ADVANCED FUNCTIONAL MATERIALS
www.afm-journal.de

Enhancing Power Conversion Efficiency of Perovskite Solar Cells Through Machine Learning Guided Experimental Strategies

Antai Yang, Yonggui Sun, Jingzi Zhang, Fei Wang, Chengquan Zhong, Chen Yang, Hanlin Hu, Jiakai Liu,* and Xi Lin**

Predicting the power conversion efficiency (PCE) using machine learning (ML) can effectively accelerate the experimental process of perovskite solar cells (PSCs). In this study, a high-quality dataset containing 2079 experimental PSCs is established to predict PCE values using an accurate ML model, achieving an impressive coefficient of determination (R^2) value of 0.76. In the 12 validation experiments with PSCs, the average absolute error between the observed and predicted PCE values is only 1.6%. Leveraging the recommended improvement solutions from the ML model, the device's PCE to 25.01% in experimental PSCs is successfully enhanced, thus truly realizing the objective of machine learning-guided experiments. In addition, by improving the PCE of specific devices, the predicted value can reach 28.19%. The ML model has provided feasible strategies for experimentally improving the PCE of PSCs, which play a crucial role in achieving PCE breakthroughs.

1. Introduction

The power conversion efficiency (PCE) of perovskite solar cells (PSCs) is generally influenced by the chemical compositions and additives of active layers, materials in hole and electron transport layers, choices of passivation layers, structural layouts of devices, processing conditions, and methods, and many other factors.[1] The intricate nature of these materials and processing choices would inevitably lead to labor-extensive trial-and-error efforts in search of the optimal PSCs toward the ultimate theoretical PCE limit of 33%.[2] In recent years, machine learning (ML) has emerged as a powerful tool in predicting the PCE of PSCs.[3] Although using ML features such as the electronic energy levels,[4] carrier mobilities,[5] and grain sizes,[6] have demonstrated reasonable predictive accuracy, these microscopic or electronic features may not be applied to guide experiments directly.[4–7] On the other hand, predicting the PCE based on device structures, materials selections, and manufacturing methods is generally not optimistic.[8]

In this work, we built a set of comprehensive data on PSCs from literature published from 2013 to 2023 and used ML methods to predict the PCE of PSCs based on the dataset. Then, we analyzed the main factors affecting PCE and extracted general rules underlying the high PCE of PSCs. To make our predictions more informative for experimental purposes, we used experimental components as input features instead of material physicochemical properties or characterization. Furthermore, the accuracy of the model was validated through various experimental verifications. Subsequently, the optimization strategy suggested by the ML model was effectively employed, resulting in a significant increase in the PCE to 25.01%. Finally, we pointed out various

A. Yang, J. Zhang, C. Zhong, X. Lin
School of Materials Science and Engineering
Harbin Institute of Technology
Shenzhen 518055, P. R. China
E-mail: linxi@hit.edu.cn

A. Yang, J. Zhang, C. Zhong, X. Lin
Blockchain Development and Research Institute
Harbin Institute of Technology
Shenzhen 518055, P. R. China

Y. Sun, F. Wang, H. Hu
Hoffmann Institute of Advanced Materials
Shenzhen Polytechnic University
7098 Liuxian Boulevard, Shenzhen 518055, China
E-mail: hanlinhu@szpu.edu.cn

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/adfm.202410419

DOI: 10.1002/adfm.202410419

C. Yang, J. Liu
Xinjiang Key Laboratory of Separation Material and Technology
Xinjiang Technical Institute of Physics and Chemistry
Chinese Academy of Sciences
Urumqi 830011, China
E-mail: liujk@ms.xjb.ac.cn

J. Liu
Center of Materials Science and Opto-electronic Technology
University of Chinese Academy of Sciences
Beijing 100049, China

J. Liu
Sunrise (Xiamen) Photovoltaic Industry Co. Ltd.
44 Huli Avenue, Huli District, Xiamen, Fujian 361000, P. R. China

X. Lin
State Key Laboratory of Advanced Welding and Joining
Harbin Institute of Technology
Harbin 150001, P. R. China

Table 1. Different ML algorithms' performance in PCE prediction.

ML algorithms	Training set				Test set			
	r	R^2	RMSE [%]	MAE [%]	*R*	R^2	RMSE [%]	MAE [%]
LR	0.7050	0.4970	3.910	3.025	0.6908	0.4772	3.840	2.951
SVR	0.8275	0.6848	3.113	2.138	0.7261	0.5272	3.664	2.700
ANN	0.9083	0.8250	2.318	1.628	0.7315	0.5350	3.703	2.719
DT	0.8546	0.7304	2.837	2.113	0.7711	0.5936	3.541	2.728
RF	0.9558	0.9135	1.701	1.279	0.8610	0.7455	2.805	2.132
LGBM	0.9205	0.8473	2.155	1.585	0.8612	0.7417	2.795	2.159
XGBoost	0.9417	0.8868	1.889	1.352	0.8718	0.7601	2.695	2.055
CatBoost	0.9396	0.8828	1.890	1.364	0.8719	0.7603	2.690	2.078

possible perspectives for further improving efficiency for the current high-efficiency device solutions in a targeted manner.

2. Results and Discussion

2.1. Prediction of PCE and Experimental Validations

A set of 2079 PCE data points were extracted from 1149 articles published between 2013 and 2023, with their measured PCE ranging from 0.1% to 26.1%. 14 experimental factors (perovskite components, ETL, ETL-2, deposition procedure, deposition method, antisolvent, precursor solution, HTL, HTL-additive, ETL-passivator, HTL-passivator, additives, add-Cl, type) were selected to generate features, with the methods detailed in "Feature generation" section of the supporting information. At last, 22 features listed in Table S1 (Supporting Information) were generated. The datasets and codes were provided at https://github.com/Dangyue338/Prediction-of-perovskite-solar-cells. For prediction, we selected 4 classic ML algorithms and 4 tree-based ensemble learning algorithms, including Linear Regression (LR),[9] Artificial Neural Network (ANN),[10] Support Vector Regression (SVR),[11] Decision Tree (DT),[12] Random Forest (RF),[13] LightGBM (LGBM),[14] XGBoost,[15] and CatBoost.[16] To evaluate the models' performance, the dataset was randomly divided into training and test sets, with the training set comprising 80% of the data and the test set comprising the remaining 20%. Hyperparameter search was conducted using grid search and tenfold cross-validation. The evaluation criteria of correlation coefficient (*r*), coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE) were employed to comprehensively assess the predictive ability and accuracy of the model.

The model fitting status of the training and test sets was visually displayed in Figures S2 and S3 (Supporting Information), with the prediction performance shown in **Table 1** and Figure S4 (Supporting Information). None of the models exhibited a scenario where the training set performed well while the test set performed poorly, indicating that the models did not overfit in the complex systems, with the best hyperparameters for different models listed in Table S2 (Supporting Information). Overall, it can be observed that the ensemble learning algorithms were generally superior to the classic algorithms, as all of them achieved high R^2 values and low RMSE and MAE values. Additionally, all *r* values of ensemble learning algorithms exceeded 0.86, indicating a robust linear correlation between the predicted and experimental values. Among them, CatBoost achieved the highest performance on the test set, as shown in **Figure 1**a. It attained the highest R^2 score of 0.76, a correlation coefficient *r* value of 0.87, a low RMSE value of 2.69%, and an MAE value of 2.08%. According to Table S3 (Supporting Information), our scheme achieved the highest prediction performance in PCE prediction tasks with large data volumes.

To further validate the accuracy of our ML models, we fabricated 12 PSC groups with the structure ITO/SnO_2/Perovskite/Spiro-OMeTAD/Au for experimental validation. The experimental and predicted PCE values for these 12 devices were presented in Table S4 (Supporting Information), and specifically represented in Figure 1b. In cases with slight discrepancies, the majority of the projected values closely align with the actual values, falling within a reasonable error range. The predictions generated by our model were deemed credible, with a mean absolute error of only 1.6%. Figure 1c presents the current density-voltage (*J*−*V*) curves for 12 distinct components, along with corresponding changes in experimental and predicted values, as well as error ranges depicted in Figure S5 (Supporting Information).

At last, the component $(FA_{0.7}MA_{0.3})_{0.87}Cs_{0.13}Pb(I_{0.985}Br_{0.015})_3$ with MACl has been identified as the best component Regular 1, with a maximum PCE reaching 22.09%. After this, the experimental data points were added to the training dataset to further correct the ML model, with the training results and performance metrics shown in Figure S6 and Table S5 (Supporting Information). Subsequently, to enhance the device's PCE, a greedy strategy and iterative local optimization were applied to optimize potential improvement solutions for the remaining experimental factors. The strategies suggested by the ML model were implemented where they can improve the PCE, which were displayed in the "Model application" section of the supporting information with the detailed process. Specifically, following the procedure shown in Figure S7a (Supporting Information), we use 2-phenylethylammonium iodide (PEAI) as an HTL-passivator, following the methodology outlined in the reference[17] named Regular 2. Subsequently, based on Regular 2, CuSCN-doped Spiro-OMeTAD was employed, following the methodology described in ref. [18] and named Regular 3. After a series of experimental trials, it has been demonstrated that employing the enhanced

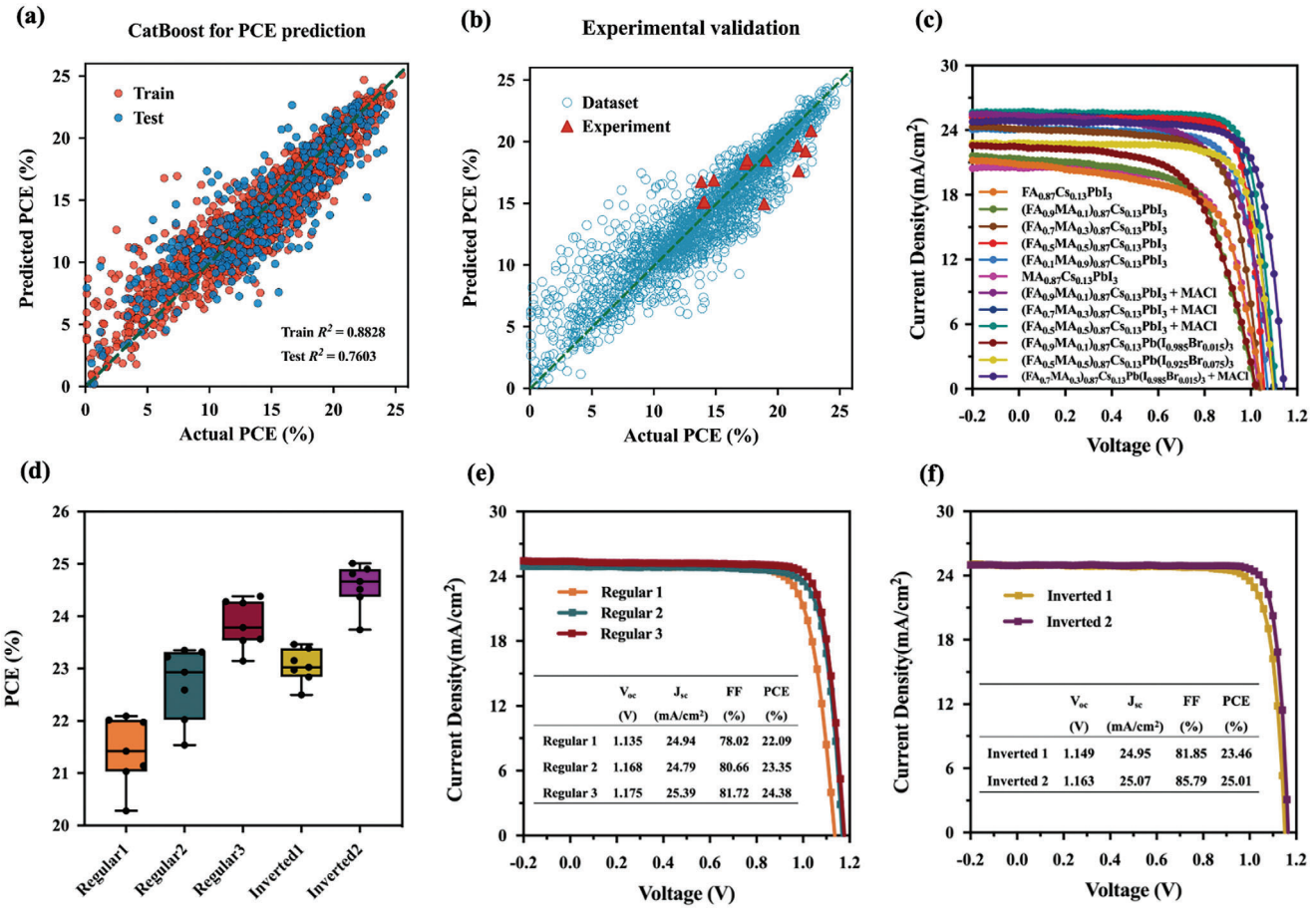

Figure 1. a) The fitting graph of PCE results by the CatBoost-based algorithm model, where red represents the training set and blue represents the test set. b) The distribution of experimental results and data points from the database in the fitting graph. c) *J–V* curves of devices with different perovskite components via a two-step spin-spin sequential deposition method. d) Statistics of the PCE of PSCs with different groups based on 7 devices. The central line represents the median, the box limits correspond to the upper and lower quartiles, and the whiskers extend to the minimum and maximum values. e) *J–V* curves of the regular devices and their champion photovoltaic performance. f) *J–V* curves of the inverted devices and their champion photovoltaic performance.

approach recommended by the model leads to a noticeable enhancement in PCE as illustrated in Figure 1d,e, with maximum PCE values reaching 23.35% for Regular 2 and 24.38% for Regular 3. This enhancement was also accompanied by improvements in other photovoltaic parameters. The open-circuit voltage (V_{oc}) increased from 1.135 V for Regular 1 to 1.168 V for Regular 2 and 1.175 V for Regular 3, while the fill factor (FF) rose from 78.02% for Regular 1 to 80.66% for Regular 2 and 81.72% for Regular 3. The short-circuit currents (J_{sc}) maintained basic stability after adding PEAI from 24.94 mA cm^{-2} for Regular 1 and 24.79 mA cm^{-2} for Regular 2 and had a improvement after employing CuSCN as HTL-additive to 25.35 mA cm^{-2} for Regular 3. The phenomenon exhibited by this photovoltaic curve is consistent with the reference source, which indicates that the strategies provided by ML are reasonable.

Additionally, the inverted structure ITO/NiO_x/Me-4PACz/perovskite/PCBM/BCP/Ag, designated as Inverted 1, was also fabricated to validate the improvement strategy recommended by the ML model. Following the procedure shown in Figure S7b (Supporting Information), 4-chlorobenzene sulfonate (4Cl-BZS) was employed as a precursor additive, using the method reported in the ref. [19] and designated as Inverted 2. This modification increased the PCE from 23.46% to 25.01%, as illustrated in Figure 1d,f. Additionally, it has been accompanied by other photovoltaic performance improvements, with V_{oc} increased from 1.149 V for Inverted 1 to 1.163 V for Inverted 2, a J_{sc} value from 24.95 mA cm^{-2} for Inverted 1 to 25.07 mA cm^{-2} for Inverted 2 and an impressive FF value from 81.85% for Inverted 1 to 85.79% for Inverted 2.

2.2. Model Explanation

The Shapley additive explanations (SHAP) were employed to analyze the model and determine the significance of various features.[20] After fitting the target using four tree-based ensemble learning models, the average contribution of each model to PCE was calculated, shown in **Figure 2a**. The estimation allowed for the comparison of contributions from different attributes to PCE and facilitated the analysis of feature importance. Then the

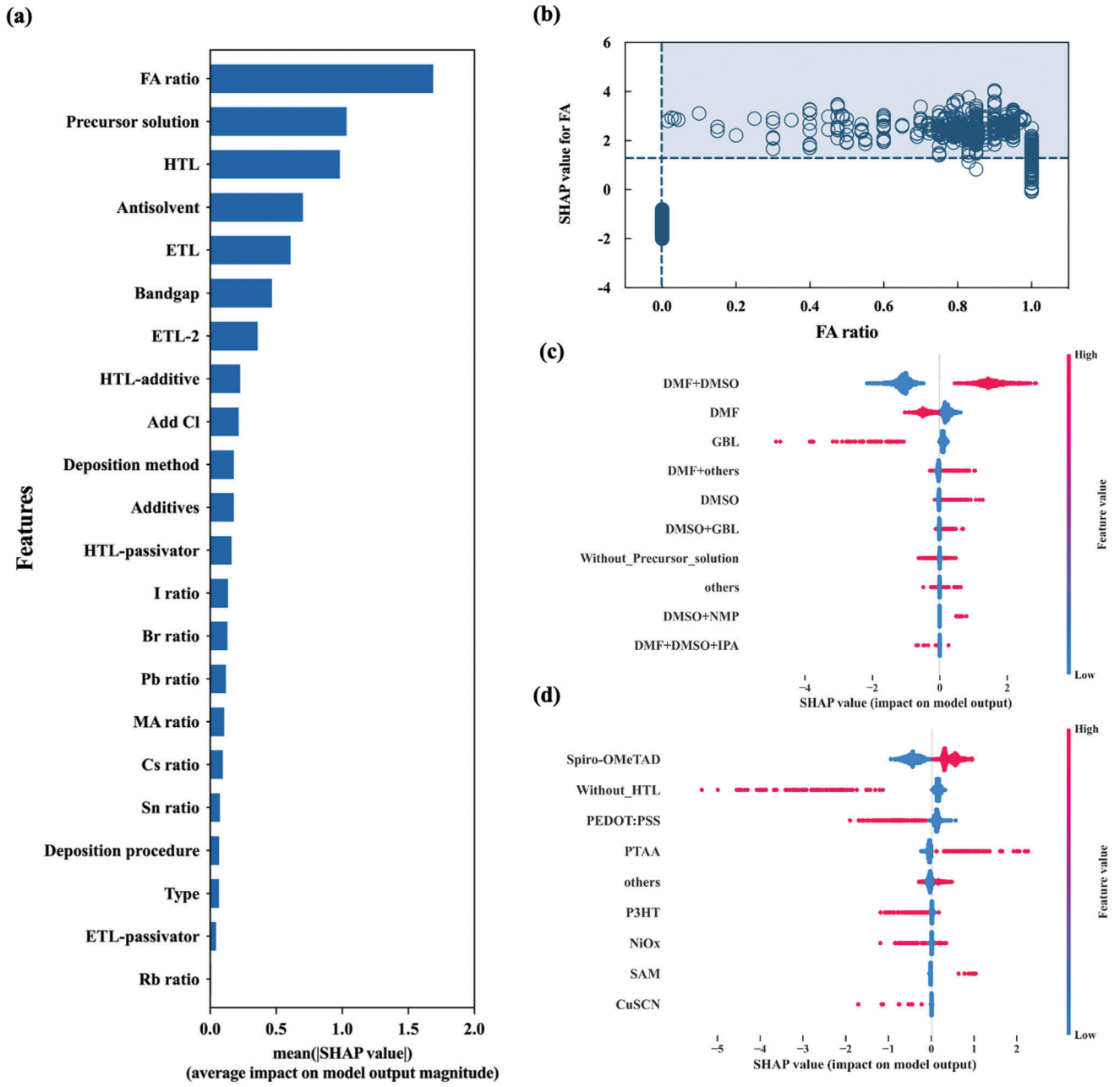

Figure 2. a) The contribution of different influencing factors to PCE, sorted by the mean absolute SHAP values (average impact on model output magnitude). Detailed SHAP for the top three importance rankings: b) The SHAP value of FA cation ratio to PCE. c) Alternatives of precursor solution and the SHAP value. d) Alternatives of HTL materials and the SHAP value.

effect of the alternatives of each layer's materials and methods on PCE was re-analyzed by one-hot encoding features as well, displayed in Figure 2c,d, and Figures S8–S15 (Supporting Information). The red dots represent data points using the relevant method or alternative, while the blue dots represent data points that do not use the method or alternative. The positive or negative effects of their SHAP values were used to determine whether they have a positive or negative impact on PCE. The establishment details are listed in **Table 2**, with the alternatives of category features organized in order of importance.

The importance of the features was demonstrated in Figure 2a, which respectively represents the severity of the impact of different factors on PCE. As can be seen, certain features significantly rank among the top contributors to PCE and even play a decisive role in determining the PCE. The proportion of FA cations was the most important factor, followed by the precursor solution and HTL materials. According to Figure 2b, adding FA will significantly lead to an increase in SHAP values. Reportedly, most efficiency records of perovskite solar cells were achieved through perovskite based on FA mixed cations, indicating the

Table 2. Factors for explaining the model and their descriptions or alternatives.

Factors	Description or alternatives for one-hot encoding (ordered by impact level)
Perovskite (MA, FA, Cs, Rb, Pb, Sn, Br, I, bandgap)	Ion ratios were distributed between 0 and 1, and the bandgap was predicted by Gok et al.'s model[7]
Deposition procedure	One-step, two-step
Deposition method	Spin-dip, spin, others, spin2-3, CVD, VASP, spin-spray
Precursor solution	DMF+DMSO, DMF, GBL, DMF+others, DMSO+GBL, DMSO, others, without precursor solution, DMSO+NMP, DMF+DMSO+IPA
Antisolvent	Chlorobenzene, toluene, diethyl ether, ethyl acetate, others, isopropanol, anisole, trifluorotoluene, without anti-solvent treatment
ETL	SnO_2, without ETL, PCBM, others, ZnO, c-TiO_2, TiO_2, PCBM+C_{60}, C_{60}, SnO_2-doped, TiO_2-doped, ZnO-doped
ETL-2	mTiO_2, others, without ETL-2, C_{60}, PCBM
HTL	Spiro-OMeTAD, without HTL, PEDOT:PSS, PTAA, others, P3HT, NiO_x, SAM, CuSCN
HTL-additive	LiTFSI+TBP, LiTFSI+TBP+FK209, others, LiTFSI+TBP+others, LiTFSI, without HTL additive
HTL-passivator	HTL passivator, without HTL passivator
ETL-passivator	ETL passivator, without ETL passivator
Additives	Additives, without additives
Add Cl	Add Cl, without Cl
Type	Regular, inverted

high-efficiency PSCs closely related to FA cation.[21] As the second factor, the selection of the precursor solution is crucial as it influences the film crystallization process and ultimately impacts the quality of film formation.[22] According to Figure 2c, DMF with DMSO was recommended as a good choice. The HTL materials ranked thirdly, which can prevent electron injection and extract photo-generated holes, while also in contact with the fragile perovskite.[23] At the same time, the HTL materials affect the matching of energy levels and the crystalline morphology of perovskite on its surface, which further impacts PCE.[24] In addition, the SHAP analysis results of the HTL materials in Figure 2d indicate that spiro-OMeTAD, poly[bis(4-phenyl)(2,4,6-trimethylphenyl)amine (PTAA), and self-assembled monolayers (SAM) are excellent choices for HTL materials.

The other SHAP analyses were provided in Figures S8–S16 (Supporting Information). To facilitate an intuitive understanding of the relationship between ions and PCE, the distribution map was displayed in Figure S17 (Supporting Information). According to the results of SHAP analyses (Figure S9, Supporting Information), the commonly used antisolvents that lead to better PCE are chlorobenzene, ethyl acetate, and IPA. For ETL (Figure S10, Supporting Information) factors, SnO_2 was considered an excellent material. For ETL-2 (Figure S11, Supporting Information), C_{60} performs better. In terms of HTL-additive (Figure S12, Supporting Information), Li-TFSI with TBP and other additives are beneficial to PCE, while Li-TFSI with only TBP was found to negatively impact PCE. For the precursor solution (Figure S13, Supporting Information), DMF with DMSO was considered a good choice. Additionally, the two-step spin-coating deposition method (Figure S14 and Table S6, Supporting Information) was found to be a more efficient approach for achieving high PCE. As can be seen in Figure S15 (Supporting Information), the additive of the Cl anion is of utmost significance in achieving high-performance devices, as it helps with crystallization.[25] Meanwhile, the SHAP conditions of all ions have been listed in Figure S16 (Supporting Information). Additionally, for a more comprehensive understanding of the device structure with high PCE and the pros and cons of various material choices, we conducted a separate analysis of the top devices of PCE > 23%, with corresponding SHAP detailed in Figure S18 (Supporting Information).

At last, to observe which factor could obtain a high PCE more effectively, the database was divided into three categories by PCE levels: low efficiency (PCE < 12%), medium efficiency (12% ≤ PCE ≤ 18%), and high efficiency (PCE > 18%), presented in **Figure 3a**. Subsequently, the average absolute improvement values of PCE resulting from changes in each factor were calculated by predicting, denoted as PCE enhancement in Figure 3b–d. It is evident that, for low-efficiency devices in Figure 3b, additive engineering, HTL-additive, and HTL-side passivator engineering are the most significant methods for enhancing efficiency. While in the medium-efficiency devices in Figure 3c, the ETL becomes the most crucial influencing factor, followed by HTL-additive and component additive engineering. In the case of PCE at the high-efficiency level in Figure 3d, ETL and HTL-additive maintain a significant position in the field, followed by precursor solution. With the improvement of basic efficiency, ETL-side passivator engineering becomes more efficient than HTL-side passivator engineering, and HTL-additive constantly maintains its advantageous position.

2.3. Potential Improvement Solutions

For each specific perovskite, the potential PCE can be predicted by altering the structure and processing of the device. Using the results from the model, several methods were identified to enhance each device by modifying their structures, including the materials and treatment techniques for each layer. For

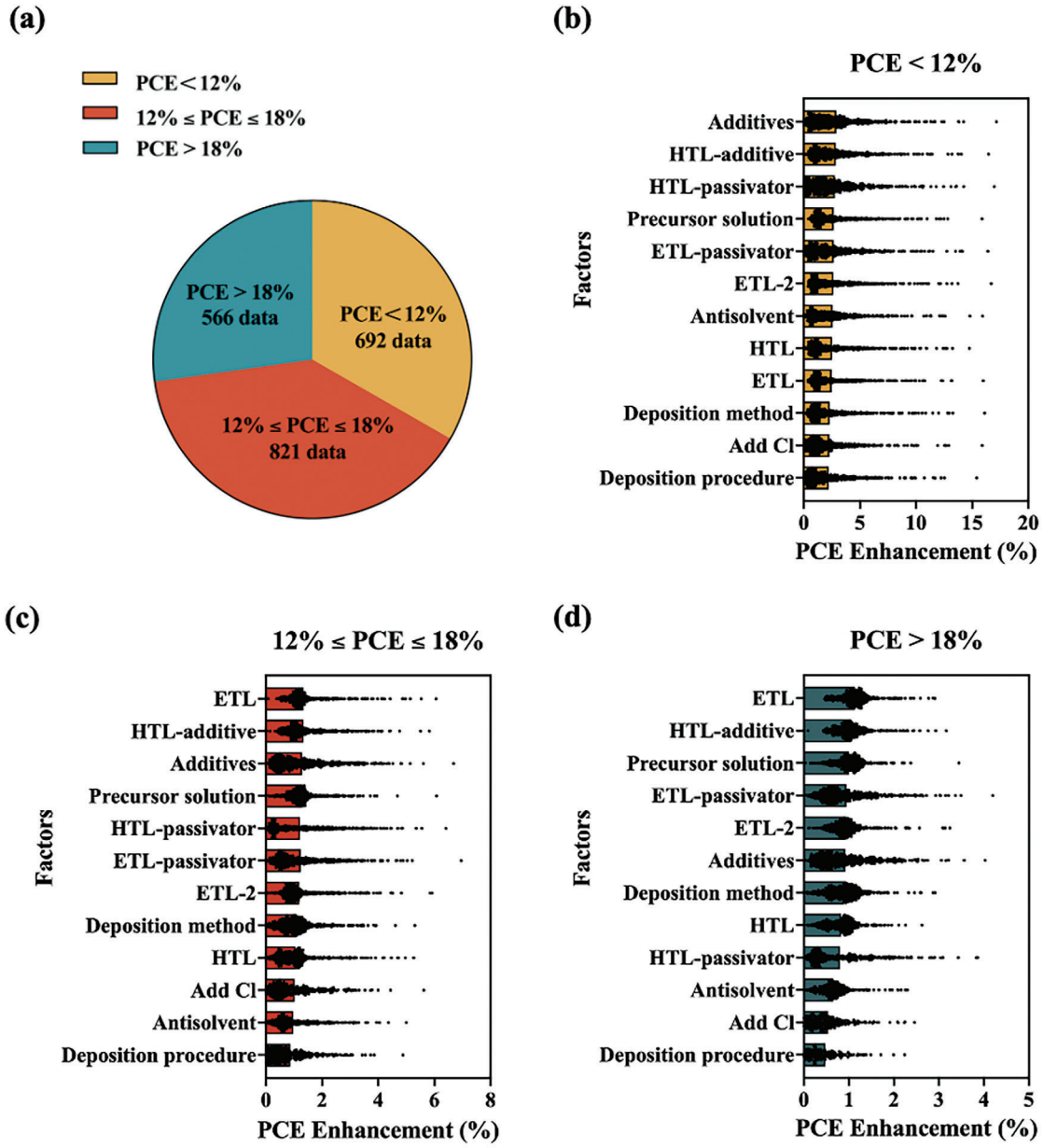

Figure 3. a) Data distribution located in high, medium, and low PCE. b) The improvement effect of different factors on PCE at low levels (PCE < 12%). c) The improvement effect of different factors on PCE at middle levels (12% ≤ PCE ≤ 18%). d) The improvement effect of different factors on PCE at high levels (PCE > 18%).

example, Yuan et al.'s device[26] was optimized through targeted improvement solutions using the ML model to enhance device performance, as shown in **Figure** 4a. The approach involved using (2-carboxyethyl) dimethyl sulfonium chloride (CDSC)[27] and DL-carnitine hydrochloride (DL)[28] as HTL-side passivation layers, CuSCN[18] and CYTFA[29] doped spiro-OMeTAD as HTL-additive, along with 2-phenylethylammonium iodide (PEAI)[17] or a combination of 4-tert-butyl-benzylammonium iodide and phenylpropylammonium iodide (tBBAI+PPAI)[30] as HTL-side passivator. This optimization increased the predicted PCE value from 26.41% to a maximum of 28.13%. Furthermore, through structural improvements, it is anticipated that the device by Sargent et al.[19] can be enhanced to a maximum of 28.19%, as shown in Figure 4b. This enhancement involves utilizing KSCN[31] and Al_2O_3[32] as an ETL-side passivator to modify the interface and incorporating HTL of SAM with NiO_x[33] and 4,4′,4″-nitrilotribenzoic acid (NA)[34] as HTL-additive on the basis. These changes can enhance the predicted PCE value to a maximum of 28.19%, representing a highly promising direction for experimental exploration. You et al.'s device,[35] depicted in Figure S19a (Supporting Information), employing Y[36] and Sb[37] doped SnO_2, CDSC[27] and KF[38] as HTL-side passivation layers, as well as CuSCN[18] and Co-salt[39] doped Spiro-OMeTAD as hole transport layers, was predicted as targeted improvement

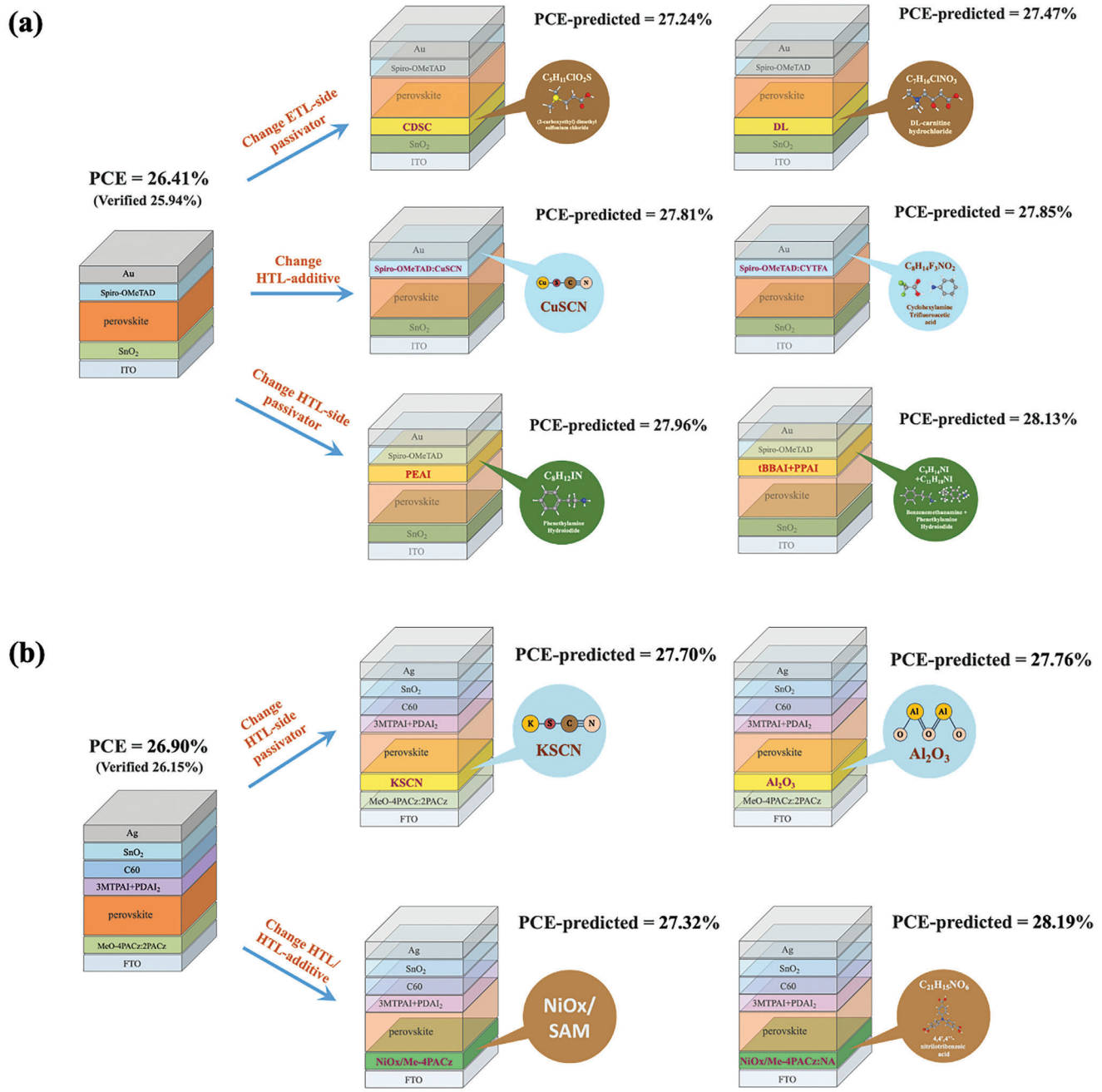

Figure 4. Machine learning solutions of PCE enhancement with different methods: a) regular device of Yuan et al.[26] b) inverted device of Sargent et al.[19]

solutions using the ML method to optimize the device's performance. This optimization increased the predicted PCE to a maximum of 27.74%. In addition, through structural improvements, it is anticipated that the device by Zhu et al.[40] shown in Figure S19b (Supporting Information) can be enhanced to a maximum of 26.88%. This enhancement involves utilizing n-butylamine hydroiodide (nBAI)[41] and KSCN[31] as a passivation layer to modify the interface and incorporating additives such as dimethylamine (DMA)[42] and LiI[43] into the precursor solution.

3. Conclusion

In this study, we curated a high-quality experimental dataset comprising 2079 cleaned data points to predict PSCs' PCE. To directly guide experiments, we exclusively utilized the device structures, materials selection, and manufacturing methods as the input for the ML model. Based on this, we established a robust ML model with a high R^2 value of 0.76 and a low RMSE value of 2.69%. After conducting 12 sets of experimental validations, the model exhibited exceptional performance with an MAE of only 1.6%.

Implementing an optimization strategy recommended by the model, we successfully increased the device's PCE to 25.01%. Additionally, we analyzed the factors influencing PCE, and we have proposed targeted improvement directions for devices at different efficiency stages. We believe that for high-efficiency devices, optimizing ETL, HTL-additive, and precursor solution is the most efficient way to improve PCE. At last, we forecasted several advanced devices that have the potential to overcome the current PCE bottleneck, with a maximum predicted value of 28.19%. In conclusion, the ML scheme employed in this study offers a novel prediction and screening methodology for investigating the manufacturing process of high-performance PSCs. The optimization strategies ML provides may bring some alternative solutions and inspiration to experimenters, holding significant importance in expediting the experimental process.

Supporting Information

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

Acknowledgements

A.Y., Y.S., and J.Z. contributed equally to this work. The authors appreciate financial support from the Guangdong Basic and Applied Basic Research Foundation (2022A1515110676, 2024A1515011845), the Shenzhen Science and Technology Program (JCYJ20220531095404009; RCBS20221008093057027; JCYJ20230807094313028, JCYJ20230807094318038), the Project Supported by Sunrise (Xiamen) Photovoltaic Industry Co., Ltd. (Development of Artificial Intelligence Technology for Perovskite Photovoltaic Materials, No. HX20230176).

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

The data and code necessary to reproduce this work can be downloaded from https://github.com/Dangyue338/Prediction-of-perovskite-solar-cells.

Keywords

machine learning, perovskite, power conversion efficiency

Received: June 14, 2024
Revised: November 19, 2024
Published online: December 17, 2024

[1] a) H. Zhu, S. Teale, M. N. Lintangpradipto, S. Mahesh, B. Chen, M. D. McGehee, E. H. Sargent, O. M. Bakr, *Nat. Rev. Mater.* **2023**, *8*, 569; b) S. N. Habisreutinger, M. O. Reese, *Science* **2022**, *377*, 265.

[2] a) W. Li, M. U. Rothmann, Y. Zhu, W. Chen, C. Yang, Y. Yuan, Y. Y. Choo, X. Wen, Y.-B. Cheng, U. Bach, J. Etheridge, *Nat. Energy* **2021**, *6*, 624; b) R. Wang, M. Mujahid, Y. Duan, Z.-K. Wang, J. Xue, Y. Yang, *Adv. Funct. Mater.* **2019**, *29*, 1808843; c) H. Zhang, X. Ji, H. Yao, Q. Fan, B. Yu, J. Li, *Sol. Energy* **2022**, *233*, 421; d) A. K. Jena, A. Kulkarni, T. Miyasaka, *Chem. Rev.* **2019**, *119*, 3036; e) J. Y. Kim, J. W. Lee, H. S. Jung, H. Shin, N. G. Park, *Chem. Rev.* **2020**, *120*, 7867.

[3] a) Z. Yao, Y. Lum, A. Johnston, L. M. Mejia-Mendoza, X. Zhou, Y. Wen, A. Aspuru-Guzik, E. H. Sargent, Z. W. Seh, *Nat. Rev. Mater.* **2023**, *8*, 202; b) H. Wang, T. Fu, Y. Du, W. Gao, K. Huang, Z. Liu, P. Chandak, S. Liu, P. Van Katwyk, A. Deac, A. Anandkumar, K. Bergen, C. P. Gomes, S. Ho, P. Kohli, J. Lasenby, J. Leskovec, T. Y. Liu, A. Manrai, D. Marks, B. Ramsundar, L. Song, J. Sun, J. Tang, P. Velickovic, M. Welling, L. Zhang, C. W. Coley, Y. Bengio, M. Zitnik, *Nature* **2023**, *620*, 47; c) Y. Liu, X. Tan, J. Liang, H. Han, P. Xiang, W. Yan, *Adv. Funct. Mater.* **2023**, *33*, 2214271.

[4] J. Li, B. Pradhan, S. Gaur, J. Thomas, *Adv. Energy Mater.* **2019**, *9*, 1901891.

[5] Y. Liu, W. Yan, S. Han, H. Zhu, Y. Tu, L. Guan, X. Tan, *Sol. RRL* **2022**, *6*, 2101100.

[6] Y. Hu, X. Hu, L. Zhang, T. Zheng, J. You, B. Jia, Y. Ma, X. Du, L. Zhang, J. Wang, B. Che, T. Chen, S. Liu, *Adv. Energy Mater.* **2022**, *12*, 2201463.

[7] E. C. Gok, M. O. Yildirim, M. P. U. Haris, E. Eren, M. Pegu, N. H. Hemasiri, P. Huang, S. Kazim, A. Uygun Oksuz, S. Ahmad, *Sol. RRL* **2021**, *6*, 2100927.

[8] a) Ç. Odabaşı, R. Yıldırım, *Nano Energy* **2019**, *56*, 770; b) C. She, Q. Huang, C. Chen, Y. Jiang, Z. Fan, J. Gao, *J. Mater. Chem. A* **2021**, *9*, 25168; c) Y. Lu, D. Wei, W. Liu, J. Meng, X. Huo, Y. Zhang, Z. Liang, B. Qiao, S. Zhao, D. Song, Z. Xu, *J. Energy Chem.* **2023**, *77*, 200; d) N. Parikh, M. Karamta, N. Yadav, M. Mahdi Tavakoli, D. Prochowicz, S. Akin, A. Kalam, S. Satapathi, P. Yadav, *J. Energy Chem.* **2022**, *66*, 74.

[9] D. V. Lindley, A. F. M. Smith, *J. R. Stat. Soc. B* **1972**, *34*, 1.

[10] T. J. Hastie, R. Tibshirani, J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York **2001**.

[11] C. Cortes, V. Vapnik, *Mach. Learn.* **1995**, *20*, 273.

[12] S. L. Salzberg, *Mach. Learn.* **1994**, *16*, 235.

[13] L. Breiman, *Mach. Learn.* **2001**, *45*, 5.

[14] G. L. Ke, Q. Meng, T. Finley, T. F. Wang, W. Chen, W. D. Ma, Q. W. Ye, T. Y. Liu, presented at 31st Annual Conf. on Neural Information Processing Systems (NIPS), Long Beach, CA, December **2017**.

[15] T. Q. Chen, C. Guestrin, M. A. Comp, presented at 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD), San Francisco, CA, August **2016**.

[16] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, A. Gulin, presented at 32nd Conf. on Neural Information Processing Systems (NIPS), Montreal, Canada, December **2018**.

[17] Q. Jiang, Y. Zhao, X. Zhang, X. Yang, Y. Chen, Z. Chu, Q. Ye, X. Li, Z. Yin, J. You, *Nat. Photonics* **2019**, *13*, 460.

[18] M. Li, Z. K. Wang, Y. G. Yang, Y. Hu, S. L. Feng, J. M. Wang, X. Y. Gao, L. S. Liao, *Adv. Energy Mater.* **2016**, *6*, 1601156.

[19] H. Chen, C. Liu, J. Xu, A. Maxwell, W. Zhou, Y. Yang, Q. Zhou, A. S. R. Bati, H. Wan, Z. Wang, L. Zeng, J. Wang, P. Serles, Y. Liu, S. Teale, Y. Liu, M. I. Saidaminov, M. Li, N. Rolston, S. Hoogland, T. Filleter, M. G. Kanatzidis, B. Chen, Z. Ning, E. H. Sargent, *Science* **2024**, *384*, 189.

[20] M. V. Garcia, J. L. Aznarte, *Ecol. Inf.* **2020**, *56*, 101039.

[21] a) X. Guo, K. Ngai, M. Qin, X. Lu, J. Xu, M. Long, *Nanotechnology* **2021**, *32*, 075406; b) W. T. M. Van Gompel, R. Herckens, G. Reekmans, B. Ruttens, J. D'Haen, P. Adriaensens, L. Lutsen, D. Vanderzande, *J. Phys. Chem. C* **2018**, *122*, 4117; c) M. Li, R. Sun, J. Chang, J. Dong, Q. Tian, H. Wang, Z. Li, P. Yang, H. Shi, C. Yang, Z. Wu, R. Li, Y. Yang, A. Wang, S. Zhang, F. Wang, W. Huang, T. Qin, *Nat. Commun.* **2023**, *14*, 573; d) Z. Huang, Y. Bai, X. Huang, J. Li, Y. Wu, Y. Chen, K. Li, X. Niu, N. Li, G. Liu, Y. Zhang, H. Zai, Q. Chen, T. Lei, L. Wang, H. Zhou, *Nature* **2023**, *623*, 531; e) 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, J. Seo, *Nature* **2021**, *590*, 587; f) E. H. Jung, N. J. Jeon, E. Y. Park, C. S. Moon, T. J. Shin, T.-Y. Yang, J. H. Noh, J. Seo, *Nature* **2019**, *567*, 511; g) J. Zhou, L. Tan, Y. Liu, H. Li, X. Liu, M. Li, S. Wang, Y. Zhang, C. Jiang, R. Hua, W. Tress, S. Meloni, C. Yi, *Joule* **2024**, *8*, 1691; h) C. Fu,

Z. Gu, Y. Tang, Q. Xiao, S. Zhang, Y. Zhang, Y. Song, *Angew. Chem., Int. Ed.* **2022**, *61*, 202117067.

[22] a) J. Jiao, C. Yang, Z. Wang, C. Yan, C. Fang, *Results Eng.* **2023**, *18*, 101158; b) M. Jung, S. G. Ji, G. Kim, S. I. Seok, *Chem. Soc. Rev.* **2019**, *48*, 2011.

[23] M. Jeong, I. W. Choi, E. M. Go, Y. Cho, M. Kim, B. Lee, S. Jeong, Y. Jo, H. W. Choi, J. Lee, J.-H. Bae, S. K. Kwak, D. S. Kim, C. Yang, *Science* **2020**, *369*, 1615.

[24] a) C. Luo, G. Zheng, F. Gao, X. Wang, C. Zhan, X. Gao, Q. Zhao, *Nat. Photonics* **2023**, *17*, 856; b) N. E. Courtier, J. M. Cave, J. M. Foster, A. B. Walker, G. Richardson, *Energy Environ. Sci.* **2019**, *12*, 396.

[25] a) Q. Chen, H. Zhou, Y. Fang, A. Z. Stieg, T. B. Song, H. H. Wang, X. Xu, Y. Liu, S. Lu, J. You, P. Sun, J. McKay, M. S. Goorsky, Y. Yang, *Nat. Commun.* **2015**, *6*, 7269; b) X. Shen, B. M. Gallant, P. Holzhey, J. A. Smith, K. A. Elmestekawy, Z. Yuan, P. Rathnayake, S. Bernardi, A. Dasgupta, E. Kasparavicius, T. Malinauskas, P. Caprioglio, O. Shargaieva, Y. H. Lin, M. M. McCarthy, E. Unger, V. Getautis, A. Widmer-Cooper, L. M. Herz, H. J. Snaith, *Adv. Mater.* **2023**, *35*, 2211742.

[26] S. Li, Y. Jiang, J. Xu, D. Wang, Z. Ding, T. Zhu, B. Chen, Y. Yang, M. Wei, R. Guo, Y. Hou, Y. Chen, C. Sun, K. Wei, S. M. H. Qaid, H. Lu, H. Tan, D. Di, J. Chen, M. Gratzel, E. H. Sargent, M. Yuan, *Nature* **2024**, *635*, 82.

[27] X. Zuo, B. Kim, B. Liu, D. He, L. Bai, W. Wang, C. Xu, Q. Song, C. Jia, Z. Zang, D. Lee, X. Li, J. Chen, *Chem. Eng. J.* **2022**, *431*, 133209.

[28] L. Yang, H. Zhou, Y. Duan, M. Wu, K. He, Y. Li, D. Xu, H. Zou, S. Yang, Z. Fang, S. Liu, Z. Liu, *Adv. Mater.* **2023**, *35*, 2211545.

[29] T. Pan, Z. Li, B. Ren, W. Yang, X. Ran, Y. Li, Y. Xu, Y. Wang, D. Li, Y. Xia, X. Gao, L. Chao, Y. Chen, *Energy Environ. Sci.* **2024**, *17*, 9548

[30] Z. Qu, Y. Zhao, F. Ma, L. Mei, X. K. Chen, H. Zhou, X. Chu, Y. Yang, Q. Jiang, X. Zhang, J. You, *Nat. Commun.* **2024**, *15*, 8620.

[31] J.-J. Cao, Y.-H. Lou, W.-F. Yang, K.-L. Wang, Z.-H. Su, J. Chen, C.-H. Chen, C. Dong, X.-Y. Gao, Z.-K. Wang, *Chem. Eng. J.* **2022**, *433*, 133832.

[32] C. Gong, X. Chen, J. Zeng, H. Wang, H. Li, Q. Qian, C. Zhang, Q. Zhuang, X. Yu, S. Gong, H. Yang, B. Xu, J. Chen, Z. Zang, *Adv. Mater.* **2024**, *36*, 2307422.

[33] Y. Zhang, Y. Zhang, B. Niu, Y. Huang, H. Wu, W. Fu, H. Chen, *Adv. Funct. Mater.* **2023**, *33*, 2307949.

[34] S. Liu, J. Li, W. Xiao, R. Chen, Z. Sun, Y. Zhang, X. Lei, S. Hu, M. Kober-Czerny, J. Wang, F. Ren, Q. Zhou, H. Raza, Y. Gao, Y. Ji, S. Li, H. Li, L. Qiu, W. Huang, Y. Zhao, B. Xu, Z. Liu, H. J. Snaith, N. G. Park, W. Chen, *Nature* **2024**, *632*, 536.

[35] Y. Zhao, F. Ma, Z. Qu, S. Yu, T. Shen, H.-X. Deng, X. Chu, X. Peng, Y. Yuan, X. Zhang, J. You, *Science* **2022**, *377*, 531.

[36] G. Yang, H. Lei, H. Tao, X. Zheng, J. Ma, Q. Liu, W. Ke, Z. Chen, L. Xiong, P. Qin, Z. Chen, M. Qin, X. Lu, Y. Yan, G. Fang, *Small* **2017**, *13*, 1601769.

[37] Y. Bai, Y. Fang, Y. Deng, Q. Wang, J. Zhao, X. Zheng, Y. Zhang, J. Huang, *ChemSusChem* **2016**, *9*, 2686.

[38] P. Xu, H. He, J. Ding, P. Wang, H. Piao, J. Bao, W. Zhang, X. Wu, L. Xu, P. Lin, X. Yu, C. Cui, *ACS Appl. Energy Mater.* **2021**, *4*, 10921.

[39] A. D. Jodlowski, C. Roldán-Carmona, G. Grancini, M. Salado, M. Ralaiarisoa, S. Ahmad, N. Koch, L. Camacho, G. de Miguel, M. K. Nazeeruddin, *Nat. Energy* **2017**, *2*, 972.

[40] Q. Jiang, J. Tong, Y. Xian, R. A. Kerner, S. P. Dunfield, C. Xiao, R. A. Scheidt, D. Kuciauskas, X. Wang, M. P. Hautzinger, R. Tirawat, M. C. Beard, D. P. Fenning, J. J. Berry, B. W. Larson, Y. Yan, K. Zhu, *Nature* **2022**, *611*, 278.

[41] M. A. Mahmud, T. Duong, Y. Yin, H. T. Pham, D. Walter, J. Peng, Y. Wu, L. Li, H. Shen, N. Wu, N. Mozaffari, G. Andersson, K. R. Catchpole, K. J. Weber, T. P. White, *Adv. Funct. Mater.* **2020**, *30*, 1907962.

[42] a) H. Chen, Q. Wei, M. I. Saidaminov, F. Wang, A. Johnston, Y. Hou, Z. Peng, K. Xu, W. Zhou, Z. Liu, L. Qiao, X. Wang, S. Xu, J. Li, R. Long, Y. Ke, E. H. Sargent, Z. Ning, *Adv. Mater.* **2019**, *31*, 1903559; b) H. Meng, Z. Shao, L. Wang, Z. Li, R. Liu, Y. Fan, G. Cui, S. Pang, *ACS Energy Lett.* **2019**, *5*, 263.

[43] a) H. Luo, J. Wu, Q. Liu, J. Huang, Y. Tu, J. Lin, M. Huang, *Energy Technol.* **2017**, *5*, 1814; b) J. Zhang, R. Chen, Y. Wu, M. Shang, Z. Zeng, Y. Zhang, Y. Zhu, L. Han, *Adv. Energy Mater.* **2017**, *8*, 1701981.

Supporting Information

Enhancing Power Conversion Efficiency of Perovskite Solar Cells through Machine Learning Guided Experimental Strategies

Antai Yang[a,c,#], Yonggui Sun[d,#], Jingzi Zhang[a,c,#], Fei Wang[d], Chengquan Zhong[a,c], Chen Yang[e], Hanlin Hu[d], Jiakai Liu[e,f,g*], and Xi Lin[a,b,c*]*

[a]*School of Materials Science and Engineering, Harbin Institute of Technology, Shenzhen 518055, P. R. China*

[b]*State Key Laboratory of Advanced Welding and Joining, Harbin Institute of Technology, Harbin 150001, P. R. China*

[c]*Blockchain Development and Research Institute, Harbin Institute of Technology, Shenzhen 518055, P.R. China*

[d]*Hoffmann Institute of Advanced Materials, Shenzhen Polytechnic University, 7098 Liuxian Boulevard, Shenzhen 518055, China.*

[e]*Laboratory of Environmental Sciences and Technology, Xinjiang Technical Institute of Physics & Chemistry, Chinese Academy of Sciences, Urumqi 830011, China*

[f]*Center of Materials Science and Optoelectronics Engineering, University of Chinese Academy of Sciences Beijing 100049, China*

[g]*Sunrise (Xiamen) Photovoltaic Industry Co. Ltd., 44 Huli Avenue, Huli Distric, Xiamen, Fujian 361000, P.R.China*

Corresponding authors

*E-mail: linxi@hit.edu.cn

*E-mail: liu.jiakai@kaust.edu.sa

*E-mail: hanlinhu@szpu.edu.cn

Dataset preparation:

The dataset for ML was composed of 2998 data points from 1149 articles published between 2013 and 2023, with PCE ranging from 0.1% to 26.1%. It encompassed both the n-i-p and p-i-n structures of single-junction perovskite solar cells. Among them, the data points from 2013 to 2020 were extracted from Odabaşı's dataset and She's dataset [1], while the data for 2021 to 2023 were newly collected from the 170 kinds of literature by conducting comprehensive searches using the keywords 'Perovskite', 'Efficiency', and 'PCE'. Furthermore, four essential factors that were previously omitted in databases have been incorporated due to their significant impact: additives, HTL-side passivator, ETL-side passivator, and anion Cl. [2] The inclusion of Cl anions was examined as a separate influencing factor, as it was often overlooked in the chemical composition due to the volatilization during the crystallization process. [3]

After going through a series of strict cleaning rules, it was further limited to 2079 data points. Specifically, the dataset underwent the cleaning process adhering to the following rules: (1) Data points measured under non-standard test conditions were eliminated. (2) Only data from single junction perovskite batteries were retained, after removing data points that were unrelated to the research focus. (3) Data points with incomplete information for one or more input features were also eliminated. (4) The stable PCE values were adjusted to retain only the highest recorded values for PSCs in the dataset. (5) The chemical formula components were re-examined, with characterized components replaced with the corresponding proportions of precursor ions. (6) In cases where data points shared identical input features but exhibited varying

reported PCE, the mean PCE was derived and preserved as a new data point. This calculation considered the experimental environment and measurement variances inherent to different laboratories.

Feature generation:

Label encoder can be used to map different categorical variables to continuous integers, which assign a unique numerical value to each class in a categorical variable.

The one-hot encoder can convert categorical variables into binary vectors, where each category is represented by a binary value (0 or 1), retaining more information about the categories. They allow us to represent the target variable in a numeric format, enabling compatibility with various ML algorithms.

The 8 ionic ratio features of MA^+ (methylammonium), FA^+ (formamidinium), Cs^+, Rb^+, Pb^{2+}, Sn^{2+}, Br^-, and I^- were extracted from the perovskite components factor. The bandgap feature was predicted using the model developed by Gok et al.,[4] with the performance shown in **Figure S1**. The remaining 13 features were encoded by label encoding method from the other 13 experimental factors (ETL, ETL-2, deposition procedure, deposition method, antisolvent, precursor solution, HTL, HTL-additive, ETL-passivator, HTL-passivator, additives, add-Cl, type), representing alternatives in sequence by integers starting from 0. Finally, 22 features listed in **Table S1** were generated through the feature engineering process.

Model selection:

Linear Regression algorithm (LR) is effective for small data volumes and simple relationships due to its simple idea, easy implementation, and fast modeling. [5]

Support Vector Regression (SVR) is a regression algorithm based on Support Vector Machines, which has superior performance in handling small samples, nonlinear regression, and high-dimensional data, and has strong robustness to outliers. [6]

Artificial Neural Networks (ANN) can self-learn and adapt to changes in data, effectively recognize complex patterns, and exhibit stability when dealing with incomplete or noisy data. Their parallel processing capability and versatility make them highly applicable in fields such as image recognition, speech recognition, and natural language processing. [7]

The Decision Tree algorithm (DT) is based on a tree structure for decision-making and has advantages in small, complex, and high-dimensional data. [8]

Random Forest (RF) is an ensemble learning algorithm based on decision trees that constructs multiple decision trees by randomly selecting subsets of data and features and makes final predictions by voting or averaging the predicted results. It has strong processing capabilities for categorical features and can automatically handle issues such as missing values, outliers, and encoding of categorical features. [9]

LightGBM (LGBM) uses a histogram-based decision tree splitting algorithm and gradient-based one-sided sampling technology to speed up the training process and reduce memory usage, and is known for its efficiency and speed. [10]

XGBoost is an optimized algorithm based on gradient-boosting decision trees that iteratively trains a series of decision trees and gradually optimizes the predictive ability of the model through gradient descent methods. It has unique advantages in feature

selection, regularization, and efficient computation, and can automatically learn the importance and interaction of features, making it perform well in various tasks. [11]

CatBoost is an open-source machine learning algorithm based on gradient boosting decision trees that introduces a learning algorithm based on symmetric random gradient boosting. Through a progressive learning strategy, it gradually increases the number of decision trees during training, effectively combating the risk of overfitting while ensuring computational speed. [12]

Model evaluation:

The correlation coefficient (r) is a statistical measure used to assess the linear correlation between predicted values and true values, which can help judge the fitting effect of the model on the data and the reliability of the prediction. The range is from -1 to 1, where the closer it is to 1 or -1, the more accurate the prediction.1 indicates a complete positive correlation, -1 indicates a complete negative correlation, and 0 indicates no linear relationship.

The coefficient of determination (R^2) represents the proportion of variance in the dependent variable that can be explained by the model and is used to evaluate the degree of model fitting. Its range is from 0 to 1, with values closer to 1 indicating better fitting effects and better explanation of variation in observed data.

Root mean square error (RMSE) is a commonly used indicator to measure model prediction errors and can indicate the average level of prediction errors of the model. Smaller values indicate higher prediction accuracy.

The mean absolute error (MAE) reflects the average level of absolute error between the predicted value and the real value. It gives equal weight to all prediction errors, making it an intuitive, interpretable, and relatively robust error measure for outliers.

The calculation formulas are given below. Where $\hat{y}_i$ is the predicted value of the model, y_i is the true value, $\bar{y}$ is the average value, and $\bar{\hat{y}}$ represents the average predicted value.

$$r=\frac{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)\left(\hat{y}_i-\bar{\hat{y}}\right)}{\sqrt{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2\sum_{i=1}^{n}\left(\hat{y}_i-\bar{\hat{y}}\right)^2}}$$

$$R^2=1-\frac{\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2}$$

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2}$$

$$MAE=\frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i-y_i\right|$$

Model application:

To optimize an existing device using machine learning methods, the steps were followed: Identify the experimental factors to be optimized. Fix the other experimental factors of the device in the model's input. Then, sequentially generate and combine alternatives for the factors to be optimized with the original device settings. Use the pre-trained model to predict the outcomes for these combinations. List the alternatives whose predicted performance exceeds that of the original device, and subsequently validate these alternatives through experimental testing.

The schemes were identified through the combination of different experimental factors and machine learning predictions. By predicting the rehearsal combinations of unexplored factors, the exploration direction was to find the scheme that can improve PCE. According to the conclusion of SHAP analysis, "Li+TBP" and "Without HTL passivator" will respectively lead to a decrease in PCE. Therefore, HTL- passivator and HTL-additive were selected as the experimental factors to be optimized. Specifically, we fixed other factors and optimized the HTL-passivator factor locally. We sequentially combined the alternatives of the HTL-passivator with the other factors of device *Control* arrangement and predicted possible PCE values. We found that when PEAI was chosen as the HTL-passivator, the predicted PCE value was increased as **Figure S7a** shows. Then, we conducted experimental verification as Regular 2. After that, using the same method, we locally optimized the HTL-additive factor based on the previous group (Regular 2) and found that when CuSCN was chosen as the HTL-additive, the predicted value of PCE further was increased. Then, we conducted experimental verification as Regular 3. Similarly, the optimization steps for inverted devices were also predicted and verified through the process shown in **Figure S7b**.

Statistical Analysis

Before training the model of LR, ANN, and SVR, the input data ought to be standardized, which was done through the StandardScaler algorithm in Scikit-learn. The test data and training data should be evaluated at the same scale to avoid data leakage and overfitting. For the statistical analysis of perovskite solar cells' experimental data, we prepared 7 cells under the same parameters to conduct statistical

analysis, then used box plots (min to max) and mean with standard deviations (mean ± SD) to statistically analyze and quantify variability and data shifts.

Materials and preparation

ITO glass substrates with a sheet resistance of ca. 9 Ω sq^{-1} were purchased from OPVTECH Inc. Formamidinium Iodide (FAI), methylammonium chloride (MACl), methylammonium bromide (MABr) were purchased from Xi'an Polymer Light Technology Crop. PbI_2 (99.8%), bis (trifluoromethane) sulfonimide lithium salt (Li-TFSI, 99%), 4-tert-butylpyridine (tBP, 96%), and CsI (99.99%) were supplied from Sigma-Aldrich. Tin (IV) oxide was purchased from Alfa Aesar. DMF, DMSO, and IPA were purchased from TCI. Spiro-OMeTAD (purity. 99.5%) was purchased from Feiming Science and Technology Co., Ltd. All chemicals were used as received without further treatment.

Device fabrication

Regular structure: ITO substrates were sequentially rinsed by sonication in detergent, deionized (DI) water, and isopropanol for 15 min, respectively, and then dried under nitrogen gas. The ITO substrates were disposed of by ultraviolet-ozone for 30 min and followed by deposition of Tin (IV) oxide solution on the substrate via spin-coating at 4000 rpm for 30 s, and subsequently annealed on a hotplate at 150 °C for 30 min. The perovskite layer was fabricated in the glovebox through a modified two-step sequential method according to the literature. First, 1.4 M PbI_2 precursor was dissolved in 950 μL DMF and 50 μL CsI solution (1.5 M, 390 mg CsI was dissolved in 1 mL DMSO). The above solution was then spin-coated on the SnO_2/ITO substrate at 1500

rpm for 30 s, and dried at 70 °C for 1 min. Thereafter, a mixture solution FAI:MAI:MABr: MACl (FAI/MAI/MABr/MACl with different ratios in 1 mL IPA) was dropped on the PbI_2 film at 1500 rpm for 30 s. The as-cast perovskite film was annealed at 150 °C for 15 min under 30-40% relative humidity. For PEAI treatment, the PEAI solution was dissolved in IPA with 30mM and spin-coated onto the perovskite surface at a spin rate of 5000 rpm. The Spiro-OMeTAD solution was composed of 72.3 mg Spiro-OMeTAD, 30 μL TBP, and 35 μL Li-TFSI solution (260 mg in 1 mL acetonitrile) in 1 mL chlorobenzene, mixing x mol% CuSCN (40 mg in 1 mL propyl sulfide, x = 0 or 33%) and then spin-coated on perovskite film at 4000 rpm for 30 s. Finally, a 100 nm Au electrode was deposited by thermal evaporation.

Inverted structure: ITO substrates were sequentially rinsed by sonication in detergent, deionized (DI) water, and isopropanol for 15 min, respectively, and then dried under nitrogen gas. Cleaned ITO substrates were treated with ultraviolet-ozone for 30 min, after which a SAM layer was deposited on the substrate by spin-coating the SAM Me-4PACz in IPA solution at 4000 rpm for 30 s and annealing it for 10 min at 100 °C on a hotplate. Perovskite precursor solution (21 mg CsI, 824 mg FAI, 14 mg of MACl, 26 mg MAI, and 824 mg PbI2 in 1 ml DMF: DMSO (4:1 in volume) mixed solvent was spin-coated on the as-prepared SAM/ITO substrates at 5,000 r.p.m. for 40 s, during which 300 μL CB was dripped onto the center of film at 10 s before the end of spin-coating and annealing it at 100 °C for 30 min. For the 4Cl-BZS treating device, 1mg 4Cl-BZS was added to the perovskite precursor. A Phenyl-C61-butyric acid methyl ester (PCBM) solution (20 mg/mL in CB) was dynamically spun onto

perovskite layers at a speed of 2000 rpm for 30 s. The samples were then annealed at 100 °C for 1 min. After cooling down to room temperature, a bathocuproine (BCP) solution (1 mg/mL in IPA) was spun onto the PCBM layer at a speed of 5000 rpm for 30 s, followed by a brief thermal annealing process at 100°C for ~1 min. Finally, an 80 nm Ag electrode was deposited by thermal evaporation.

Characterization

The *J–V* characteristics of the PSCs devices were carried out using an IVS-KA6000 Enlitech sunlight simulator equipped with an AM 1.5 filter at 100 mW cm^{-2} and Keithley SMU source after correcting the light intensity with a standard calibration cell.

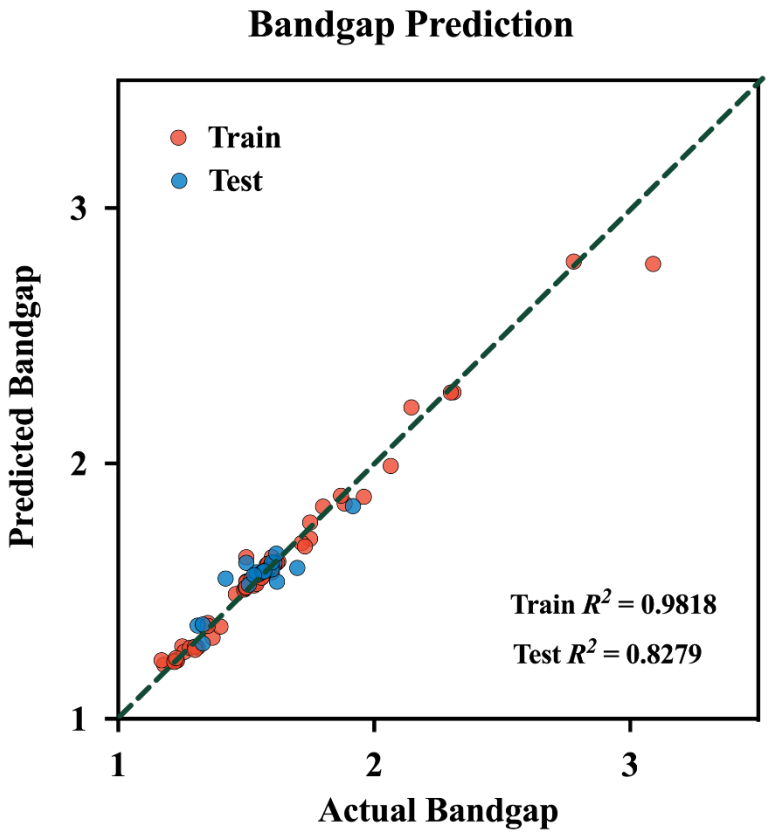

Figure S1 The prediction effect of bandgaps using random forests, with data and methods provided by Gok et al. [4]

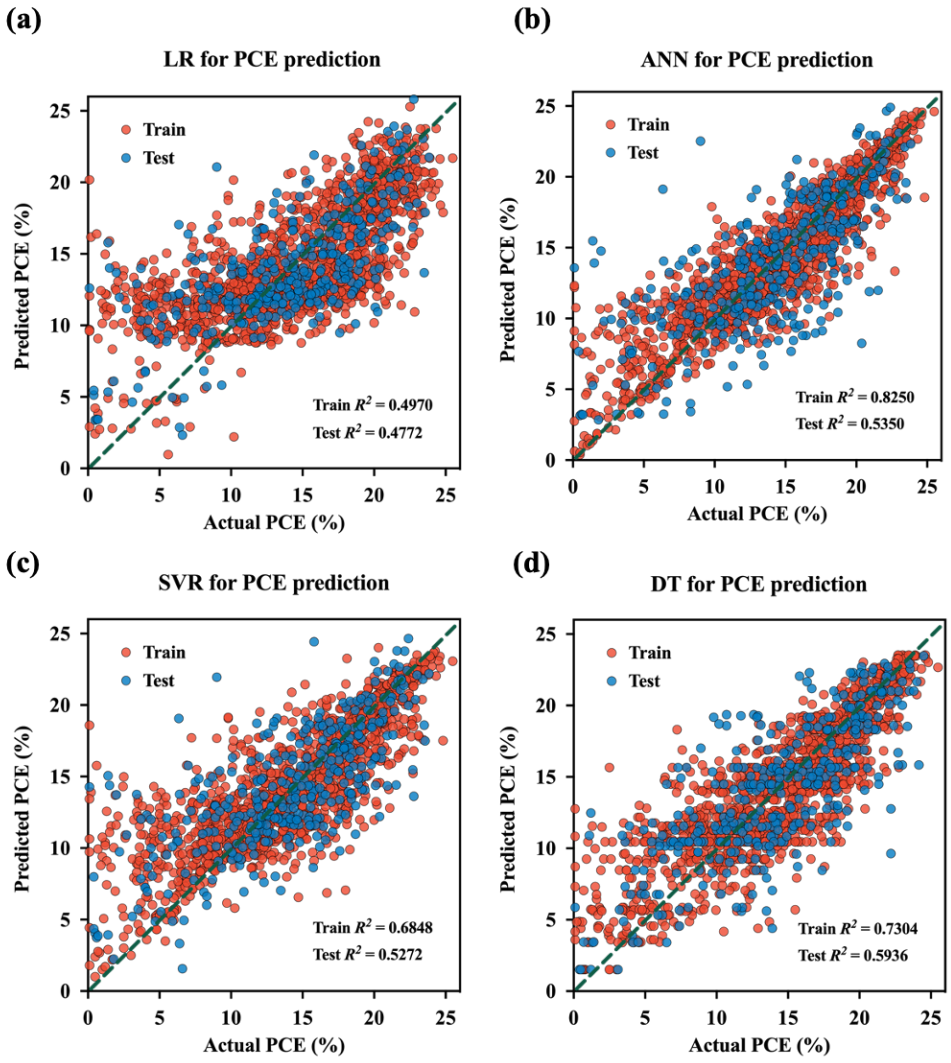

Figure S2 The fitting graph of PCE results using four different classic algorithms for PCE prediction, where red represents the training set and blue represents the test set: (a) Linear Regression (LR), (b) Artificial Neural Network (ANN), (c) Support Vector Regression (SVR), (d) Decision Tree (DT).

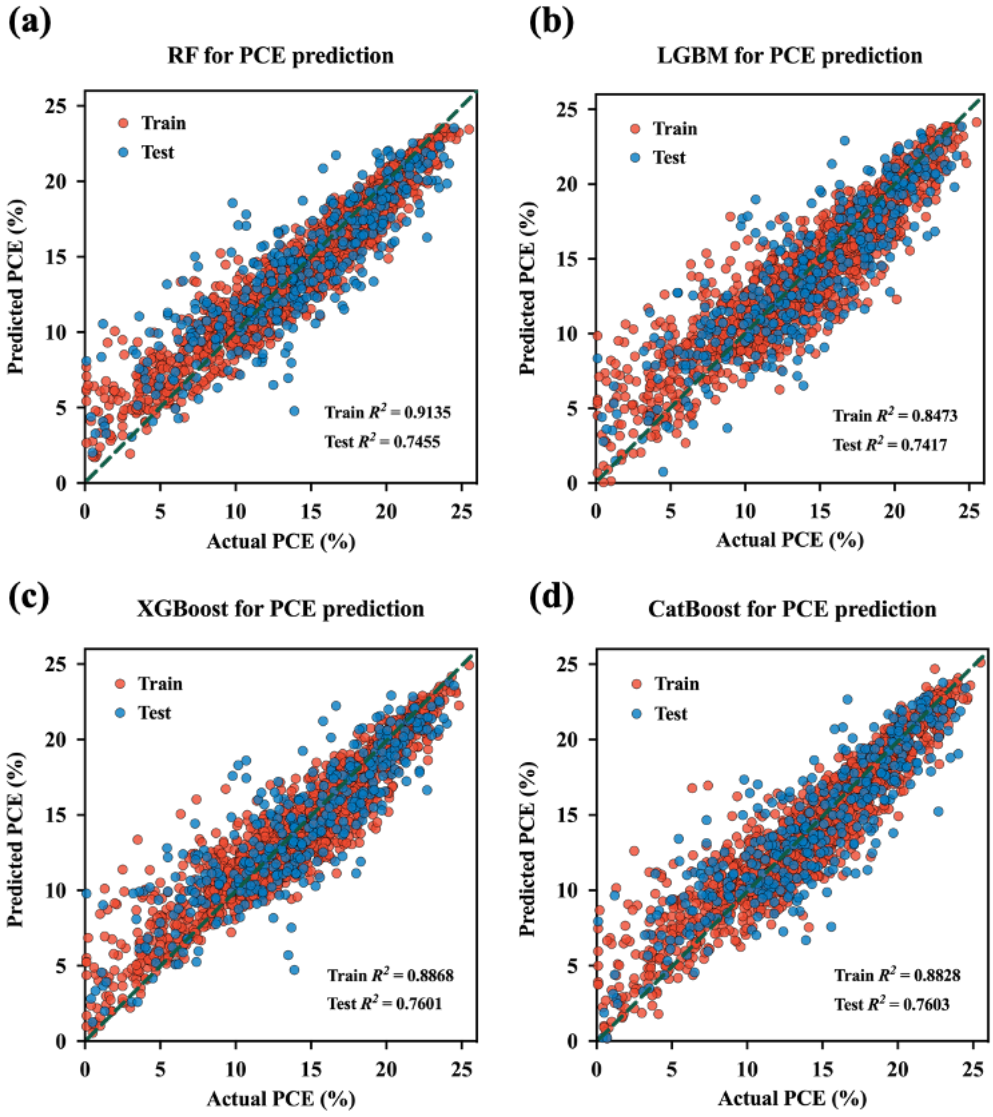

Figure S3 The fitting graph of PCE results using different ensemble learning algorithms, where red represents the training set and blue represents the test set: (a) Random Forest (RF), (b) LightGBM (LGBM), (c) XGBoost, and (d) CatBoost.

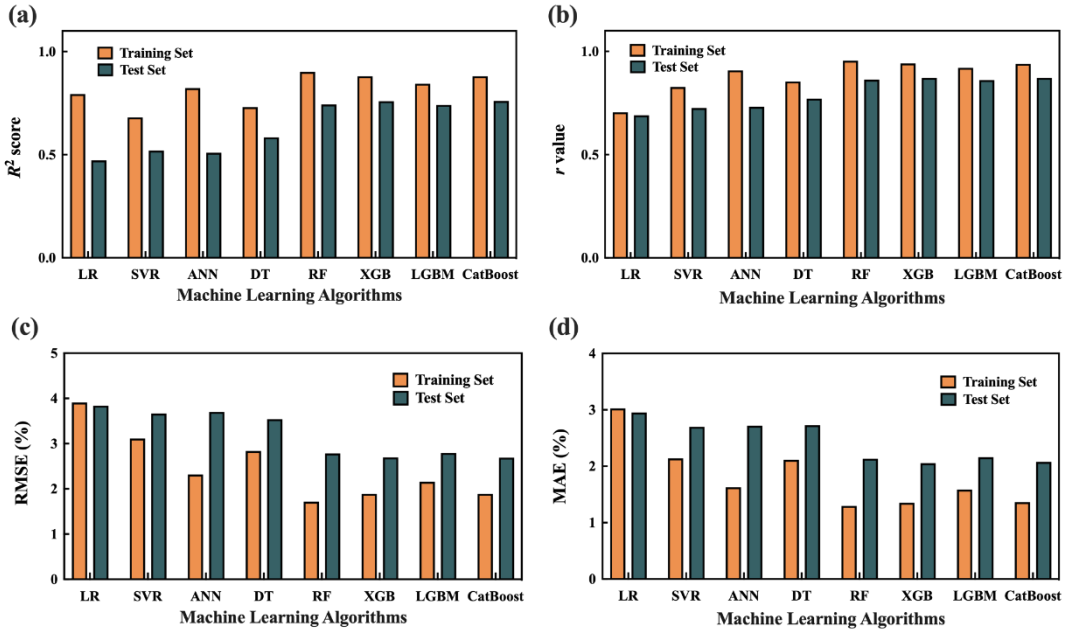

Figure S4 Performance of 8 ML models: (a) R^2 score, (b) *r* value, (c) RMSE, (d) MAE.

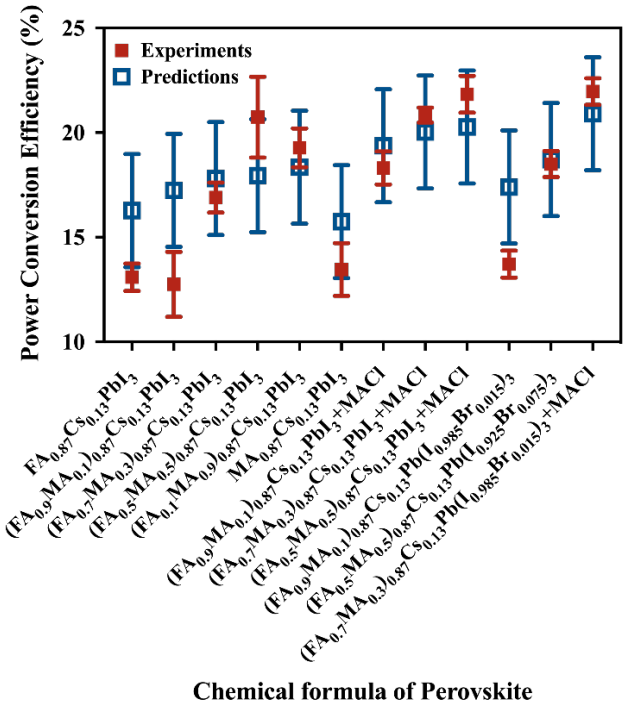

Figure S5 Comparison between experimental PCE values (mean ± SD) and predicted PCE values from ML Model (predict value ± RMSE)

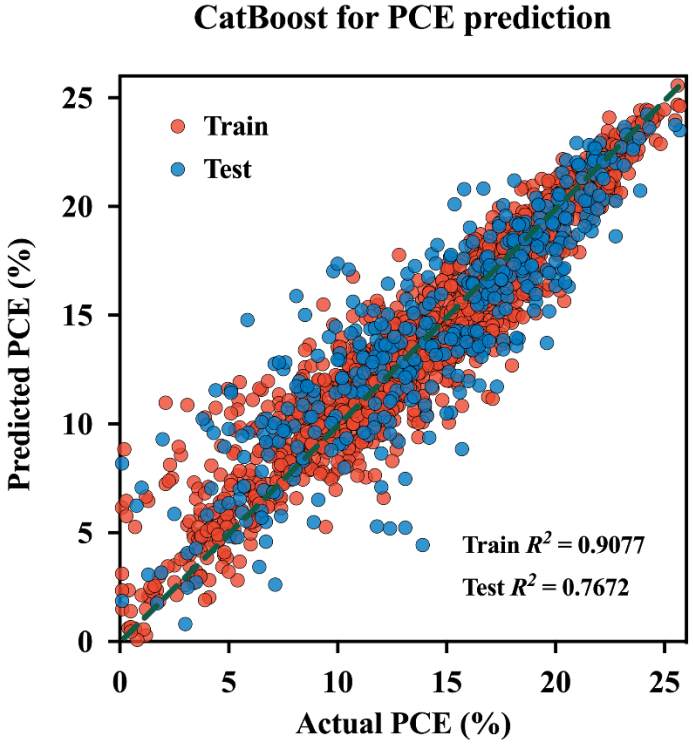

Figure S6 The performance of the CatBoost model corrected by experimental data on the training and testing sets.

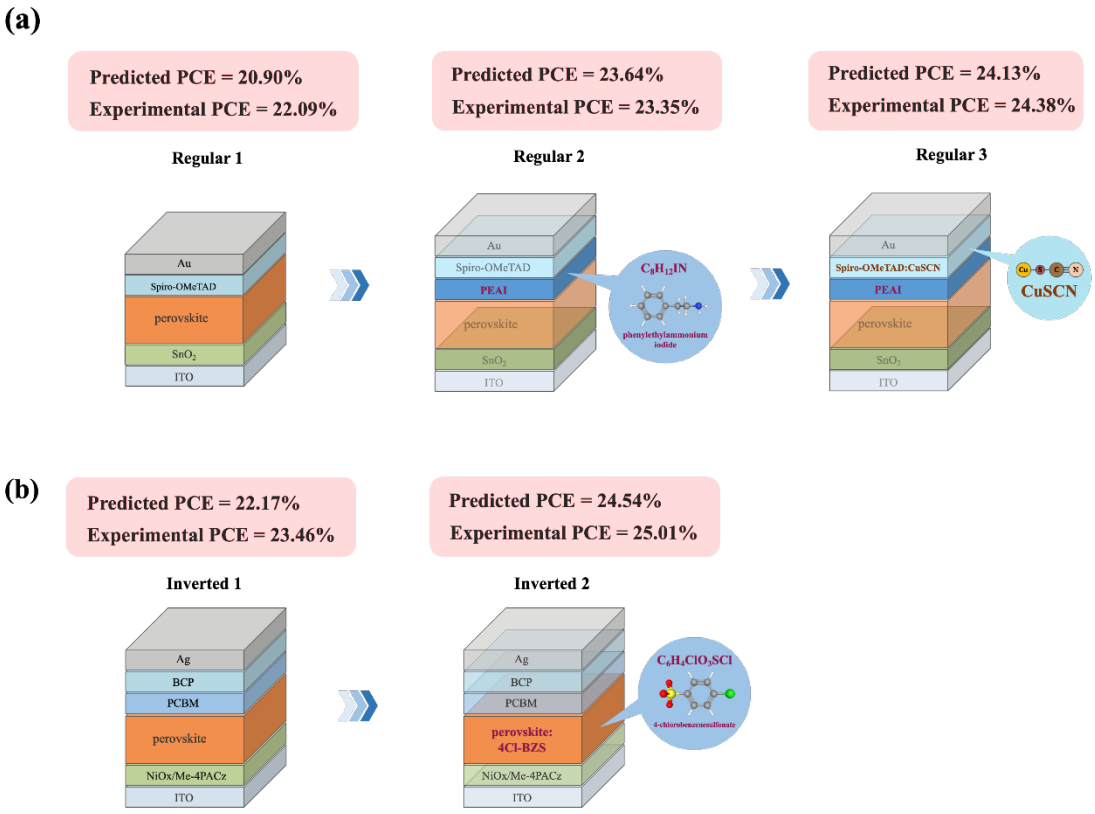

Figure S7 Strategies and predicted PCE values for improving our device's PCE provided by the ML model: (a) Regular device. (b) Inverted device.

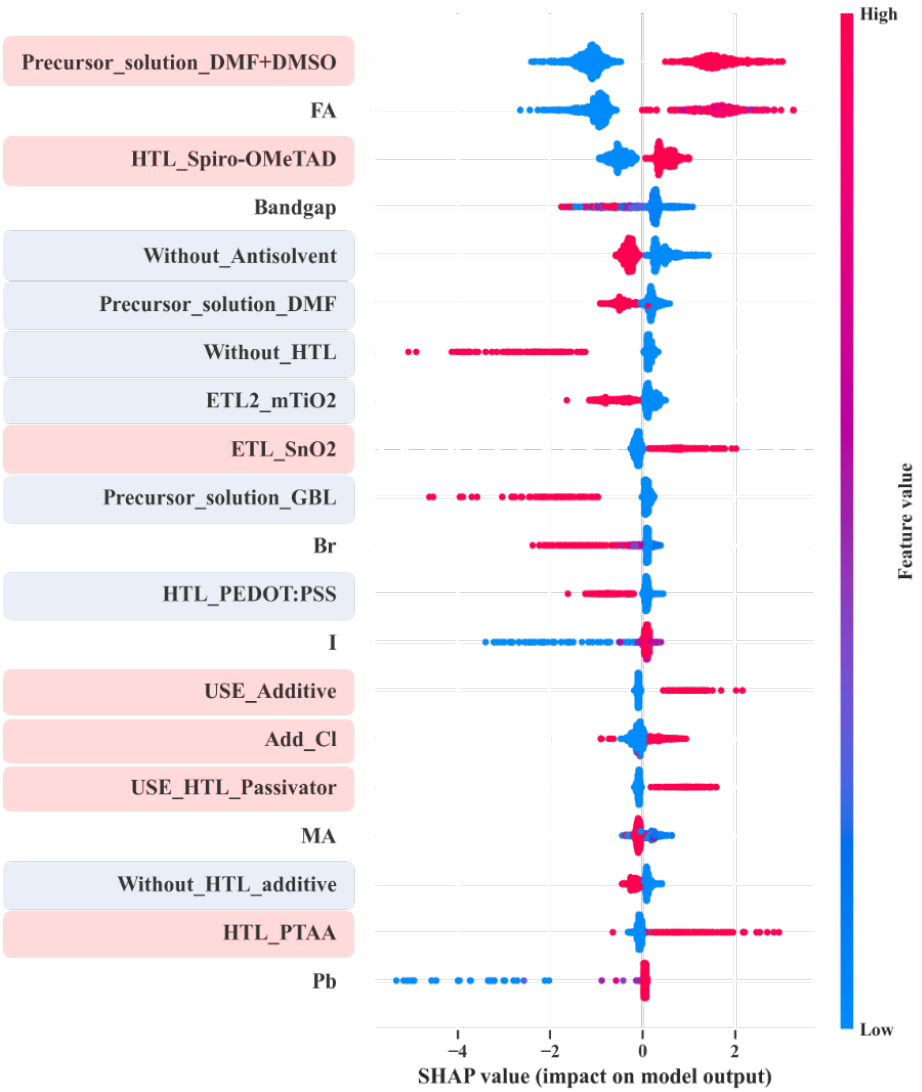

Figure S8 The average SHAP values of the four algorithms using one hot feature encoding represent the impact of the top 20 features on the contribution of PCE, with a pink background representing positive contribution and a light blue background representing negative contribution. The red dots represent the data using this option, while the blue dots indicate not using it.

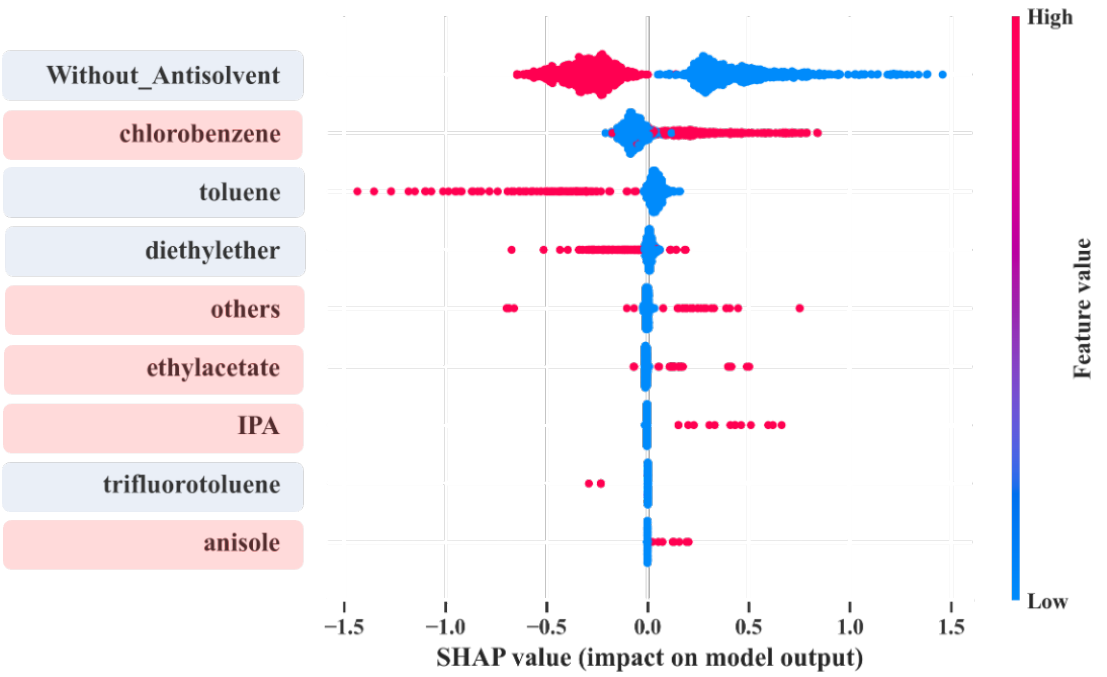

Figure S9 The ranked importance of anti-solvent and the impact of alternatives, with a pink background representing positive contribution and a light blue background representing negative contribution. The red dots represent the data using this option, while the blue dots indicate not using it.

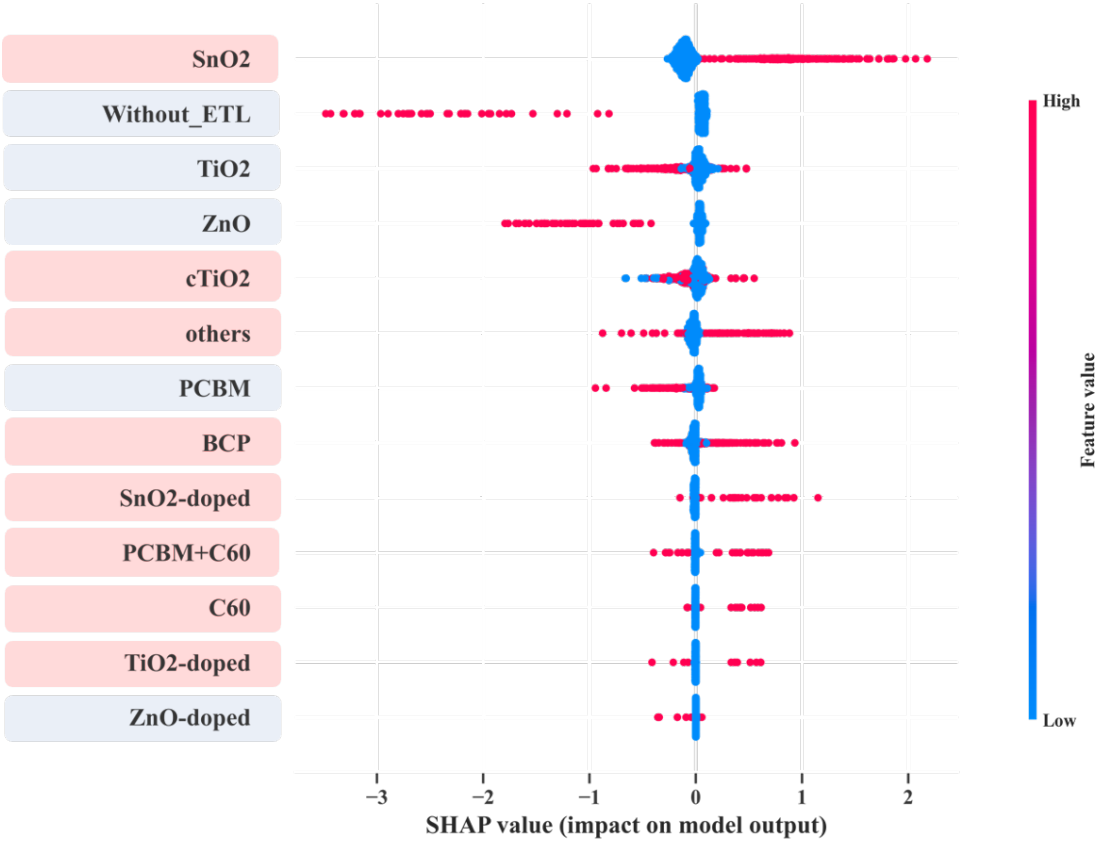

Figure S10 The ranked importance of ETL and the impact of alternatives, with a pink background

representing positive contribution and a light blue background representing negative contribution. The red dots represent the data using this option, while the blue dots indicate not using it.

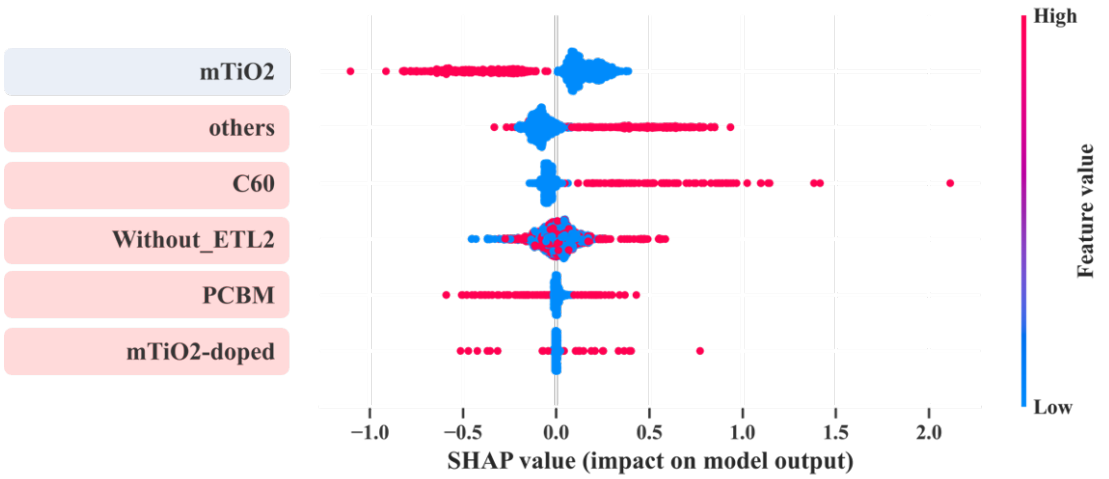

Figure S11 The ranked importance of ETL-2 and the impact of alternatives, with a pink background representing positive contribution and a light blue background representing negative contribution. The red dots represent the data using this option, while the blue dots indicate not using it.

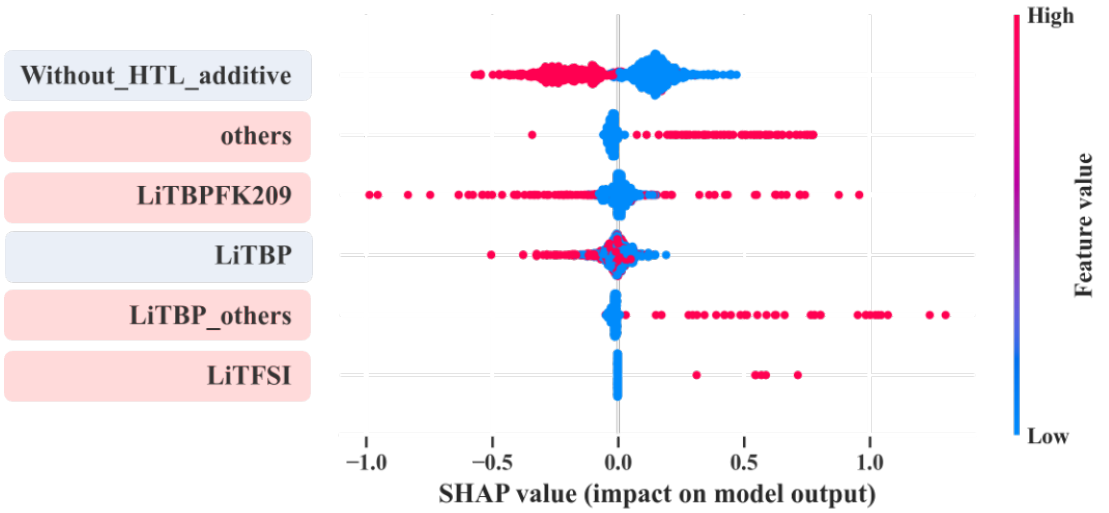

Figure S12 The ranked importance of HTL-additive and the impact of alternatives, with a pink background representing positive contribution and a light blue background representing negative contribution. The red dots represent the data using this option, while the blue dots indicate not using it.

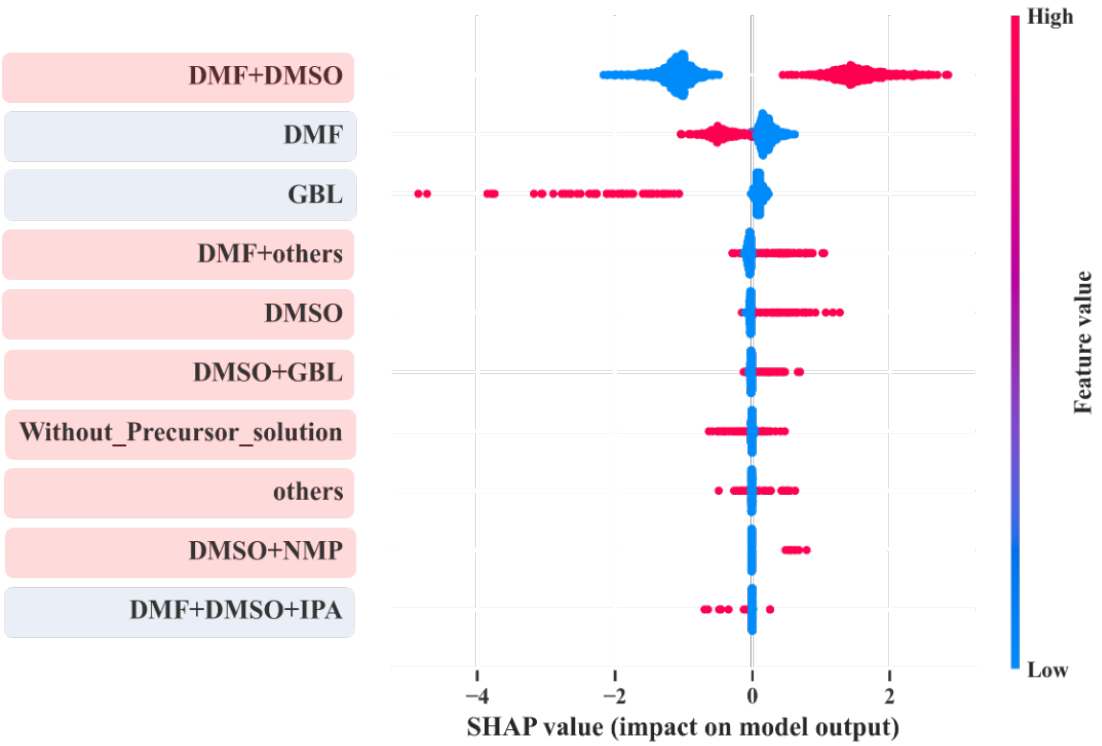

Figure S13 The ranked importance of precursor solution and the impact of alternatives, with a pink background representing positive contribution and a light blue background representing negative contribution. The red dots represent the data using this option, while the blue dots indicate not using it.

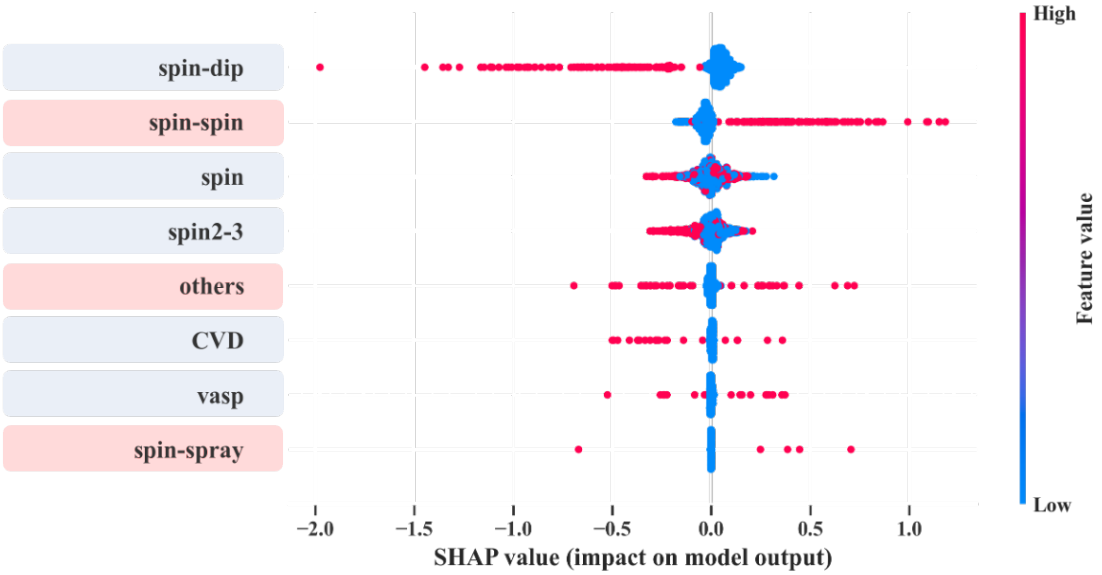

Figure S14 The ranked importance of the deposition method and the impact of alternatives, with a pink background representing positive contribution and a light blue background representing negative contribution. The red dots represent the data using this option, while the blue dots indicate not using it.

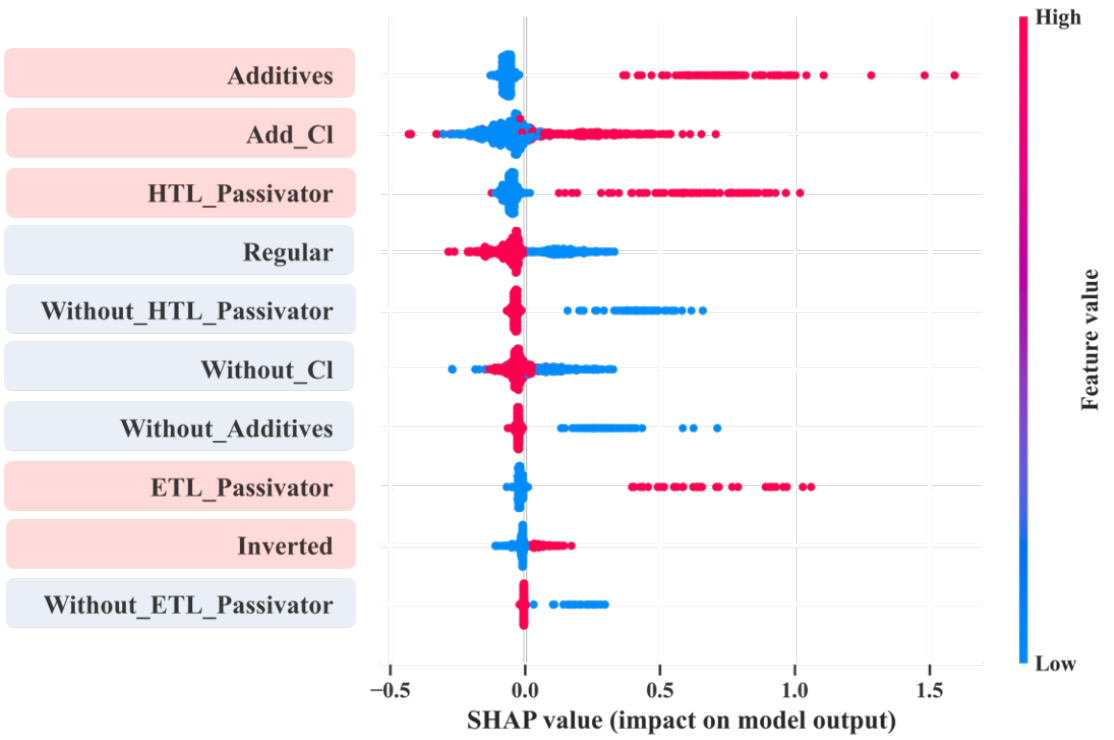

Figure S15 The ranked importance of the ETL-passivator, HTL-passivator, additives, add-Cl, and type, with a pink background representing positive contribution and a light blue background representing negative contribution. The red dots represent the data using this option, while the blue dots indicate not using it.

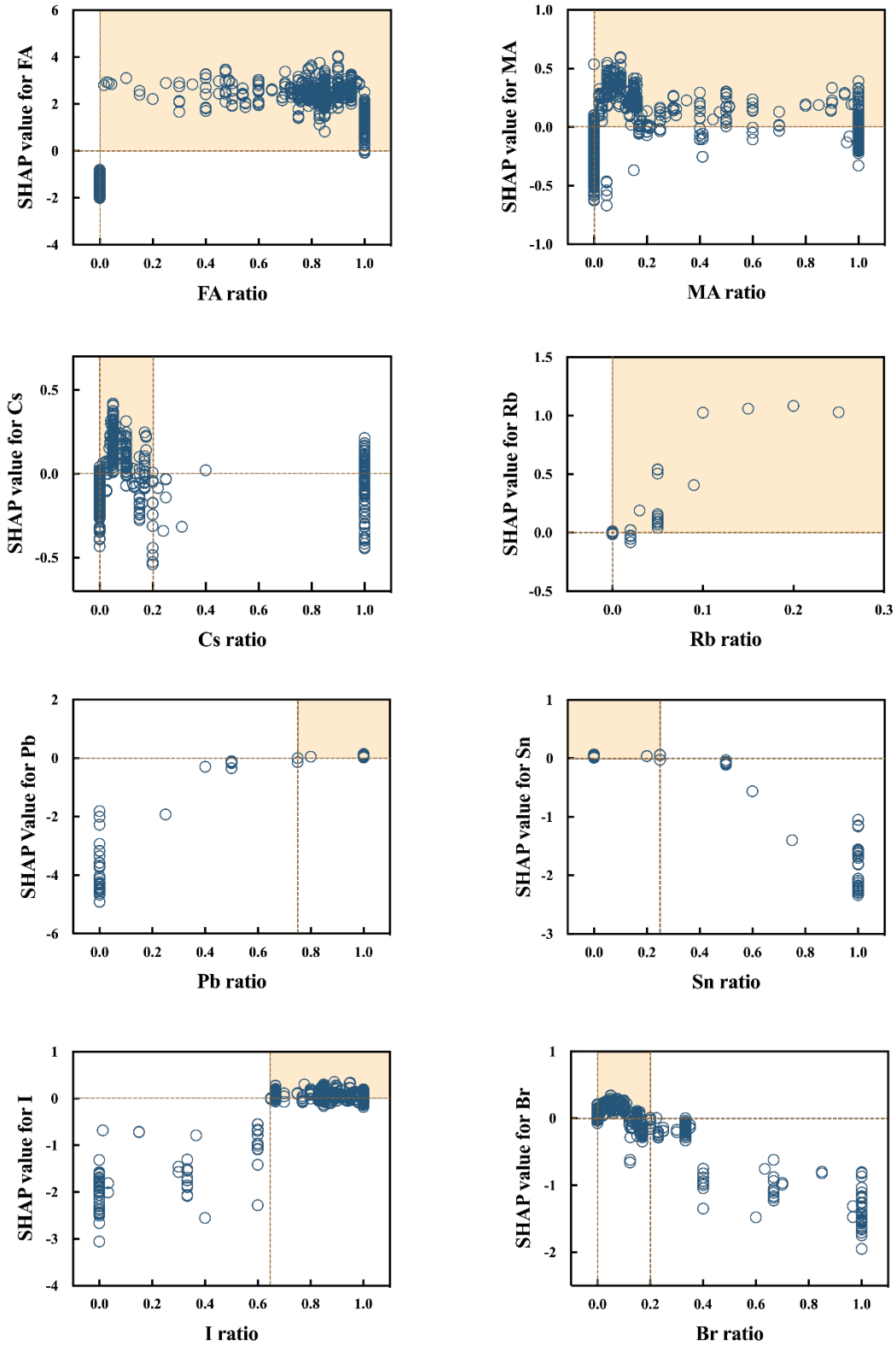

Figure S16 Contributions of different ion ratios to PCE. The horizontal axis in the figure represents the content ratio of different ions, while the vertical axis represents their SHAP values, which represents the contribution to PCE of each data point in the dataset.

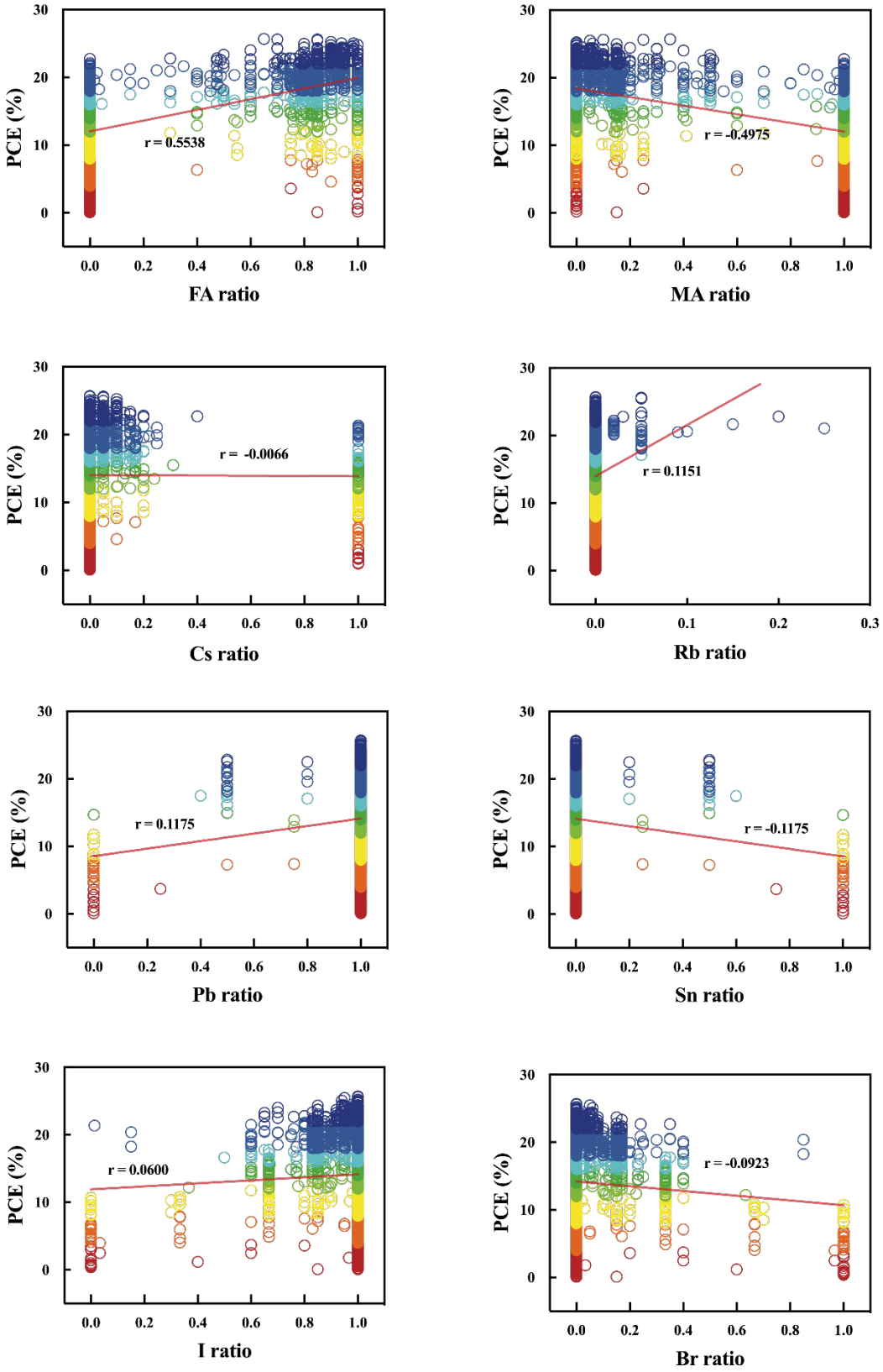

Figure S17 The correlation between different ions and PCE in the dataset. Similar PCE values are represented with similar colors.

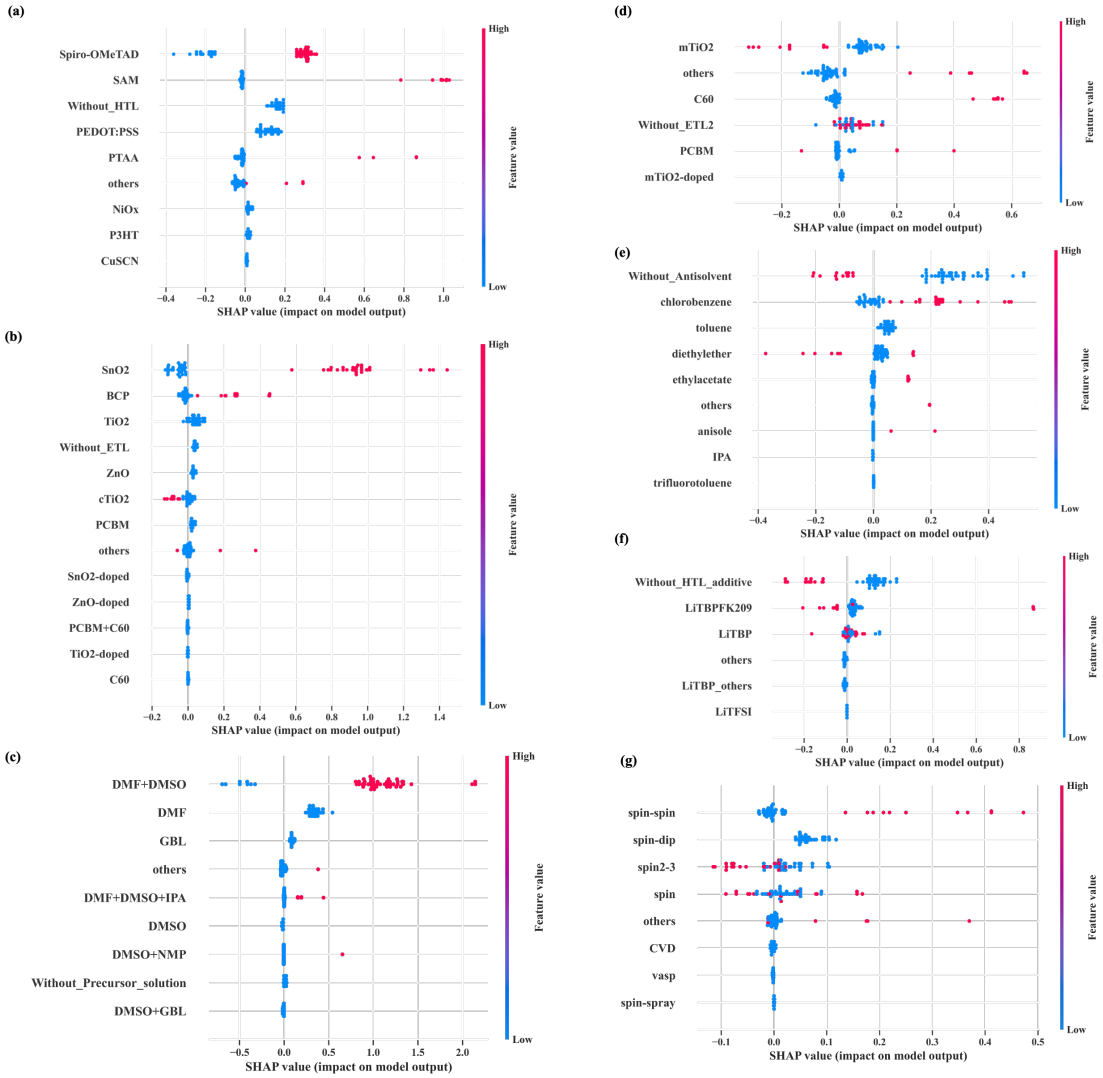

Figure S18 The ranked importance and the impact of alternatives among the top efficiency devices (PCE > 23%). (a) HTL, (b) ETL, (c) Precursor solution, (d) ETL-2, (e) Anti-solvent treatment, (f) HTL additive, and (g) Deposition method.

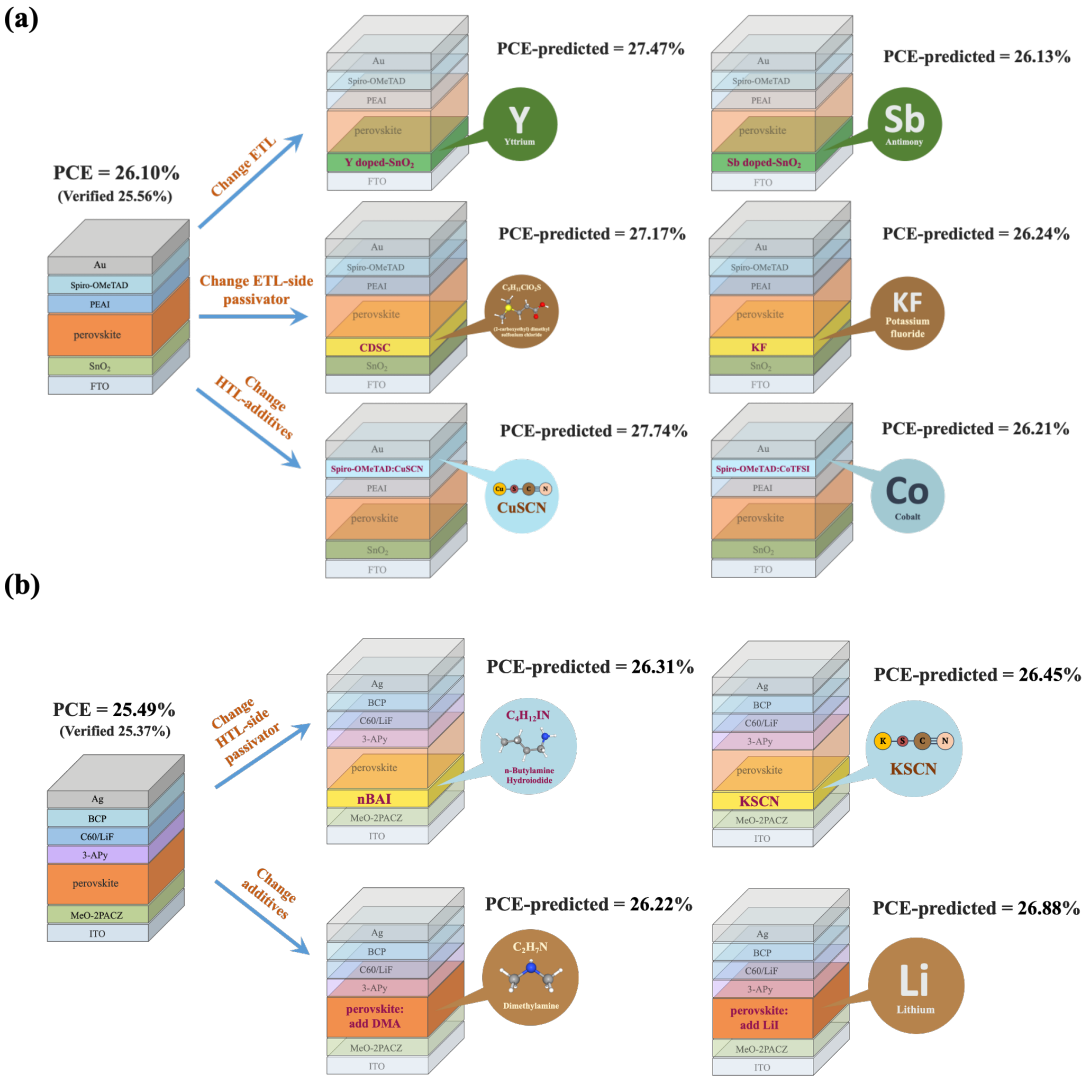

Figure S19 Machine learning solutions of PCE enhancement with different methods: (a) a regular device of You et al.[13], (b) an inverted device of Zhu et al.[14]

Table S1. The feature encoding method and description for 22 features in the dataset.

Num	Feature name and encoding method	Feature description
1	MA ratio (extracted from perovskite components)	Continuous numbers from 0 to 1
2	FA ratio (extracted from perovskite components)	Continuous numbers from 0 to 1
3	Cs ratio (extracted from perovskite components)	Continuous numbers from 0 to 1

4	Rb ratio (extracted from perovskite components)	Continuous numbers from 0 to 1
5	Pb ratio (extracted from perovskite components)	Continuous numbers from 0 to 1
6	Sn ratio (extracted from perovskite components)	Continuous numbers from 0 to 1
7	Br ratio (extracted from perovskite components)	Continuous numbers from 0 to 1
8	I ratio (extracted from perovskite components)	Continuous numbers from 0 to 1
9	Bandgap (predicted from perovskite components by Gok et al.'s model [1])	Continuous numbers
10	ETL (encoded by label-encoder)	Contains 141 alternatives
11	ETL-2 (encoded by label-encoder)	Contains 111 alternatives
12	Perovskite_deposition_procedure (encoded by label-encoder)	Contains 2 alternatives
13	Perovskite_deposition_method (encoded by label-encoder)	Contains 39 alternatives
14	Antisolvent (encoded by label-encoder)	Contains 34 alternatives
15	Precursor_solution (encoded by label-encoder)	Contains 88 alternatives
16	HTL (encoded by label-encoder)	Contains 244 alternatives
17	HTL_additive (encoded by label-encoder)	Contains 61 alternatives
18	ETL_Passivator (encoded by label-encoder)	Contains 33 alternatives
19	HTL_Passivator (encoded by label-encoder)	Contains 76 alternatives
20	Additives (encoded by label-encoder)	Contains 98 alternatives

21	Add_Cl (encoded by label-encoder)	Contains 2 alternatives
22	Type (encoded by label-encoder)	Contains 2 alternatives

Table S2 Different ML algorithms' best parameters in PCE prediction tasks

ML algorithms	Best parameters
LR	/
SVR	'C': 10, 'degree': 2, ' epsilon': 0.5, 'kernel': 'rbf'
ANN	'activation': 'relu', 'alpha': 0.01, 'hidden_layer_sizes': (200,), 'learning_rate': 'constant', 'solver': 'sgd'
DT	'criterion': 'friedman_mse', 'max_depth': 10, 'min_samples_leaf': 4, 'min_samples_split': 2, 'splitter': 'best'
RF	'max_depth': 20, 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 800
XGBoost	'n_estimators': 200, 'learning_rate': 0.05, 'max_depth': 5
LGBM	'n_estimators': 500, 'learning_rate': 0.1, 'num_leaves': 10, 'max_depth': 10,
CatBoost	'depth': 7, 'iterations': 5000, 'l2_leaf_reg': 3, 'learning_rate': 0.09, 'early_stopping_rounds': 1500

Table S3 Comparison of the PCE prediction task.

Year	Features	Data volume	r	R^2	RMSE(%)
2022 [15]	Perovskite family, Device structure, and HTL descriptors	269	0.72	-	3.00
2019 [16]	Perovskite composition, and Device descriptors (ΔH, ΔL, Bandgap)	333	0.80	-	3.23
2022 [17]	Perovskite composition, and Device descriptors (ΔH, ΔL, Hole mobility, Electron mobility)	248	0.86	-	1.58
2023 [18]	Perovskite composition, Material selection, Device structure, Manufacturing methods, Anti-solvent descriptors	1072	0.77	-	1.28
2019 [1a]	Perovskite family, Material selection, Device structure, Manufacturing methods	1408/515	-	-	3.56/3.38
2021 [1b]	Perovskite family, Material selection, Device structure, Manufacturing methods	1820	0.66	0.43	3.81
This work	*Perovskite composition, Material selection, Device structure, Manufacturing methods*	***2079***	***0.87***	***0.76***	***2.63***

Table S4 Comparison of experimental and predicted PCE values for different groups of devices

Perovskite	Predicted PCE (%)	Experimental verification (%)	Absolute Error(%)
$FA_{0.87}Cs_{0.13}PbI_3$	16.27	13.97	2.30
$(FA_{0.9}MA_{0.1})_{0.87}Cs_{0.13}PbI_3$	17.24	13.83	3.41
$(FA_{0.7}MA_{0.3})_{0.87}Cs_{0.13}PbI_3$	17.80	17.40	0.40
$(FA_{0.5}MA_{0.5})_{0.87}Cs_{0.13}PbI_3$	17.94	21.67	3.73
$(FA_{0.1}MA_{0.9})_{0.87}Cs_{0.13}PbI_3$	18.35	19.13	0.78
$MA_{0.87}Cs_{0.13}PbI_3$	15.74	14.19	1.55
$(FA_{0.9}MA_{0.1})_{0.87}Cs_{0.13}PbI_3$+MACl	19.37	19.05	0.32
$(FA_{0.7}MA_{0.3})_{0.87}Cs_{0.13}PbI_3$+MACl	20.03	21.61	1.58
$(FA_{0.5}MA_{0.5})_{0.87}Cs_{0.13}PbI_3$+MACl	20.27	21.81	1.54
$(FA_{0.9}MA_{0.1})_{0.87}Cs_{0.13}Pb(I_{0.985}Br_{0.015})_3$	17.40	14.84	2.56

$(FA_{0.5}MA_{0.5})_{0.87}Cs_{0.13}Pb(I_{0.925}Br_{0.075})_3$	18.71	18.90	0.19
$(FA_{0.7}MA_{0.3})_{0.87}Cs_{0.13}Pb(I_{0.985}Br_{0.015})_3$+MACl	20.90	22.09	1.19

Table S5 The performance of CatBoost model corrected by experimental data on the training and testing sets

Training Set				Test Set			
r	***R^2***	**RMSE (%)**	**MAE (%)**	***r***	***R^2***	**RMSE (%)**	**MAE (%)**
0.9527	0.9063	1.651	1.191	0.8759	0.7672	2.685	2.046

Table S6 Presentation and detailed explanations of deposition methods

Deposition methods	**Explanations**
spin	one-step spin-coating method without changing the rotational speed
spin2-3	one-step spin-coating method changing the speed two or three times within 60 seconds
spin-spin	two-step spin-coating method of depositing inorganic salts first before organic salts
spin-dip	two-step deposition method of depositing inorganic salts first before soaking them in an organic salt solution
spin-spray	two-step deposition method for spin-coating inorganic salts before organic salts with spray assisted deposition
vasp	short of low-pressure vapor-assisted solution process, a two-step deposition method of depositing inorganic salts first and then reacting to form perovskite in a gas atmosphere of the organic salts
CVD	short of chemical vapor deposition, a two-step deposition method of firstly vacuum vapor depositing inorganic salts, followed by vapor depositing organic salts

Reference

[1] a)Ç. Odabaşı, R. Yıldırım, *Nano Energy* **2019**, 56, 770; b)C. She, Q. Huang, C. Chen, Y. Jiang, Z. Fan, J. Gao, *Journal of Materials Chemistry A* **2021**, 9, 25168.

[2] a)J. Y. Kim, J. W. Lee, H. S. Jung, H. Shin, N. G. Park, *Chem Rev* **2020**, 120, 7867; b)A. Hassan, Z. Wang, Y. H. Ahn, M. Azam, A. A. Khan, U. Farooq, M. Zubair, Y. Cao, *Nano Energy* **2022**, 101.

[3] a)Y. Fan, X. Wang, Y. Miao, Y. Zhao, *J Phys Chem Lett* **2021**, 12, 11636; b)H. Wang, J. Zhang, *International Journal of Photoenergy* **2022**, 2022, 5288400.

[4] E. C. Gok, M. O. Yildirim, M. P. U. Haris, E. Eren, M. Pegu, N. H. Hemasiri, P. Huang, S. Kazim, A. Uygun Oksuz, S. Ahmad, *Solar RRL* **2021**, 6, 2100927.

[5] D. V. Lindley, A. F. M. Smith, *Journal of the royal statistical society series b-methodological* **1972**, 34, 1.
[6] C. Cortes, V. Vapnik, *Machine Learning* **1995**, 20, 273.
[7] T. J. Hastie, R. Tibshirani, J. H. Friedman, presented at *Springer Series in Statistics*, **2001**.
[8] S. L. Salzberg, *Machine Learning* **1994**, 16, 235.
[9] L. Breiman, *Machine Learning* **2001**, 45, 5.
[10] G. L. Ke, Q. Meng, T. Finley, T. F. Wang, W. Chen, W. D. Ma, Q. W. Ye, T. Y. Liu, presented at *31st Annual Conference on Neural Information Processing Systems (NIPS)*, Long Beach, CA, Dec 04-09, **2017**.
[11] T. Q. Chen, C. Guestrin, M. Assoc Comp, presented at *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, San Francisco, CA, Aug 13-17, **2016**.
[12] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, A. Gulin, presented at *32nd Conference on Neural Information Processing Systems (NIPS)*, Montreal, CANADA, Dec 02-08, **2018**.
[13] Y. Zhao, F. Ma, Z. Qu, S. Yu, T. Shen, H.-X. Deng, X. Chu, X. Peng, Y. Yuan, X. Zhang, J. You, Science **2022**, 377, 531.
[14] Q. Jiang, J. Tong, Y. Xian, R. A. Kerner, S. P. Dunfield, C. Xiao, R. A. Scheidt, D. Kuciauskas, X. Wang, M. P. Hautzinger, R. Tirawat, M. C. Beard, D. P. Fenning, J. J. Berry, B. W. Larson, Y. Yan, K. Zhu, Nature **2022**, 611, 278.
[15] M. Del Cueto, C. Rawski-Furman, J. Arago, E. Orti, A. Troisi, *J Phys Chem C Nanomater Interfaces* **2022**, 126, 13053.
[16] J. Li, B. Pradhan, S. Gaur, J. Thomas, *Advanced Energy Materials* **2019**, 9, 1901891.
[17] Y. Liu, W. Yan, S. Han, H. Zhu, Y. Tu, L. Guan, X. Tan, *Solar RRL* **2022**, 6, 2101100.
[18] Y. Lu, D. Wei, W. Liu, J. Meng, X. Huo, Y. Zhang, Z. Liang, B. Qiao, S. Zhao, D. Song, Z. Xu, *Journal of Energy Chemistry* **2023**, 77, 200.