Solar RRL
www.solar-rrl.com

How Machine Learning Predicts and Explains the Performance of Perovskite Solar Cells

Yiming Liu, Wensheng Yan, Shichuang Han, Heng Zhu, Yiteng Tu, Li Guan,* and Xinyu Tan**

Characterizing the electrical parameters of perovskite solar cells (PSCs) usually requires a lot of time to fabricate complete devices. Here, machine learning (ML) is used to reduce the device fabrication process and predict the electrical performance of PSCs. Using ML algorithms and 814 valid data cleaned from 2735 peer-reviewed publications, ML prediction models are built for bandgap, conduction band minimum, valence band maximum of perovskites, and electrical parameters of PSCs. These prediction models have excellent accuracy, and the root mean square error of the prediction models for bandgap and power conversion efficiency (PCE) reaches 0.064 eV and 1.58%, respectively. Among the many factors that affect the performance of PSCs, those factors play a major role in the lack of comprehensive explanation. Through the prediction model of electrical parameters and Shapley Additive explanations theory, the factors affecting the PCE of PSCs are explained and analyzed. It can not only verify the objective physical laws from the perspective of ML, but also conclude that among the 13 features, the content of formamidinium/$NH_2CHNH_2^+$ plays the most important role in improving the PCE of PSCs. These results show that ML has great application possibilities in the PSC field.

1. Introduction

As a representative of the third generation of solar cells, perovskite solar cells (PSCs) have experienced rapid development in the past 12 years, and the certified power conversion efficiency (PCE) of single-junction PSC has now reached 25.5%.[1] The PCE of PSCs can be improved so quickly in a short time, which is inseparable from the excellent properties of the perovskite material itself (tunable bandgap, faster carrier mobility, high absorption coefficients, lower exciton binding energy, etc.).[2–6] Of course, it is even more inseparable from the optimization design of PSCs by many scholars, such as the optimization of energy-level alignment; the development of new transport layer materials; the passivation of material defects.[7–9] However, without exception, these require a lot of time and resources for trial and error experiments. To overcome this problem, some simulation software based on density functional theory (DFT) have been developed, which can simulate the structure and chemical composition of materials.[10–12] Nevertheless, researchers need to have a lot of computing resources and rich knowledge of quantum chemistry, and most simulation calculations are aimed at the material itself. As the studied system becomes complex, such as devices (systems) composed of various chemical materials, these methods are no longer applicable.

The behavior that computers use data to analyze, learn and summarize is called artificial intelligence,[13] and ML belongs to an important branch of artificial intelligence.[14] It can infer potential rules and relationships among materials, and between materials and the features of the complex system composed of materials only through the data itself without knowing the physical laws. Machine learning (ML) has broad prospects in the field of materials science, such as searching for virtual materials; reducing the amount of calculation required for DFT; reducing experimental procedures and time, etc.[15–17] Although ML has been used in the field of materials, the research in the field of PSCs is still in the preliminary exploration stage,[18,19] such as the prediction of the formability and stability of perovskites,[20] the screening of lead-free perovskite materials suitable for cells,[21] and the judgment of the features affecting the stability of PSCs,[22] and so on. According to the way of data acquisition, the research of ML in the field of perovskite can be divided into two types: ML analysis based on simulated data generated by DFT theory or experimental data from real experiments.[23,24] In which, Saidi et al. used structural information of 380 perovskite compositions, which were generated by DFT

Y. Liu, W. Yan, H. Zhu, Y. Tu, X. Tan
College of Electrical Engineering & New Energy
Hubei Provincial Collaborative Innovation Center for New Energy Microgrid
China Three Gorges University
Yichang 443002, China
E-mail: tanxin@ctgu.edu.cn

W. Yan
Electronics and Information College
Hangzhou Dianzi University
Hangzhou 310018, China
E-mail: wensheng.yan@hdu.edu.cn

S. Han, L. Guan
Department of Physics Science and Technology
Hebei University
Baoding 071000, China
E-mail: lguan@hbu.edu.cn

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/solr.202101100.

DOI: 10.1002/solr.202101100

and convolutional neural networks (CNN) to predict bandgaps, lattice constants, and octahedral angles.[25] Shahzada. Ahmad et al used the random forest (RF) algorithm to predict the bandgap and PCE of PSCs based on UV–vis absorption and *J*–*V* spectra data[26] The aforementioned articles mostly use ML tools to screen unknown materials or predict their performance. The explanations of ML models and the mining of potential physical laws between features are rarely involved.

Our contributions mainly include the following three aspects: 1) We cleaned 814 data with bandgap, CBM, VBM, perovskite composition, and carrier mobility from 4812 published research articless and further established device-level database of bandgap and electrical parameters with 8/11/13 dimensional features; 2) Through the aforementioned 2 databases and 7 ML algorithms, we have built 49 prediction models and selected 7 prediction models with the best performance ($Bandgap_{XGBoost}$, $CBM_{XGBoost}$, $VBM_{XGBoost}$, PCE_{RF}, V_{ocRF}, $J_{scRF,}$ and FF_{RF}, where the subscript is the ML algorithm used in the prediction model) based on root mean square error (RMSE) and Pearson correlation coefficient. The features required by the electrical parameter prediction models can also be input through $Bandgap_{XGBoost}$, $CBM_{XGBoost,}$ and $VBM_{XGBoost}$ to reduce the amount of features that need to be actually characterized; and 3) Exploring the predictive mechanisms of ML models may lead to some unexpected findings that can be used to guide further research directions. Global and local explanations of samples and prediction models are carried out through the SHapley Additive exPlanations (SHAP) explainer. At the global level, the SHAP values of each feature under different algorithms and data structures are calculated, and the feature importance ranking affecting the PCE of ABX_3-type PSCs is obtained. At the local level, the samples with low and high efficiency in the database are explained, respectively, and the reasons for low and high PCE are found. Finally, it is found that the content of FA is essential for single-junction high-performance ABX_3-type PSCs.

2. Build Models

As shown in **Figure 1**, the present work establishes and explains the prediction model according to the following process. It mainly includes model explanation, establishment of database, feature extraction, model selection, model evaluation, and model application.

2.1. Model Explanation

As the complexity of ML models increases, some ML models gradually lose explainability while gaining superior predictive ability, and these models are called "black-box models." For example, RF, gradient boosting decision tree (GBDT), Xtreme Gradient boosting (XGBoost). SHAP originated from cooperative game theory. In 2017, Lundberg and Lee's article used SHAP values to explain various ML models, making the ML models explicable.[27] SHAP is a cumulative explainable tool. Suppose the ith sample is x_i, the jth feature of the ith sample is $x_{i,j}$, the predicted value of the model for the ith sample is $f(x_i)$, and the base value of the model (the mean of the dependent variable for all samples) is f_{base}, then the SHAP value obeys the following equation

$$f(x_i) = f_{base} + \sum_{j=1}^{n} \varphi(x_{i,j}) \tag{1}$$

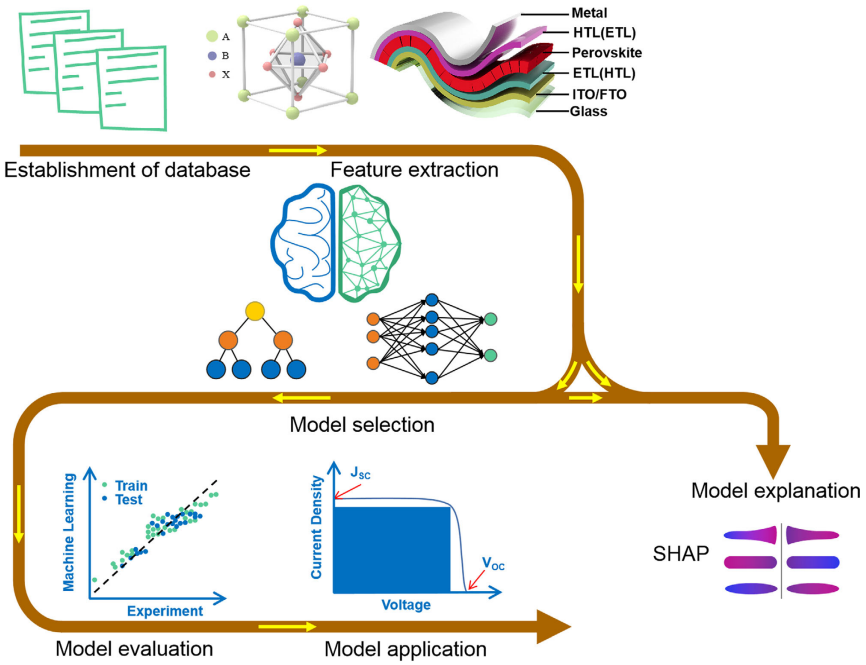

Figure 1. The flowchart of establishing and explaining for the prediction model.

$\varphi(x_{i,j})$ is the contribution of the j_{th} feature in the i_{th} sample to the predicted value, namely, SHAP value.

As shown in **Figure 2**, when $\varphi(x_{i,j}) > 0$, it means that the feature boosts the predictive value, corresponding to the red area; conversely, it means that the feature lowers the predictive value, corresponding to the blue area. the greatest advantage of SHAP is that it reflects the influence of the feature of each sample, and also shows the negative and positive contribution of this influence. We make two kinds of explanations by SHAP. 1) Global explanations. Rank the importance of all features and visually represent the contribution of each feature when it changes; and 2) Local features and sample explanations. Explaining the influence of features on predicted values under a single feature or two-feature interactions. Interpretable dimensionality reduction was performed using principal component analysis (PCA) and SHAP values to analyze the features that had the greatest impact on the first and second principal components. The explanatory analysis is performed for the main features that affect the predicted value of a single sample.

2.2. Establishment of Database

2.2.1. Data Preparation

Articles that contain the keywords "perovskite energy level," "perovskite bandgap," and "perovskite mobility" were downloaded from Wiley, Elsevier, Springer Nature, and other publishers, totaling 4812 articles. Only articles published after 2014 were retained, totaling 3224 articles. To ensure a sufficient amount of data and the practicability of the ML model, we only retained articles about the perovskite solar cell that has ABX_3-type structure. In the end, there were only 2735 articles left. The monovalent cation at the A position is usually cesium (CS^+), methylammonium ($MA^+/CH_3NH_3^+$), formamidinium ($FA^+/NH_2CHNH_2^+$); the divalent metal ion at the B position is usually lead (Pb^{2+}), Tin (Sn^{2+}); the halogen anion at the X position is iodide (I^-), bromide (Br^-), chloride (Cl^-). For the structure of single-junction PSCs, we chose the most common structure (with photoanode, single-electron transport layer, single-perovskite layer, single-hole transport layer, and metal cathode). The active areas of PSCs are about 0.1 cm^2.

2.2.2. Data Cleaning

The first step is to delete the duplicate and abnormal data, such as the duplicate bandgap data under the same chemical composition and the data with too low PCE and fill factor (FF) due to immature technological means.[28] The second step is to ensure the standardization of data sources that the bandgap, CBM, and VBM of perovskite materials are obtained by experimental means. The electrical parameters of PSCs were measured under AM1.5G sunlight at 100 mW cm^{-2}. The carrier mobility is uniformly measured by the space-charge-limited current method (SCLC), and considering the influence of doping on the carrier mobility of the material, ensure that the state of the material used in the device (whether doped) is consistent with the state of the material during SCLC test. The third step is data supplement. As shown in Table S1, Supporting Information, common material features that are not mentioned in the literature but are used in modeling are uniformly supplemented.

2.3. Feature Extraction

The properties of materials or devices are determined by some factors, which are also called features. Feature extraction usually follows three principles: highly relevant to the output, easy to obtain, and minimal number.[29,30] For ABX_3-type perovskites, the information of bandgap, CBM, and VBM are mainly determined by the chemical compositions due to the fixed structure. Therefore, we select eight chemical compositions (MA, FA, CS, Pb, Sn, Br, Cl, I) as the features of the bandgap, CBM, and VBM prediction models. The electrical performances of PSCs are determined by numerous features such as the high absorption coefficient and high carrier mobility of perovskite, the transport and blocking effect of transport layer for holes and electrons, the effect of energy offsets alignment at the interface for carrier transfer, etc. However, considering the principles of feature extraction, cell structure, and energy conversion theory (exciton generation, carrier separation, and collection), we constructed a

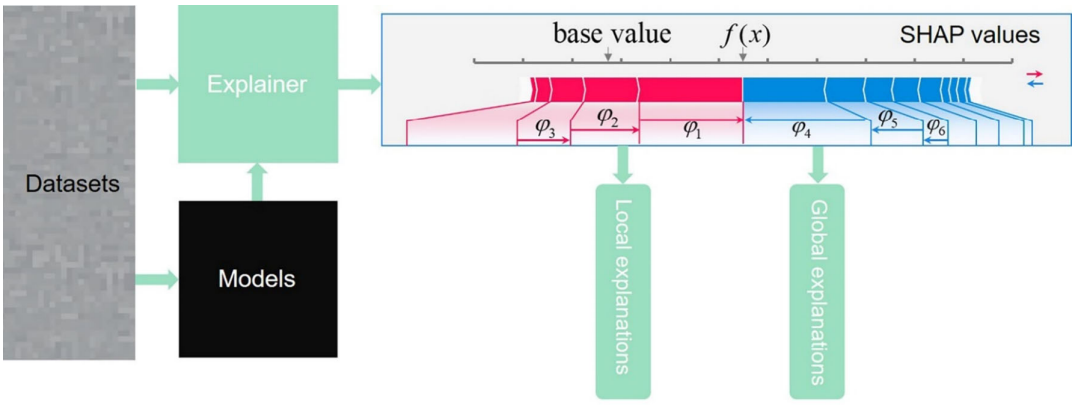

Figure 2. Schematic representation of SHAP explanation. $f(x)$ is the final predicted value of a sample, base value is the average of the predicted values of all samples. The red and blue areas represent the positive and negative contributions to the final predicted value of the sample, φ_x is the contribution value of the X_{th} feature.

set of representative features. Taking into account the properties of perovskite itself, we choose eight chemical compositions and bandgap of ABX_3-type perovskites as the features of absorber layer properties(MA, FA, Cs, Pb, Sn, Br, Cl, I, bandgap). The energy-level alignment at the interface is considered to act as a barrier and energy loss for the transfer of electrons and holes. We use energy-level alignment H ($H = \mathrm{HTL_{HOMO}} - \mathrm{Perovskite_{VBM}}$) and L ($L = \mathrm{Perovskite_{CBM}} - \mathrm{ETL_{LOMO}}$) as a feature representing the energy-level alignment at the interface (HTL/perovskite, ETL/perovskite, highest occupied molecular level for HOMO, and lowest unoccupied molecular level for LOMO). The transport layer can promote carrier transport and inhibit recombination. Taking into account the role of the transport layer on carrier transport, we use electron and hole mobility to represent the features of the transport layer material. Therefore, we constructed a set of 13-dimensional features as the inputs of the electrical parameter prediction models (MA, FA, Cs, Pb, Sn, Br, Cl, I, bandgap, H, L, electron mobility, hole mobility). Finally, we established device-level database, including 103 and 248 sets of data, respectively. The databases are shown in Table S2 and S3 (Supporting Information). The 11-dimensional feature database of electrical parameters (except electron mobility and hole mobility) used only for data dimension comparison and model explanation contains 463 groups of data, as shown in Table S4, Supporting Information. The 13 features were chosen for an explanation as shown in Table S19, Supporting Information.

2.4. Model Selection

Prediction models and SHAP explainer are built using scikit-learn and SHAP toolkit on PyCharm platform.[31] Each prediction model is optimized by seven ML algorithms,[32] which are linear regression (LR), K-nearest neighbor (KNN),[33] support vector regression (SVR),[34] random forest (RF),[35] multilayer perceptron (MLP), gradient boosting decision tree (GBDT)[36] and XGBoost.[37] Linear regression (LR) is the simplest regression algorithm based on supervised learning, which uses the best linear function between the independent variable and the dependent variable for fitting. KNN is often used for classification problems, and its principle is simply summarized as "those who are close to each other are red, and those who are close to each other are black." The K known "neighbors" closest to the unknown sample are used to vote, and then the unknown sample is predicted according to the majority-voting rule and the information of K "neighbors." The K value of the KNN algorithm is regulated by the n_neighbors parameter in scikit-learn. SVR is an application of SVM for nonlinear regression problems that maps the data to the required feature space using kernel function. SVR creates a "spacing band" on both sides of the linear function, with a spacing of ε (an empirical value set manually). The loss is not calculated for all samples falling into the spacing band. Finally, the optimized model is obtained by minimizing the total loss and maximizing the spacing. MLP is also known as a feedforward neural network, which consists of an input layer, hidden layer, and output layer. Each layer is fully connected to the next layer. It has three basic elements: weights, biases, and activation functions. RF is an ensemble of independent tree learners. First, multiple training sets are generated by the bootstrap method, and then, for each training set, a decision tree is constructed. When the nodes of the tree need to be split, a part of the features is randomly extracted from the total features, and the optimal solution is found among the extracted features, which are applied to the nodes and split. Due to the idea of bagging, its over-fitting and anti-interference ability are relatively strong. For regression problems, the final result depends on the average of the prediction results of all tree learners. Unlike RF, GBDT is an iterative algorithm that uses decision trees as learners, where all decision trees are regression trees. Its core is to use the residuals left by K-1 regression trees before fitting the Kth regression tree, thus continuously reducing the error of the model. The basic idea of XGBoost is the same as that of GBDT, but some optimizations are made, such as default missing value processing, adding second-order derivative information, regular term, and parallel computing. Among them, RF, GBDT, XGBoost are originated from ensemble learning ideas.

2.5. Model Evaluation

The relevant hyper-parameter search and model evaluation was performed by the "GridSearchCV" and "Fivefold cross-validation." The evaluation metrics are RMSE and Pearson's correlation coefficient (r value). RMSE is the most popular regression evaluation indicator, which can be used to judge the accuracy of the ML model. To prevent overfitting,[17,38] the RMSE of the training set shall not be less than 70% of the RMSE of the test set. Pearson's correlation coefficient is used to judge the correlation between the predicted values output by the ML model and the collected experimental values. The r value between 0.8 and 1 is a very strong correlation, and the correlation decreases every 0.2 for strong correlation, moderate correlation, and weak correlation, respectively.

2.6. Model Application

Based on the established database and ML algorithm, we selected 7 models from 49 prediction models to predict parameters of PSCs (Bandgap, CBM, VBM, PCE, J_{sc}, V_{oc}, FF). The PCE prediction model is used to perform SHAP explanations and analysis, and then find the main factors affecting PCE of PSCs among 13 influencing factors.

3. Results and Discussion

Prediction performance of bandgap, CBM, and VBM. First, 20% of the total data in Table 2 is randomly selected as the test set while the remaining 80% is used as the training set (the training set and the test set are mutually exclusive), and then use 7 ML algorithms for model training. As shown in Table S5–S7, Supporting Information, we list the optimal algorithm, hyperparameters, r-value of the test set, RMSE of the training set and test set under each prediction model. The RMSE of the training set represents the learning performance of the model. The lower the RMSE value of the training set, the better the learning performance of the model. However, too much difference from the RMSE of the test set may lead to over-fitting. The r-value

Table 1. Comparison of the performance for different prediction models and their parameters under the XGBoost algorithm (using Table S2, Supporting Information, as the database).

Models	Hyper-parameters	Train RMSE [eV]	Test RMSE [eV]	*r* value (Test)
Band $gap_{XGBoost}$	'colsample_bytree':0.7,'gamma':0,'learning_rate':0.1,'max_depth':2,'min_child_weight':3, 'n_estimators':300, 'reg_alpha':0, 'reg_lambda':1, 'subsample': 0.8	0.044	0.064	0.96
$CBM_{XGBoost}$	'colsample_bytree':0.5,'gamma':0.1,'learning_rate':0.15,'max_depth':4,'min_child_weight':3, 'n_estimators':5000,'reg_alpha':0, 'reg_lambda':1,'subsample': 0.9	0.134	0.133	0.78
$VBM_{XGBoost}$	'colsample_bytree':0.7, 'gamma':0.1,'learning_rate': 0.15,'max_depth':4,'min_child_weight':4, 'n_estimators':5000, 'reg_alpha':0, 'reg_lambda':2,'subsample': 0.9	0.149	0.178	0.71

and RMSE of the test set represent the prediction ability of the model for unknown materials. High *r*-value and low RMSE indicate that the model has strong prediction ability. It can be seen from Table S5–S7, Supporting Information, that as the ML algorithm becomes more and more complex, the fitting situation and predictive ability of the model become more perfect. Among them, the top three prediction performances are three algorithms (RF, GBDT, XGBoost) under the ensemble learning framework. Because ensemble learning builds multiple "learners" to make combined predictions together, which has a strong accuracy for some tabular data that are not strongly related in the spatial and temporal dimensions. Of course, the best performing algorithm is XGBoost for the database we built. **Table 1** shows the three models with the best prediction performance for Bandgap, CBM, and VBM ($Bandgap_{XGBoost}$, $CBM_{XGBoost}$, $VBM_{XGBoost}$). As shown in Table 1, the training and test sets RMSE of Band $gap_{XGBoost}$, $CBM_{XGBoost}$, $VBM_{XGBoost}$ are 0.044 and 0.064 eV, 0.134 and 0.133 eV, 0.149, and 0.178 eV. This shows that these models have a perfect fit and satisfactory predictive ability. The RMSE of a similar training set and test set shows that the model is not over-fitting. The *r*-value of each model is above 0.7, indicating that the predicted value of our model has a strong correlation with the experimental value. In other words, the predictive model we build has a better predictive ability for the bandgaps of unknown composition. **Figure 3** shows the fitting performance of our ML model. The prediction values of the three prediction models are very close to the experimental values in published articles, which represents the superiority of the prediction model we built.

Prediction performance of four electrical parameters. As shown in Table S3, Supporting Information, 13 features (MA, FA, Cs, Pb, Sn, Br, Cl, I, Bandgap, H, L, electron mobility, hole mobility) have been used to predict four electrical parameters (PCE, J_{sc}, V_{oc}, FF). The division of training and test sets is the same as aforementioned. The optimal hyper-parameters, algorithms, and training results under the four electrical parameters are shown in Table S8–S11, Supporting Information. Because the nature of the data is similar to that of Table S2, Supporting Information, the best performance is also the three algorithms under the ensemble learning. As shown in Table S8, Supporting Information, the RMSE of the training set and test set are decreased from 2.34% and 2.96% to 1.17% and 1.58%, in other words, the accuracy of the ML model increased by more than 40%, and the *r*-value of the test set increased from 0.61 to 0.86, which further illustrates the superiority of the complex ML models. As shown in Table S9, Supporting Information, the RMSE of V_{oc} predicted value are also significantly reduced through the filtering of the algorithm, the RMSE of the training set is reduced from 0.077 to 0.036 V, the RMSE of the test set is reduced from 0.093 to 0.051 V, and the *r* value is increased from 0.76 to 0.93. As shown in Table S10, Supporting Information, the *r*-value under each algorithm is above 0.9, which indicates that the features we built can well correlate with the changes of J_{sc}. Of course, the RMSE of training and test sets can also be reduced

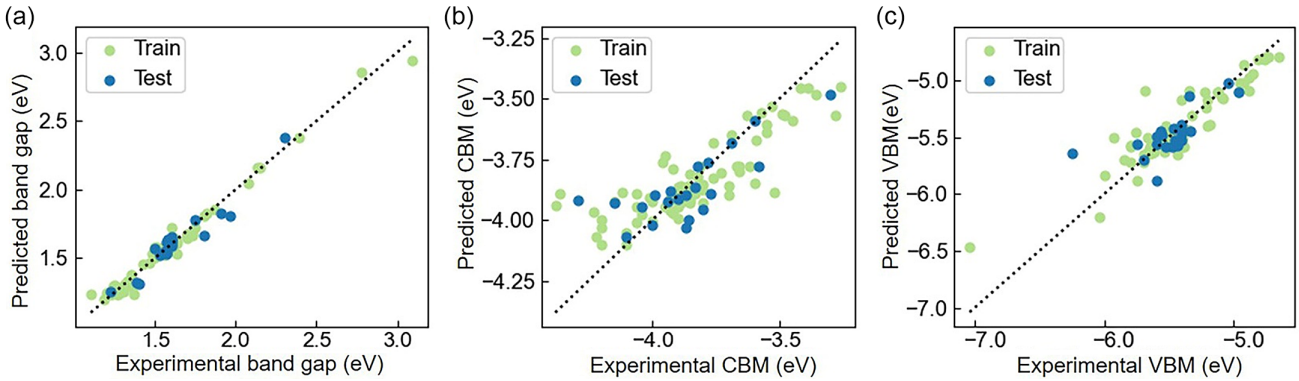

Figure 3. Comparison of experimental values and ML prediction values using XGBoost algorithm. The horizontal axis represents the experimental data collected from published articles, and the vertical axis represents the predicted value output by our prediction model. The green and blue dots are the training and test dataset. a) Predicted bandgap versus experimental bandgap. b) Predicted CBM versus experimental CBM. c) Predicted VBM versus experimental VBM.

Table 2. Comparison of the performance for different prediction models and their parameters under the RF algorithm (using Table S3, Supporting Information, as the database).

Models	Hyper-parameters	Train RMSE	Test RMSE	*r* value (Test)
PCE_{RF}	n_estimators = 50,max_features = 'auto',max_depth = 10,min_samples_split = 3,min_samples_leaf = 1,	1.15%	1.58%	0.86
Voc_{RF}	n_estimators = 500,max_features = 'sqrt',max_depth = 10,min_samples_split = 3,min_samples_leaf = 1,	0.036 V	0.051 V	0.93
Jsc_{RF}	n_estimators = 50,max_features = 'auto',max_depth = 10,min_samples_split = 3,min_samples_leaf = 1,	0.85 mA cm^{-2}	1.04 mA cm^{-2}	0.96
FF_{RF}	n_estimators = 20,max_features = 'sqrt',max_depth = 110,min_samples_split = 5,min_samples_leaf = 3,	0.034	0.046	0.63

with this highly correlated data. As shown in Table S11, Supporting Information, although the RMSE of the training set and the test set are also low (from training set: 0.049 to 0.033, test set: 0.073 to 0.046) and the r-value has increased considerably from 0.18 to 0.63, the *r*-value is the lowest of the four-parameter predictions. There are two reasons for this: We artificially deleted some data with low FF to prevent the influence of defects and fabrication processes on the prediction; Due to the difficulty of collecting some features that can represent FFs, we chose to give up, resulting in our built features being insensitive for predicting FF values. For this point, we can also use the three parameters predicted and the formula of photoelectric conversion efficiency to calculate. The best predictive models of electrical parameters (PCE_{RF}, V_{ocRF}, J_{scRF}, FF_{RF}) are shown in **Table 2**. The optimal algorithm is RF algorithm. The training and test sets RMSE of PCE_{RF}, V_{ocRF}, Jsc_{RF} and FF_{RF} are 1.15% and 1.58%, 0.036 V and 0.051 V, 0.85 and 1.04 mA cm^{-2}, 0.034 and 0.046, respectively. **Figure 4** shows the fitting performance of the training set and the test set of four electrical parameters prediction models, and the predicted values of the four electrical parameters are very close to the experimental values without over-fitting. Based on Table 2 and Figure 4, it can be demonstrated that our prediction models not only fit the known training data well but also predicts the electrical parameters for massive unknown materials of ABX3 perovskite solar cells. Despite the exciting prediction performance of the prediction models, it has to be admitted that there is still some error in the predicted values from the experimental values. We consider the differences of the cell fabrication process as one main reason for the error, although we use some means to

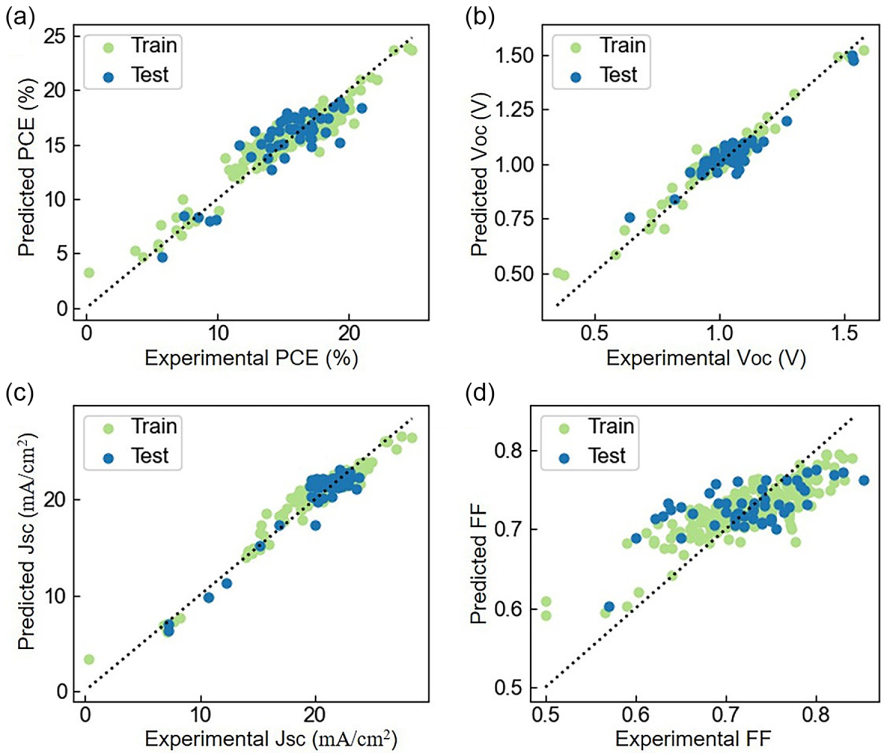

Figure 4. Comparison of experimental values and ML prediction values using RF algorithm. The horizontal axis represents the experimental data collected from published articles, and the vertical axis represents the predicted value output by our prediction model. The green and blue dots are the training and test dataset. a) Predicted PCE versus experimental PCE. b) Predicted V_{oc} versus experimental V_{oc}. c) Predicted J_{sc} versus experimental J_{sc}. d) Predicted FF versus experimental FF.

avoid it as much as possible (collecting the latest published data; avoiding data under immature processes; collecting data with detailed experimental details, etc.).

$Bandgap_{XGBoost}$, $CBM_{XGBoost}$, $VBM_{XGBoost}$ are the inputs of PCE_{RF}, V_{ocRF}, J_{scRF}, FF_{RF}. Using $Bandgap_{XGBoost}$, $CBM_{XGBoost}$, $VBM_{XGBoost}$ predicted bandgap, CBM, VBM as the inputs of four electrical parameter prediction models (PCE_{RF}, V_{ocRF}, J_{scRF}, FF_{RF}), to test the feasibility of combining the models. We randomly select 27 sets of experimental data from Table S3, Supporting Information, then replace the three features of Bandgap, H, and L with predicted values, and finally, predict four electrical parameters. Note that we use the same prediction model as in Table 1 and 2, making only changes in the data. The fitting performance is shown in Figure S1, Supporting Information, and the specific prediction parameters are shown in Table S13, Supporting Information. Table S13, Supporting Information, includes the comparison between the predicted and experimental values of the bandgap, PCE, J_{sc}, $V_{oc,}$ and FF. We found that the accuracy of the model when the two models are combined to predict electrical parameters is not much different from that when 13 experimental values are used as features. When the model is combined, the RMSE and *r*-value of the predicted and experimental values of PCE, V_{oc}, J_{sc}, FF are 1.06% and 0.96, 0.059 V, and 0.96, 1.03 mA cm^{-2} and 0.98, 0.041, and 0.62, respectively. This shows the feasibility of the model combination. The combination of models can significantly reduce the collection of experimental values (features). For devices where the transport layer is a material with unknown properties, we only need to collect the electron mobility and LOMO of the electron transport layer, and the hole mobility and HOMO of the hole transport layer to predict the electrical parameters of the device. When the composition of the ABX_3 perovskite and the material of the transport layer are known in the device, the collection of experimental values is not even required. The prediction model we built can also assist in the experimental work. For example, for the development of new materials of the transport layer, if the properties of the materials meet the expectations, but there are problems in the device PCE, this may be caused by energy level misalignment. At this time, the change of perovskite compositions and energy level alignment can be carried out through the above ML model to find the ABX_3 perovskite compositions most suitable for this material. To verify the accuracy of the model for bandgap and PCE prediction, we collected 7 latest reports. The five bandgap components and two hole transport materials have never existed in our database. As shown in Table S20 and S21, Supporting Information, our model is able to make excellent predictions.

Unlike traditional analysis methods, we can use the SHAP tool to find the main factors that influence the PCE of PSCs in a context when the underlying physical laws are not clear. To ensure that the factors affecting the PCE of PSCs found by SHAP theory are consistent with the real experimental conditions. First, we choose an electrical parameter prediction model (PCE_{RF}) with 13 experimental values as features for SHAP explanation, and the algorithm of PCE_{RF} used is the RF algorithm with the highest accuracy. Then, we verify whether the explanation obtained through the SHAP theory conforms to the physical facts.

As shown in **Figure 5**a,b, Supporting Information, the red dashed frames are the area where the SHAP value is positive (positive contribution area). When the energy-level difference is 0 eV, the SHAP value of H and L is the largest, representing the maximum gain effect on PCE of PSCs at this time. With the increase of energy-level difference, the SHAP value changes from positive to negative and shows an obvious downward trend, which shows that the increase of energy-level difference has a significant negative effect on the PCE of PSCs, and further shows that good energy level alignment can effectively increase the PCE of PSCs. As shown in Figure 5c,d, the SHAP value of electron and hole mobility changes sharply from a negative value to a positive value within a range. Most samples with high mobility have a positive range of SHAP value, but the SHAP value changes little with the increase of mobility. This indicates that too low mobility will significantly reduce the PCE of PSCs. The increase of mobility can appropriately improve the PCE of PSCs, but when the mobility increases more than a certain level, it has little effect on the change of PCE for PSCs. Figure 5 proves that the model explanation obtained by the SHAP explanation conforms to the objective physical facts.

As shown in **Figure 6**a, the change trend of FA content with bandgap is not obvious. The samples with a band gap in the range of 1.5–1.55 eV and high FA content have the highest SHAP value. With the increase or decrease of bandgap, the SHAP value has an obvious downward trend, indicating that high FA content and appropriate bandgap have an obvious gain contribution to the PCE of PSCs. As shown in Figure 6b, bandgap has an obvious decreasing trend with the increase of Sn content. The PSC with bandgap in the range of 1.5–1.55ev and low Sn content has the highest SHAP value, which indicates that the content of Sn can significantly affect the change of bandgap, and low bandgap and high content of Sn are the killers of high PCE. Therefore, it is necessary to have an appropriate bandgap, increase the addition of FA and reduce the content of Sn as much as possible for PSC with high PCE. See the following section for specific analysis.

Global explanations for factors affecting PCE of PSCs by SHAP theory. As shown in **Figure 7**, the PCE_{RF} is explained and 13 features that affect the PCE of PSCs are analyzed. The top five features with greater impact are FA, hole mobility, Sn, bandgap, and I in order. The decrease of hole mobility has a greater negative effect than the positive effect when it increases on the PCE of PSCs, which shows that the low PCE of PSCs are generally caused by low hole mobility, but this factor is not the most important factor leading to high PCE of PSCs. FA content has the most important influence among the 13 features. The decrease of FA content has a little negative impact on the PCE of PSCs, but the increase of FA content has an obvious gain trend on the PCE of PSCs, indicating that FA content plays a major role in the PSCs with high PCE. The increase of Sn content and the decrease of I content can significantly reduce the PCE of PSCs, which may be due to the content of Sn and I having an impact on the bandgap of perovskites, as well as the change of bandgap has an impact on both J_{sc} and V_{oc} of PSCs, and finally affects the PCE of PSCs. This indicates that PSCs with high PCE should increase the content of FA and select the appropriate bandgap as much as possible. We further explained different data and ML models by changing the algorithm and feature dimension to see whether similar findings could be explored. As algorithm comparison, we chose the prediction models

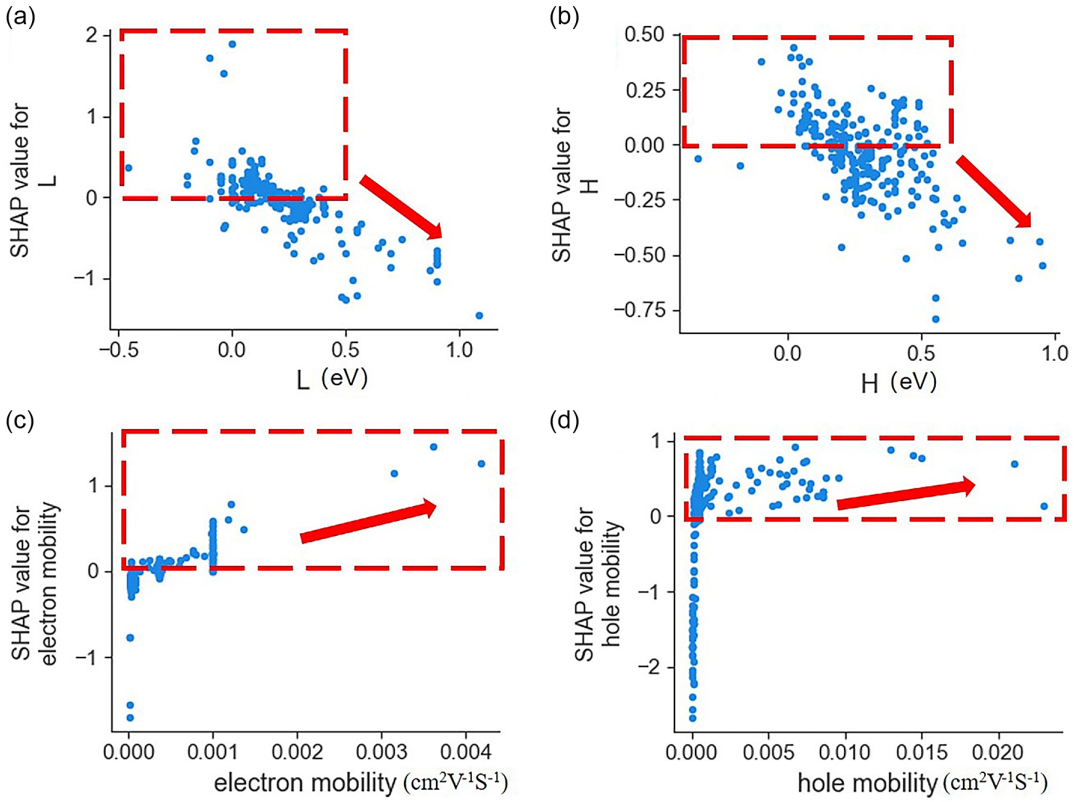

Figure 5. Dependence plot of a single feature versus its SHAP value in the PCE_{RF}. SHAP value represents the contribution of each feature to make a predicted value, the SHAP value increases as the contribution increases. the positive and negative SHAP values represent the gain and deduction contribution to the predicted value. a) L versus its SHAP value. b) H versus its SHAP value. c) Electron mobility versus its SHAP value. d) Hole mobility versus its SHAP value.

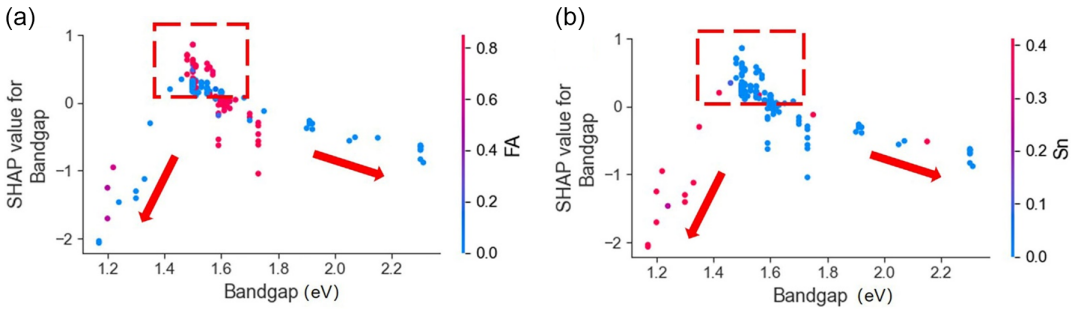

Figure 6. SHAP dependence plot of bandgap versus its SHAP value in the PCE_{RF}. a) Variation with FA. b) Variation with Sn.

(PCE_{GBDT}, $PCE_{XGBoost}$) whose prediction effect is second only to PCE_{RF} for an explanation. As shown in Figure S2 and S3, Supporting Information, the top five features that have the greatest impact when using PCE_{GBDT} are FA, I, hole mobility, bandgap, I. When the $PCE_{XGBoost}$ is used, the top five features with greater impact are FA, hole mobility, Pb, bandgap, MA. As a data comparison, we have established an electrical parameter database with 11 features, which contains 463 sets of data (Table S4, Supporting Information). The PCE prediction model is built by Table S4, Supporitng Information. The specific parameters and algorithm comparisons are shown in Table S12, Supporting Information. The top three models with the best accuracy are PCE_{RF11}, $PCE_{XGBoost11}$ and PCE_{GBDT11}. We perform a SHAP explanation on these three models. As shown in Figure S4, Supporting Information, when the PCE_{RF11} is used, the top 5 features with greater impact are FA, bandgap, Sn, Pb, H. As shown in Figure S5, Supporting Information, when using $PCE_{XGBoost11}$, the top five features that have great influence are FA, Pb, L, H and bandgap in turn. As shown in Figure S6, Supporting Information, FA, Pb, H, L, and bandgap are the top five features that have great influence when using PCE_{GBDT11}. By explaining the different data and algorithm

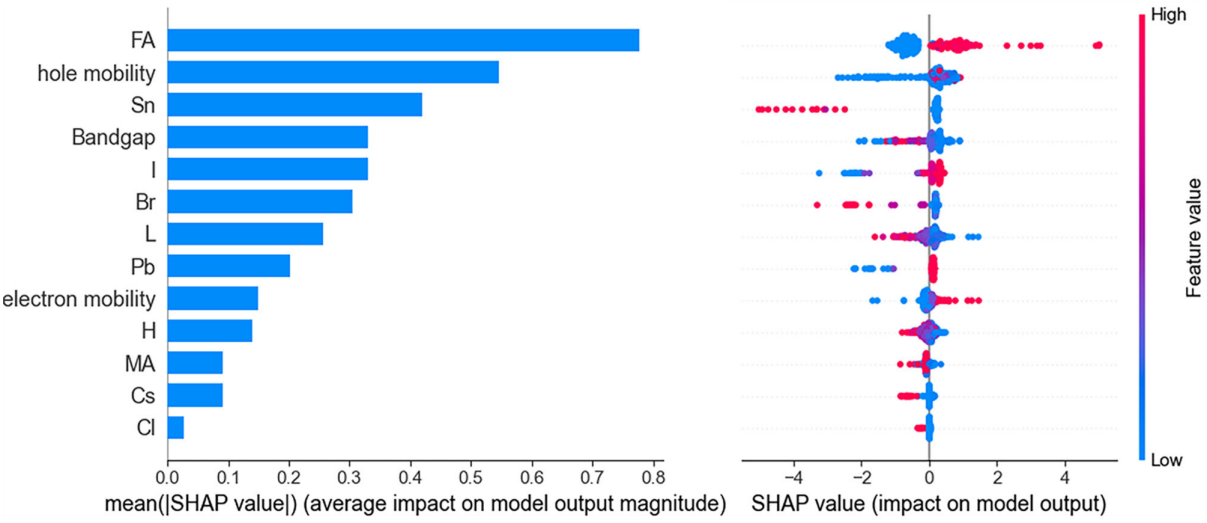

Figure 7. Feature importance ranking plot based on SHAP value (the model used is PCE_{RF}). Left: bar chart of the average absolute value of the SHAP value magnitude. Right: each point represents a sample, and each row represents a feature. The order of the features is in descending order of the average absolute value of the SHAP value. Crowded places indicate a large number of samples accumulated. The color indicates the size of the feature value (red indicates high feature value, blue indicates low feature value), and the horizontal axis represents positive and negative SHAP values.

models, we can get some similar findings, for example, FA has the most important influence, bandgap exists in the top five features, three of which represent the perovskite material itself. This indicates that more attention should be paid to the composition engineering of perovskite, especially to the proper addition of FA, to improve the PCE of PSCs. Because the above comparative analysis proves that FA has the greatest impact on the PCE of PSCs, in other words, the content of FA is essential for PSCs with high PCE.

Interpretable dimensionality reduction for PCE_{RF}. PCA is used to extract the 13 features affecting PCE, which can retain most of the information of the original influence factor matrix and greatly reduce the feature dimension. As shown in Table S14, Supporting Information, the variance contribution rates of the first principal component and the second principal component are 34% and 26%, respectively. Among them, the original feature with the largest weight of the first principal component is Sn, and the original feature with the largest weight of the second principal component is FA. As shown in **Figure 8**, the top two principal components of the explanation embedding highlight two distinct features for the difference with an average value of PCE (i.e., the difference between the PCE value of a single sample and the average PCE value of samples). In the interpretation space, the distribution of all samples is roughly "V"-shaped as shown in Figure 8. At the same time, through color rendering, it can be found that the difference with the average value of PCE increases from bottom to top along the "V"-shaped left direction, while the right direction decreases from bottom to top (Figure 8a). The difference with the average value of PCE can be regarded as the clustering of the SHAP values of FA, Sn, and 11 other features. Figure 8b shows that the SHAP value of FA (i.e., the contribution of the FA value of a single sample to the predictive value of its PCE) is almost unchanged from bottom to top along the "V"-shaped right direction, while the left direction increases from bottom to top; Figure 8c shows that the SHAP value of Sn (i.e., the contribution of the Sn value of a single sample to the predictive value of its PCE) decreases along the "V"-shaped right direction from

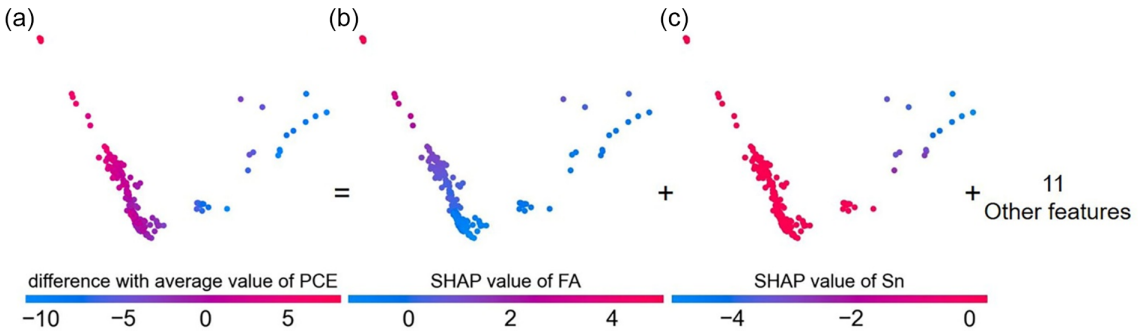

Figure 8. Interpretable dimensionality reduction (the model used is PCE_{RF}). A local explanation embedding of PCE cast onto two principal components. Local feature SHAP values can be viewed as an embedding of the samples into a space where each dimension corresponds to a feature. a–c) shows the distribution of the samples after dimensionality reduction by PCA. a) Difference with an average value of PCE. b)SHAP value of FA. c) SHAP value of Sn.

bottom to top, while the left direction has almost no change from bottom to top. The combination of Figure 8b,c can partially explain the distribution of difference with the average value of PCE, that is, the change along the left direction follows the FA feature, and the change along the right direction follows the Sn feature. Figure 8 reveals the two features that have a major impact on the difference with the average value of PCE, and also supports that FA and Sn have greater weights on the top two principal components than other features in Table S14, Supporting Information.

Sample explanations for factors affecting PCE of PSCs by SHAP theory. **Figure 9** is a sample with lower PCE in the database. The perovskite composition is $MAPb_{0.25}Sn_{0.75}I_3$. The predicted PCE value is 5.16%, and the experimental value is 3.74%. By analyzing the SHAP values of each feature, Sn, Pb, bandgap and L have the greater contribution to the loss (with a large negative SHAP), which indicates that the main reasons for the low PCE are excessive Sn content, low bandgap, and energy-level misalignment. Figure S8, Supporting Information, is a sample with higher PCE in the database. The perovskite composition is $FAPbI_3$. The predicted value of PCE is 23.53% and the experimental value is 24.82%. It can be seen that the gain contribution of FA, hole mobility and bandgap is the greatest (with a larger positive SHAP value). This indicates that the reason for higher PCE is the high content of FA, higher hole mobility, and suitable bandgap.

Explaining the influence of FA on PSCs by DFT calculations. The above SHAP explanation shows that FA plays a crucial role for high-performance PSCs, and for further physical explanation we need to use DFT calculations. As shown in Table S15, Supporting Information, we used the Vienna ab initio simulation package (VASP) for the calculation of the effective masses of holes and electrons, and exciton binding energies. Compared with $MAPbI_3$, the effective masses of holes and electrons in $MA_xFA_{1-x}PbI_3$ are significantly lower, where the exciton binding energy tends to decrease significantly as the doping ratio of FA increases (from19.50 to 4.57 meV). This indicates that the doping of FA facilitates the transport and separation of holes and electrons. As shown in Figure S7, Supporting Information, we proceeded to calculate the optical properties of $MA_xFA_{1-x}PbI_3$. With the doping of FA, the absorption spectrum has been significantly improved, and the absorption intensity also increases with the increase in the proportion of FA. This is consistent with the reference.[39] Details of DFT calculations can be found in Supporting Information.

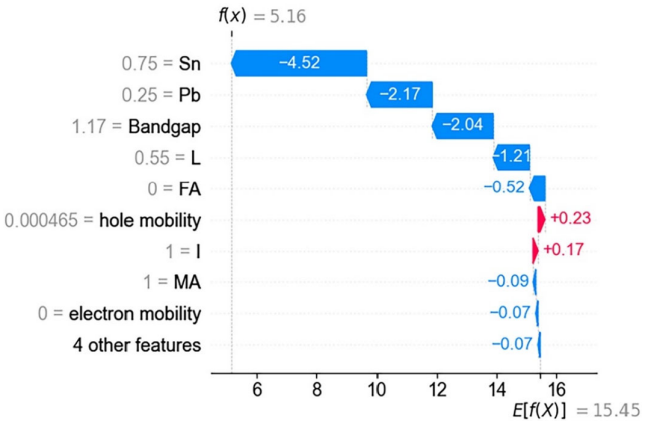

Figure 9. Evaluation process of a single sample. Perovskite composition is $MAPb_{0.25}Sn_{0.75}I_3$.

In conclusion, we built 49 ML models using 814 real experimental data from the literature and 7 ML algorithms in order to predict the performance of PSCs, and finally selected 7 optimal models. The experimental PCE and predicted PCE obtained from the RF algorithm have relatively high correlation and low RMSE, with r value of 0.86, RMSE of 1.58%. These ML models show excellent prediction performance that can assist the fabrication of PSCs, and reduce the probability of experimental errors and the computational resources required for simulation calculations. For the first time, we propose a series of new explainable strategies for PCE, including the global feature importance ranking, the influence of single feature and two feature interaction on PCE, interpretable dimensionality reduction for PCE, etc. Through these explanations, we found that the most important influence on PCE is the features of the perovskite material itself, especially the addition of FA. The prediction models can be widely applied to the prediction of the performance of ABX_3 perovskite solar cells. the SHAP explanation strategies can yield generalized and common findings which explore the physical laws and chemical meaning behind the features and further inspires the discovery of new materials.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

Thanks to the support of the National Natural Science Foundation of China(U1765105, 52007104) and the 111 Project(D20015) of China.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

The data that support the findings of this study are available in the supplementary material of this article.

Keywords

machine learning, perovskites, power conversion efficiencies, SHAP

Received: December 29, 2021
Revised: February 13, 2022
Published online: March 7, 2022

[1] National Renewable Energy Lab, *Best research-cell efficiency chart. Photovoltaic Research*, NREL, https://www.nrel.gov/pv/cell-efficiency.html (accessed: August 2021).

[2] R. Azmi, S. Y. Namb, S. Sinaga, Z. A. Akbara, C. L. Lee, S. C. Yoon, I. H. Jung, S. Y. Jang, *Nano Energy* **2017**, *44*, 191.
[3] S. A. Veldhuis, P. P. Boix, N. Yantara, M. J. Li, T. Sum, N. Mathews, S. G. Mhaisalkar, *Adv. Mater.* **2016**, *28*, 6804.
[4] Y. N. Wang, Y. Tang, J. Z. Jiang, Q. Zhang, J. Sun, Y. F. Hu, Q. H. Cui, F. Teng, Z. D. Lou, Y. B. Hou, *J. Mater. Chem. C.* **2020**, *8*, 5399.
[5] J. S. Huang, Y. B. Yuan, Y. C. Shao, Y. F. Yan, *Nat. Rev. Mater.* **2017**, *2*, 17042.
[6] J. S. Huang, Y. C. Shao, Q. F. Dong, *J. Phys. Chem. Lett.* **2015**, *6*, 3218.
[7] B. Chen, P. N. Rudd, S. Yang, Y. B. Yuan, J. S. Huang, *Chem. Soc. Rev.* **2019**, *48*, 3842.
[8] S. Wang, T. Sakurai, W. J. Wen, Y. B. Qi, *Adv. Mater. Interface* **2020**, *7*, 2000423.
[9] W. Fei, Y. H. Shan, J. H. Qiao, C. Zhong, R. Wang, Q. L. Song, L. N. Zhu, *ChemSusChem* **2017**, *10*, 3833.
[10] A. Jain, G. Hautier, S. P. Ong, K. Persson, *J. Mater. Res.* **2016**, *31*, 977.
[11] C. Kim, T. D. Huan, S. Krishnan, R. Ramprasad, *Sci. Data* **2017**, *4*, 170057.
[12] X. Zhang, J. X. Shen, M. E. Turiansky, C. G. Walle, *Nat. Mater.* **2021**, *20*, 971.
[13] A. Agrawal, A. Choudhary, *APL. Mater.* **2016**, *4*, 053208.
[14] M. I. Jordan, T. M. Mitchell, *Science* **2015**, *349*, 255.
[15] W. Sun, Y. J. Zheng, K. Yang, Q. Zhang, A. A. Shah, Z. Wu, Y. Y. Sun, L. Feng, D. Y. Chen, Z. Xiao, S. R. Lu, Y. Li, K. Sun, *Sci. Adv.* **2019**, *5*, eaay4275.
[16] V. Stanev, C. Oses, A. G. Kusne, E. Rodriguez, J. Paglione, S. Curtarolo, I. Takeuchi, *npj. Comput. Mater.* **2018**, *4*, 29.
[17] J. Schmidt, M. R. G. Marques, S. Botti, M. A. L. Marques, *npj. Comput. Mater.* **2019**, *5*, 83.
[18] N. Hartono, J. Thapa, A. Tiihonen, F. Oviedo, C. Batali, J. J. Yoo, Z. Liu, R. Li, D. F. Marrón, M. G. Bawendi, T. Buonassisi, S. J. Sun, *Nat. Commun.* **2020**, *11*, 5675.
[19] M. Saliba, *Adv. Energy. Mater.* **2019**, *9*, 1803754.
[20] Z. Z. Li, Q. C. Xu, Q. D. Sun, Z. F. Hou, W. J. Yin, *Adv. Funct. Mater.* **2019**, *29*, 1807280.
[21] S. H. Lu, Q. H. Zhou, Y. X. Ouyang, Y. L. Guo, Q. Li, J. L. Wang, *Nat. Commun.* **2018**, *9*, 3405.
[22] Ç. Odabas, R. Yıldırım, *Sol. Energy Mater. Sol. Cells* **2020**, *205*, 110284.
[23] M. L. Agiorgousis, Y. Y. Sun, D. H. Choe, D. West, S. B. Zhang, *Adv. Theor. Simul.* **2019**, *2*, 1800173.
[24] F. Oviedo, Z. Ren, S. J. Sun, C. Settens, Z. Liu, N. T. P. Hartono, S. Ramasamy, B. L. DeCost, S. I. P. Tian, G. Romano, A. G. Kusne, T. Buonassisi, *npj Comput Mater.* **2019**, *5*, 60.
[25] W. A. Saidi, W. Shadid, I. E. Castelli, *npj. Comput. Mater.* **2020**, *6*, 36.
[26] E. C. Gok, M. O. Yildirim, M. P. U. Haris, E. Eren, M. Pegu, N. H. Hemasiri, P. Huang, S. Kazim, A. U. Oksuz, S. Ahmad, *Sol. RRL.* **2021**, 2100927.
[27] S. M. Lundberg, S. I. Lee, *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4765.
[28] C. Chen, Y. X. Zuo, W. K. Ye, X. G. Li, Z. Deng, S. Y. Ong, *Adv. Energy. Mater.* **2020**, *10*, 1903242.
[29] R. Ramprasad, R. Batra, G. Pilania, A. M. Kanakkithodi, C. Kim, *npj. Comput. Mater.* **2017**, *3*, 54.
[30] T. Zhou, Z. Song, K. Sundmacher, *Engineering* **2019**, *5*, 1017.
[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, *J. Mach. Learn. Res.* **2011**, *12*, 2825.
[32] N. V. Orupattur, S. H. Mushrif, V. Prasad, *Comput. Mater. Sci.* **2020**, *174*, 109474.
[33] N. S. Altman, *Am. Stat.* **1992**, *46*, 175.
[34] C.-C. Chang, C.-J. Lin, *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 1.
[35] T. K. Ho, *IEEE. Trans. Pattern Anal. Mach. Intell.* **1998**, *20*, 832.
[36] J. H. Friedman, *Ann. Stat.* **2001**, *29*, 1189.
[37] T. Chen, C. Guestrin, in *Proc. of the 22nd ACM SIGKDD Inter. Conf. on Knowledge Discovery and Data Mining*, Association for Computing Machinery, USA **2016**, p. 785, https://doi.org/10.1145/2939672.2939785.
[38] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, A. Walsh, *Nature* **2018**, *559*, 547.
[39] W. Li, M. U. Rothmann, Y. Zhu, W. J. Chen, C. Q. Yang, Y. B. Yuan, Y. Y. Choo, X. Wen, Y. B. Cheng, U. Bach, J. Etheridge, *Nat. Energy* **2021**, *6*, 624.