearson correlation coefficient (r). The RMSE is the most commonly used regression evaluation metric, which can be used to judge the accuracy of the ML model. The r-value is used to determine the correlation between the predicted values output by the ML model and the collected experimental data. The smaller the value of RMSE, the higher the predictive accuracy of the model. The r-values range from 0 to 1, and the closer the value is to 1, the higher the correlation of the model [40]

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{N}}\sum_{i=1}^{N}\left(y_{(i)} - y_{p(i)}\right)^2} \tag{1}$$

$$r = \frac{\sum_{i=1}^{N}\left(y_{(i)} - \overline{y}\right)\left(y_{p(i)} - \overline{y_p}\right)}{\sqrt{\sum_{i=1}^{N}\left(y_{(i)} - \overline{y}\right)^2 \times \sum_{i=1}^{N}\left(y_{p(i)} - \overline{y_p}\right)^2}} \tag{2}$$

where N is the number of samples, $y_{(i)}$ and $y_{p(i)}$ are the ith individual sample points of true and predicted value, $\overline{y}$ and $\overline{y_p}$ are the mean value of y and y_p, respectively.

3. Result and discussion

Firstly, we utilized seven supervised ML methods, including LR, KNN, SVR, RF, GBRT, XGBRT, and ANN, to sequentially train models for the target properties PCE, V_{oc}, J_{sc}, and FF. We used Z-score normalization for data preprocessing before inputting the data. This will reduce the error caused by the size of the features [38]. Bayesian optimization was employed to optimize parameters for each model, with the parameter optimization in table S2 in the supplemental material. To ensure the reliability of the experiments and avoid chance effects, we employed a five-fold cross-validation method. The choice to use five-fold cross-validation allows for better utilization of limited data resources while maintaining a reasonable assessment of model performance. Several studies similar to ours have used a five-fold cross-validation approach [28, 29], which also facilitates experimental comparisons, which we analyze in the following sections. The experimental data was divided into five sets, and during each experiment, one set was used as the testing dataset, while the remaining data served as the training dataset. It was ensured that each set of data was used once as a testing dataset [41].

During the training process of ANN, an additional 10% of the training dataset was set aside as a validation set to ensure the convergence of the training process. For the ANN model, we configured four hidden layers (64, 50, 32, 16), and the batch size and initial learning rate as 32 and 0.01. The loss function and optimizer were set as L2 and Adamax, respectively. RMSE and r were adopted as performance metrics for the models.

Table 1. Predictions error for bandgap (eV) on the test sets. (Bolded values are optimal results).

	LR	KNN	SVR	RF	GBRT	XGBRT	ANN
RMSE	0.066	0.064	0.060	**0.053**	0.055	**0.053**	0.057
r	0.926	0.926	0.944	0.955	0.951	**0.956**	0.947

We first use seven supervised ML methods to predict the bandgap of perovskite materials, as the results shown in table 1. The features are based on the material composition of the perovskites. On the test set, the RF model and the GBRT model have the highest prediction accuracy with an RMSE of 0.053 eV, which is higher than that of Li *et al* (0.060 eV) [28] and Liu *et al* (0.064 eV) [29]. The r-value of the GBRT model is 0.955, indicating that the model has strong correlation, which is close to that of Li *et al* (0.97) [28] and Liu *et al* (0.96) [29]. It can be seen that the RMSE of the bandgap task decreases and the accuracy improves when the dataset is enlarged. Whereas the r-value is flat or decreases, we infer that this is due to the different bandgap values for the same perovskite composition in the dataset. In our recently published work, we used deep learning to predict properties such as bandgap of perovskite materials, and since the data sources were calculated by density functional theory simulations, the accuracy of the predictions was slightly less accurate due to intrinsic errors in the dataset [42]. However, all the bandgap values herein are obtained experimentally, contributing to the accuracy of predictions and making the error relatively low [43].

Then, we use the ML model to train and predict the performance of the PSCs, as the results shown in table 2, where the bold values are the optimal value for the same group. Among the predictions for PCE and J_{sc}, the designed ANN demonstrated notable performance on the test sets, achieving RMSE of 2.632% and 2.244 mA cm^{-2}. The highest r-values of 0.742 and 0.824 were also obtained, respectively. We compared the RF and ANN models before and after parameter optimization for PCE prediction, and the prediction accuracy was significantly improved, as shown in table S3 in the supplemental material. Obviously, the model prediction is more accurate through parameter optimization. As mentioned above, the RMSE shows the error-value in prediction and the r-value reflects the relevance of the model. Specifically, the r-value is commonly used to measure the linear correlation between model predictions and actual observations while ignoring non-linear relationships. Therefore, combined considerations are needed when evaluating prediction accuracy.

Meanwhile, the RF algorithm exhibited superior performance in predicting V_{oc} and FF, with RMSE of 0.078 and 0.057. The r-values were respectively 0.865 and 0.391. We try to

Table 2. Prediction error statistics for each target property on the test sets. (Bolded values are optimal results for the same group).

	PCE (%)		V_{oc} (V)		J_{sc} (mA cm^{-2})		FF	
	RMSE	r	RMSE	r	RMSE	r	RMSE	r
LR	2.825	0.695	0.084	0.846	2.480	0.777	0.059	0.341
KNN	2.790	0.702	0.084	0.858	2.576	0.750	0.060	0.287
SVR	2.871	0.692	0.116	0.715	2.547	0.771	0.063	0.273
RF	2.648	0.738	**0.078**	**0.865**	2.387	0.802	**0.057**	**0.391**
GBRT	2.657	0.734	0.083	0.855	2.412	0.787	0.058	0.360
XGBRT	2.715	0.724	0.081	0.857	2.426	0.794	0.059	0.348
ANN	**2.632**	**0.742**	0.087	0.833	**2.244**	**0.824**	0.059	0.329

understand that ANN performs well in predicting PCE and J_{sc}. This may be attributed to the range of data for the target properties. The data range for PCE is between 0.11 and 22.51%; J_{sc} is between 0. 35 and 30.75 mA cm^{-2}. While V_{oc} ranges between 0.15 and 1.602 V and FF ranges between 0.278 and 0.902. ANN performs well when trained on data with a large data range and could be at risk of overfitting when the data range is small [44]. The RF model tends to perform well when predicting data with a small data range [45]. Overall, ANN and several ensembled learning methods performed well in the test set. Despite the insufficient amount of data, ANN still equaled or even surpassed the other algorithms in predicting the PSC performance. It should be noted that the predicted RMSE of PCE is lower than the experimental results of Li *et al* [28] with an RMSE of 3.23%. However, there is a decrease in the model correlation metric r, which has a value of 0.80 compared to the experiment of Li *et al.* We infer that the data with the same composition but different bandgap were removed from the experiment of Li *et al* resulting in an increase in the r-value. Our experimental results are in between compared to the r-value of 0.7 obtained in the experiment using 11 features by Liu *et al* [29]. Meanwhile, the experimental RMSE values of Liu *et al* (2.35%) are slightly lower than ours. Although we increased the dataset, the RMSE values were reduced but did not obtain the optimal values due to the large time span of data collection and differences in device fabrication conditions. Besides, the sparse number of features limited the performance of the model. In the study by Lu *et al* [46], 17 features including perovskite fractions, fabrication process and device structure were used. The XGBoost model predicted an RMSE of 1.28% for PCE in the test set. We also calculated the MAE and R^2 of the model as well, as shown in table S4 in the supporting material. The only difference is that the MAE of ANN (2.072%) is higher than RF (2.038%) in predicting PCE, which is due to more outliers and MAE is more tolerant to outliers. One should point out that the R^2 value is small in predicting FF, which is caused by the small range of variation in the label values of FF (0.278–0.902) and the relatively large amount of variation in the values of the features.

Scatter plots in figure 4 illustrate the predictions of the best models for each property on the test sets. These scatter plots vividly display the errors between the actual values and predicted values for each data point in the test set. The closer the data points are to the dotted line, the closer the predicted value is to the true value, so the model fits better. It is evident that V_{oc} has the best prediction performance, with the data points clustered at the dotted line. In figure 4(a), most data points are concentrated within the range of 12%–19%, with the best predictions falling between 15%–19% for PCE. Similarly, in figures 4(b) and (c), the data are clustered around 0.8–1.2 V and 20–25 mA cm^{-2}, respectively. Figure 4(d) illustrates the prediction results for FF, mainly distributed within the range of 0.65–0.8, demonstrating a consistent pattern of distribution.

The selection of features plays a crucial role in model training. For the chosen 11 features, the r-value among these features can be observed in figure 5. Notably, Pb exhibits a strong negative correlation with Sn (−0.99), while Br displays a strong negative correlation with I (−0.97). To illustrate, attempting to enhance predictions for PCE by eliminating less important features from the highly correlated features. The features ranked in descending order of importance were Pb, Sn, I, and Br. Experiments are conducted using GBRT, RF, XGBRT due to high prediction accuracy of integrated learning. ANN is also used for training and prediction process. The prediction results on the test sets are shown in table 3.

Utilizing all 11 features for PCE tasks resulted in an RMSE of 2.632% and an r-value of 0.742 using ANN models. After removing the features of Sn and Br, the RMSE and r-values were 2.824% and 0.705, respectively. Specifically removing the Br feature led to an RMSE of 2.784% and an r-value of 0.712. It can be seen that the accuracy as well as the relevance of the model predictions decreases after removing the features. Unfortunately, deleting features with high correlation but low feature ranking did not improve prediction accuracy. We believe this could be due to the sparsity of features and training data. These findings indicate that in scenarios with a small number of features, removing highly correlated features with low feature importance may not necessarily improve prediction accuracy [47].

For ML models, the predictions are a result of the mapping of input features. Therefore, understanding the internal prediction logic of the model is crucial. The SHAP facilitates interpretable analysis based on statistical values derived from the trained model [48]. Utilizing SHAP can aid in a better

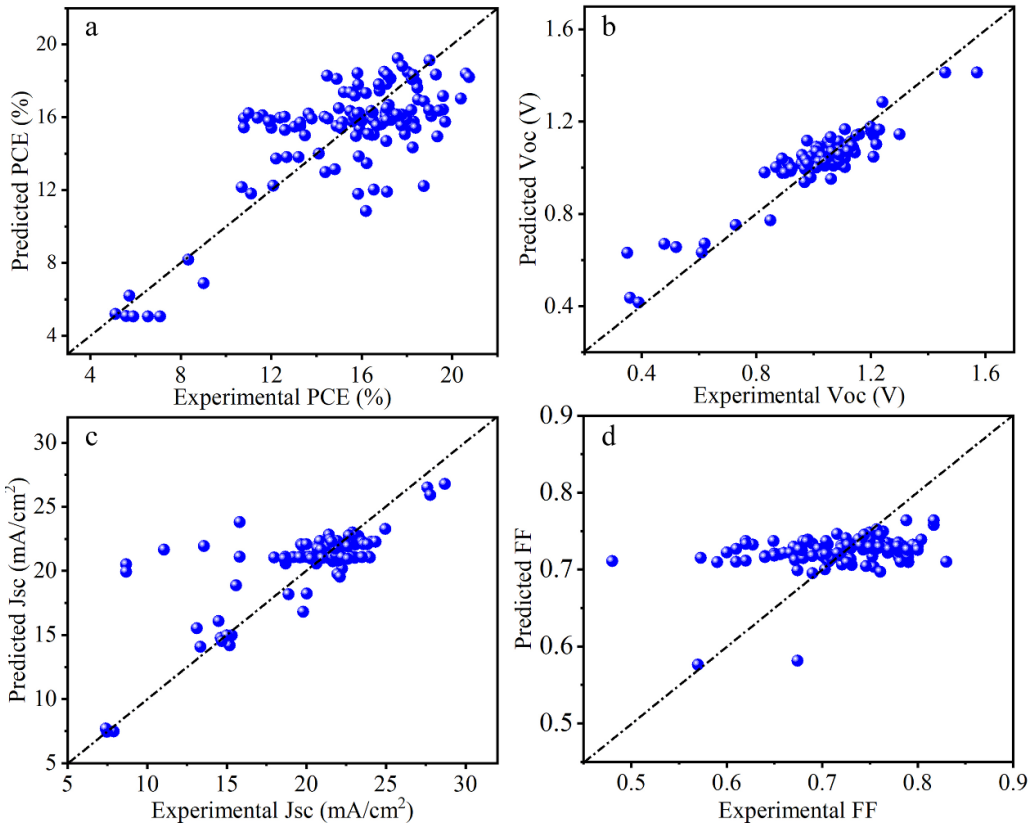

Figure 4. Scatter plot of test sets for the optimal model for each property. (a) PCE, (b) V_{oc}, (c) J_{sc}, (d) FF.

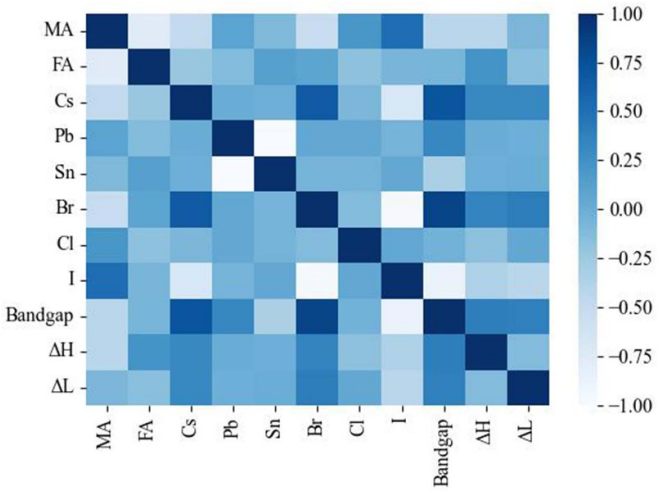

Figure 5. Heat map of the Pearson correlation coefficients between features.

understanding of how ML models utilize input features for predictions and reveal the interrelationships between different features. However, SHAP is not absolute and it offers a relative contribution explanation regarding features to predictions. Different models and datasets might result in different interpretative outcomes [49].

Due to the strong interpretability of the ensembled learning methods, we combined the RF model with the SHAP tool to conduct an interpretability analysis on the ML models trained previously. The main purpose of feature importance ranking is to perform interpretive analysis of ML models in order to better understand the prediction process and results of the models.

Table 3. Feature selection results using ensembled learning and ANN models.

	RF		GBRT		XGBRT		ANN	
	RMSE	r	RMSE	r	RMSE	r	RMSE	r
All features	2.645	0.738	2.657	0.734	2.655	0.736	2.632	0.742
Without Sn	2.702	0.726	2.683	0.729	2.665	0.736	2.649	0.736
Without Br	2.651	0.737	2.718	0.723	2.669	0.731	2.784	0.712
Without Sn and Br	2.668	0.733	2.725	0.719	2.672	0.733	2.824	0.705

Thereby it also helps in experimental guidance and insights. Figure 6 presents feature importance and contribution plots for the four target properties: PCE, V_{oc}, J_{sc}, and FF. Figures 6(a), (c), (e) and (g) offer a clear visualization of feature importance rankings via histograms. Each bar chart depicts the average SHAP values of the features for the predicted property, with each row representing a distinct feature. In figures 6(b), (d), (f) and (h), individual points denote samples, and the SHAP values of these samples are depicted using color cards. Notably, the order of various features is organized in descending order of average SHAP values. In the prediction of FF properties, bandgap ranked first among the features. It is ranked second among the other three properties. This shows the importance of bandgap in the light-absorbing layer. The bandgap of the light-absorbing layer (perovskite material) influences the spectral absorption range and its absorption and electron behavior of PSCs [50]. The trend of the influence of features on property prediction can also be summarized. In figure 6(b), there is evidence of the effect of large FA features on increasing PCE predictions. A larger FA value (indicated by a redder color, denoting a higher feature value) shows a more significant positive impact on PCE (the X-axis represents the SHAP values). This implies that higher FA values enhance the PCE of PSCs, while lower FA values might decrease it. This conclusion similarly applies to J_{sc} prediction, as depicted in figure 6(f). Figures 6(d) and (h) suggest that higher-values of Pb and bandgap similarly contribute positively to V_{oc} and FF values.

In addition to the overall feature analysis, which includes the previously mentioned feature importance ranking and contribution, SHAP can also analyze the impact of a single feature on predicted targets. Feature importance and contribution cannot infer the effect of a specific range of features on the property gain, which can be solved by using single/multiple feature dependency graph approach. Here, we take the RF model to predict the PCE of PSCs as an example. First, we analyze the effect of single features on the properties and visualize the effect of important features such as FA and bandgap on the PCE, respectively, as shown in figure 7. The X-axis in the graph represents the magnitude of the feature, while the Y-axis represents the SHAP value of the feature, denoting its positive or negative influence on the PCE prediction. In figure 7(a), as the FA value increases, its positive effect on PCE also grows, but there is a decline when the FA value reaches 1. When the FA value ranges between 0.7–1, its positive impact on PCE is most prominent, potentially resulting in a higher PCE. This is similar to the findings of Li *et al* [51] where the absorption spectra increased with increasing FA ratio. Similarly, a trend of initially increasing and then decreasing SHAP values of the bandgap is observed in figure 7(b). When the bandgap value lies between 1.5 and 1.7 eV, a higher PCE could be achievable. With a range of bandgap values between 1.6 and 1.7 eV, the gain in PCE is the maximum. In fact, a smaller bandgap means that the material can absorb higher energy (shorter wavelength) light, thus absorbing more energy across the solar spectrum and increasing the PCE. It also means that electrons can easily move through the conduction band, thus reducing electron transport losses through the material and contributing to the PSCs [52].

The above presentation illustrates the relationship between the size of single features and their SHAP values. Furthermore, it can explore the relationship between multiple features and the target property. Here, we utilize the SHAP tool to study the synergistic effects of two different features on the PCE. Figure 8 displays the predictive impact of different features on PCE. In figure 8(a), it can be observed that the joint influence of FA and bandgap on PCE shows a similar distribution to the individual influence of FA. Compared to single-feature graphs, the double-feature graph demonstrates how the combined effects of bandgap and FA sizes impact the prediction of PCE. When FA increases, a larger SHAP value is achieved when the bandgap is between 1.5–1.7 eV, thereby increasing the PCE of PSCs. However, excessively high FA can reduce PCE. Figure 8(b) illustrates the impact of Pb and Sn features on PCE, showing that with an increase in Pb value, smaller Sn values have a positive impact on PCE. For B-site metal ions, replacing Pb with smaller Sn will decrease the bandgap of the perovskite layer [53], and the higher PCE can generally be obtained with relatively small bandgap. In figures 8(c) and (d), we can observe a similar trend of increase and decrease, peaking when bandgap ranges between 1.5 and 1.7 eV. However, the impact of Pb and Sn sizes on PCE is opposite, with higher Pb content yields higher PCE and lower Sn content leads to higher PCE. This demonstrates the effectiveness of ML in photovoltaic parameters prediction and also provides valuable insights that can help in the design and optimization of future PSCs.

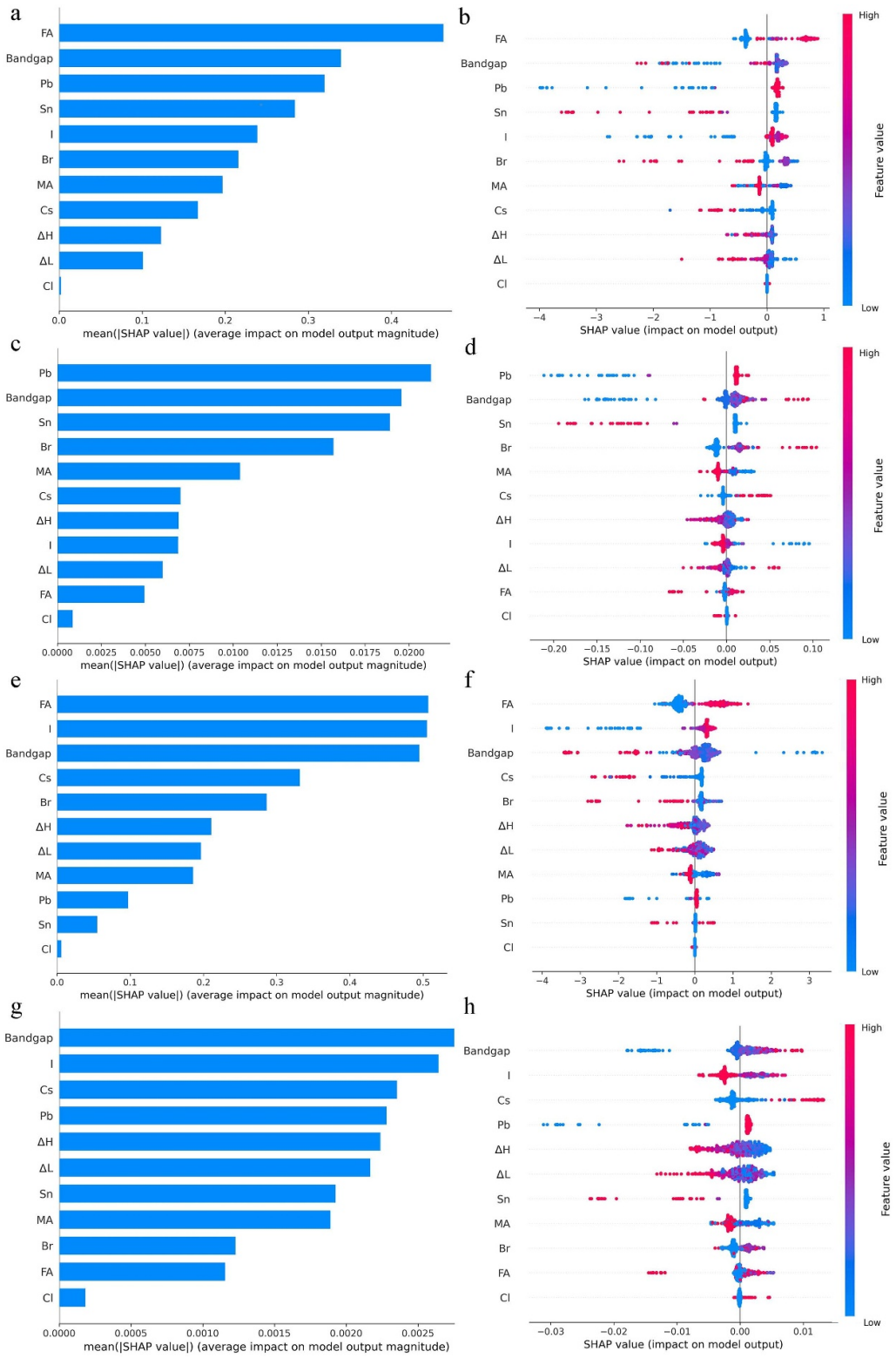

Figure 6. (a) Bar chart of feature importance and (b) SHAP values of each sample for PCE. (c) Bar chart of feature importance and (d) SHAP values of each sample for V_{oc}. (e) Bar chart of feature importance and (f) SHAP values of each sample for J_{sc}. (g) Bar chart of feature importance and (h) SHAP values of each sample for FF.

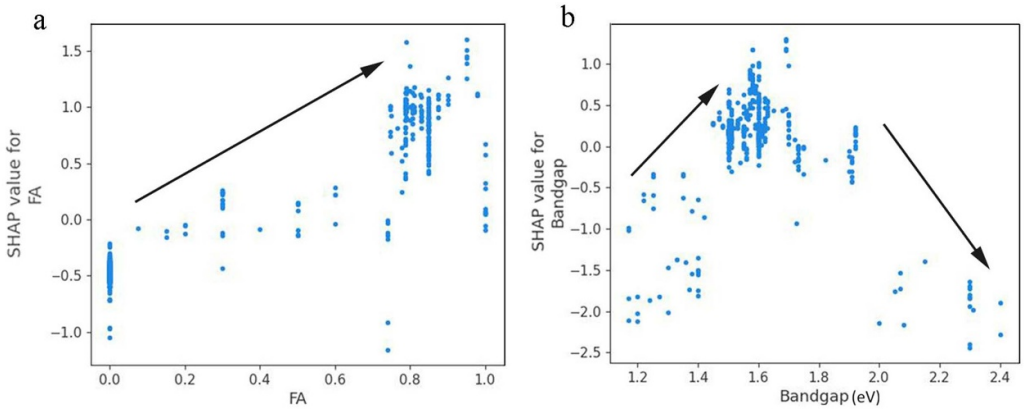

Figure 7. (a) Contribution statistics of FA and (b) bandgap features.

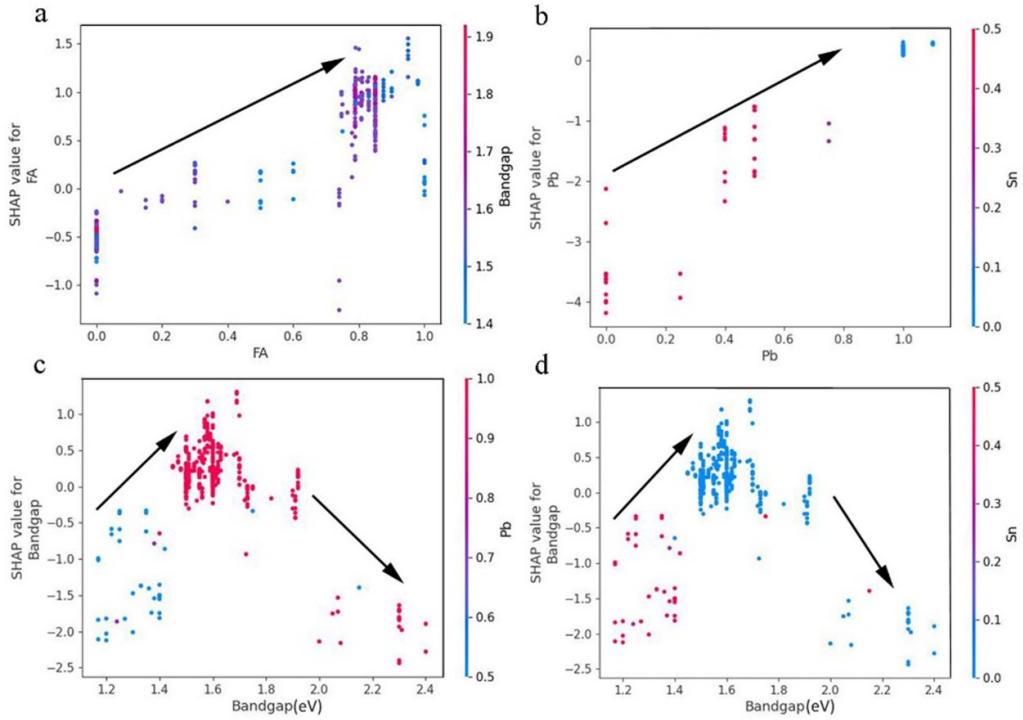

Figure 8. (a) Contribution statistics of FA with bandgap features (b) Pb with Sn features (c) bandgap with Pb features (d) bandgap with Sn features.

4. Conclusion

In conclusion, this study employed seven supervised ML algorithms to investigate the photovoltaic parameters of PSCs. Through rigorous data preprocessing of literature-collected datasets, we ensured the quality and suitability of the data, enabling the ML models to learn and predict more effectively. Subsequently, ML methods were applied to analyze device photovoltaic parameters, including PCE, J_{sc}, V_{oc}, and FF of PSCs. Notably, we enhanced the performance of the ANN model by introducing LayerNorm and residual connections to extract feature information, resulting in improved model fitting. The designed ANN model demonstrated superior performance in predicting PCE and J_{sc}, with RMSE values of 2.632% and 2.244 mA cm^{-2}, respectively. On the other hand, the RF model exhibited the best performance in predicting V_{oc} and FF, with RMSE values of 0.078 V and 0.057, respectively. Additionally, we employed the SHAP method to investigate the relative importance of features and provide insights into the factors influencing the performance of PSCs. This comprehensive analysis sheds light on the underlying mechanisms governing PSC performance and aids in the prediction of photovoltaic parameters. Overall, the findings of this study offer guidance for predicting the photovoltaic parameters of PSCs and are helpful in the fabrication of high-performance PSCs.

Data availability statement

All data that support the findings of this study are included within the article (and any supplementary files).

Funding

This research received support from the Natural Science Foundation of Jiangsu Province (Grant No. BK20190878), the Universities Natural Science Research Project of Jiangsu Province (Grant No. 19KJB510062), and the Lv Yang Jinfeng Project (Grant No. YZLYJF2020PHD073).

ORCID iDs

Zhan Hui https://orcid.org/0009-0005-2183-8675
Yunliang Yue https://orcid.org/0000-0002-6379-1888

References

[1] Kung P-K, Li M-H, Lin P-Y, Jhang J-Y, Pantaler M, Lupascu D C, Grancini G and Chen P 2020 Lead-free double perovskites for perovskite solar cells *Sol. RRL* **4** 1900306
[2] Kojima A, Teshima K, Shirai Y and Miyasaka T 2009 Organometal halide perovskites as visible-light sensitizers for photovoltaic cells *J. Am. Chem. Soc.* **131** 6050–1
[3] Chen H, Maxwell A, Li C, Teale S, Chen B, Zhu T, Ugur E, Harrison G, Grater L and Wang J 2023 Regulating surface potential maximizes voltage in all-perovskite tandems *Nature* **613** 676–81
[4] Kumar A, Singh S, Mohammed M K and Sharma D K 2022 Accelerated innovation in developing high-performance metal halide perovskite solar cell using machine learning *Int. J. Mod. Phys.* B **37** 2350067
[5] She C, Huang Q, Chen C, Jiang Y, Fan Z and Gao J 2021 Machine learning-guided search for high-efficiency perovskite solar cells with doped electron transport layers *J. Mater. Chem.* A **9** 25168–77
[6] Kumar N S and Naidu K C B 2021 A review on perovskite solar cells (PSCs), materials and applications *J. Mater.* **7** 940–56
[7] Kim D, Muckley E S, Creange N, Wan T H, Ann M H, Quattrocchi E, Vasudevan R K, Kim J H, Ciucci F and Ivanov I N 2021 Exploring transport behavior in hybrid perovskites solar cells via machine learning analysis of environmental-dependent impedance spectroscopy *Adv. Sci.* **8** 2002510
[8] Roy P, Sinha N K, Tiwari S and Khare A 2020 A review on perovskite solar cells: evolution of architecture, fabrication techniques, commercialization issues and status *Sol. Energy* **198** 665–88
[9] Ansari M I H, Qurashi A and Nazeeruddin M K 2018 Frontiers, opportunities, and challenges in perovskite solar cells: a critical review *J. Photochem. Photobiol.* C **35** 1–24
[10] Liu Y, Tan X, Liang J, Han H, Xiang P and Yan W 2023 Machine learning for perovskite solar cells and component materials: key technologies and prospects *Adv. Funct. Mater.* **33** 2214271
[11] Hu Y *et al* 2022 Machine-learning modeling for ultra-stable high-efficiency perovskite solar cells *Adv. Energy Mater.* **12** 2201463
[12] Kumar A, Singh S, Srivastava K, Sharma A and Sharma D K 2022 Performance and stability enhancement of mixed dimensional bilayer inverted perovskite (BA_2PbI_4/$MAPbI_3$) solar cell using drift-diffusion model *Sustain. Chem. Pharm.* **29** 100807
[13] Correa-Baena J-P, Saliba M, Buonassisi T, Grätzel M, Abate A, Tress W and Hagfeldt A 2017 Promises and challenges of perovskite solar cells *Science* **358** 739–44
[14] Chen J and Park N-G 2018 Inorganic hole transporting materials for stable and high efficiency perovskite solar cells *J. Phys. Chem.* C **122** 14039–63
[15] Wang Z, Yang M, Xie X, Yu C, Jiang Q, Huang M, Algadi H, Guo Z, Zhang H J A C and Materials H 2022 Applications of machine learning in perovskite materials *Adv. Compos. Hybrid Mater.* **5** 2700–20
[16] Li J, Peng Y, Zhao L, Chen G, Zeng L, Wei G and Xu Y 2022 Machine-learning-assisted discovery of perovskite materials with high dielectric breakdown strength *Mater. Adv.* **3** 8639–46
[17] Liu Z, Rolston N, Flick A C, Colburn T W, Ren Z, Dauskardt R H and Buonassisi T 2022 Machine learning with knowledge constraints for process optimization of open-air perovskite solar cell manufacturing *Joule* **6** 834–49
[18] Lemm D, von Rudorff G F and von Lilienfeld O A 2023 Improved decision making with similarity based machine learning: applications in chemistry *Mach. Learn.: Sci. Technol.* **4** 045043
[19] Lundberg S M and Lee S-I 2017 A unified approach to interpreting model predictions *Proc. 31st Int. Conf. on Neural Information Proc.* pp 4768–77
[20] Hui Z, Wang M, Yin X and Yue Y 2023 Machine learning for perovskite solar cell design *Comput. Mater. Sci.* **226** 112215
[21] Bansal N K, Mishra S, Dixit H, Porwal S, Singh P and Singh T 2023 Machine learning in perovskite solar cells: recent developments and future perspectives *Energy Technol.* **11** 2300735
[22] Gok E C, Yildirim M O, Haris M P, Eren E, Pegu M, Hemasiri N H, Huang P, Kazim S, Uygun Oksuz A and Ahmad S 2022 Predicting perovskite bandgap and solar cell performance with machine learning *Sol. RRL* **6** 2100927
[23] Del Cueto M, Rawski-Furman C, Arago J, Orti E and Troisi A 2022 Data-driven analysis of hole-transporting materials for perovskite solar cells performance *J. Phys. Chem.* C **126** 13053–61
[24] Mishra S, Gaikwad S B and Singh T 2024 Machine learning guided strategies to develop high efficiency indoor perovskite solar cells *Adv. Theory Simul.* **7** 2301193
[25] Yan W, Liu Y, Zang Y, Cheng J, Wang Y, Chu L, Tan X, Liu L, Zhou P and Li W 2022 Machine learning enabled development of unexplored perovskite solar cells with high efficiency *Nano Energy* **99** 107394
[26] Tao Q, Xu P, Li M and Lu W 2021 Machine learning for perovskite materials design and discovery *npj Comput. Mater.* **7** 23
[27] Rumman A H, Sahriar M A, Islam M T, Shorowordi K M, Carbonara J, Broderick S and Ahmed S 2023 Data-driven design for enhanced efficiency of Sn-based perovskite solar cells using machine learning *APL Mach. Learn.* **1** 046117
[28] Li J, Pradhan B, Gaur S and Thomas J 2019 Predictions and strategies learned from machine learning to develop high-performing perovskite solar cells *Adv. Energy Mater.* **9** 1901891
[29] Liu Y, Yan W, Han S, Zhu H, Tu Y, Guan L and Tan X 2022 How machine learning predicts and explains the performance of perovskite solar cells *Sol. RRL* **6** 2101100
[30] Cai X, Liu F, Yu A, Qin J, Hatamvand M, Ahmed I, Luo J, Zhang Y, Zhang H and Zhan Y 2022 Data-driven design of high-performance $MASn_xPb_{1-x}I_3$ perovskite materials by machine learning and experimental realization *Light Sci. Appl.* **11** 234

[31] Vakharia V, Castelli I E, Bhavsar K and Solanki A J P L A 2022 Bandgap prediction of metal halide perovskites using regression machine learning models *Phys. Lett.* A **422** 127800
[32] Zhang L, He M and Shao S 2020 Machine learning for halide perovskite materials *Nano Energy* **78** 105380
[33] Yılmaz B and Yıldırım R 2021 Critical review of machine learning applications in perovskite solar research *Nano Energy* **80** 105546
[34] Wang B, Fan Q and Yue Y 2022 Study of crystal properties based on attention mechanism and crystal graph convolutional neural network *J. Phys.: Condens. Matter* **34** 195901
[35] Probst P, Wright M N and Boulesteix A L 2019 Hyperparameters and tuning strategies for random forest *Wiley Interdiscip. Rev.* **9** e1301
[36] Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado G S, Davis A, Dean J and Devin M 2016 TensorFlow: large-scale machine learning on heterogeneous distributed systems (arXiv:1603.04467)
[37] Cherukara M J and Mannodi-Kanakkithodi A 2022 Deep learning the properties of inorganic perovskites *Modelling Simul. Mater. Sci. Eng.* **30** 034005
[38] Huang L, Qin J, Zhou Y, Zhu F, Liu L and Shao L 2023 Normalization techniques in training DNNs: methodology, analysis and application *IEEE Trans. Pattern Anal. Mach. Intell.* **45** 10173–96
[39] He K, Zhang X, Ren S and Sun J 2015 Deep residual learning for image recognition *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* pp 770–8
[40] Sivaraman G, Jackson N E, Sanchez-Lengeling B, Vázquez-Mayagoitia Á, Aspuru-Guzik A, Vishwanath V and De Pablo J J 2020 A machine learning workflow for molecular analysis: application to melting points *Mach. Learn.: Sci. Technol.* **1** 025015
[41] Schleder G R, Padilha A C, Acosta C M, Costa M and Fazzio A 2019 From DFT to machine learning: recent approaches to materials science–a review *J. Phys. Mater.* **2** 032001
[42] Hui Z, Wang M, Wang J, Chen J, Yin X and Yue Y 2024 Predicting the properties of perovskite materials by improved compositionally restricted attention-based networks and explainable machine learning *J. Appl. Phys.* **57** 315303
[43] Xie T and Grossman J C 2018 Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties *Phys. Rev. Lett.* **120** 145301
[44] Li W, Jacobs R and Morgan D 2018 Predicting the thermodynamic stability of perovskite oxides using machine learning models *Comput. Mater. Sci.* **150** 454–63
[45] Yang J, Tao L, He J, McCutcheon J R and Li Y 2022 Machine learning enables interpretable discovery of innovative polymers for gas separation membranes *Sci. Adv.* **8** eabn9545
[46] Lu Y, Wei D, Liu W, Meng J, Huo X, Zhang Y, Liang Z, Qiao B, Zhao S and Song D 2023 Predicting the device performance of the perovskite solar cells from the experimental parameters through machine learning of existing experimental results *J. Energy Chem.* **77** 200–8
[47] Chen J, Ooi L Q R, Tan T W K, Zhang S, Li J, Asplund C L, Eickhoff S B, Bzdok D, Holmes A J and Yeo B T 2023 Relationship between prediction accuracy and feature importance reliability: an empirical and theoretical study *NeuroImage* **274** 120115
[48] Van den Broeck G, Lykov A, Schleich M and Suciu D 2022 On the tractability of SHAP explanations *J. Artif. Intell. Res.* **74** 851–86
[49] Zhang S, Lu T, Xu P, Tao Q, Li M and Lu W 2021 Predicting the formability of hybrid organic–inorganic perovskites via an interpretable machine learning strategy *J. Phys. Chem. Lett.* **12** 7423–30
[50] Gong S, Wang S, Xie T, Chae W H, Liu R, Shao-Horn Y and Grossman J C 2022 Calibrating DFT formation enthalpy calculations by multifidelity machine learning *JACS Au* **2** 1964–77
[51] Li W, Rothmann M U, Zhu Y, Chen W, Yang C, Yuan Y, Choo Y Y, Wen X, Cheng Y-B and Bach U 2021 The critical role of composition-dependent intragrain planar defects in the performance of $MA_{1-x}FA_xPbI_3$ perovskite solar cells *Nat. Energy* **6** 624–32
[52] Das B, Aguilera I, Rau U and Kirchartz T 2022 Effect of doping, photodoping, and bandgap variation on the performance of perovskite solar cells *Adv. Opt. Mater.* **10** 2101947
[53] Mishra S, Boro B, Bansal N K and Singh T 2023 Machine learning-assisted design of wide bandgap perovskite materials for high-efficiency indoor photovoltaic applications *Mater. Today Commun.* **35** 106376