Supporting Information

Machine Learning Assisted Screening of the Effective Passivation Material for Perovskite Solar Cells with P-I-N Type

Di Huang, [a] Chaorong Guo, [a] Zhennan Li, [a] Haixin Zhou, [a] Xiaojie Zhao, [a] Zhimin Feng, [a] Rui Zhang, [a] Menglong Liu, [a] Jiaojiao Liang, [a,b] * Ling Zhao, [c] * Juan Meng[d] *

a, Hunan University of Technology, Zhuzhou 412008, China
b, Qinghai Provincial Key Laboratory of Nanomaterials and Nanotechnology, Qinghai Minzu University, Qinghai 810007, P.R. China
c, Shandong Provinical Key Laboratory of Optical Communication Science and Technology, School of Physical Science and Information Technology, Liaocheng University, Liaocheng 252059, China
d, Department of physics, Guangxi Minzu University, Nanning,530006, China
**Corresponding author: liangjiaojiao@hut.edu.cn, zhaoling9966@163.com, juanmeng420@gmail.com*

Data collection and prepossessing:

We collected a total of 95 sets of data based on the passivation material for passivating the perovskite/ETL interface of perovskite solar cells with p-i-n type. Each group of data includes photovoltaic parameters of the device (PCE, V_{oc} and J_{sc}), rA, AX, PX, and Molecular SMILES code as shown in **Table S1**. The performance of the device is relatively concentrated, therefore the prediction accuracy in these segments is high.

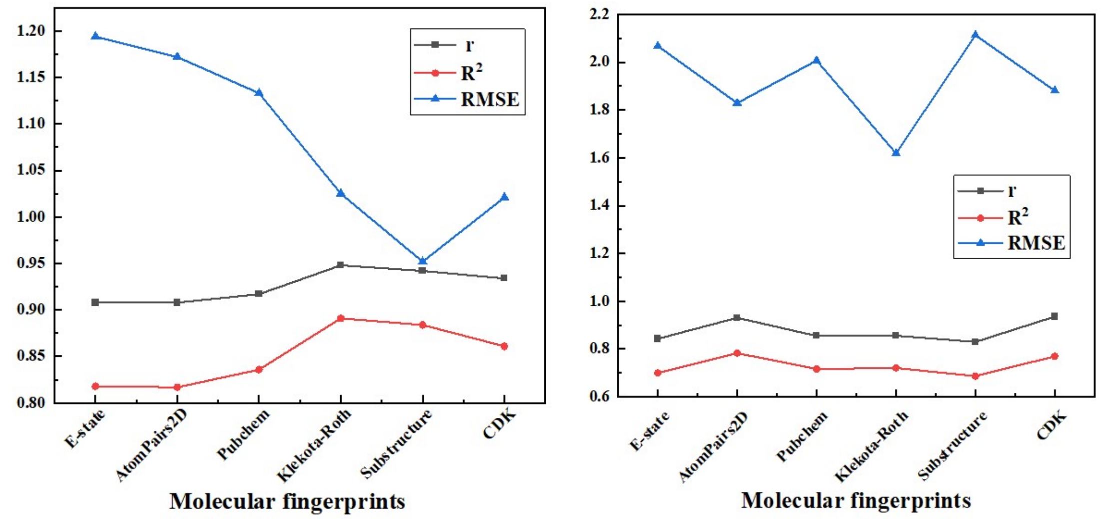
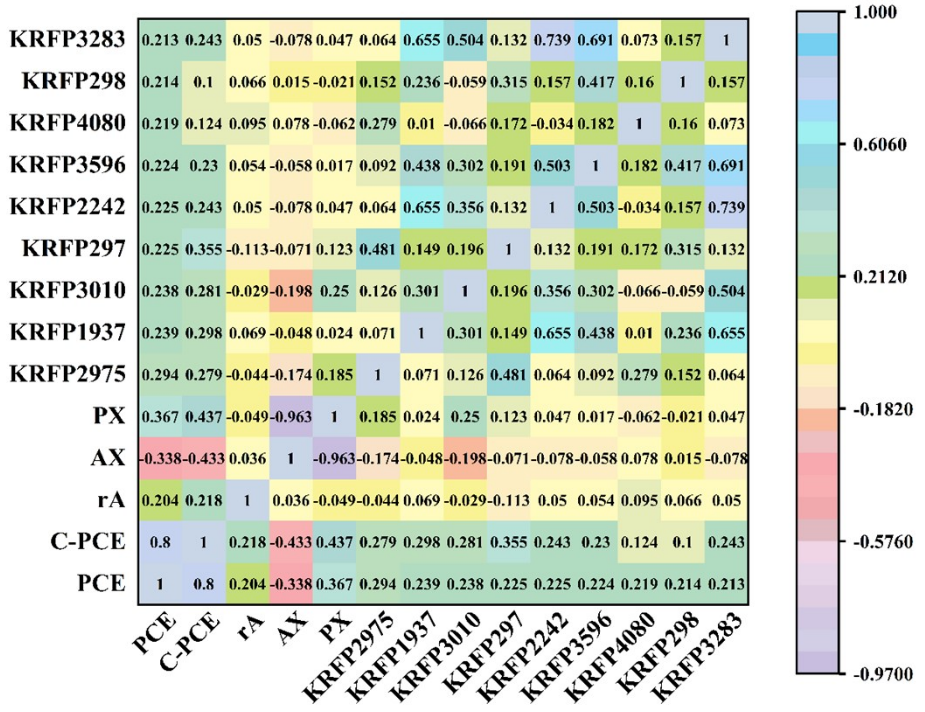


Fig. S1 Evaluation index for six types of the molecular fingerprints with RF model



	PCE	C-PCE	rA	AX	PX	KRFP2975	KRFP1937	KRFP3010	KRFP297	KRFP2242	KRFP3596	KRFP4080	KRFP298	KRFP3283
KRFP3283	0.213	0.243	0.05	-0.078	0.047	0.064	0.655	0.504	0.132	0.739	0.691	0.073	0.157	1
KRFP298	0.214	0.1	0.066	0.015	-0.021	0.152	0.236	-0.059	0.315	0.157	0.417	0.16	1	0.157
KRFP4080	0.219	0.124	0.095	0.078	-0.062	0.279	0.01	-0.066	0.172	-0.034	0.182	1	0.16	0.073
KRFP3596	0.224	0.23	0.054	-0.058	0.017	0.092	0.438	0.302	0.191	0.503	1	0.182	0.417	0.691
KRFP2242	0.225	0.243	0.05	-0.078	0.047	0.064	0.655	0.356	0.132	1	0.503	-0.034	0.157	0.739
KRFP297	0.225	0.355	-0.113	-0.071	0.123	0.481	0.149	0.196	1	0.132	0.191	0.172	0.315	0.132
KRFP3010	0.238	0.281	-0.029	-0.198	0.25	0.126	0.301	1	0.196	0.356	0.302	-0.066	-0.059	0.504
KRFP1937	0.239	0.298	0.069	-0.048	0.024	0.071	1	0.301	0.149	0.655	0.438	0.01	0.236	0.655
KRFP2975	0.294	0.279	-0.044	-0.174	0.185	1	0.071	0.126	0.481	0.064	0.092	0.279	0.152	0.064
PX	0.367	0.437	-0.049	-0.963	1	0.185	0.024	0.25	0.123	0.047	0.017	-0.062	-0.021	0.047
AX	-0.338	-0.433	0.036	1	-0.963	-0.174	-0.048	-0.198	-0.071	-0.078	-0.058	0.078	0.015	-0.078
rA	0.204	0.218	1	0.036	-0.049	-0.044	0.069	-0.029	-0.113	0.05	0.054	0.095	0.066	0.05
C-PCE	0.8	1	0.218	-0.433	0.437	0.279	0.298	0.281	0.355	0.243	0.23	0.124	0.1	0.243
PCE	1	0.8	0.204	-0.338	0.367	0.294	0.239	0.238	0.225	0.225	0.224	0.219	0.214	0.213

Fig. S2 Correlation matrix between the 13 input features and the PCE. The values are the Pearson correlation coefficient, and the sign (±) of the value means positive (+) or negative (-) correlation.

Model performance evaluation:

Using the coefficient of determination (R^2), the root means square error (RMSE) and Pearson’s coefficient (r) judge the pros and cons of the algorithm. The calculation formulas are shown in (1) (2) and (3).

$$R^2 = \frac{\left[\sum_{i=1}^{n}\left(x_i - \hat{x}\right)\left(x_i^{'} - \hat{x}^{'}\right)\right]}{\sum_{i=1}^{n}\left(x_i - \hat{x}\right)^2 \cdot \sum_{i=1}^{n}\left(x_i^{'} - \hat{x}^{'}\right)^2} \qquad (1)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i^{'} - x_i\right)^2}{n}} \qquad (2)$$

$$r = \frac{\sum_{i=1}^{n}\left(x_i - \hat{x}\right)\left(x_i^{'} - \hat{x}^{'}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \hat{x}\right)^2}\sqrt{\sum_{i=1}^{n}\left(x_i^{'} - \hat{x}^{'}\right)^2}} \qquad (3)$$

where n is the total number of data; $x_i$and $x_i^{'}$ represent the original and predicted values, respectively; $\hat{x}$ and $\hat{x}^{'}$ stand for the average of the original and predicted values, respectively.More importantly r, R2, and RMSE in each algorithm is measured 5 times and the average is used to to increase the reliability of algorithm.

Machine learning settings:

The algorithm network is all completed by python. After reading the data with pandas, the data set is divided into training set and test set, normalized and standardized. According to the characteristics of the data, the Scikit Learn class is called to initially obtain random forest (RF), K-nearest neighbors (KNN), support vector machine (SVM), extreme gradient boosting (XGBoost), and gradient boosting decision tree (GBDT) to establish the algorithm network model. We divide the data set based on 7:3, which means the training set with 70% of the data is applied for training to obtain the network model parameters of the five algorithms, and then we use the cross-validation method to evaluate the performance for the network model via the remaining 30% of the data.

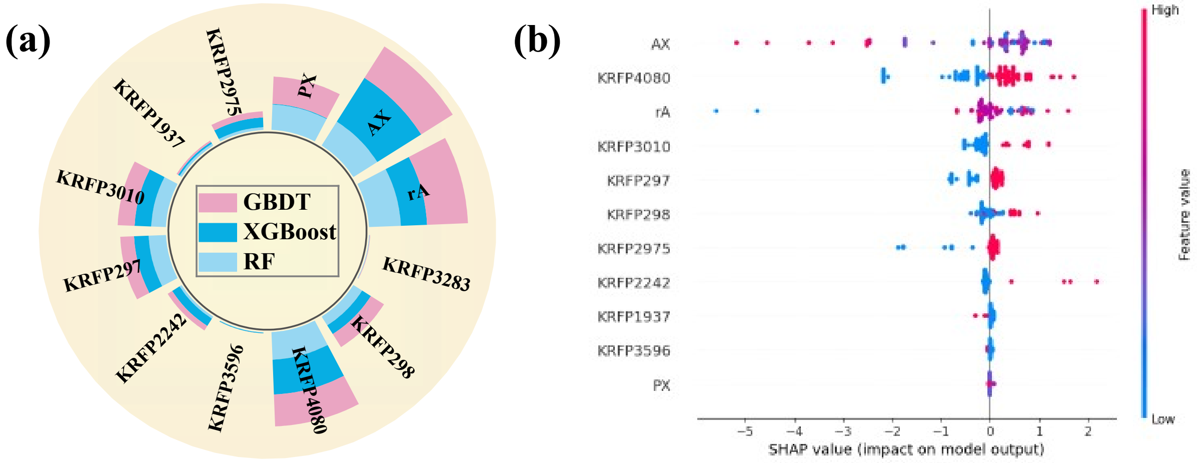


Fig. S3 (a)Polar plots of the feature importance derived from tree-based ML algorithms (GBDT, XGBoost, and RF) and (b)Input feature importance analysis based on XGBoost model.

Device fabrication:

For the device with ITIC treatment

The indium tin oxide (ITO) substrates with a nominal sheet resistance of 15 Ω/square were cleaned with detergent and ultrasonicated in special glass lotion, deionized water, and ethanol. The cleaned ITO substrates were dried by nitrogen gas and rested for 15 minutes in the petri dish before UV-ozone treatment. The PEDOT:PSS (Clevios Al4083) was filtered through polytetrafluoroethylene (PTFE) filters (0.45μm) prior to the layers fabrication. The PEDOT:PSS was spin-coated onto ITO glass and then thermally annealed at 150 °C for 15 min in ambient atmosphere. The substrates with PEDOT:PSS were transferred into a N_2-filled glove box, where 1.8M $CH_3NH_3PbI_3$ solution(the molar ratio of PbI_2 and CH_3NH_3I is 1:1 in DMF with 10 wt.% $PbCl_2$, and stirred overnight at 65 °C.) was applied onto the ITO/PEDOT:PSS substrates by one spin-coating step at 7500 rpm for 25s. After about 3.5s of the spin-coating, 400 μL chlorobenzene (CB) was quickly dripped onto the rotating substrate. After drying the substrate at 100 °C for 5 min, phenyl-C61-butyric acid methyl ester ($PC_{61}BM$) solution (30 mg/mL^{-1} in CB) was deposited onto the perovskite films. Finally, 5 nm of BCP (Luminescence Technology Corp.) and 100 nm of silver (Ag) were thermally evaporated at ~10^{-7} torr to form the contact electrodes. The active area of perovskite devices is about 1.8 mm^2.

For the device with ITIC-M treatment

All the perovskite solar cells with ITIC-M treatment were fabricated onto indium tin oxide (ITO)-coated glass substrates. The ITO substrates with sheet resistance of 15 Ω/□ were consecutively cleaned with glass lotion, de-ionized water, and alcohol. The cleaned ITO glass substrates were dried by nitrogen gas and then were treated by UV-Ozone for 15 min to further clean the substrates and improve the work function. We use modified PEDOT:PSS (m-PEDOT:PSS) as HTL, the m-PEDOT:PSS was prepared by combining 1 ml filtered PEDOT:PSS , 5 ml de-ionized water and 60 mg sodium polystyrenesulfonate (PSS-Na). The m-PEDOT:PSS was deposited onto ITO substrates at 5000 rounds per minute (rpm) for 20s. Then the substrates were dried in air at 150 ℃ for 10 min. After that, the substrates were transferred into N2 glove box. To prepare perovskite films, the MAPbIxCl3-x precursor solution(809.25mgof PbI_2, 20.75mg of $PbCl_2$， 300mg of MAI in 1 ml DMF.) was deposited onto m-PEDOT:PSS/ITO substrates at 7000 rpm for 25 s. And the perovskite coated substrates were thermally annealed at 100 ^{o} C for 10 min. The PCBM layer was deposited onto the perovskite layer at 1000 rpm for 40 s. Following a layer of BCP with a thickness about 8 nm and a silver (Ag) cathode layer of about 100 nm was deposited under 4×10^{-4} Pa vacuum conditions. The device area is defined by be the overlap of the Ag and ITO electrodes, which is 4 mm^2.

Characterization:

The current-voltage characteristics of the devices were measured using a Keithley 2400 source meter. The devices were illuminated under 1 sun, AM1.5G from Abet Technologies using a calibrated silicon diode. Steady-state photoluminescence (PL) was measured using a fluorescence spectrometer (Epsilon 3XLE) and Time-resolved photoluminescence spectra (TR-PL) were carried out using a time-correlated single photon counting measurement system with 470 nm excitation wavelength and 770 nm probing wavelength. XRD patterns of the perovskite films were obtained by X-ray diffractometer (XRD) (UItima IV X-ray diffractometer). AFM measurements were conducted with a Bruker Dimension Fastscan model in tapping mode with reflective probes resonating at 150 kHz frequency. X-ray photoelectron spectroscopy (XPS) measured by Thermo Scientific K-Alpha+ (monochromatic Al Ka, vacuum below 2×10^{-7}mbar, Beam spot 30-400um is continuously adjustable with a step size of 5um, High performance data acquisition at low power (72 W)). And the surface potential images were performed by MFP-3D Infinity of Asylum Research. The perovskite layer was deposited on Si substrate. There was no buffer layer between Si substrate and perovskite layer. The scanning area and rate were 2.0 μm×2.0 μm and 1 HZ, respectively. The lift height for KPFM measurements was 5 nm for all samples. The measurement was carried out under dark condition.

DFT calculation settings and analysis:

The structural optimization and electronic structure calculations were carried out by Cambridge Serial Total Energy Package (CASTEP) in Materials studio. The generalized gradient approximation (GGA) of the Perdew-Burke-Ernzerhof (PBE) functional was employed. Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm was used for the structural optimization of the model with the following optimization parameters: the calculation was expanded by using the ultrasoft pseudopotential with a cutoff energy of 435 eV, and the total energy was converged to 2×10^{-5} eV.

The structural optimization was optimized until the force tolerance on each atom was smaller than 0.05eVÅ^{-1}, the stress tolerance was smaller than 0.1 GPa, and the displacement tolerance was smaller than 0.002 Å. The Monkhorst-Pack grids with the actual spacing of 0.041Å^{-1} and SCF tolerance of 2×e^{-6}eV/atom was used in all DFT simulations.

$CH_3NH_3PbI_3$ possesses a cubic structure, with the space group Pm-3m at room temperature. A 5×4×1 supper cell and a 15 Å vacuum slab were employed to investigate the adsorption of different additives. The additive was placed on the supper cell surface to optimize to convergence. And the solubility of Pbl$_2$ is generally lower compared to MAI. Pbl$_2$ with slightly beyond the stoichiometric ratio is used to address this issue.However, this also results in the exposure of Pbl$_2$ surfaces in $MAPbI_3$ perovskite. Therefore, in order to approximate the actual surface state as closely as possible, we established a calcium iron perovskite model based on the Pbl$_2$ surface for studying the interaction between this surface and ITIC in DFT calculations.

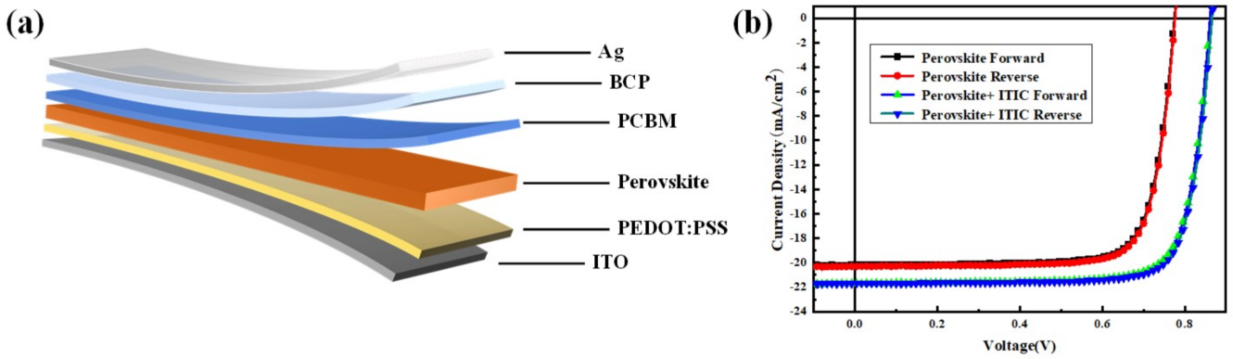


Fig. S4 (a) Structure diagram of PSCs, (b) J-V curves of the devices based on the pristine and treated perovskite films.

Table S1 the device performance of perovskite solar cells with different ITIC concentrations.

Anti-solvent	J_{sc}(mA/cm^{-2})	V_{oc}(V)	FF(%)	PCE(%)
CB	18.67	0.82	0.80	12.24
CB+ITIC(1mg/ml)	19.03	0.90	0.81	13.87
CB+ITIC(2mg/ml)	19.87	0.92	0.80	14.62
CB+ITIC(4mg/ml)	17.57	0.92	0.82	13.25

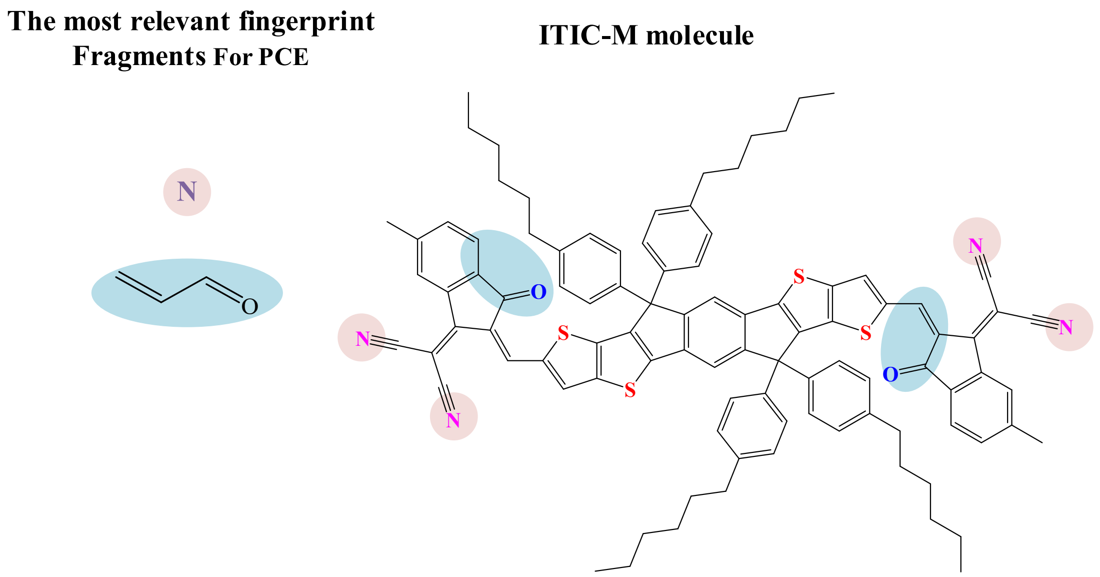


Fig. S5 ITIC-M molecule structure with the effective screened fingerprint fragments.

Table S2 the device performance of perovskite solar cells with different ITIC-M concentrations.

Anti-solvent	J_{sc}(mA/cm^{-2})	V_{oc}(V)	FF(%)	PCE(%)
CB	19.51	0.95	0.79	14.61
CB+ITIC-M(1mg/ml)	20.44	1.00	0.81	16.59
CB+ITIC-M(2mg/ml)	20.71	1.02	0.82	17.33
CB+ITIC-M(4mg/ml)	20.30	0.96	0.81	15.88

Moreover, planar structured perovskite solar cells can be treated as a single junction diode, it’s I-V characteristics can be described by:

$$J = J_{SC} - J_0[\exp\left(\frac{q\left(V + JR_s\right)}{mK_BT}\right) - 1] \qquad (4)$$

where J is the current density flow through the external load, J_{SC} is the short current density under light, J_0 is the reverse saturation current density, q is the electron charge, V is the applied voltage, m is the ideality factor, K_B is the Boltzmann constant, and T is the temperature. When the current flowing through the external circuit is 0, we can obtain V_{oc} according to Equation (4):

$$V_{OC} = \frac{mK_BT}{q} ln(\frac{J_{SC}}{J_0} + 1) \qquad (5)$$

From Equation (5), it can be seen that a higher V_{oc} corresponds to a lower $J_{0.}$ The series resistance (R_s) and J_0 of the cells can be calculated according to the diode equation (6) and (7) which come from equation (4):

$$\frac{dV}{dJ} = \frac{mk_BT}{q}(J_{SC} + J)^{-1} + R_s \qquad (6)$$

$$\ln(J_{SC} + J) = \frac{q}{mK_BT}(V - J \times R_S)\ln J_0 \qquad (7)$$

After the introduction of ITIC and ITIC-M, the V_{oc} of corresponding devices has been significantly improved, taking the ITIC-M modified device as an example, by linearly fitting the $\ln(J_{sc}+J)$ vs $(V+R_s \cdot J)$*,* the values of J_0 are calculated for the control and ITIC-M modified device are 6.51×10^{-8} and 3.90×10^{-9} mA/cm^2, respectively, as shown in Fig. S6. The reduced J_0 will contribute to the improvement of V_{oc}.

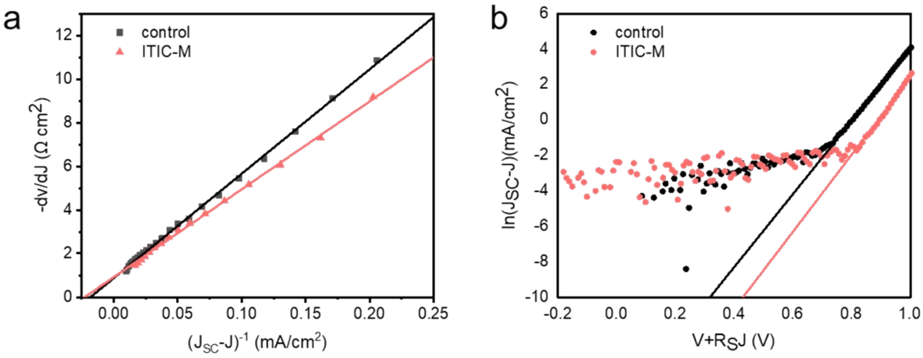


Fig. S6 (a) Plots of dV/dJ versus (J_{sc} + J) -1and the linear fitting curves, (b) plots of ln (J_{sc} + J) versus (V – R_s·J) and the linear fitting curves.

J_0 reflects the carrier recombination in the perovskite solar cells, according to the scanning electron microscope (SEM) results (Fig. 7Sa-7Sb), the average grain size of ITIC-M modified perovskite film is improved from 349 nm (control film) to 450 nm, as shown in Fig. 7Sc-7Sd, which means there are less grain boundaries in the ITIC-M modified perovskite film. The less grain boundaries will contribute to reduce the non-radiative recombination in the ITIC-M modified perovskite film. On the other hand, the ideality factor m of control and ITIC-M modified PSCs are calculated according to the slope of curves in Fig. S6(a), which are 1.86 and 1.57, respectively. A perfect solar cell has a m of unity, if m approaches 2, charge recombination processes are dominated by non-radiative trap-assisted recombination. An ideality factor m closer to unity infers that the solar cells operate with less trap-assisted recombination. In the PSCs with ITIC-M, the reduced m means the less trap-assisted recombination, which is evidence of improvements in V_{oc}.

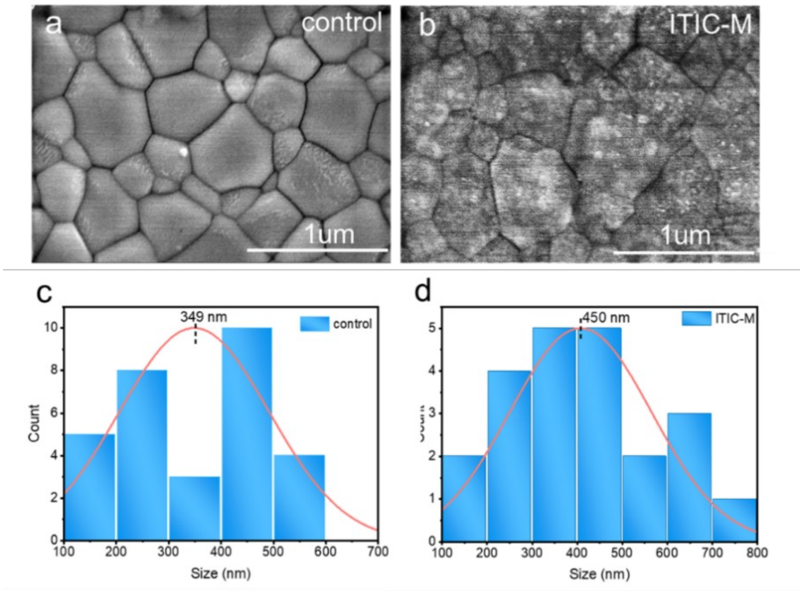


Fig. S7 (a-b) The SEM images of control and ITIC-M modified perovskite films, respectively; (c-d) the distribution of grain size of control and ITIC-M modified perovskite films, respectively.

Table S3 Datasets on the interface passivation at the perovskite/ETL interface of the PSCs

ID	**Passivation materials**	**Device PCE**	**Control device PCE**	**rA**	**AX**	**PX**	**References DOI**
1	TPPO	14.9	13.7	2.7	1:3	1:3	10.1016/j.optmat.2022.112264
2	capsaicin	21.88	19.16	2.7	0.949:3	1.025:3	10.1016/j.joule.2020.12.009
3	Y6	20.2	17.39	2.7	0.843:3	1.079:3	10.1021/acsami.1c13447
4	4,4'-Bipyridine	16.67	15.92	2.720	1.045:3	0.935:3	10.1016/j.solener.2019.08.026
5	2,2'-BiPy	9.31	15.92	2.720	1.045:3	0.935:3	10.1016/j.solener.2019.08.026
6	BCP	13.11	7.05	2.7	1:3	1:3	10.1109/JPHOTOV.2017.2651108
7	PFNOX	14	11.4	2.7	1.8:3	0.6:3	10.1002/advs.201500353
8	C3AI	20.3	17.7	2.7	1:3	1:3	10.1016/j.dyepig.2021.109385
9	C3A	19.6	17.7	2.7	1:3	1:3	10.1016/j.dyepig.2021.109385
10	C4	19.6	17.7	2.7	1:3	1:3	10.1016/j.dyepig.2021.109385
11	PVK	19.65	16.54	2.7	1:3	1:3	10.1021/acsaem.1c00219
12	2-HI-PVK	12.88	12.23	2.7	1:3	1:3	10.1021/acsaem.9b00757
13	4-HI-PVK	13.76	12.23	2.7	1:3	1:3	10.1021/acsaem.9b00757
14	P4VP	20.02	17.46	2.7	1:3	1:3	10.1039/C9TC06578D

15	PS-PAN	22.02	18.18	2.742	0.95:3	1.059:3	10.1016/j.nanoen.2020.105731
16	PHMT	21.11	18.11	2.7	1:3	1:3	10.1016/j.jechem.2020.12.035
17	ETMT	19.36	18.11	2.7	1:3	1:3	10.1016/j.jechem.2020.12.035
18	PRMT	18.57	18.11	2.7	1:3	1:3	10.1016/j.jechem.2020.12.035
19	DPSI	21.1	19.1	2.777	1:3	1:3	10.1002/adma.201803428
20	TMTA	20.22	19.08	2.7	1:3	1:3	10.1038/s41467-018-06204-2
21	PU	18.7	16.4	2.7	1:3	1:3	10.1002/adfm.201703061
22	SP1	18.75	18.19	2.7	1:3	1:3	10.1002/aenm.201803766
23	SP2	19.27	18.19	2.7	1:3	1:3	10.1002/aenm.201803766
24	SP3	20.43	18.19	2.7	1:3	1:3	10.1002/aenm.201803766
25	PBTI	20.67	18.89	2.712	0.989:3	1.005:3	10.1002/adfm.201808855
26	PEA	20.9	20.5	2.712	0.989:3	1.005:3	10.1038/s41560-019-0538-4
27	BA	20.8	20.5	2.712	0.989:3	1.005:3	10.1038/s41560-019-0538-4
28	Oam	23	20.5	2.712	0.989:3	1.005:3	10.1038/s41560-019-0538-4
29	BMIMBF4	19.8	18.5	2.727	0.953:3	1.049:3	10.1038/s41586-019-1357-2
30	F-PDI	18.28	15.37	2.7	1:3	1:3	10.1002/aenm.201900198
31	PFN-2TNDI	16.7	12.9	2.7	1.8:3	0.6:3	10.1002/aenm.201501534
32	HATNASOC7-Cs	17.62	15.91	2.7	0.949:3	1.025:3	10.1002/anie.201604399
33	HATNASO2C7-Cs	14.42	15.91	2.7	0.949:3	1.025:3	10.1002/anie.201604399
34	HATNAS3C4	11.59	15.91	2.7	0.949:3	1.025:3	10.1002/anie.201604399
35	HATNAS3C7	13.49	15.91	2.7	0.949:3	1.025:3	10.1002/anie.201604399
36	HATNAS3C7-C3h	13.38	15.91	2.7	0.949:3	1.025:3	10.1002/anie.201604399
37	HATNAS3C7-C3	13.95	15.91	2.7	0.949:3	1.025:3	10.1002/anie.201604399
38	ITIC-Th	22.87	21.85	2.716	0.989:3	1.005:3	10.1002/adma.202202100
39	IT-Cl	23.74	21.85	2.716	0.989:3	1.005:3	10.1002/adma.202202100
40	NAP	13.4	10.3	2.7	1.5:3	0.75:3	10.1016/j.isci.2018.11.003
41	EVA	19.01	17.15	2.674	0.871:3	1.112:3	10.1002/adfm.201902629
42	AIA	15.7	13.6	2.7	1:3	1:3	10.1002/adsu.202000078
43	HIA	17.29	13.6	2.7	1:3	1:3	10.1002/adsu.202000078
44	CA	19.06	13.6	2.7	1:3	1:3	10.1002/adsu.202000078
45	SubPc	13.6	9.96	2.7	1.8:3	0.6:3	10.1109/JPHOT.2016.2608619
46	PNDI-2T	21.13	19	2.7	1:3	1:3	10.1002/solr.202100236
47	thiazole	17.98	14.34	2.7	1:3	1:3	10.1021/acsami.8b16124
48	PEGDA	21.03	18.73	2.7	1:3	1:3	10.1021/acsami.0c11468
49	PLL	19.45	16.72	2.7	1:3	1:3	10.1016/j.jechem.2020.05.040
50	PVA	17.28	16.46	2.7	1:3	1:3	10.1021/acsami.1c08539
51	PMA	19.05	16.46	2.7	1:3	1:3	10.1021/acsami.1c08539

52	PAA	20.29	16.46	2.7	1:3	1:3	10.1021/acsami.1c08539
53	2FEABr	21.06	19.44	2.7	0.949:3	1.025:3	10.1007/s40820-022-00854-0
54	AIBN	19.56	16.92	2.7	1:3	1:3	10.1002/solr.202200238
55	AIBME	19.69	16.92	2.7	1:3	1:3	10.1002/solr.202200238
56	ACVA	19.21	16.92	2.7	1:3	1:3	10.1002/solr.202200238
57	BDAI2	23.1	20.8	2.643	0.85:3	1.15:3	10.1002/adfm.202205009
58	PDAI2	22.2	20.8	2.643	0.85:3	1.15:3	10.1002/adfm.202205009
59	TFBA	20.39	19.09	2.574	1:3	1:3	10.1002/inf2.12307
60	HA	19.5	18.4	2.612	0.824:3	1.176:3	10.1016/j.nanoen.2022.107193
61	BA	19.77	18.4	2.612	0.824:3	1.176:3	10.1016/j.nanoen.2022.107193
62	PA	20	18.4	2.612	0.824:3	1.176:3	10.1016/j.nanoen.2022.107193
63	PHA	20.72	18.4	2.612	0.824:3	1.176:3	10.1016/j.nanoen.2022.107193
64	TEAI	19.19	18.62	2.744	0.924:3	1.053:3	10.1002/ange.202202346
65	TEASCN	21.26	18.62	2.744	0.924:3	1.053:3	10.1002/ange.202202346
66	QA	15.6	13.4	2.7	1:3	1:3	10.1016/j.cej.2022.135107
67	CH3	17.91	17.49	2.744	0.824:3	1.176:3	10.1016/j.apsusc.2021.151740
68	CHO	18.68	17.49	2.744	0.824:3	1.176:3	10.1016/j.apsusc.2021.151740
69	COCH3	18.86	17.49	2.744	0.824:3	1.176:3	10.1016/j.apsusc.2021.151740
70	BHF	20.3	16.2	2.643	0.85:3	1.15:3	10.1002/solr.202200296
71	DMAII	13.14	10.53	1.81	0.793:3	1.103:3	10.1021/acsami.1c23637
72	BABr	20.3	18.5	2.7	1:3	1:3	10.1039/d2ee00759b
73	Trometamol	17.91	16.25	2.7	1:3	1:3	10.1021/acs.jpclett.2c01089
74	HaHc	9.18	4.75	2.79	1:3	1:3	10.1016/j.cej.2021.133745
75	PEAI	19.58	17.07	2.772	1:3	1:3	10.1002/adma.202110241
76	EDAI	20.67	17.07	2.772	1:3	1:3	10.1002/adma.202110241
77	PEAI-EDAI	22.51	17.07	2.772	1:3	1:3	10.1002/adma.202110241
78	TFAA	20.1	15.08	2.7	1:3	1:3	10.1021/acsaem.1c02984
79	FPA	21.28	17.87	2.7	1:3	1:3	10.1002/solr.202101101
80	2-TPAA	14.23	12.94	1.81	1:3	1:3	10.1016/j.cej.2022.136242
81	KBF4	23.04	21.13	2.604	0.888:3	1.135:3	10.1002/adfm.202204880
82	dtdn	18.34	16.76	2.7	1:3	1:3	10.1039/d1se01892b
83	coumarin343	19.8	18	2.643	0.85:3	1.15:3	10.1016/j.nanoen.2022.106935
84	3-HBA	23.25	21.56	2.741	0.917:3	1.065:3	10.1002/ange.202206914
85	SA	19.48	18.32	2.7	1:3	1:3	10.1016/j.apsusc.2022.152670
86	PTA	20.3	18.32	2.7	1:3	1:3	10.1016/j.apsusc.2022.152670
87	GuaBF4	20.87	18.09	2.7	1:3	1:3	10.1016/j.solmat.2022.111682
88	BCP-3N	20.9	18.7	2.729	1.151:3	0.924:3	10.1002/solr.202200559
89	BCP-Oam	17.3	18.7	2.729	1.151:3	0.924:3	10.1002/solr.202200559
90	BCP-3N-I	14.8	18.7	2.729	1.151:3	0.924:3	10.1002/solr.202200559

91	DMAI-TFMPHC	21.4	18.3	2.755	0.941:3	1.039:3	10.1016/j.cej.2022.135974
92	DMAI	20.7	18.3	2.755	0.941:3	1.039:3	10.1016/j.cej.2022.135974
93	TFMPHC	19.9	18.3	2.755	0.941:3	1.039:3	10.1016/j.cej.2022.135974
94	COTIC-4F	20.69	20.52	2.736	0.935:3	1.052:3	10.1002/aenm.202200005
95	PTB7-Th	21.4	20.52	2.736	0.935:3	1.052:3	10.1002/aenm.202200005

The specific code:

```
import numpy as np
import pandas as pd
import shap
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.model_selection import KFold
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import cross_val_score, GridSearchCV
from sklearn import metrics
from pandas import read_csv
import xgboost as xgb
from sklearn.ensemble import GradientBoostingRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.svm import SVR
from sklearn.neighbors import KNeighborsRegressor
from scipy import stats
from sklearn.metrics import mean_squared_error
from sklearn.metrics import r2_score
from matplotlib.pyplot import savefig
#Read data
filename = ' .csv'
```

```
data0 = read_csv(filename)
X = data0.iloc[:, :]
y = data0.iloc[:, ]
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=test_size, random_state=10)
scaler = StandardScaler()
scaler.fit(X_train)
X_train_std = scaler.transform(X_train)
X_test_std = scaler.transform(X_test)
#Select model
#XGBoost
param_grid = {'learning_rate': [0.1,0.15,0.35],'max_depth': [3],'min_child_weight': [3,4,5],'colsample_bytree': [0.8],'gamma': [0, 0.1, 0.2, 0.3],'reg_alpha': [0, 0.1, 0.2, 0.3],'reg_lambda': [1,3,5]}
#GBDT
param_grid = {'n_estimators': [5,10,20,30,40,50,60,70,80,90,100,500],'max_features': ['auto', 'sqrt'],'max_depth': range(1,10),'min_samples_leaf': [1, 2],}
#RF
param_grid = {'n_estimators': [10,20,30,40,50,60,70,80,100,500],'max_features': ['auto', 'sqrt'],'max_depth': [3,5,8],'min_samples_split': [2,3,5],'min_samples_leaf': [1,2],'bootstrap': [True]}
#SVM
param_grid={'kernel':['linear','poly','rbf','sigmoid'],'degree':[1,2,3],'gamma':[0.01,0.05,0.1,0.2,0.5,0.6,0.8],'C':[0.1,0.2,0.3,0.4,0.5,1],'epsilon':[0.5,1,2]}
#kNN
param_grid = {'weights': ['uniform'],'n_neighbors': range(2, 20),'algorithm': ['ball_tree','kd_tree','brute']}
```

```
#Fit and predict
best_model = model.best_estimator_
y_train_hat = best_model.predict(X_train_std)
y_test_hat = best_model.predict(X_test_std)
#SHAP
xgb_model = xgb.XGBRegressor(random_state=42)
xgb_model = RandomForestRegressor(random_state=42)
xgb_model = GradientBoostingRegressor(random_state=42)
xgb_model.fit(X_train, Y_train)
shap.initjs()
explainer = shap.TreeExplainer(xgb_model)
shap.initjs()
shap_value = explainer.shap_values(X)
print(shap_value)
shap.summary_plot(shap_value, X)
shap.summary_plot(shap_value, X, plot_type="bar")
abs_shapvalue = abs(shap_value)
shap_average = np.average(abs_shapvalue,axis=0)
print(shap_average)
```