Bayesian reverse design of high-efficiency perovskite solar cells based on experimental knowledge constraints

Cite as: Appl. Phys. Lett. **125**, 063901 (2024); doi: 10.1063/5.0216447
Submitted: 29 April 2024 · Accepted: 22 July 2024 ·
Published Online: 5 August 2024

Hongyu Liu,[1] Zhengxin Chen,[1] Yaping Zhang,[1] Jiang Wu,[2] Lin Peng,[1] Yanan Wang,[1] Xiaolin Liu,[1,a)] Xianfeng Chen,[3,4,a)] and Jia Lin[1,a)]

AFFILIATIONS

[1] College of Mathematics and Physics, Shanghai University of Electric Power, Shanghai 200090, China
[2] College of Energy and Mechanical Engineering, Shanghai University of Electric Power, Shanghai 200090, China
[3] State Key Laboratory of Advanced Optical Communication Systems and Networks, School of Physics and Astronomy, Shanghai Jiao Tong University, Shanghai 200240, China
[4] Collaborative Innovation Center of Light Manipulation and Applications, Shandong Normal University, Jinan 250358, China

[a)] Authors to whom correspondence should be addressed: xlliu@shiep.edu.cn; xfchen@sjtu.edu.cn; and jlin@shiep.edu.cn

ABSTRACT

To alleviate high costs and lengthy trial-and-error periods associated with traditional optimization methods for perovskite solar cells (PSCs), we developed a data-driven reverse design framework for high-efficiency PSCs. This framework integrates machine learning and Bayesian optimization (BO) to accelerate the optimization process of PSCs by intelligently recommending the most promising parameter configurations for PSCs, such as device structure and fabrication processes. To improve the robustness of the framework, we first designed a two-stage sampling strategy to alleviate the issue of imbalanced dataset classes. Subsequently, by integrating "experimental knowledge constraints" into the BO process, we achieved precise parameter configurations, thus avoiding discrepancies between predicted and actual results due to parameter mismatches. Finally, using SHapley Additive exPlanations, we unveiled key factors influencing the power conversion efficiency (PCE), such as the composition of perovskite solvents. Our framework not only precisely predicted the PCE of PSCs with an area under the curve of 0.861 but also identified the optimal parameter configurations, achieving a high probability of 0.981. This framework offers substantial support for minimizing redundant experiments and characterizations, effectively accelerating the optimization process of PSCs.

During the development of perovskite solar cells (PSCs), their excellent power conversion efficiency (PCE), low-cost manufacturing process, and sensitive response to a broad spectrum have collectively established their leading position in the future development of the photovoltaic field.[1,2] Over the past decade, the PCE of single-junction PSCs has surged from 3.8% to 26.1%[3–5] through continuous optimization in device structure and fabrication processes.[6–18] As research progressively deepens, an undeniable fact gradually emerges: the PCE of PSCs is highly sensitive to the configuration and fabrication process parameters of each functional layer. Whether it is the strategic choice of charge transport layer (CTL) materials or the precise input of additives, every step of adjustment is directly related to the final PCE. Therefore, the rational design of device structure and fabrication processes is crucial for achieving high-efficiency PSCs.

Faced with the increasing number of potential structural configurations and fabrication process parameter combinations, as well as the complex interactions between device performance and parameters, finding the optimal configuration for each layer through experiments to effectively optimize PSC often means a long trial-and-error process and substantial consumable costs. This method not only heavily relies on the extensive experience and intuitive judgment of researchers but also struggles to systematically cover enough parameter combinations and their corresponding impacts on performance due to the complexity and variability of experimental conditions. An efficient, accurate, and easy-to-operate design method for high-efficiency PSCs is urgently needed. Fortunately, a substantial amount of experimental data have been generated during the research process, particularly for $MAPbI_3$ devices, providing an opportunity to use artificial intelligence (AI) technology to promote the design of high-efficiency PSCs through a data-driven approach.

In recent years, machine learning (ML), an AI technology with powerful pattern recognition capabilities, has been primarily applied in the form of forward prediction in the field of PSCs, where the input is the parameters of PSCs and the output is the performance of PSCs.[19–27] ML can learn hidden patterns between various influencing factors and PSC performance from vast data, enabling precise predictions. For example, Li *et al.*,[25] Gok *et al.*,[26] and Lu *et al.*[27] predicted the performance of PSCs based on different input parameters such as perovskite bandgap, perovskite composition, and fabrication process. However, in the design of PSCs, what is more crucial is designing the appropriate device structure and fabrication processes based on specific requirements, such as high PCE, which embodies a concept of reverse design. It is worth noting that Bayesian optimization (BO), an efficient global optimization technique (as shown in Fig. S1), is a crucial technology to achieve performance-oriented reverse design process, and has made significant progress in accelerating material performance improvement and experimental process optimization.[28–30] For example, MacLeod *et al.* utilized BO to enhance the optoelectronic performance of thin films by optimizing the cobalt doping ratio and annealing time of the spiro-OMeTAD.[28] The powerful ability of BO in handling complex parameter spaces and achieving precise objective optimization is very suitable for the reverse design of PSCs. Nevertheless, resolving the issue of parameter mismatch is an imminent challenge within this process.

Different from the research focused on improving specific materials, our work shifts the optimization focus to the device level, including all functional layers. While this strategy broadens the optimization horizon, it inevitably introduces a series of hard constraints, such as the compatibility between materials and additives, resulting in a high interdependency among certain parameters. BO typically does not automatically consider these hard constraints, which may lead to parameter mismatch issues in optimization results, thereby affecting the consistency between predicted results and actual applications. This explains why the application of BO to optimize systems with highly interdependent parameters is more intricate. To address the issue, we introduce "experimental knowledge constraints." This approach is based on existing experimental parameter configurations, constructing multiple dictionaries containing relevant feature combinations to guide and constrain the BO process. Through this method, we can achieve precise and reasonable parameter configurations, thereby significantly enhancing the accuracy and practicality of reverse design.

In this study, we combine the powerful pattern recognition capability of ML with the efficient global optimum identification ability of BO to construct a reverse design framework for designing high-efficiency PSCs. Initially, we established the $MAPbI_3$ device dataset based on an experimental database and designed a two-stage sampling strategy to more effectively alleviate class imbalance issues in the dataset. Subsequently, the voting model achieved accurate predictions of device PCE and assessed the potential for the devices to become high-efficiency PSCs. In addition, by employing BO combined with "experimental knowledge constraints," we intelligently recommend the device parameter configuration with the most potential to achieve high-efficiency PSCs. Finally, utilizing SHAP (SHapley Additive exPlanations), we revealed the key factors affecting the PCE.

Figure 1 shows the workflow diagram of our framework. The $MAPbI_3$ dataset we constructed, containing 3526 device data points, is derived from the open database of PSCs created by Jacobsson *et al.*, which compiles peer-reviewed experimental data up until February 2020.[31] As shown in Table S1, we used the device structure and fabrication process parameters in the data as the input features of the ML model.

Based on recent literature and considering the proportion of PSCs with high PCE in the dataset [as shown in Fig. 2(a)], we classify the devices into two categories, high efficiency (positive category) and low efficiency (negative category), based on whether the PCE value reaches 17%, and use them as labels for the model output.[2] Figures 2(b)–2(d) show the distribution of device architecture (p-i-n or n-i-p) and CTL materials in the $MAPbI_3$ dataset. It is worth noting that in Figs. 2(c) and 2(d), we provide lists of electron transport layer (ETL)

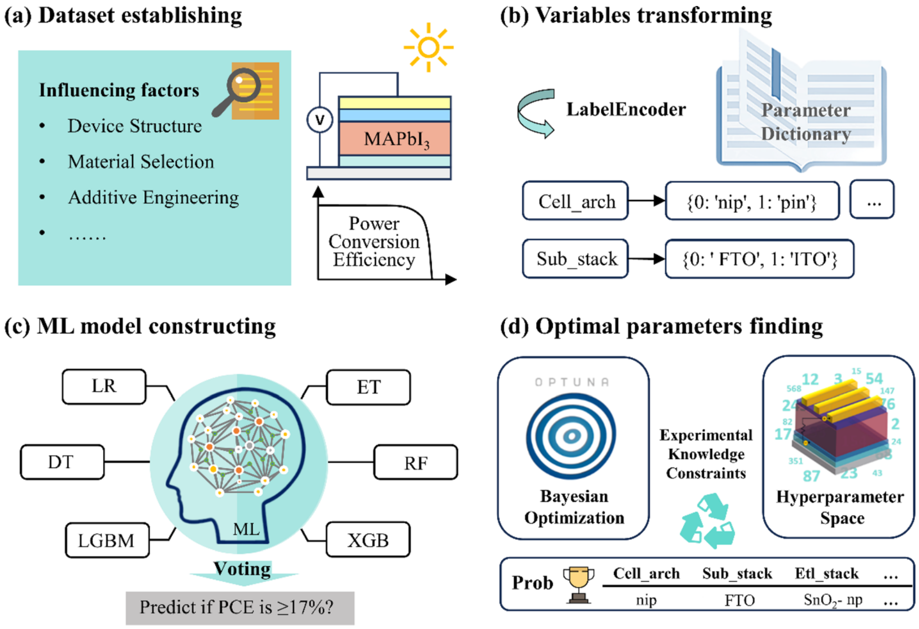


FIG. 1. The reverse design framework for PSCs, where the procedure includes (a) dataset establishment, (b) variable transformation, (c) ML model construction, and (d) optimal parameters identification.

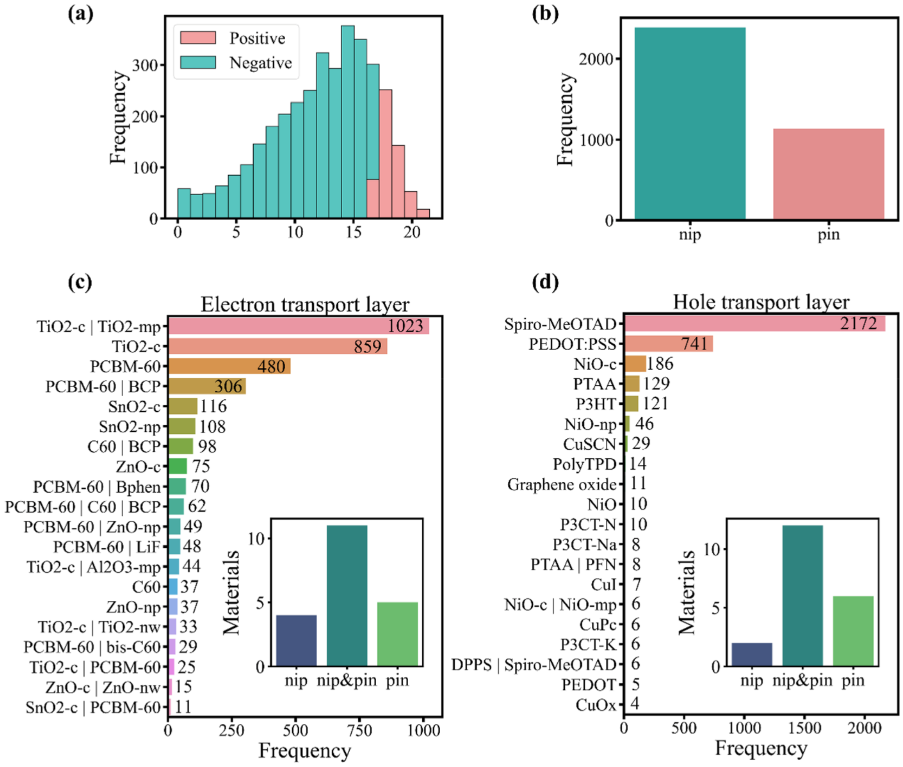


FIG. 2. The distribution of PCE, device architecture, and CTL materials in each functional layer in the $MAPbI_3$ dataset: (a) PCE of the device ("Jv_PCE"), (b) device architecture ("Cell_arch"), (c) electron transport layer ("Etl_stack"), and (d) hole transport layer ("Htl_stack"). In the subplots for (c) and (d), the counts of materials used in the two device architectures are tabulated: "nip" represents exclusive use in nip devices, "nip&pin" indicates usage in both structures, and "pin" signifies exclusive use in inverted devices.

and hole transport layer (HTL) materials by frequency, revealing the differences in ETL and HTL materials used in different device architectures. Additionally, due to the significant difference in thickness between carbon electrodes and metal electrodes, this dataset only includes devices employing metal electrodes.

Given the $MAPbI_3$ dataset encompasses a large number of textual features related to materials and fabrication processes, we employed the LabelEncoder method to convert these textual features into numerical labels, thereby transforming device information into ML recognizable low-dimensional matrices. This approach not only avoids the challenges and limitations associated with obtaining descriptors of material physical properties and their lack of universality but also avoids the dimensionality explosion caused by using one-hot encoding.

During the ML model construction process, we employed six algorithms to construct classification models. These included logistic regression (LR), decision tree (DT), and four tree-based ensemble learning models, namely random forest (RF), Light Gradient-Boosting Machine (LGBM), Extremely Randomized Trees (ET), and eXtreme Gradient Boosting (XGB).[32–36] Furthermore, three classification evaluation indicators were selected to comprehensively assess the performance of the models: accuracy, F1 score, and area under the curve (AUC).[37] Additionally, we employed the tenfold cross-validation results as model scores to reduce performance fluctuations caused by data segmentation and to enhance the models' generalization capability.[38] Figure 3(a) compares the performance of six models on the dataset, and we choose the more robust ensemble models RF, LGBM, ET, and XGB for subsequent work.

Observing Fig. 2(a), we discovered a severe class imbalance issue in the dataset, manifested as a ratio of positive to negative classes of 1:6. This imbalance led models to tend to predict the majority classes (negative samples), resulting in high accuracy but low F1 scores, as shown in Fig. 3(a). We devised a two-stage sampling strategy to mitigate class imbalance, which first involves RandomUnderSampler (RUS) sampling followed by Synthetic Minority Over-sampling Technique (SMOTE). As illustrated in Fig. 3(b) and Fig. S2, we utilized BO to determine the optimal undersampling rate $\alpha_{us} = 0.38$ and oversampling rate $\alpha_{os} = 0.93$, resulting in a peak F1 score of 0.633, although at the cost of a slight accuracy reduction of about 0.07. Figure 3(c) displays the distribution of classes before and after the two-stage sampling. The use of this method not only effectively alleviates the class imbalance issue within the dataset but also reduces the potential for information loss or noise introduced by employing a single sampling method (the comparison of sampling strategies can be found in Table S2).

To ensure outstanding model performance while minimizing computational resource consumption, as shown in Fig. 3(d) and Fig. S3, we selected 16 key features to construct the optimal feature subset (refer to Table S3) for subsequent modeling processes. Additionally, to further improve the robustness of the model, we used BO to fine-tune the hyperparameters of four models, as shown in Table S4. In the process of identification of optimal parameters, we focused more on the model's ability to rank output probability scores rather than solely on classification accuracy. Voting algorithm is an ensemble learning algorithm that follows the "majority rules" principle.[39] This method combines predictions from multiple models according to specific rules to reach a final collective decision, thereby enhancing the overall robustness of the model. For this purpose, we constructed an additional voting model based on high-performance RF and XGB models (with

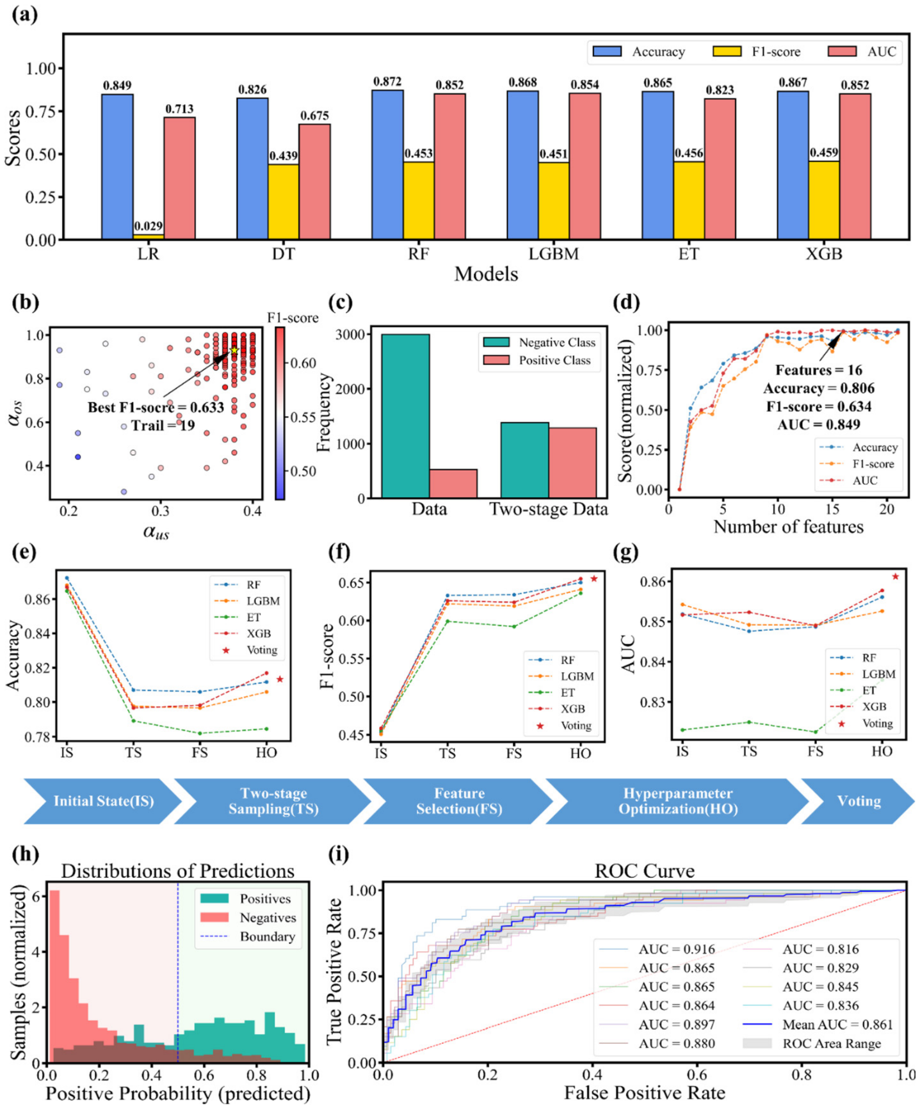


FIG. 3. The performance visualization during ML model modeling process. (a) The performance comparison of six classification models on the dataset. (b) The F1 score chart of the two-stage sampling strategy under various sampling ratios. (c) Comparison of sample distribution before and after implementing the two-stage sampling strategy. (d) Influence of the number of features on the performance of the RF model, with scores normalized. Select corresponding features from high to low based on the importance determined by the model. The evaluation metric scores of models in the following stages: initial state, two-stage sampling, feature selection, hyperparameter optimization, and voting. (e) Accuracy, (f) F1 score, and (g) AUC. (h) Visualization of the voting model output probability distribution. The legend colors reflect the true class of the samples. (i) The ROC curve of the voting model.

weights set to 0.47 and 0.53, respectively), achieving a peak AUC of 0.861. Figures 3(e)–3(g) delineate the influence of each stage strategy throughout the modeling process on the model's performance, confirming the effectiveness of our optimization efforts. The performance of the optimized models is shown in Table S5.

In order to more intuitively present the excellent performance of the voting model, we visualized it and presented the results in Fig. 3. The distribution of model prediction probabilities and its confusion matrix are clearly presented in Fig. 3(h) and Fig. S4, respectively. The proportion of false positive (FP) samples is minimal and significantly decreases as the positive probability increases. Notably, all samples with a probability exceeding 0.927 were accurately classified, indicating a high level of reliability for samples predicted as positive class with probabilities exceeding this threshold. The model exhibits excellent performance in each round of tenfold cross-validation, with a maximum AUC value of 0.916 and an average value of 0.861, as shown in Fig. 3(i).

In the process of optimal parameter identification, the Optuna, a hyperparameter optimization framework, was used to perform the BO process.[40] During this process, we limit the parameter search space to a range defined by the features in the optimal feature subset, as shown in Fig. 4(a), ensuring that the optimization process focuses on the factors most critical to PCE. We selected the indicator of the device's potential for high-efficiency (the probability of the device being classified as positive by the voting model) as the objective function. This intended to continuously update the probability surrogate model through iteration to fit the objective function in order to identify the most promising combination of device structures and fabrication process parameters.

The proper configuration of parameters is crucial for achieving high PCE, yet BO often overlooks this aspect. For example, the nip and pin architectures exhibit differences in charge transfer and collection mechanisms, directly impacting the selection of materials for the CTL [as illustrated in Figs. 2(c) and 2(d)]. Failure to consider and

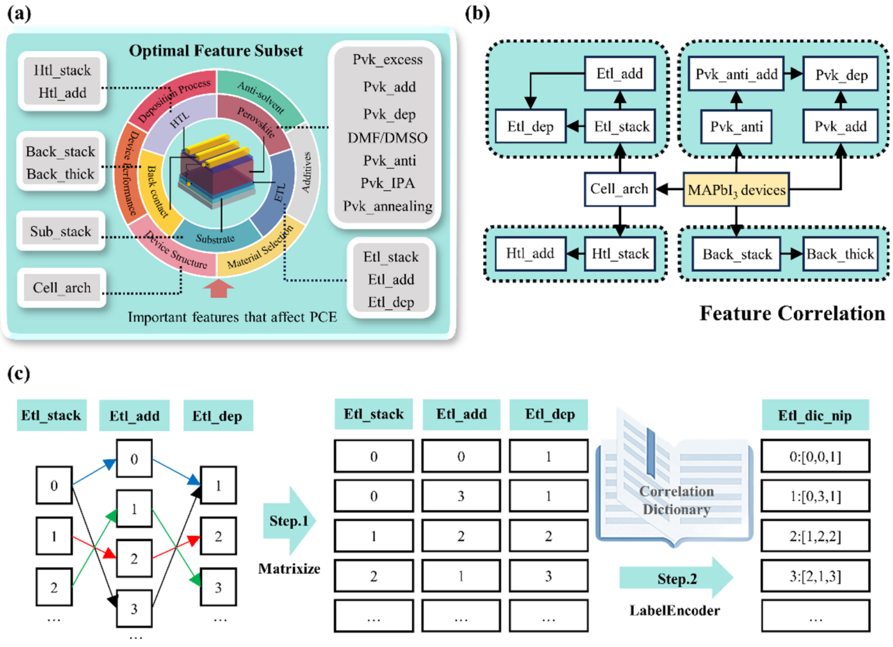
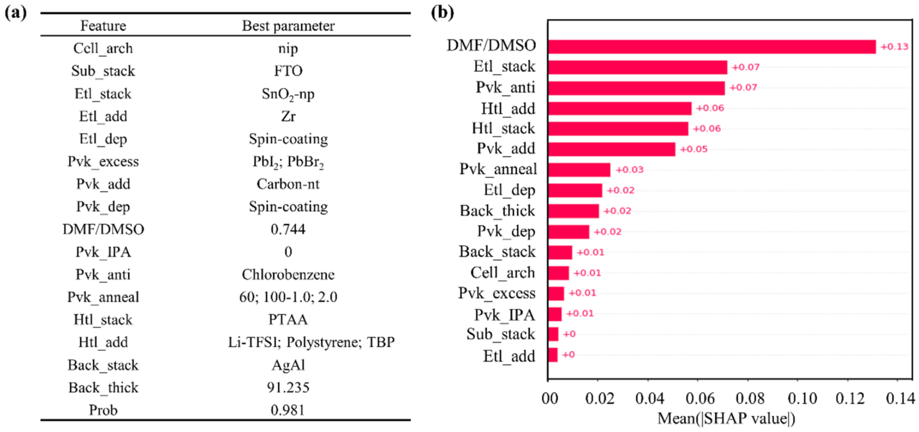


FIG. 4. The process of establishing "experimental knowledge constraints." (a) The optimal subset of features. (b) The correlation between features. (c) The process of establishing "Etl_dic_nip" feature association dictionary.

constrain these differences may lead to mismatches between the device architecture and CTL materials, resulting in significant discrepancies between predicted and actual results. Similar situations include the matching of materials and additives. Faced with parameter configuration issues, we innovatively introduced "experimental knowledge constraints" into the BO process. This strategy utilizes prior knowledge from specific experimental configurations to avoid undesirable parameter settings, ensuring that the reverse design results align with physical principles and experimental facts while efficiently approaching target performance metrics. Figures 4(b) and 4(c) show the introduction process of "experimental knowledge constraints." First, it establishes the correlation between features based on experimental knowledge and the functional layer to which they belong. Then, it utilizes matrix transformation and the LabelEncoder method to convert all parameter configurations of relevant features into multiple feature association dictionaries. The parameter spaces of these dictionaries were used to replace the original parameter spaces of the related features, thereby imposing constraints on parameter selection during the BO process. The adoption of this strategy enables precise parameter configuration during the BO process, significantly enhancing the accuracy and practicality of reverse design. For detailed information, please refer to Tables S6 and S7, and Fig. S5.

Figure 5(a) displays the optimal parameter configuration results for the $MAPbI_3$ device, derived through BO. The device structure is FTO|SnO_2-np|$MAPbI_3$|PTAA|AgAl, with a remarkably high positive probability of 0.981. To gain a deeper understanding of the positive impact of these parameters on the device PCE, we reviewed the relevant literature. Previous studies have indicated that Zr-doping can enhance the electron transport properties and adjust the energy-level structure of the SnO_2-np layer, thereby improving the performance of PSCs.[41] Excessive introduction of PbI_2 and $PbBr_2$ into the precursor of $MAPbI_3$ perovskite can improve the surface quality of the film, making


(a)

Feature	Best parameter
Cell_arch	nip
Sub_stack	FTO
Etl_stack	SnO_2-np
Etl_add	Zr
Etl_dep	Spin-coating
Pvk_excess	PbI_2; $PbBr_2$
Pvk_add	Carbon-nt
Pvk_dep	Spin-coating
DMF/DMSO	0.744
Pvk_IPA	0
Pvk_anti	Chlorobenzene
Pvk_anneal	60; 100-1.0; 2.0
Htl_stack	PTAA
Htl_add	Li-TFSI; Polystyrene; TBP
Back_stack	AgAl
Back_thick	91.235
Prob	0.981

FIG. 5. The results of reverse design and model analysis. (a) The optimal parameter configuration for $MAPbI_3$ devices. (b) The feature importance of the voting model based on SHAP values.

it smoother with fewer defects, thus improving the device performance.[42] Additionally, carbon nanotubes (Carbon-nt) have been employed to improve the charge extraction capability of PSCs.[43,44] There is research suggesting that the optimal DMF:DMSO ratio for nip devices is 0.7:0.3, which is a mere 0.044 variance from our suggested ratio.[45] This minor discrepancy is primarily attributed to differences in other parameter settings. Furthermore, the use of an appropriate amount of chlorobenzene as anti-solvent can improve the morphology and crystallinity of the $MAPbI_3$ film, promoting the accumulation of photogenerated carriers at the interface, thereby accelerating the electron transport rate and enhancing the PCE.[46] Conventional two-step annealing processes are also widely employed to improve film quality. Although PTAA is a commonly used HTL material in the pin device, it is also applied in the nip device. Doping PTAA with Li-TFSI, polystyrene, and TBP can improve the electrical properties of PTAA, thereby enhancing the PCE of PSCs.[47] Finally, AgAl-based electrode demonstrates higher PCE and stability compared to single Ag or Al electrodes.[48] In summary, through the coordinated optimization and precise control of the parameters for the device structure and fabrication processes, the potential for this device to become a high-efficiency PSC has been significantly enhanced.

SHAP is a sophisticated tool for model interpretability, which helps researchers understand the decision-making process of models and reveal the complex mechanisms between device parameters and their performance.[49] To determine the crucial factors affecting the PCEs of PSCs, we employed SHAP values to quantify the contribution of each feature to model predictions. In this section, we focus on analyzing the top three features that have the greatest impact on PCE. As shown in Fig. 5(b), the ratio of DMF to DMSO in perovskite solvents has a significant impact on the performance of PSCs (with a SHAP value of 0.13), consistent with recent research.[45,50,51] The mixture of DMF and DMSO solvents, with their stronger ability to form intermediate phases, can promote the formation of larger grain perovskite films, thereby enhancing light absorption and charge transfer efficiency. However, an inappropriate ratio of DMF to DMSO can severely affect the crystallization rate and PCEs. Additionally, the ETL is located between the perovskite layer and the electrode, responsible for the effective transfer of electrons and preventing the migration of holes toward the cathode. The matching energy levels between the ETL and the perovskite layer facilitate the extraction of electrons, while mismatched energy levels can create an interfacial barrier, reducing the efficiency of electron transport and the overall PCEs. Therefore, the selection of ETL materials is crucial for enhancing the PCEs of PSCs.[52,53] Furthermore, anti-solvent treatment significantly improves film quality and interface characteristics by promoting crystallization of the perovskite film, thereby enhancing the PCEs of PSCs.[54,55] Improper selection of anti-solvent can lead to a decrease in film quality and even the formation of cracks, significantly impacting device performance. Additionally, the selection of materials and additives for the HTL can also notably impact the final PCEs of PSCs.[47,56]

In this work, we have developed a data-driven framework for reverse designing PSCs, aiming to overcome the inefficiency and high cost issues associated with traditional experimental optimization methods. By applying the two-stage sampling strategy and the "experimental knowledge constraint" method, our framework has not only accurately predicted the PCE of PSCs with an AUC of 0.861 but also identified the optimal parameter configuration, achieving a high probability of 0.981. The corresponding device structure is FTO|SnO_2-np|$MAPbI_3$|PTAA|AgAl. Additionally, we have identified key factors that significantly impact PCE, such as the composition of the perovskite solvents. Although our work is primarily focused on single-junction PSCs, the method we propose is also applicable to the design of other photovoltaic devices with similar structures, such as perovskite/silicon tandem solar cells (TSCs).[57–59] This approach allows for the various components of TSCs (such as nanotextures) to be considered as a unified system for the synergistic optimization of device configuration and manufacturing processes, rather than being optimized in isolation. The framework provides substantial support for reducing redundant experiments, thereby effectively accelerating the optimization process of photovoltaic devices.

See the supplementary material for modeling details, including a comparison of effectiveness of sampling strategies, ranking of 16 features and their importance in feature selection, the impact of feature selection on the performance of four models, parameter space and optimal hyperparameters for model hyperparameter optimization, detailed establishment process of the feature correlation dictionary and the included feature information, and the parameter space for $MAPbI_3$ devices in the process of finding optimal parameters.

This work was partially sponsored by the National Natural Science Foundation of China (Grant Nos. 12174246 and 52076126), the National Key Research and Development Program (Grant No. 2021YFE0109800), the Key Laboratory of Clean Power Generation and Environmental Protection Technology in Mechanical Industry, Science and Technology Commission of Shanghai Municipality (Grant No. 21010501300), and the Shuguang Program of Shanghai Education Development Foundation and Shanghai Municipal Education Commission (Grant No. 22SG51).

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Hongyu Liu and Zhengxin Chen contributed equally to this work.

Hongyu Liu: Data curation (equal); Formal analysis (equal); Validation (equal); Writing – original draft (equal). **Zhengxin Chen:** Formal analysis (equal); Investigation (equal); Methodology (equal); Validation (equal). **Yaping Zhang:** Investigation (equal); Methodology (equal). **Jiang Wu:** Resources (equal); Validation (equal). **Lin Peng:** Formal analysis (equal); Methodology (equal). **Yanan Wang:** Data curation (equal); Writing – review & editing (equal). **Xiaolin Liu:** Formal analysis (equal); Validation (equal); Visualization (equal). **Xianfeng Chen:** Funding acquisition (equal); Resources (equal). **Jia Lin:** Conceptualization (equal); Supervision (equal); Writing – review & editing (equal).

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding authors upon reasonable request.

REFERENCES

[1] N.-G. Park, "Perovskite solar cells: An emerging photovoltaic technology," Mater. Today **18**(2), 65 (2015).

[2] J. Y. Kim, J. W. Lee, H. S. Jung, H. Shin, and N. G. Park, "High-efficiency perovskite solar cells," Chem. Rev. **120**(15), 7867 (2020).

[3] A. Kojima, K. Teshima, Y. Shirai, and T. Miyasaka, "Organometal halide perovskites as visible-light sensitizers for photovoltaic cells," J. Am. Chem. Soc. **131**(17), 6050 (2009).

[4] H. Chen, C. Liu, J. Xu, A. Maxwell, W. Zhou, Y. Yang, Q. Zhou, A. S. R. Bati, H. Wan, Z. Wang *et al.*, "Improved charge extraction in inverted perovskite solar cells with dual-site-binding ligands," Science **384**(6692), 189 (2024).

[5] M. A. Green, E. D. Dunlop, M. Yoshita, N. Kopidakis, K. Bothe, G. Siefer, and X. Hao, "Solar cell efficiency tables (Version 63)," Prog. Photovoltaics **32**(1), 3 (2023).

[6] J. W. Lee, D. J. Seol, A. N. Cho, and N. G. Park, "High-efficiency perovskite solar cells based on the black polymorph of $HC(NH_2)_2PbI_3$," Adv. Mater. **26**(29), 4991 (2014).

[7] S. Liu, W. Huang, P. Liao, N. Pootrakulchote, H. Li, J. Lu, J. Li, F. Huang, X. Shai, X. Zhao *et al.*, "Correction: 17% efficient printable mesoscopic PIN metal oxide framework perovskite solar cells using cesium-containing triple cation perovskite," J. Mater. Chem. A **6**(9), 4220 (2018).

[8] P. W. Liang, C. Y. Liao, C. C. Chueh, F. Zuo, S. T. Williams, X. K. Xin, J. Lin, and A. K. Jen, "Additive enhanced crystallization of solution-processed perovskite for highly efficient planar-heterojunction solar cells," Adv. Mater. **26**(22), 3748 (2014).

[9] H. Min, M. Kim, S. U. Lee, H. Kim, G. Kim, K. Choi, J. H. Lee, and S. I. Seok, "Efficient, stable solar cells by using inherent bandgap of alpha-phase formamidinium lead iodide," Science **366**(6466), 749 (2019).

[10] Y.-J. Kang, S.-N. Kwon, S.-P. Cho, Y.-H. Seo, M.-J. Choi, S.-S. Kim, and S.-I. Na, "Antisolvent additive engineering containing dual-function additive for triple-cation p–i–n perovskite solar cells with over 20% PCE," ACS Energy Lett. **5**(8), 2535 (2020).

[11] M. Hou, H. Zhang, Z. Wang, Y. Xia, Y. Chen, and W. Huang, "Enhancing efficiency and stability of perovskite solar cells via a self-assembled dopamine interfacial layer," ACS Appl. Mater. Interfaces **10**(36), 30607 (2018).

[12] W. Kong, W. Li, C. Liu, H. Liu, J. Miao, W. Wang, S. Chen, M. Hu, D. Li, A. Amini *et al.*, "Organic monomolecular layers enable energy-level matching for efficient hole transporting layer free inverted perovskite solar cells," ACS Nano **13**(2), 1625 (2019).

[13] C. Wang, Y. Zhao, T. Ma, Y. An, R. He, J. Zhu, C. Chen, S. Ren, F. Fu, D. Zhao *et al.*, "A universal close-space annealing strategy towards high-quality perovskite absorbers enabling efficient all-perovskite tandem solar cells," Nat. Energy **7**(8), 744 (2022).

[14] L. Huang, Y. H. Lou, and Z. K. Wang, "Buried interface passivation: A key strategy to breakthrough the efficiency of perovskite photovoltaics," Small **19**(38), e2302585 (2023).

[15] B. A. Al-Asbahi, S. M. H. Qaid, M. Hezam, I. Bedja, H. M. Ghaithan, and A. S. Aldwayyan, "Effect of deposition method on the structural and optical properties of $CH_3NH_3PbI_3$ perovskite thin films," Opt. Mater. **103**, 109836 (2020).

[16] B. Li, S. Li, J. Gong, X. Wu, Z. Li, D. Gao, D. Zhao, C. Zhang, Y. Wang, and Z. Zhu, "Fundamental understanding of stability for halide perovskite photovoltaics: The importance of interfaces," Chem **10**(1), 35 (2024).

[17] Z. Huang, Y. Bai, X. Huang, J. Li, Y. Wu, Y. Chen, K. Li, X. Niu, N. Li, G. Liu *et al.*, "Anion-pi interactions suppress phase impurities in $FAPbI_3$ solar cells," Nature **623**(7987), 531 (2023).

[18] Z. Liang, Y. Zhang, H. Xu, W. Chen, B. Liu, J. Zhang, H. Zhang, Z. Wang, D. H. Kang, J. Zeng *et al.*, "Homogenizing out-of-plane cation composition in perovskite solar cells," Nature **624**(7992), 557 (2023).

[19] Z. Chen, H. Wang, B. Liu, H. Zhou, Y. Cao, Y. Wang, L. Peng, X. Liu, J. Lin, X. Chen *et al.*, "Cationic perturbation strategy to solve the information drought in material explainable machine learning," Phys. Rev. B **109**(8), 085306 (2024).

[20] S. Lu, Q. Zhou, Y. Ouyang, Y. Guo, Q. Li, and J. Wang, "Accelerated discovery of stable lead-free hybrid organic-inorganic perovskites via machine learning," Nat. Commun. **9**(1), 3405 (2018).

[21] C. Zhi, S. Wang, S. Sun, C. Li, Z. Li, Z. Wan, H. Wang, Z. Li, and Z. Liu, "Machine-learning-assisted screening of interface passivation materials for perovskite solar cells," ACS Energy Lett. **8**(3), 1424 (2023).

[22] N. T. P. Hartono, J. Thapa, A. Tiihonen, F. Oviedo, C. Batali, J. J. Yoo, Z. Liu, R. Li, D. F. Marron, M. G. Bawendi *et al.*, "How machine learning can help select capping layers to suppress perovskite degradation," Nat. Commun. **11**(1), 4172 (2020).

[23] Y. Yu, X. Tan, S. Ning, and Y. Wu, "Machine learning for understanding compatibility of organic–inorganic hybrid perovskites with post-treatment amines," ACS Energy Lett. **4**(2), 397 (2019).

[24] J. Hu, Z. Chen, Y. Chen, H. Liu, W. Li, Y. Wang, L. Peng, X. Liu, J. Lin, X. Chen *et al.*, "Interpretable machine learning predictions for efficient perovskite solar cell development," Sol. Energy Mater. Sol. Cells **271**, 112826 (2024).

[25] J. Li, B. Pradhan, S. Gaur, and J. Thomas, "Predictions and strategies learned from machine learning to develop high-performing perovskite solar cells," Adv. Energy Mater. **9**(46), 1901891 (2019).

[26] E. C. Gok, M. O. Yildirim, M. P. U. Haris, E. Eren, M. Pegu, N. H. Hemasiri, P. Huang, S. Kazim, A. U. Oksuz, and S. Ahmad, "Predicting perovskite bandgap and solar cell performance with machine learning," Sol. RRL **6**(2), 2100927 (2022).

[27] Y. Lu, D. Wei, W. Liu, J. Meng, X. Huo, Y. Zhang, Z. Liang, B. Qiao, S. Zhao, D. Song *et al.*, "Predicting the device performance of the perovskite solar cells from the experimental parameters through machine learning of existing experimental results," J. Energy Chem. **77**, 200 (2023).

[28] B. P. MacLeod, F. G. L. Parlane, T. D. Morrissey, F. Hase, L. M. Roch, K. E. Dettelbach, R. Moreira, L. P. E. Yunker, M. B. Rooney, J. R. Deeth *et al.*, "Self-driving laboratory for accelerated discovery of thin-film materials," Sci. Adv. **6**(20), eaaz8867 (2020).

[29] Z. Liu, N. Rolston, A. C. Flick, T. W. Colburn, Z. Ren, R. H. Dauskardt, and T. Buonassisi, "Machine learning with knowledge constraints for process optimization of open-air perovskite solar cell manufacturing," Joule **6**(4), 834 (2022).

[30] W. Li, J. Hu, Z. Chen, H. Jiang, J. Wu, X. Meng, X. Fang, J. Lin, X. Ma, T. Yang *et al.*, "Performance prediction and optimization of perovskite solar cells based on the Bayesian approach," Sol. Energy **262**, 111853 (2023).

[31] T. J. Jacobsson, A. Hultqvist, A. García-Fernández, A. Anand, A. Al-Ashouri, A. Hagfeldt, A. Crovetto, A. Abate, A. G. Ricciardulli, A. Vijayan *et al.*, "An open-access database and analysis tool for perovskite solar cells based on the FAIR data principles," Nat. Energy **7**(1), 107 (2021).

[32] D. W. Hosmer, Jr., S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression* (John Wiley & Sons, 2013).

[33] W. Loh, "Fifty years of classification and regression trees," Int. Stat. Rev. **82**(3), 329 (2014).

[34] G. Biau and E. Scornet, "A random forest guided tour," Test **25**(2), 197 (2016).

[35] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," Artif. Intell. Rev. **54**(3), 1937 (2020).

[36] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," Mach. Learn. **63**, 3 (2006).

[37] Ž. Vujović, "Classification model evaluation metrics," Int. J. Adv. Comput. Sci. Appl. **12**(6), 599 (2021).

[38] T. Fushiki, "Estimation of prediction error by using K-fold cross-validation," Stat. Comput. **21**, 137 (2011).

[39] R. Polikar, "Ensemble based systems in decision making," IEEE Circuits Syst. Mag. **6**(3), 21 (2006).

[40] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, paper presented at the Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019.

[41] Y. W. Noh, J. H. Lee, I. S. Jin, S. H. Park, and J. W. Jung, "Tailored electronic properties of Zr-doped SnO nanoparticles for efficient planar perovskite solar cells with marginal hysteresis," Nano Energy **65**, 104014 (2019).

[42] J. Zhang, X. Li, L. Wang, J. Yu, S. Wageh, and A. A. Al-Ghamdi, "Enhanced performance of $CH_3NH_3PbI_3$ perovskite solar cells by excess halide modification," Appl. Surf. Sci. **564**, 150464 (2021).

[43] P. Schulz, A. M. Dowgiallo, M. Yang, K. Zhu, J. L. Blackburn, and J. J. Berry, "Charge transfer dynamics between carbon nanotubes and hybrid organic metal halide perovskite films," J. Phys. Chem. Lett. **7**(3), 418 (2016).

[44] S. Agbolaghi, "Efficacy beyond 17% via engineering the length and quality of grafts in organic halide perovskite/CNT photovoltaics," New J. Chem. **43**(26), 10567 (2019).

[45] Y.-H. Seo, E.-C. Kim, S.-P. Cho, S.-S. Kim, and S.-I. Na, "High-performance planar perovskite solar cells: Influence of solvent upon performance," Appl. Mater. Today **9**, 598 (2017).

[46] L. Yang, Y. Gao, Y. Wu, X. Xue, F. Wang, Y. Sui, Y. Sun, M. Wei, X. Liu, and H. Liu, "Novel insight into the role of chlorobenzene antisolvent engineering for highly efficient perovskite solar cells: Gradient diluted chlorine doping," ACS Appl. Mater. Interfaces **11**(1), 792 (2019).

[47] E. J. Juarez-Perez, M. R. Leyden, S. Wang, L. K. Ono, Z. Hawash, and Y. Qi, "Role of the dopants on the morphological and transport properties of spiro-MeOTAD hole transport layer," Chem. Mater. **28**(16), 5702 (2016).

[48] X. Jia, Z. Jiang, X. Chen, J. Zhou, L. Pan, F. Zhu, Z. Sun, and S. Huang, "Highly efficient and air stable inverted polymer solar cells using LiF-modified ITO cathode and MoO_3/AgAl alloy anode," ACS Appl. Mater. Interfaces **8**(6), 3792 (2016).

[49] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S. I. Lee, "From local explanations to global understanding with explainable AI for Trees," Nat. Mach. Intell. **2**(1), 56 (2020).

[50] N. J. Jeon, J. H. Noh, Y. C. Kim, W. S. Yang, S. Ryu, and S. I. Seok, "Solvent engineering for high-performance inorganic-organic hybrid perovskite solar cells," Nat. Mater. **13**(9), 897 (2014).

[51] N. Ahn, D.-Y. Son, I.-H. Jang, S. M. Kang, M. Choi, and N.-G. Park, "Highly reproducible perovskite solar cells with average efficiency of 18.3% and best efficiency of 19.7% fabricated via Lewis base adduct of lead(II) iodide," J. Am. Chem. Soc. **137**(27), 8696 (2015).

[52] G. Yang, H. Tao, P. Qin, W. Ke, and G. Fang, "Recent progress in electron transport layers for efficient perovskite solar cells," J. Mater. Chem. A **4**(11), 3970 (2016).

[53] W. Ke, G. Fang, Q. Liu, L. Xiong, P. Qin, H. Tao, J. Wang, H. Lei, B. Li, J. Wan *et al.*, "Low-temperature solution-processed tin oxide as an alternative electron transporting layer for efficient perovskite solar cells," J. Am. Chem. Soc. **137**(21), 6730 (2015).

[54] K. Wang, T. You, R. Yin, B. Fan, J. Liu, S. Cui, H. Chen, and P. Yin, "Precise nucleation regulation and defect passivation for highly efficient and stable carbon-based $CsPbI_2Br$ perovskite solar cells," ACS Appl. Energy Mater. **4**(4), 3508 (2021).

[55] A. D. Taylor, Q. Sun, K. P. Goetz, Q. An, T. Schramm, Y. Hofstetter, M. Litterst, F. Paulus, and Y. Vaynzof, "A general approach to high-efficiency perovskite solar cells by any antisolvent," Nat. Commun. **12**(1), 1878 (2021).

[56] L. Calió, S. Kazim, M. Grätzel, and S. Ahmad, "Hole-transport materials for perovskite solar cells," Angew. Chem. Int. Ed. **55**(47), 14522 (2016).

[57] G. Ayvazyan, F. Gasparyan, and V. Gasparian, "Optical simulation and experimental investigation of the crystalline silicon/black silicon/perovskite tandem structures," Opt. Mater. **140**, 113879 (2023).

[58] F. Sahli, J. Werner, B. A. Kamino, M. Bräuninger, R. Monnard, B. Paviet-Salomon, L. Barraud, L. Ding, J. J. Diaz Leon, and D. Sacchetto, "Fully textured monolithic perovskite/silicon tandem solar cells with 25.2% power conversion efficiency," Nat. Mater. **17**(9), 820 (2018).

[59] K. A. Bush, A. F. Palmstrom, Z. J. Yu, M. Boccard, R. Cheacharoen, J. P. Mailoa, D. P. McMeekin, R. L. Z. Hoye, C. D. Bailie, and T. Leijtens, "23.6%-efficient monolithic perovskite/silicon tandem solar cells with improved stability," Nat. Energy **2**(4), 1 (2017).

Supplementary Material

Bayesian reverse design of high-efficiency perovskite solar cells based on experimental knowledge constraints

Hongyu Liu[1,#], Zhengxin Chen[1,#],Yaping Zhang[1], Jiang Wu[2], Lin Peng[1], Yanan Wang[1], Xiaolin Liu[1,*], Xianfeng Chen[3,4,*], and Jia Lin[1,*]

1. College of Mathematics and Physics, Shanghai University of Electric Power, Shanghai 200090, China

2. College of Energy and Mechanical Engineering, Shanghai University of Electric Power, Shanghai 200090, China.

3. State Key Laboratory of Advanced Optical Communication Systems and Networks, School of Physics and Astronomy, Shanghai Jiao Tong University, Shanghai 200240, China

4. Collaborative Innovation Center of Light Manipulation and Applications, Shandong Normal University, Jinan 250358, China

# These authors contributed equally to this work.

*Corresponding authors: xlliu@shiep.edu.cn (Xiaolin Liu); xfchen@sjtu.edu.cn (Xianfeng Chen); jlin@shiep.edu.cn (Jia Lin)

Method

Classification evaluation indicators

Accuracy is the proportion of correctly predicted samples out of the total samples, which is a commonly used evaluation indicator for classification tasks. However, in the case of class imbalance, it may mislead the assessment of the model. Formula provided in Eq. (A1):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{A1}$$

Where true positive (TP) represents the number of samples correctly predicted as positive by the model; True negative (TN) represents the number of samples correctly predicted as negative; False positive (FP) represents the number of negative samples incorrectly predicted as positive by the model; False negative (FN) represents the number of positive samples incorrectly predicted as negative by the model.

F1 score is the harmonic mean of precision and recall, suitable for problems with class imbalance, especially when there are more negative instances than positive ones. Formula provided in Eq. (A2):

$$F1 = \frac{2 \times Precision \times Recall}{Precision+Recall} \tag{A2}$$

Where $Precision = \frac{TP}{TP+FP}$, indicates the proportion of correctly predicted positive samples out of all samples predicted as positive; $Recall = \frac{TP}{TP+FN}$, indicates the proportion of correctly predicted positive samples out of all actual positive samples.

Area Under the Curve (AUC) is the area under the receiver operating characteristic (ROC) curve, assessing the model's ability to distinguish between positive and negative instances in terms of ranking quality. Its advantage lies in its robustness to imbalanced samples. Formula provided in Eq. (A3):

$$AUC = \int_0^1 TPR[FPR^{-1}(t)]dt \tag{A3}$$

Where true positive rate (TPR)is the ratio of actual positive instances that the model accurately identifies as positive. False positive rate (FPR) is the ratio of actual negative instances that the model mistakenly identifies as positive.

Bayesian optimization

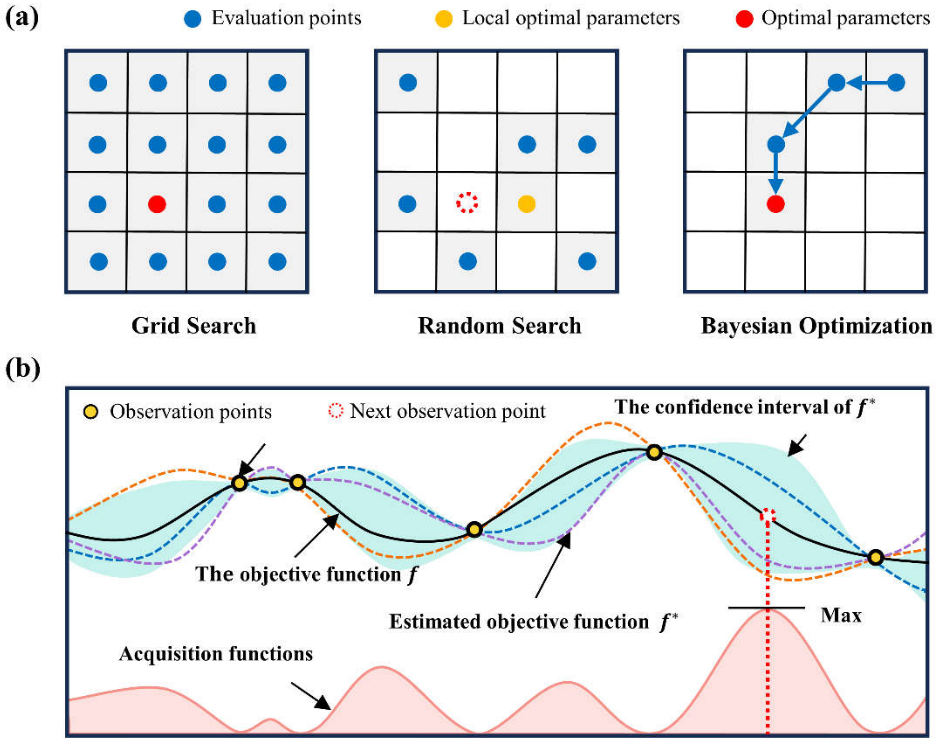


Figure S1. The advantages and principles of BO. (a) Comparison of BO with grid search and random search. (b) Visualization of the principles of BO.

The bayesian optimization (BO) is widely recognized as the current state-of-the-art method in the field of hyperparameter optimization[1,2]. As shown in Figure S1a, compared to grid search and random search methods, BO can efficiently find the optimal hyperparameter configuration with relatively few iterations and avoid falling into local optimal parameters. The efficacy of the algorithm stems from its foundation in bayesian theory. It continuously updates prior information and observed data to obtain the posterior probability distribution of the objective function, thereby guiding the next steps in the optimization decision-making process.

The bayesian theorem formula is provided in Eq. (A4):

$$p(f|D_{1:t}) = \frac{p(D_{1:t}|f)p(f)}{p(D_{1:t})} \tag{A4}$$

$p(f|D_{1:t})$ is the posterior distribution of the unknown target function f given the observed data $D_{1:t}$, where $D_{1:t} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_t, y_t)\}$. Here, x_t is the decision vector, and y_t is the observed value corresponding to x_t. $p(D_{1:t}|f)$ is the likelihood distribution of the observed values, representing the probability distribution of the observed data given the target function f . $p(f)$ is the prior probability distribution of the target function f, indicating assumptions about the state of the target function before observing the data. $p(D_{1:t})$ is the marginal likelihood distribution, representing the overall probability distribution of the observed data.

The following is the process of BO (the meanings of some parameters depicted in Figure S1b): (i) Define the objective function f to be optimized and the range of decision variables x; (ii) Randomly select initial observation points to construct a set of observation points; (iii) Build a probabilistic surrogate model, which can estimate the objective function (f^*) based on the observation points set. Common probabilistic surrogate models include Gaussian processes and Gaussian mixture models; (iv) Utilize the current probabilistic surrogate model to calculate an acquisition function, which measures the impact of observing points on fitting f^*; (v) Select the point with the maximum value of the acquisition function for actual observation, and add it to the observation point set to update the probabilistic surrogate model; (vi) Iterate through steps (iv)-(v) until the stopping criterion is met. Finally, based on the results of Bo, return the estimated minimum or maximum value of f^* and the corresponding decision variables.

Sampling strategy

RandomUnderSampler (RUS) sampling is the process of randomly selecting a certain number of samples from a majority class (i.e. a class with a large number of samples), in order to achieve a balance between the sample sizes of the majority and minority classes; Synthetic Minority Over-sampling Technique (SMOTE) sampling can generate new minority class samples that are similar to the original samples based on distance, thereby increasing the number of minority class samples[3].The formulas for undersampling rate α_{us} and oversampling rate α_{os} are provided in Eq. (A4) and Eq. (A5), respectively.

$$\alpha_{us} = \frac{N_m}{N_{rM}} \tag{A4}$$

where N_m is the number of samples in the minority class and N_{rM} is the number of samples in the majority class after resampling.

$$\alpha_{os} = \frac{N_{rm}}{N_M} \tag{A5}$$

where N_{rm} is the number of samples in the minority class after resampling and N_M is the number of samples in the majority class.

Dataset Establishment

Table S1. Descriptions of 21 features in the $MAPbI_3$ dataset.

Input feature	Feature description
Cell_arch	The device architecture with respect to the direction of current flow and the order in which layers are deposited(nip/pin).
Sub_stack, Etl_stack, Htl_stack, Back_stack	The stack sequence, containing material selection information.
Etl_add, Pvk_add, Pvk_anti_add, Htl_add	List of the dopants and additives.
Etl_dep, Pvk_dep, Htl_dep, Back_dep	The deposition procedures.
Pvk_ABC3	TRUE if the photo-absorber has a perovskite structure.
Pvk_excess	Components that are in excess in the perovskite synthesis.
DMF/DMSO	The ratio of DMF in the entire DMF to DMSO mixed solvent.
Pvk_IPA	TRUE if there has been a IPA step.
Pvk_anti	The solvents used in the antisolvent treatment.
Pvk_solv_anneal	TRUE if there has been a separate solvent annealing step.
Pvk_anneal	The temperature and time during the solvent annealing step.
Back_thick	The list of thicknesses in the stack.

The thickness of the perovskite layer is a key parameter affecting the performance of perovskite solar cells (PSCs), significantly influencing the power conversion efficiency of the PSCs[4]. Choosing the appropriate perovskite layer thickness is crucial for device optimization. However, after performing a statistical analysis of the missing values in the initial dataset, we found that the missing value proportion for the perovskite layer thickness was as high as 68.5%. Using missing data imputation methods (such as multiple imputation) to fill in these missing values could introduce significant bias, leading to data distortion and substantially reducing reliability. We prefer to use more accurate and reliable data to train the model, since studies have shown that the thickness of the perovskite layer has a significant impact on power conversion efficiency. Additionally, we considered using parameters closely related to the perovskite layer thickness, such as the concentration of the perovskite precursor solution, as an indirect substitute for thickness data. However, the missing value proportions for these substitute parameters were still too high to meet the requirements for model training. Therefore, to ensure the accuracy and reliability of the framework, we have carefully decided not to include this feature in the model training.

ML model constructing

See the link for the main code: https://github.com/wthappy/BO_Reverse_design.

Two-stage sampling

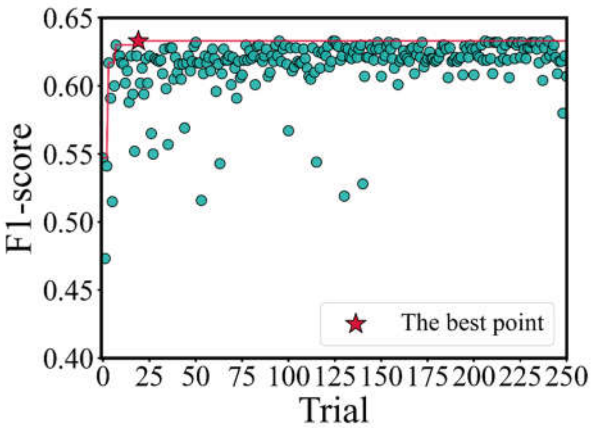


Figure S2. Experimental trial chart based on BO to select the sampling ratio for achieving the best F1 score. Peak value reached at the 19th iteration.

Table S2. The comparison of three sampling strategies.

	RandomUnderSampler sampling		
Model	Accuracy	f1 score	AUC
RF	0.746	0.743	0.829
LGBM	0.758	0.760	0.834

ET	0.734	0.728	0.808
XGB	0.754	0.755	0.827
	SMOTE sampling		
Model	Accuracy	f1 score	AUC
RF	0.868	0.509	0.849
LGBM	0.862	0.500	0.853
ET	0.859	0.488	0.819
XGB	0.867	0.513	0.848
	Two-stage sampling		
Model	Accuracy	f1 score	AUC
RF	0.807	0.633	0.848
LGBM	0.798	0.622	0.849
ET	0.789	0.599	0.825
XGB	0.797	0.626	0.852

Feature selection

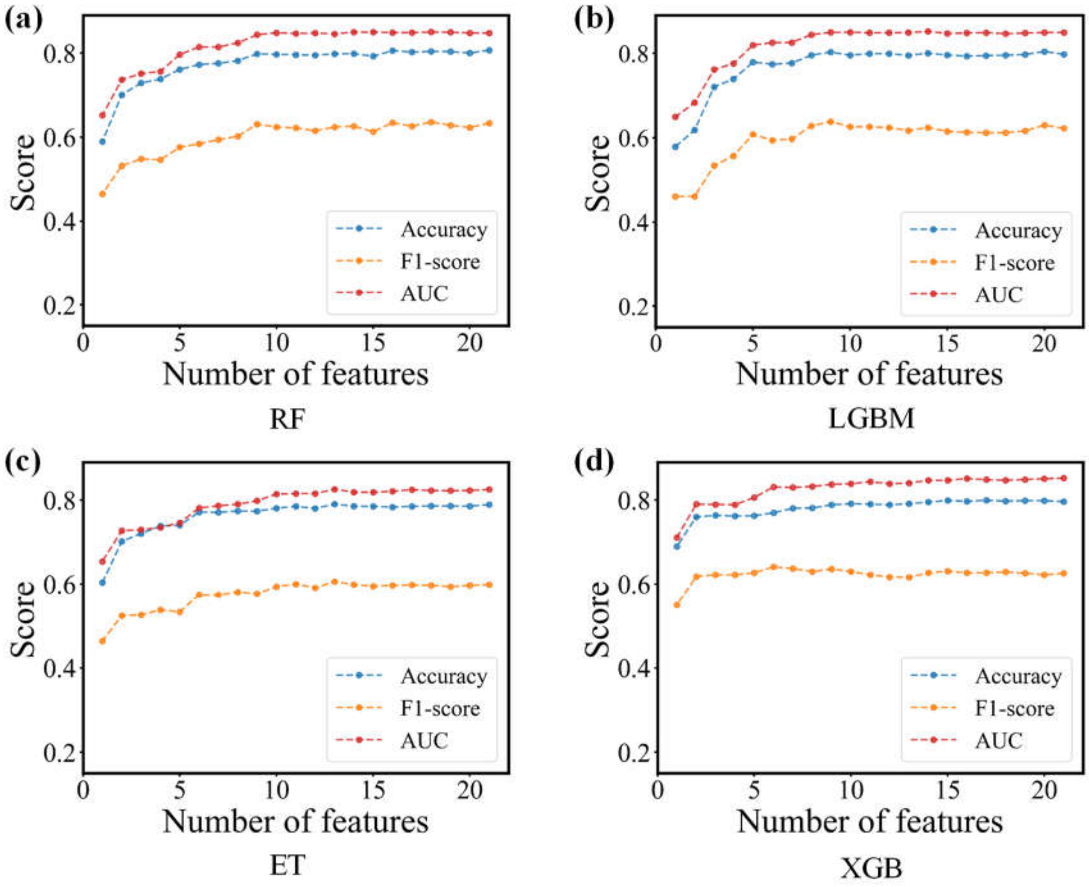


Figure S3. Influence of the number of features on the performance of (a) RF, (b) LGBM, (c) ET, and (d) XGB models.

Table S3. The lists of the optimal feature subset, sorted according to the features importance.

No.	Feature	No.	Feature
1	Pvk_anneal	9	Htl_stack
2	DMF/DMSO	10	Pvk_anti
3	Pvk_add	11	Back_stack
4	Pvk_dep	12	Etl_add
5	Etl_stack	13	Pvk_excess
6	Back_thick	14	Sub_stack
7	Etl_dep	15	Pvk_IPA
8	Htl_add	16	Cell_arch

Hyperparameter optimization

Table S4. The hyperparameters, optimization range, and optimal hyperparameter values of the binary classification algorithm in this work.

	RF	
Hyperparameter	Optimization range	Optimal hyperparameter
n_estimators	(100, 300)	184
max_depth	(1,30)	18
min_samples_split	(2,20)	2
min_samples_leaf	(1,20)	2
	LGBM	
Hyperparameter	Optimization range	Optimal hyperparameter
n_estimators	(100,300)	148
max_depth	(5,30)	12
learning_rate	[0.01,0.015,0.025,0.05,0.1,0.15,0.2]	0.05
min_split_gain	(0,1.0)	0.45
subsample	(0.5,1.0)	0.61
colsample_bytree	(0.5,1.0)	0.62
	ET	
Hyperparameter	Optimization range	Optimal hyperparameter
n_estimators	(100,300)	213
max_depth	(1,30)	10
min_samples_split	(2,30)	3
min_samples_leaf	(1,20)	3
	XGB	
Hyperparameter	Optimization range	Optimal hyperparameter
eta	(0.001,0.1)	0.388
max_depth	(2,30)	10
gamma	(0.01,1.0)	0.223
min_child_weight	(1,10)	1
subsample	(0.5,1.0)	0.948
colsample_bytree	(0.5,1.0)	0.538

Voting

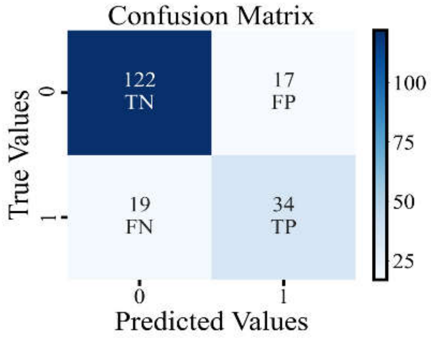


Figure S4. The confusion matrix of voting model.

Table S5. Evaluation indicators of the five classification models.

Model	Accuracy	F1 score	AUC
RF	0.812	0.650	0.856
LGBM	0.806	0.641	0.853
ET	0.785	0.636	0.835
XGB	0.817	0.661	0.858

Optimal parameters finding

Table S6. The feature elements included in the feature association dictionary.

Dictionary	Feature elements
Etl_nip_dic Etl_pin_dic	"Etl_stack", "Etl_add","Etl_dep"
Htl_nip_dic Htl_pin_dic	"Htl_stack","Htl_add"
Pvk_list_dic	"Pvk_stack", " Pvk_add"," Pvk_anti "," Pvk_dep"

Table S7. The hyperparameter space for the materials and fabrication processes of $MAPbI_3$ devices in BO.

Hyperparameter	Hyperparameter space	Hyperparameter	Hyperparameter space
Etl_nip_dic	(0,167)	Htl_nip_dic	(0,87)
Etl_pin_dic	(0,51)	Htl_pin_dic	(0,81)
Pvk_list_dic	(0,1386)	Cell_architecture	(0,1)
Pvk_excess	(0,12)	Sub_stack	(0,1)
DMF/DMSO	(0,1.0)	Back_stack	(0,8)
Pvk_IPA	(0,1)	Back_thick	(0,550.0)
Pvk_anneal	(0,455)		

Due to the limited amount of $MAPbI_3$ dataset, we conducted a statistical and analytical study of the "Back_thick" parameter across the PSC dataset[5] (as shown in Figure S5). We found that the thickness distribution of various metal electrodes is relatively balanced, mostly concentrated between 0-300 nm, without the phenomenon of being significantly higher by an order of magnitude as seen with carbon electrodes. In addition, some data have a higher thickness of metal electrodes. After inquiring about the data source, these data are not incorrectly counted but are real data. Based on the above reasons, we did not establish a thickness association dictionary for specific electrodes, but instead used the maximum thickness of the metal electrode in the $MAPbI_3$ dataset as the threshold for "Back_thick" space exploration.

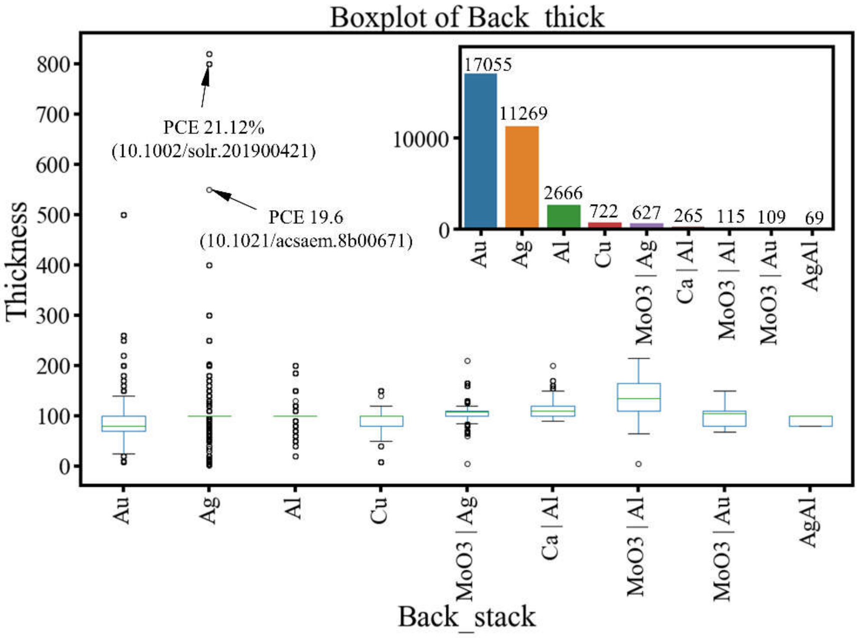


Figure S5. The Box plot of "Back_thick" in the PSC dataset.

References

[1]Jasper Snoek, Hugo Larochelle, and Ryan P Adams, in Advances in neural information processing systems (2012), Vol. 25.

[2]Stewart Greenhill, Santu Rana, Sunil Gupta, Pratibha Vellanki, and Svetha Venkatesh,“Bayesian Optimization for Adaptive Experimental Design: A Review,” IEEE Access 8, 13937 (2020).

[3]N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer,“SMOTE: Synthetic Minority Over-sampling Technique,” Journal of Artificial Intelligence Research 16, 321 (2002).

[4] A Mortadi, E El Hafidi, M Monkade, and R El Moznine,"Investigating the influence of absorber layer thickness on the performance of perovskite solar cells: A combined simulation and impedance spectroscopy study," Materials Science for Energy Technologies 7, 158 (2024).

[5] T. Jesper Jacobsson, Adam Hultqvist, Alberto García-Fernández, Aman Anand, Amran Al-Ashouri, Anders Hagfeldt, Andrea Crovetto, Antonio Abate, Antonio Gaetano Ricciardulli, Anuja Vijayan et al.,"An open-access database and analysis tool for perovskite solar cells based on the FAIR data principles," Nat. Energy 7 (1), 107 (2021).