Journal of Materials Chemistry A

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: J. Mater. Chem. A, 2021, 9, 25168

Machine learning-guided search for high-efficiency perovskite solar cells with doped electron transport layers†

Chenglong She, ‡[ab] Qicheng Huang,‡[a] Cong Chen,[ac] Yue Jiang,[a] Zhen Fan *[a] and Jinwei Gao *[a]

Received 23rd September 2021
Accepted 24th October 2021

DOI: 10.1039/d1ta08194b

rsc.li/materials-a

The experimental search for high-efficiency perovskite solar cells (PSCs) is an extremely challenging task due to the vast search space comprising the materials, device structures, and preparation methods. Herein, using a two-step machine learning approach and 2006 PSC experimental data points extracted from 880 articles published between 2013 to 2020, we develop some heuristics for high-efficiency PSC and power conversion efficiency (PCE) improvement induced by doping of the electron transport layer (ETL). We show that the utilizations of SnO_2 and TiO_2 ETLs, mixed-cation perovskites, dimethyl sulfoxide and dimethylformamide perovskite precursor solvents and anti-solvent treatment are the most significant factors that lead to the high PCEs of PSCs. The PCE can be further improved by ETL doping for tuning the conduction band minimum, Fermi level, and conductivity of the ETL. Moreover, we predict that a FA–MA based PSC with a Cs-doped TiO_2 ETL and a Cs–FA–MA based PSC with S-doped SnO_2 ETL exhibit PCEs of as high as 30.47% and 28.54%, respectively. This study provides insightful guidance for the development of high-efficiency PSCs.

1. Introduction

Perovskite solar cells (PSCs) are complex devices with power conversion efficiencies (PCEs) that are dependent on not only the optical and electrical properties of the perovskite layer itself, but also the electron transport layer (ETL), hole transport layer (HTL), and other layers that constitute the device. For each layer, there are also many variables, such as the deposition technique (spin coating, spin-dip, slot-die, CVD *etc.*), conditions during the deposition process (temperature, humidity, spinning rate and time, annealing time, *etc.*), and the species of the solvents and anti-solvents. All these factors co-determine the PCE of a PSC. Among them, the ETL is of great significance for the following reasons. First, an ETL with a favourable energy band that matches the perovskite layer can effectively extract electrons while preventing the reverse transfer of holes. In addition, it can reduce the density of defect states at the interface to prevent harmful interface recombination.[1] Careful engineering of the ETL is therefore beneficial for achieving a high PCE.[2] For example, Yoo *et al.*[3] achieved a record high PCE of 25.2% in a SnO_2-based PSC by tuning the chemical bath deposition (CBD) of the SnO_2 ETL.

Doping has become a popular and efficient approach by which to modulate the physical and chemical properties of an ETL so as to improve the performance of a PSC. Hereafter, we focus mainly on the doping effect in inorganic ETLs, while that in organic ETLs can be referred to in ref. 4 and 5. The role of doping in ETL includes, but is not limited to, adjusting the energy level, improving electron transport, reducing the density of defect states, and modulating film morphology.[6] To date, various elements have been used as dopants in ETLs, such as Ni^{2+},[7] Nb^{5+},[8] Sb^{5+},[9] Li^{+},[10] Mg^{2+},[11] Al^{3+},[12] Ga^{3+} (ref. 13) and Y^{3+}.[14] However, previous research on doped-ETL-based high-efficiency PSCs mostly involved empirical experimental investigations that used a trial-and-error method. This makes it a high-cost, time-consuming, and unreliable task to study all the effects of dopants in ETLs, materials of ETLs/perovskite layers/other layers, and the deposition methods on the PCEs of PSCs. Therefore, there is an urgent need to determine a general rule for device optimization to guide the experimental development of doped-ETLs-based high-efficiency PSCs.

In this work, we demonstrate a two-step machine learning (ML) approach to extract general rules underlying the high PCEs of PSCs and further predict some high-efficiency PSCs with doped ETLs. ML has emerged recently as a powerful tool that

[a]Institute for Advanced Materials, South China Academy of Advanced Optoelectronics, Guangdong Provincial Key Laboratory of Optical Information Materials and Technology, South China Normal University, Guangzhou 510006, China. E-mail: gaojinwei@m.scnu.edu.cn; fanzhen@m.scnu.edu.cn

[b]School of Information and Optoelectronic Science and Engineering, South China Normal University, Guangzhou 510006, China

[c]Department of Mechanical Engineering, The University of Hong Kong, Pokfulam Rd., Pokfulam, Hong Kong, China

† Electronic supplementary information (ESI) available: Experimental, characterization and synthesis details. See DOI: 10.1039/d1ta08194b

‡ These authors contributed equally to this work.

can be used to accelerate the discovery of new materials and devices.[15] It can be used to learn hidden knowledge from existing data, establish relationships between material/device descriptors and targeted performance, and further predict the performance of an unexplored material/device. We first conducted a comprehensive review of 880 articles on PSCs published between 2013 to 2020, established a dataset containing 2006 PSC samples, and trained ML models based on this dataset. Then, we demonstrated that PSCs with high PCEs are closely related to the use of SnO_2 and TiO_2 ETLs, the conduction band minimum (CBM) and Fermi level of ETLs, perovskite materials, perovskite precursor solvents and anti-solvent treatment. What is more, we predicted that a Cs–FA–MA based PSC with Cs-doped TiO_2 ETL (24.53% Cs doping) and a FA–MA based PSC with S-doped SnO_2 ETL (32.85% S doping) can exhibit PCEs of as high as 30.47% and 28.54%, respectively. The predicted results may be useful for guiding the future development of high-efficiency PSCs.

2. Results and discussion

2.1 Construction of the dataset

We established two datasets that were used for the first- and second-step ML, respectively. The first dataset was composed of 1820 PSC performance data points extracted from 795 articles published in the period of 2013–2020 (see Tables S1 and S2 in the ESI†). Notably, the data points in the years 2013–2018 were borrowed from the dataset established by Çağla Odabaşı *et al.*,[16] while those from 2019 to 2020 were newly collected from the literature. The PCE, as the key performance metric of a PSC, was chosen as the target attribute. Nine major factors influencing the PCE were chosen as features, which were the perovskite type, ETL and second layer of the ETL/interfacial layer (ETL-2) materials, deposition procedure and method, perovskite precursor solvent, anti-solvent, HTL materials and additive.

To visualize the first dataset, we divided the PSCs in the dataset into TiO_2-based, SnO_2-based and other ETLs-based ones according to the ETL materials. TiO_2 has a suitable band alignment,[17] while SnO_2 possesses high electron mobility, a small number of defects, high transmittance, a low charge recombination rate, and good photostability.[18,19] TiO_2 and SnO_2 have therefore become the mainstream ETL materials of PSCs. There have been some PSCs using other ETL materials, but their number is small and therefore they will not be shown together with the TiO_2- and SnO_2-based PSCs. Fig. 1a shows the evolution trend of the average efficiency of TiO_2- and SnO_2-based PSCs from 2013 to 2020. The relative size of the circular symbol indicates the number of samples reported in that year. It can be seen that the average efficiency increases rapidly before 2017, while the PCE increase slows down after 2017. SnO_2-based PSCs exhibit higher average efficiencies than TiO_2-based PSCs in all years. In addition, SnO_2 has become more widely used than TiO_2 since 2019. Fig. 1b shows the quantity distribution of TiO_2- and SnO_2-based PSCs in different efficiency intervals. The TiO_2-based PSCs are mainly distributed in a medium efficiency range of 9–18%, showing a Gaussian-like distribution. By contrast, the number of SnO_2-based PSCs increases in the higher efficiency range, suggesting that SnO_2 is beneficial to achieving a high PCE. The reason for this can be attributed to SnO_2 showing higher electron mobility and better band alignment with perovskite materials.[20] In addition, SnO_2 is low-temperature solution processable (in contrast, TiO_2 needs a quite high temperature of ~400 °C for annealing), which allows the large-scale roll-to-roll production of PSCs.[21] Therefore, SnO_2 has become the mainstream ETL material for PSCs in recent years.

Based on the first database, a ML model was built to predict the PCEs of PSCs with undoped ETLs. Note that the ETLs in the first dataset (including TiO_2, SnO_2 and other ETLs) are all undoped. To further improve the PCEs of the PSCs, we considered the ETL doping effect (mainly for TiO_2 and SnO_2) and hence constructed a second dataset containing PSCs with doped ETLs. This dataset contains 90 data points of doped-SnO_2-based PSCs collected from 36 articles and 96 data points of doped-TiO_2-based PSCs collected from 49 articles (see Tables S4–S6 in the ESI†). To study the effect of ETL doping, we took the efficiency improvement ratio (EIR) as the target attribute. An EIR induced by the ETL doping was obtained by dividing the PCE of a doped-ETL-based PSC by that of the corresponding undoped-ETL-based PSC. We set the doping element and concentration, the physical and chemical properties of the doping element such as atomic number (AN), ionic radius (IR/[pm]), ionic charge (IC), ionization energy (IE/[kJ mol^{-1}]), electron affinity (EA/[kJ mol^{-1}]), electronegativity (Pauling scales), electron numbers of the s, p, d, f orbitals and the sum of the s and p orbital radii (SP orbital/[pm]), and the optical and electrical properties of ETL after doping, like the Fermi level (eV), CBM (eV), band gap (eV) and conductivity (S cm^{-1}), as the features. These data were obtained from the website https://www.webelements.com/.

2.2 Model framework

The workflow of the proposed two-step ML approach is shown in Fig. 2. The first step of ML used the first dataset containing undoped-ETL-based PSCs, which was divided into a training set and a test set with 80% and 20% of the samples, respectively. The training set was used to train the ML models, including decision tree (DTree), extra tree (ETree), random forest (RF), adaboost (ABoost), gradient boost (GBoost), xgboost (XGBoost), multi-layer perceptron (MLP) and support vector machine (SVM) (see those codes in the ESI†). Then, the predictive capabilities of the models were assessed using the test set. Both classification and regression models were established. Using the classification models, the features most related to the high PCE were identified. Using the regression models and the genetic algorithm, which can generate high-quality solutions to optimization and search problems,[22] a series of PSCs potentially exhibiting high PCEs were predicted. The second step of ML used the second dataset containing PSCs with doped ETLs. Feature engineering was performed for the second dataset. This eliminated redundant features and improved the model performance using the most relevant knowledge.[23] The second dataset was also divided into a training set and a test set with 80% and 20% of the samples, respectively. In this step, the

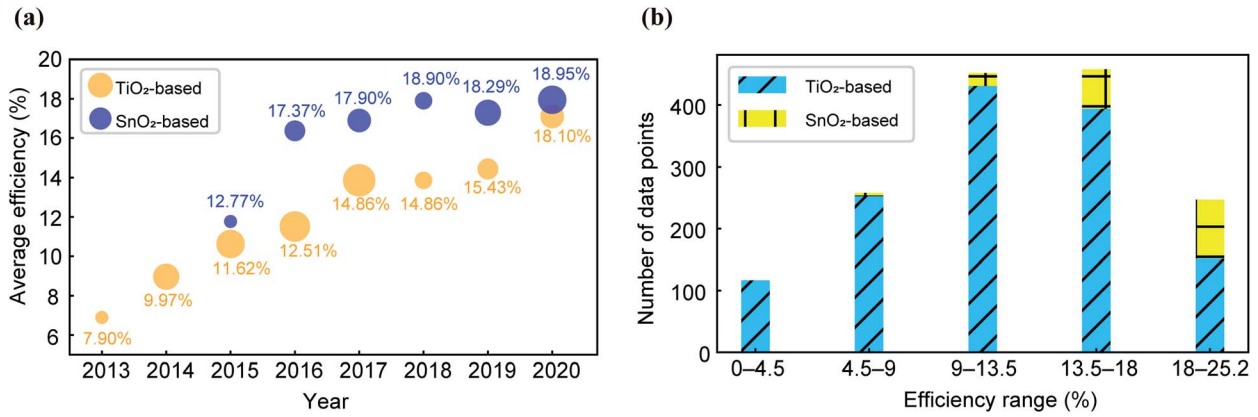

Fig. 1 Comparison of TiO_2 and SnO_2-based PSCs. (a) Average efficiencies for TiO_2- and SnO_2-based PSCs in different years. (b) Quantity distribution of TiO_2- and SnO_2-based PSCs in different efficiency ranges.

training set was used to train the RF model. Then the model was evaluated on the test set, exhibiting excellent prediction ability. Using the RF model and a genetic algorithm, the doped ETLs potentially resulting in high EIRs were predicted. Assuming that the PCEs of the undoped-ETL-based PSCs (predicted in the first step of ML) can be directly improved by the corresponding EIRs (predicted in the second step of ML) after the ETL doping, the PCEs of the doped-ETL-based PSCs were obtained.

2.3 First-step ML for undoped-ETL-based PSCs

Using the first dataset containing undoped-ETL-based PSCs, ML-based classification was performed firstly to reveal the most

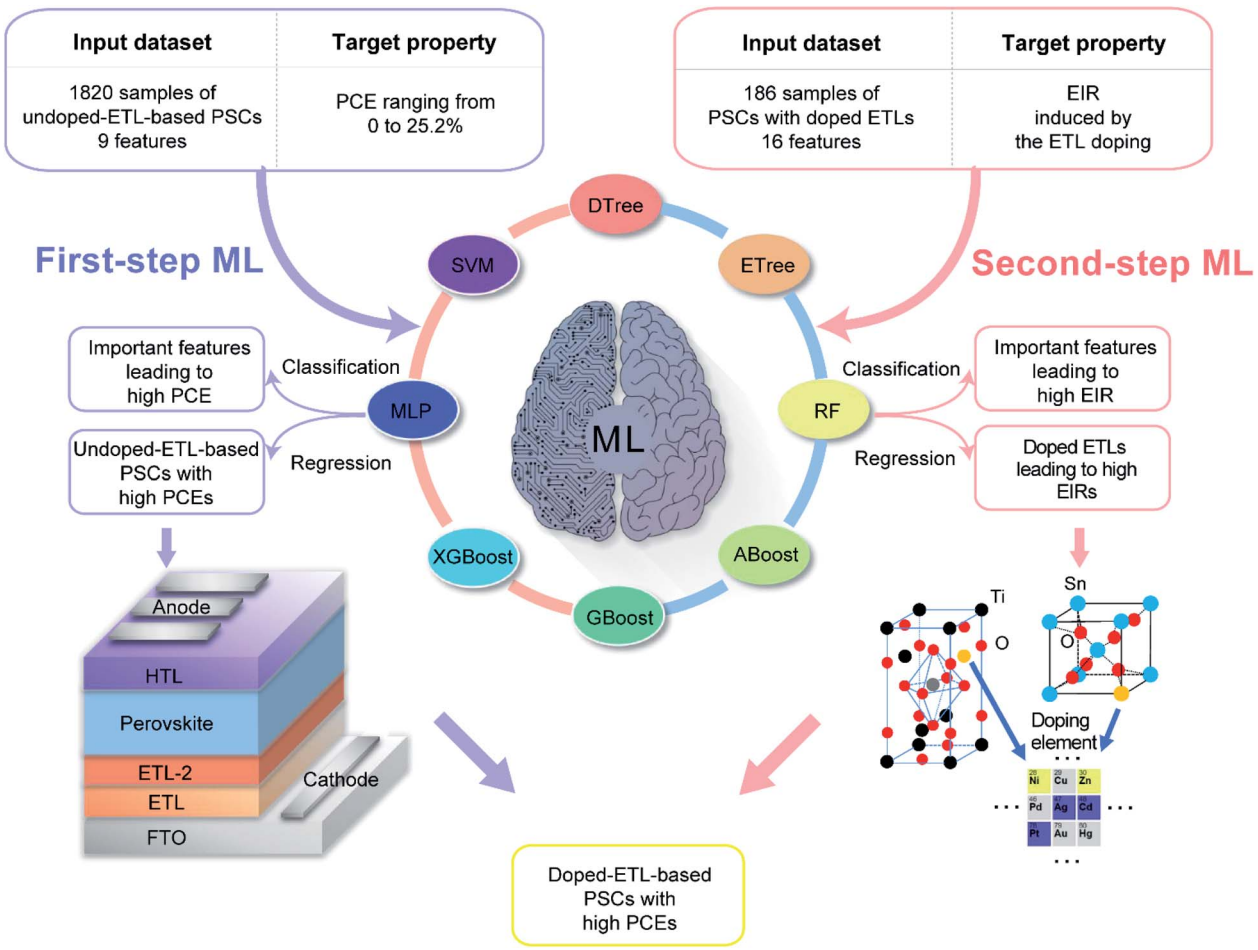

Fig. 2 Flow chart of the two-step ML in this work, which includes two steps, first-step ML (left side) and second-step ML (right side).

important features influencing the PCEs of PSCs. To avoid sample imbalance issues, all the samples were divided into three classes: class A with efficiencies lower than 9%, class B with efficiencies in the medium range of 9–18%, and class C with efficiencies above 18%. The numbers of samples of the three classes were 445, 1079 and 296 respectively, the quantitative proportions of which are shown in Fig. 3a.

Eight typical classification models were used, and their accuracies on the training and test sets are shown in Fig. 3c. GBoost and XGBoost achieve the highest accuracy (83.72%) on the training set, while the RF classification model achieves the highest accuracy (66.48%) on the test set. The accuracy on the test set is of more significance because it reflects the generalization ability of the model. In terms of this, the RF classification model performs the best and was selected for further study. Fig. 3b shows the confusion matrix for the RF classification model on the test set. 176 out of 210 samples in class B were correctly classified, giving rise to an accuracy of 83.81%. However, the classification accuracies for class A and C were much lower, at 30.77% and 60.32%, respectively. Observing the misclassified class A and C samples, we found that most of them can be classified into class B. Specifically, most of the misclassified class A samples featured a HTL additive (Li + TBP and Li + TBP + FK209), while most of the misclassified class C samples used TiO_2 as the ETL material, and the used perovskite materials were mainly single-cation perovskites (*e.g.*, $MAPbI_3$, $FAPbI_3$ and $CsPbI_3$).

Based on the RF classification model, we further investigated the feature importance to identify the features most associated with a high PCE. We focused mainly on ETL-related features because in this work we aimed to improve the PCE through ETL doping. Fig. 3d shows the importance of different materials for both ETL and ETL-2. ETL-2 represents the second layer of the ETL, or the interfacial layer. Immediately following the underline of ETL or ETL-2 is the material of the layer. "Non" implies the absence of the related layer (one of ETL or ETL-2), while "others" means that several options are available (see Table S2 in the ESI for more detail†).

TiO_2 and SnO_2 emerged as the top two ETL materials with high importance. This suggests that using TiO_2 or SnO_2 as the ETL material is likely to result in a high-performance PSC due to the merits of a wide band gap, suitable band alignment, high conductivity, high optical transmittance, and low charge recombination rate. The detailed values of these properties are presented in Table S3 in the ESI,† and can be considered as the major criteria for good ETL materials.

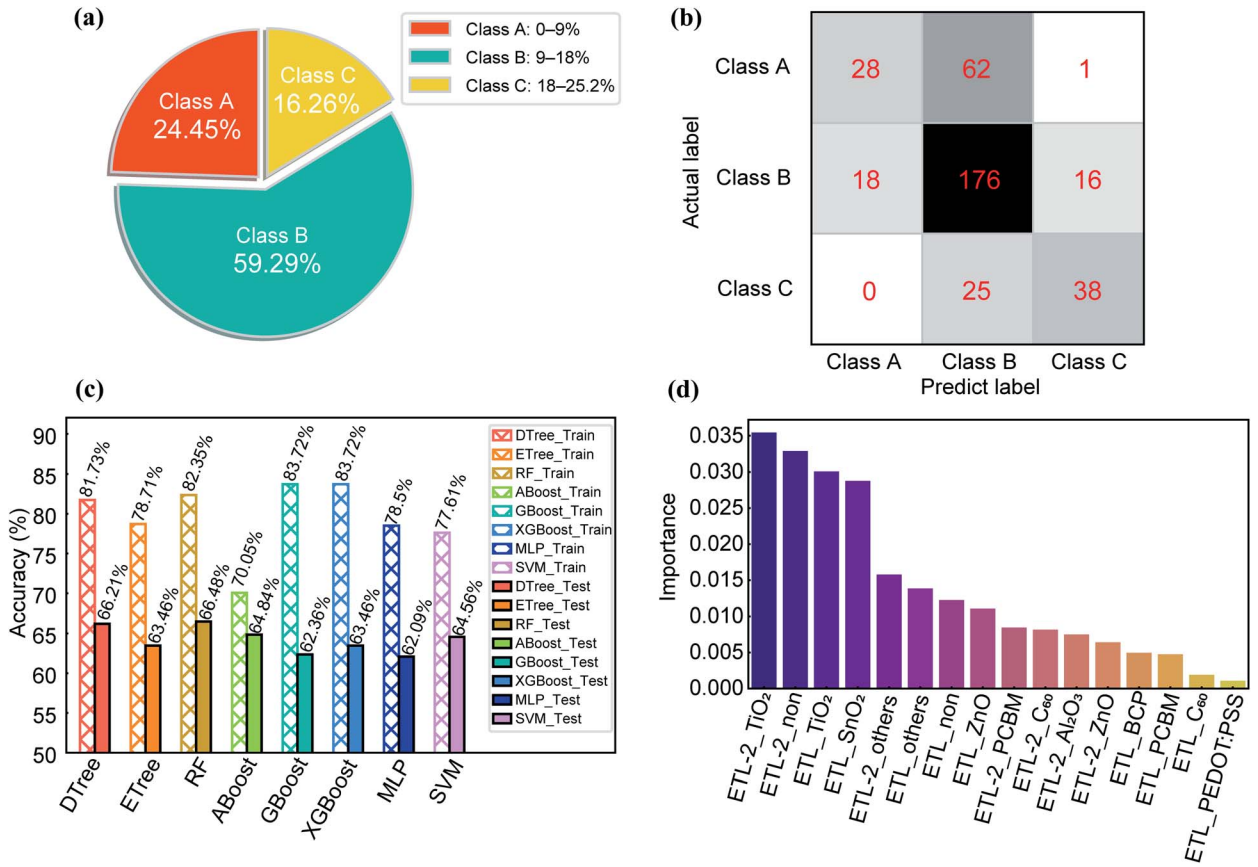

Fig. 3 Classification results in the first-step ML. (a) Number of samples in class A, B, and C (corresponding efficiency ranges: 0–9%, 9–18%, and 18–25.2%, respectively) for the first dataset. (b) Confusion matrix of the RF classification model on the test set. (c) Comparison of the classification accuracies of various models on the training and test sets. (d) Relative importance of the ETL-related features given by the RF classification model. The descriptor "ETL-2" represents the second layer of ETL/interfacial layer, and "non" implies the absence of the related layer (one of ETL or ETL-2) while "others" means that several options are available.

In terms of ETL-2 materials, TiO_2 is followed by "non", "others" and phenyl-C61-butyric acid methyl ester (PCBM). The high rank of "non" suggests that the ETL-2 may not be very necessary for a high-performance PSC. In terms of "others" (such as CdS) and PCBM, their band structures are slightly different from those of TiO_2 and SnO_2 (Table S3 in the ESI†), which can be used to tune the band alignment in a PSC. Other functions of the ETL-2 materials include filling the pinhole to reduce recombination, and hydrophilic treatment.[24–26]

In addition to the good ETL materials represented by TiO_2 and SnO_2, mixed-cation perovskites, dimethyl sulfoxide (DMSO) and dimethylformamide (DMF) perovskite precursor solvents, and anti-solvent treatment have also been found to be features whose importance rank top among all the features (results not shown). Previous studies have also shown that these features can lead to high PCEs for PSCs.[16,17,20,27,28]

Having identified the important features using the RF classification model, a RF regression model was further established to predict the PCEs of PSCs. Plots of predicted PCEs *versus* actual PCEs for the training and test sets are shown in Fig. 4a. Both root mean square errors (RMSE) and R^2 (coefficient of determination) were used as the indicators of the predictive capability of the model. The RMSE reflects the discrepancy between the value predicted by the regression model and the actual value. The lower the RMSE, the better the agreement between the predicted and actual results. On the other hand, the R^2 is the proportion of the variance of a dependent variable that can be explained using an independent variable (or variables). Here, the dependent variable is the predicted PCE and the independent variable is the actual PCE. A R^2 value is in the range of 0 to 1, and the closer it is to 1, the better the RF regression model fits the actual PCE values. As can be seen from Fig. 4a, the RMSE values of the training and test sets are 2.70 and 3.81 respectively, while the corresponding R^2 values are 0.73 and 0.43 respectively. These results are comparable to those reported by Çağla Odabaşı *et al.*[16] and Masanori Kaneko *et al.*[29] The relatively large RMSE and relatively small R^2 values of the test set suggest that the predictive capability of the RF regression model is limited. Nevertheless, it can be observed from Fig. 4a that the discrepancies between the predicted and actual PCEs are much greater in the low-efficiency region (<10%) than in the high-efficiency region (>18%). Therefore, it may still be reasonable to use the RF regression model to predict high PCEs.

Prior to using the RF regression model for PCE prediction, the feature importance was probed using this model. As shown in Fig. 4b, TiO_2 and SnO_2 were identified as the top two ETL materials that have significant impacts on the PCE. This result is in good agreement with that revealed by the above classification model (see Fig. 3d). Therefore, the RF regression model appears to be a suitable predictor for the prediction of undoped-ETL-based PSCs with high PCEs. The predicted PSCs along with the corresponding features are shown in Section 2.5.

2.4 Second-step ML for ETL doping

Having predicted undoped-ETL-based PSCs with high PCEs, we proceeded to further improve the PCEs through ETL doping. The second dataset (see Tables S4–S6 in the ESI†) was used to train a RF regression model to study the effect of the ETL doping on the PCE. As demonstrated in Section 2.3, TiO_2 and SnO_2 are the top two most important ETL materials. We therefore focused only on the doping effect of TiO_2 and SnO_2 ETLs-based PSCs. Fig. S1 in the ESI† shows all the doping elements in the TiO_2 and SnO_2 ETLs. Some elements are doped only in SnO_2 (such as S, Sb, and In) or TiO_2 (such as Ga, Ag, and Cs), while some other elements can be doped in both (such as Li, Mg, and Zn). In total, there are 90 samples of doped-SnO_2-based PSCs and 96 samples of doped-TiO_2-based PSCs. The number of samples does not far exceed the number of features (16; as described in Section 2.1). Therefore, to avoid curse of dimensionality,[30] redundant features need to be eliminated.

Feature engineering was thus performed as follows. First, the RF regression model was trained using all 16 features and

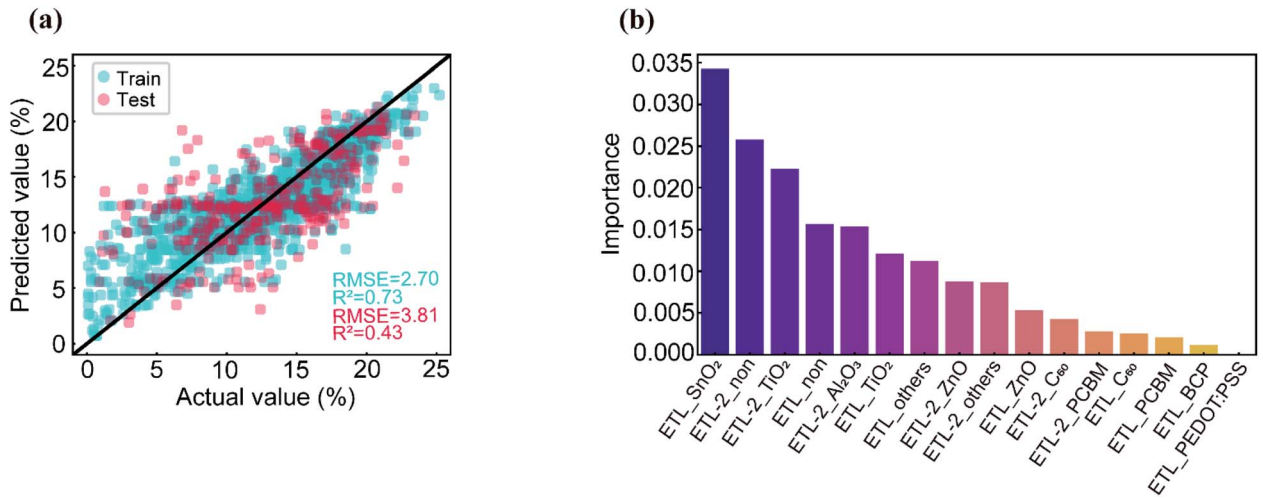

Fig. 4 Regression results of the first-step ML. (a) Actual *versus* predicted PCE computed using the RF regression model on the training and test sets. (b) Relative importance of the ETL-related features computed using the RF regression model.

a target attribute, *i.e.*, the EIR after ETL doping. The features were ranked according to feature importance. Then, the Pearson correlation coefficients (P) between any two features were calculated. If $|P|$ is larger than 80%, the two features are highly correlated and one of them can be represented by the other. The one with lower importance was removed, while the other with higher importance was kept. The only exception was the atomic number, which was always kept because it indicates the type of doping element.

The above feature engineering was performed for doped-TiO_2-based and doped-SnO_2-based PSCs separately. Fig. 5a and b show the feature importance rankings for TiO_2 and SnO_2, respectively. Among the first five features with high importance for both SnO_2 and TiO_2, there are four common features, namely, the CBM, Fermi level, band gap, and conductivity. These features are in good agreement with the widely acknowledged factors for an ETL to realize a high PCE of a PSC. During doping, especially n-type doping, the extra valence electrons from dopants can enhance the electrical conductivity of ETLs.[31] Meanwhile, the increased carrier concentration due to doping enables the upward shift of the Fermi level.[32,33] The above two features directly enhance the PCEs of PSCs. In addition, the introduction of dopants may lead to a downward shift in the CBM, and better band alignment with the absorption layer, which then improves the carrier transmission ability and inhibits carrier recombination,[10,34,35] thus enhancing the PCEs of PSCs.

Besides the above features that describe the optical and electrical properties of the ETL, the features describing the physical and chemical properties of the doping element are also worth investigating. As shown in Fig. 5a and b, the top two most important features of the doping elements are the s and p orbital radii and ionic radii (IR) for TiO_2, while those for SnO_2 are the IR and atomic number (AN). The IR is a common feature, suggesting that the IR of the dopant plays an important role in influencing the performance of the ETL and

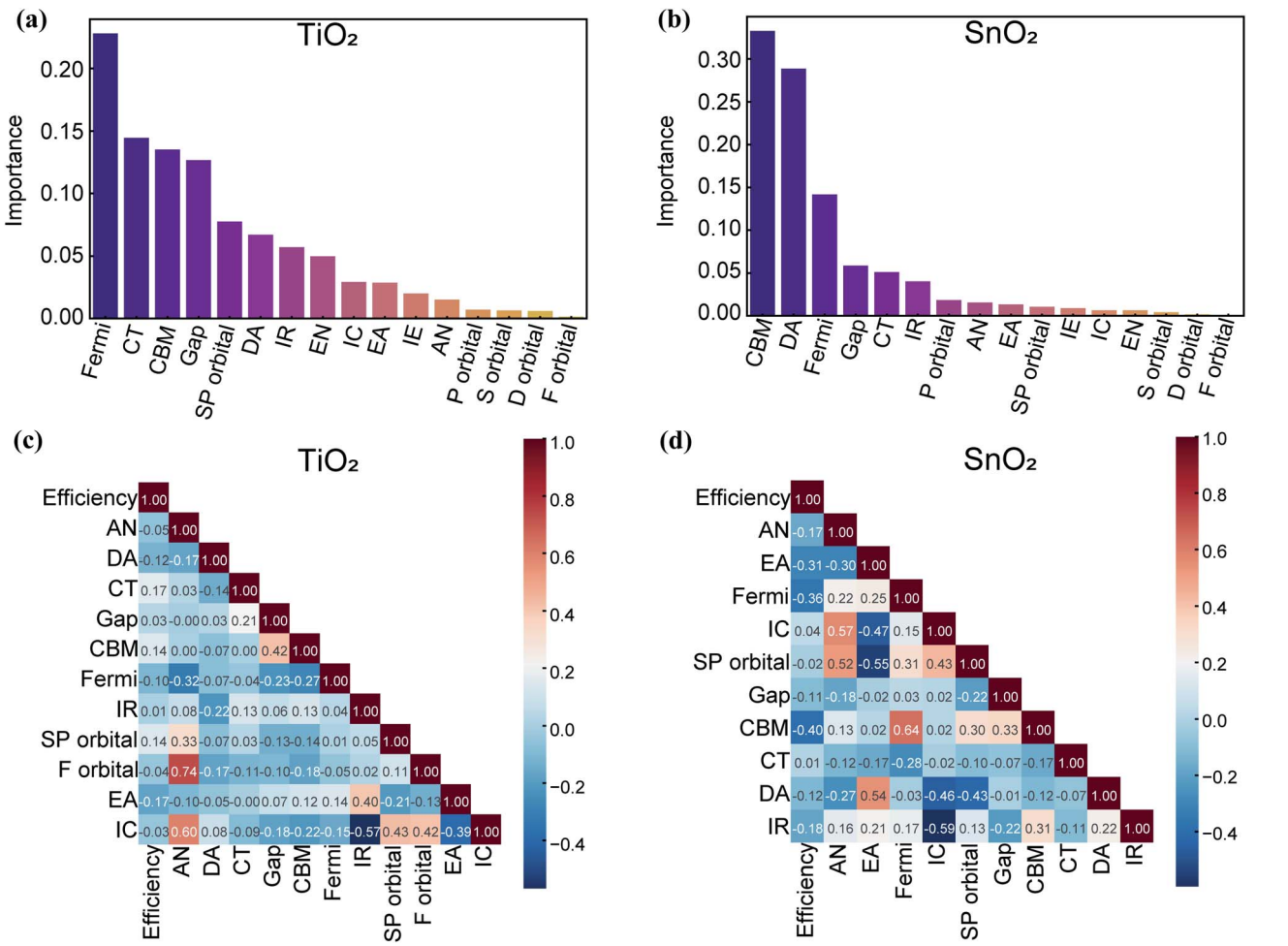

Fig. 5 Feature engineering for the second-step ML. Feature importance rankings for the doped (a) TiO_2 and (b) SnO_2 ETLs computed using the RF regression model. Heat maps of Pearson correlation coefficients between the selected features for the doped (c) TiO_2 and (d) SnO_2 ETLs. The features include atomic number (AN), ionic radius (IR/[pm]), ionic charge (IC), ionization energy (IE/[kJ mol^{-1}]), electron affinity (EA/[kJ mol^{-1}]), electronegativity (Pauling scales), electron numbers of the s, p, d, f orbitals and sum of the s and p orbital radii (SP orbital/[pm]), which represent the properties of the doping element. In addition, the features represent the optical and electrical properties of ETL after doping and the Fermi level (eV), CBM (eV), band gap (eV) and conductivity (S cm^{-1}) are also included.

consequently the PCE. It has been well investigated that the IRs of dopants are typically compatible to those of the host ions of oxides. The mismatch in the IR may increase lattice distortion, induce the formation of defect dipoles, and even generate the secondary phase.[36] This can degrade the properties of host materials, and finally reduce the efficiencies of PSCs.

After analysing the feature importance, the paired Pearson correlation coefficient matrices were calculated. The results after eliminating the redundant features are presented in Fig. 5c and d for TiO_2 and SnO_2, respectively. The Pearson correlation coefficients between two of the remaining features are mostly below 0.5, suggesting that the redundant features have been successfully removed. The number of features for TiO_2 is reduced to 11, while that for SnO_2 is reduced to 10.

After feature engineering, the RF regression model was re-trained with the retained features. Fig. 6a and b present the predicted EIR *versus* the actual EIR for doped-TiO_2- and doped-SnO_2-based PSCs, respectively. For TiO_2, the RMSE values of the training and test sets are 0.07 and 0.14 respectively, and the R^2 values are 0.84 and 0.33 respectively. For SnO_2, the RMSE values of the training and test sets are 0.05 and 0.04 respectively, and the R^2 values are 0.90 and 0.92 respectively. Such performances compare favourably with those reported previously,[29] suggesting that the RF regression model is capable of predicting the EIR induced by the TiO_2 or SnO_2 doping. Nevertheless, the model performance on the test set for TiO_2 is not as good as that for SnO_2, probably because the doping effect is less regular in the TiO_2-based PSCs than in the SnO_2-based PSCs.

2.5 Prediction of the PCEs of doped-ETL-based PSCs

Having established the two predictors for the PCE of the undoped-ETL-based PSC and the EIR induced by the ETL doping, respectively, we then predicted the PCE of the doped-ETL-based PSC. As described in Section 2.2, a genetic algorithm was employed to search the optimal solutions of the feature values that lead to high PCE or EIR.

Table 1 shows the high-performance undoped-ETL-based PSCs predicted by the model established in the first-step ML. Among the undoped-SnO_2-based PSCs, the 6th sample exhibits the highest PCE of 22.47%. This PSC features an FA–MA perovskite as the light-absorbing layer, and only one ETL layer of SnO_2; meanwhile, as the anti-solvent, diethyl ether was used to reduce the solubility of the perovskite in the precursor solution and to speed up the crystallization and nucleation. On the other hand, the 12th sample emerged as the best undoped-TiO_2-based PSC, showing a PCE of as high as 20.73%. This PSC features a Cs–FA–MA perovskite as the light-absorbing layer, and two ETL layers, which are both TiO_2. Significantly, the anti-solvent in this sample is "others", which means that several options are available (see Table S2 in the ESI for more detail†). It can be seen that the perovskite light-absorbing layers of high-efficiency samples are mostly binary and ternary mixed-cation systems, such as FA–MA and CS–FA–MA. This is consistent with previous experimental results. It was demonstrated that the PCE of the PSC using an FA–MA light-absorbing layer was higher than that using a single-cation perovskite layer.[37] It was also revealed that the introduction of Cs to FA–MA (*i.e.*, the Cs–FA–MA ternary system) improved the photostability and moisture resistance of the perovskite layer, leading to a higher PCE.[38,39] In addition, Table 1 shows that high-efficiency PSCs commonly use diethyl ether as an anti-solvent, DMF and DMSO as perovskite precursor solutions, and spiro-OMeTAD as the HTL. All these predicted materials and preparation methods are consistent with those used in recent emerging high-efficiency PSCs.

Next, the EIR induced by the ETL doping was predicted using the model established in the second step of ML. It was found that doping 24.53% Cs into TiO_2 can lead to an EIR of 1.47. The reason for this may be that the Cs doping shifts the CBM and Fermi level of TiO_2 upward,[40,41] thus reducing the energy loss of photo-excited electrons and promoting the electron extraction without causing too much interface recombination. On the other hand, doping 32.85% S into SnO_2 can result in an EIR of 1.27. Such a high EIR may originate from the interface passivation of SnO_2. An undoped SnO_2 film typically contains a high density of defect states and a large number of Sn dangling bonds on the surface. It may adsorb O_2 and H_2O in the air, capture electrons in the conduction band for recombination,

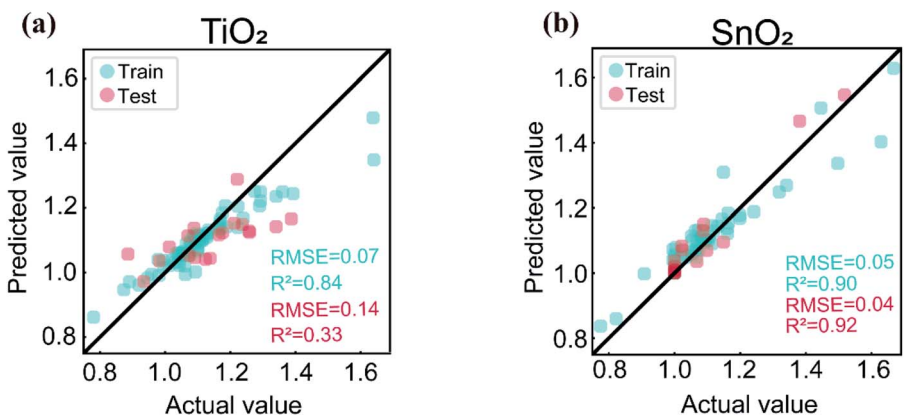

Fig. 6 Regression results of the second-step ML. Actual *versus* predicted EIR computed by the RF regression model for the doped (a) TiO_2- and (b) SnO_2-based PSCs.

Table 1 PSCs with potentially high PCEs predicted using the RF regression model combined with the genetic algorithm in the first-step ML

Sample number	ETL	ETL-2	Perovskite	Deposition procedure	Deposition method	Anti-solvent treatment	Precursor solution	HTL	HTL additive	PCE [%]
1	SnO_2	Non	Cs–FA	One-step	Spin	Others	DMF + DMSO + others	PCBM	Li + TBP	12.83
2	SnO_2	Non	FA–MA	One-step	Spin	Trifluorotoluene	DMF + DMSO + others	Spiro-OMeTAD	Li + TBP	21.14
3	SnO_2	Others	Cs–FA	One-step	Spin	Others	DMF + DMSO	Spiro-OMeTAD	Li + TBP	19.49
4	SnO_2	C60	FA–MA	One-step	Spin	Diethyl ether	DMF + DMSO	Spiro-OMeTAD	Li + TBP + FK209	21.82
5	SnO_2	Others	FA–MA	One-step	Spin	Diethyl ether	DMF + DMSO	Spiro-OMeTAD	Li + TBP + FK209	21.27
6	SnO_2	Non	FA–MA	One-step	Spin	Diethyl ether	DMF + DMSO	Spiro-OMeTAD	Li + TBP + FK209	22.47
7	TiO_2	TiO_2	FA–MA	One-step	Spin	Others	Others	PTAA	Li + TBP	17.47
8	TiO_2	TiO_2	Cs–FA–MA	One-step	Spin	Diethyl ether	DMF + DMSO + others	PCBM	Li + TBP + FK209	10.80
9	TiO_2	TiO_2	Cs–FA–MA	One-step	Spin	Diethyl ether	DMF + DMSO + others	PCBM	Li + TBP	10.20
10	TiO_2	TiO_2	Cs–FA–MA	One-step	Spin	Others	DMF + DMSO + others	Spiro-OMeTAD	Li + TBP	16.89
11	TiO_2	TiO_2	Cs–FA–MA	One-step	Spin	Diethyl ether	DMF + DMSO + others	Spiro-OMeTAD	Li + TBP	17.00
12	TiO_2	TiO_2	Cs–FA–MA	One-step	Spin	Others	DMF + DMSO	Spiro-OMeTAD	Li + TBP	20.73

and form a potential barrier that hinders electron transport.[42,43] The addition of S can fill the oxygen vacancies in the film, and the dangling Sn bond on the surface can be saturated by S ions. This in turn upshifts the CBM of SnO_2 to better match that of the perovskite light-absorbing layer. Due to the reduced potential barrier of electron transport, both the electron collection efficiency of the ETL and the photovoltaic performance of PSC are improved.[43] In addition, S atoms interact with poorly coordinated lead (Pb) ions in perovskite films to passivate the interface trap states, inhibit charge recombination, and promote the transmission of electrons across the ETL/perovskite interface.[44]

Based on the two ML steps, the PCE of the doped-ETL-based PSC can be obtained by multiplying the PCE of undoped-ETL-based PSC and the EIR induced by the ETL doping. For PSCs with TiO_2-based ETL, we predict that the 12th sample in Table 1 with a 24.53% Cs-doped TiO_2 ETL may produce an unprecedented PCE of 30.47%; for PSCs with a SnO_2-based ETL, we predict that the 6th sample in Table 1 with a 32.85% S-doped SnO_2 may produce a PCE of as high as 28.54%. Experimental verification of the predicted results is of great interest, which is currently on going.

3. Conclusions

Based on a comprehensive literature survey and a two-step ML approach, we have investigated the high-efficiency PSCs with doped ETLs. It was identified that the utilization of SnO_2 and TiO_2 ETLs, mixed-cation perovskites, DMSO and DMF perovskite precursor solvents and anti-solvent treatment are the factors that are the most relevant to achieving high PCEs in PSCs. The PCE can be further improved by ETL doping for tuning the CB minimum, the Fermi level, and conductivity of the ETL. It was further predicted that a Cs–FA–MA based PSC with Cs-doped TiO_2 ETL (24.53% Cs doping) and a FA–MA based PSC with S-doped SnO_2 ETL (32.85% S doping) could exhibit PCEs of as high as 30.47% and 28.54%, respectively. It has therefore been shown that ML can greatly accelerate the search for materials, device structures and preparation methods for high-efficiency PSCs, as well as dopants for ETLs, suggesting its usefulness in the development of functional materials and devices.

4. Experimental

4.1 Machine learning model

RF is an ensemble method combining multiple decision trees. Each tree is trained with bootstrapped samples from the training set. When growing a tree, a subset of randomly selected features are used for each node split. The final prediction can be voted by aggregating all of the tree's predictions.

GBoost iteratively constructs an ensemble of decision tree learners through gradient boosting. In each iteration, a new tree is created to fit the current residual, namely, the gradient of loss function. This new tree is then added to the ensemble to improve the final prediction.

XGBoost is essentially a GBoost, but it strives to maximize the speed. GBoost is not designed to process missing values, while XGBoost can automatically learn the processing strategy of missing values. GBoost uses CART as the base classifier, while XGBoost supports various base classifiers such as linear classifiers. When using CART, XGBoost explicitly adds a regular

term to control the complexity of the model. GBoost uses all data in each iteration, while XGBoost supports the data sampling. GBoost only uses the first derivative information of the loss function. XGBoost carries out the second-order Taylor expansion of the loss function.

ETree is basically similar to RF. Their major differences are that ETree uses all the samples to train a tree, while RF uses the bootstrapped samples. In addition, ETree randomly selects a split at each node, while RF selects the best split.

SVM is a supervised learning algorithm, which performs the classification by finding a hyper-plane (*i.e.*, the decision boundary) in the feature space that best separates the data points from two classes. For nonlinearly separable data, a kernel function is used to map the data points in the low-dimension space to a higher-dimension space.

MLP is an ANN model mimicking how the human brain processes information, which consists of many neurons arranged in different layers: input, hidden, and output layers. Neurons in a layer are connected with those in the neighbouring layers. Each neuron computes a weighted sum of its inputs and then feeds the weighted sum to an activation function to produce an intermediate output. The intermediate output then propagates to the next layer and eventually reaches the output layer. The outputs are used to calculate the errors that are backpropagated. Based on the backpropagated errors the weights are adjusted, which is the training process for MLP.

ABoost is an iterative algorithm. A classifier is added in each round until the error rate reduces to a predetermined value. Each training sample has a weight that indicates the probability that the sample is selected into the next training iteration. Initially, the weight of each sample is equal. For the k-th iteration, the samples will be selected determined by their weights to train the classifier C_k. Then, one can improve the weight of the incorrectly classified sample and reduce the weight of the correctly classified sample. The samples with updated weights are used to train the next classifier C_{k+1}. The whole training process goes on iteratively.

4.2 Genetic algorithm (GA)

GA is a metaheuristic inspired by the process of natural selection that is commonly used to generate high-quality solutions to optimize and search problems by relying on biologically inspired operators such as mutation, crossover, and selection. GA starts with a population representing the possible potential solution. The population is composed of a certain number of individuals encoded by genes. Each individual has a collection of multiple genes. In the beginning, it is necessary to realize the mapping from phenotype to genotype, namely, coding. In each generation, individuals with excellent fitness are selected to generate the population by combining crossover and mutation with the help of genetic operators of natural genetics.

Author contributions

C. L. S. and Q. C. H. constructed the datasets and model. C. L. S., Q. C. H., Z. F. and J. W. G. wrote the manuscript. All authors discussed the results and made contributions to the manuscript. Z. F. and J. W. G. directed the research.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

We acknowledge the financial support from the NSFC-Guangdong Joint Fund (No. U1801256), the NSFC (No. 51803064, U1932125, and 52172143), the Guangdong Innovative Research Team Program (No. 2013C102), and the Natural Science Foundation of Guangdong Province (No. 2020A1515010996, 2020A0505100054). We are also thankful for the support from the Guangdong Provincial Engineering Technology Research Center for Transparent Conductive Materials.

References

1 M. Stolterfoht, V. M. Le Corre, M. Feuerstein, P. Caprioglio, L. J. A. Koster and D. Neher, *ACS Energy Lett.*, 2019, **4**, 2887–2892.
2 Q. Jiang, L. Wang, C. Yan, C. Liu, Z. Guo and N. Wang, *Eng. Sci.*, 2018, **1**, 64–68.
3 J. J. Yoo, G. Seo, M. R. Chua, T. G. Park, Y. Lu, F. Rotermund, Y. K. Kim, C. S. Moon, N. J. Jeon, J. P. Correa-Baena, V. Bulović, S. S. Shin, M. G. Bawendi and J. Seo, *Nature*, 2021, **590**, 587–593.
4 A. A. Said, J. Xie and Q. Zhang, *Small*, 2019, **15**, 1–23.
5 P. Y. Gu, N. Wang, C. Wang, Y. Zhou, G. Long, M. Tian, W. Chen, X. W. Sun, M. G. Kanatzidis and Q. Zhang, *J. Mater. Chem. A*, 2017, **5**, 7339–7344.
6 Z. K. Wang and L. S. Liao, *Adv. Opt. Mater.*, 2018, **6**, 1–13.
7 Y. Guo, T. Liu, N. Wang, Q. Luo, H. Lin, J. Li, Q. Jiang, L. Wu and Z. Guo, *Nano Energy*, 2017, **38**, 193–200.
8 X. Ren, D. Yang, Z. Yang, J. Feng, X. Zhu, J. Niu, Y. Liu, W. Zhao and S. F. Liu, *ACS Appl. Mater. Interfaces*, 2017, **9**, 2421–2429.
9 Y. Bai, Y. Fang, Y. Deng, Q. Wang, J. Zhao, X. Zheng, Y. Zhang and J. Huang, *ChemSusChem*, 2016, **9**, 2686–2691.
10 M. Park, J. Y. Kim, H. J. Son, C. H. Lee, S. S. Jang and M. J. Ko, *Nano Energy*, 2016, **26**, 208–215.
11 L. Xiong, M. Qin, C. Chen, J. Wen, G. Yang, Y. Guo, J. Ma, Q. Zhang, P. Qin, S. Li and G. Fang, *Adv. Funct. Mater.*, 2018, **28**, 1–10.
12 H. Chen, D. Liu, Y. Wang, C. Wang, T. Zhang, P. Zhang, H. Sarvari, Z. Chen and S. Li, *Nanoscale Res. Lett.*, 2017, **12**, 2–7.
13 B. Roose, C. M. Johansen, K. Dupraz, T. Jaouen, P. Aebi, U. Steiner and A. Abate, *J. Mater. Chem. A*, 2018, **6**, 1850–1857.
14 J. Song, W. Zhang, D. Wang, K. Deng, J. Wu and Z. Lan, *Sol. Energy*, 2019, **185**, 508–515.
15 M. Umehara, H. S. Stein, D. Guevarra, P. F. Newhouse, D. A. Boyd and J. M. Gregoire, *npj Comput. Mater.*, 2019, **5**, 34.

16 Ç. Odabaşı Özer and R. Yıldırım, *Nano Energy*, 2019, **56**, 770–791.
17 M. M. Tavakoli, P. Yadav, R. Tavakoli and J. Kong, *Adv. Energy Mater.*, 2018, **8**, 1–9.
18 Y. Chen, Q. Meng, L. Zhang, C. Han, H. Gao, Y. Zhang and H. Yan, *J. Energy Chem.*, 2019, **35**, 144–167.
19 N. Zhou, Q. Cheng, L. Li and H. Zhou, *J. Phys. D: Appl. Phys.*, 2018, **51**, 394001.
20 Q. Jiang, X. Zhang and J. You, *Small*, 2018, **14**, 1–14.
21 Q. Jiang, L. Zhang, H. Wang, X. Yang, J. Meng, H. Liu, Z. Yin, J. Wu, X. Zhang and J. You, *Nat. Energy*, 2017, **2**, 1–7.
22 W. Zhong, J. Liu, M. Xue and L. Jiao, *IEEE Trans. Syst. Man Cybern. B Cybern.*, 2004, **34**, 1128–1141.
23 S. Lu, Q. Zhou, Y. Ouyang, Y. Guo, Q. Li and J. Wang, *Nat. Commun.*, 2018, **9**, 1–8.
24 T. Leijtens, G. E. Eperon, S. Pathak, A. Abate, M. M. Lee and H. J. Snaith, *Nat. Commun.*, 2013, **4**, 1–8.
25 M. Hou, H. Zhang, Z. Wang, Y. Xia, Y. Chen and W. Huang, *ACS Appl. Mater. Interfaces*, 2018, **10**, 30607–30613.
26 Y. Li, Y. Zhao, Q. Chen, Y. Yang, Y. Liu, Z. Hong, Z. Liu, Y. T. Hsieh, L. Meng, Y. Li and Y. Yang, *J. Am. Chem. Soc.*, 2015, **137**, 15540–15547.
27 Y. H. Seo, E. C. Kim, S. P. Cho, S. S. Kim and S. I. Na, *Appl. Mater. Today*, 2017, **9**, 598–604.
28 S. Paek, P. Schouwink, E. N. Athanasopoulou, K. T. Cho, G. Grancini, Y. Lee, Y. Zhang, F. Stellacci, M. K. Nazeeruddin and P. Gao, *Chem. Mater.*, 2017, **29**, 3490–3498.
29 M. Kaneko, M. Fujii, T. Hisatomi, K. Yamashita and K. Domen, *J. Energy Chem.*, 2019, **36**, 7–14.
30 N. M. Nasrabadi, *J. Electron. Imag.*, 2007, **16**, 049901.
31 J. Bahadur, A. H. Ghahremani, B. Martin, T. Druffel, M. K. Sunkara and K. Pal, *Org. Electron.*, 2019, **67**, 159–167.
32 Y. Xiang, Z. Ma, J. Zhuang, H. Lu, C. Jia, J. Luo, H. Li and X. Cheng, *J. Phys. Chem. C*, 2017, **121**, 20150–20157.
33 Z. Xu, J. Wu, T. Wu, Q. Bao, X. He, Z. Lan, J. Lin, M. Huang, Y. Huang and L. Fan, *Energy Technol.*, 2017, **5**, 1820–1826.
34 Z. Cao, C. Li, X. Deng, S. Wang, Y. Yuan, Y. Chen, Z. Wang, Y. Liu, L. Ding and F. Hao, *J. Mater. Chem. A*, 2020, **8**, 19768–19787.
35 J. H. Heo, M. S. You, M. H. Chang, W. Yin, T. K. Ahn, S. J. Lee, S. J. Sung, D. H. Kim and S. H. Im, *Nano Energy*, 2015, **15**, 530–539.
36 N. J. Jeon, J. H. Noh, W. S. Yang, Y. C. Kim, S. Ryu, J. Seo and S. Il Seok, *Nature*, 2015, **517**, 476–480.
37 N. Pellet, P. Gao, G. Gregori, T. Y. Yang, M. K. Nazeeruddin, J. Maier and M. Grätzel, *Angew. Chem., Int. Ed.*, 2014, **53**, 3151–3157.
38 H. Choi, J. Jeong, H. B. Kim, S. Kim, B. Walker, G. H. Kim and J. Y. Kim, *Nano Energy*, 2014, **7**, 80–85.
39 J. W. Lee, D. H. Kim, H. S. Kim, S. W. Seo, S. M. Cho and N. G. Park, *Adv. Energy Mater.*, 2015, **5**, 1501310.
40 W. Wang, H. Zheng, Y. Liu, J. Sun and L. Gao, *J. Nanosci. Nanotechnol.*, 2016, **16**, 12768–12772.
41 J. Liu, L. Zhu, S. Xiang, H. Wang, H. Liu, W. Li and H. Chen, *ACS Sustain. Chem. Eng.*, 2019, **7**, 16927–16932.
42 C. Chen, K. Byrappa, C. S. Oakes, W. Sushanek, R. E. Riman, M. Senna, K. Brown, K. S. Tenhuisen and V. F. Janas, *MRS Proceedings*, 2001, **662**, 2–7.
43 Y. Ai, W. Liu, C. Shou, J. Yan, N. Li, Z. Yang, W. Song, B. Yan, J. Sheng and J. Ye, *Sol. Energy*, 2019, **194**, 541–547.
44 X. Zhao, S. Liu, H. Zhang, S. Y. Chang, W. Huang, B. Zhu, Y. Shen, C. Shen, D. Wang, Y. Yang and M. Wang, *Adv. Funct. Mater.*, 2019, **29**, 1–8.