# Machine Learning for Screening Small Molecules as Passivation Materials for Enhanced Perovskite Solar Cells

*Xin Zhang, Bin Ding, Yao Wang, Yan Liu, Gao Zhang, Lirong Zeng, Lijun Yang, Chang-Jiu Li, Guanjun Yang,* Mohammad Khaja Nazeeruddin,* and Bo Chen**

**Utilization of small molecules as passivation materials for perovskite solar cells (PSCs) has gained significant attention recently, with hundreds of small molecules demonstrating passivation effects. In this study, a high-accuracy machine learning model is established to identify the dominant molecular traits influencing passivation and efficiently screen excellent passivation materials among small molecules. To address the challenge of limited available dataset, a novel evaluation method called random-extracted and recoverable cross-validation (RE-RCV) is proposed, which ensures more precise model evaluation with reduced error. Among 31 examined features, dipole moment is identified, hydrogen bond acceptor count, and HOMO-LUMO gap as significant traits affecting passivation, offering valuable guidance for the selection of passivation molecules. The predictions are experimentally validate with three representative molecules: 4-aminobenzenesulfonamide, 4-Chloro-2-hydroxy-5-sulfamoylbenzoic acid, and Phenolsulfonphthalein, which exhibit capability to increase absolute efficiency values by over 2%, with a champion efficiency of 25.41%. This highlights its potential to expedite advancements in PSCs.**

## 1. Introduction

Perovskite solar cells (PSCs) have rapidly evolved as a cutting-edge photovoltaic technology, undergoing substantial advancements in recent years.[1] This progress can be attributed to the excellent properties of perovskite materials, including an appropriate band gap,[2,3] large absorption coefficient,[4] high carrier mobility,[5,6] long carrier diffusion length,[7] and low exciton binding energy.[8] However, the presence of defects within perovskite films, including both point defects (such as vacancy, interstitial, and anti-site defects) and extended defects (such as surface and grain boundaries),[9,10] inevitably leads to nonradiative recombination, which reduces the power conversion efficiency (PCE) of PSCs.[11,12] Therefore, passivating these defects in the perovskite film has become a focal point of recent research efforts, and numerous studies have emphasized the significance of defect passivation in boosting photovoltaic performance of PSCs.[10] Diverse passivation materials have been reported, categorized as small molecules,[13,14] polymers,[15,16] and ionic compounds.[17,18] These materials affect by bonding with uncoordinated $Pb^{2+}$ or $I^-$ ions in the perovskite or forming low-dimensional passivation layer.[19,20]

Traditional experimental screening of effective passivation materials requires substantial time and resources. In an era of data-driven insights, artificial intelligence, particularly machine learning (ML), has emerged as a pivotal tool in scientific research due to its robust data processing capabilities.[21,22] ML empowers researchers to scrutinize existing databases, uncover hidden relationships among molecular characteristics, and anticipate uncharted outcomes. In the field of PSCs, ML has found broad applications.[23,24] For instance, Hartono et al.[25] successfully employed ML to explore the relationship between organic halide salt properties and the stability of capping layers on perovskite films, leading to the prediction of optimal performance with phenyltriethylammonium iodide. ML methodologies have also been applied to understand the link between ionic compounds and passivation effects.[26–29] In comparison to polymers and ionic compounds, there exist millions of small molecules in molecular library and still being expanded. While a few hundred molecules have been applied within the realm of perovskite passivation, the vast majority of molecules remain largely unexplored, offering substantial untapped potential. Comprehensive molecular databases such as PubChem and ChemSpider offer diverse molecular traits that can be readily transformed into digital datasets.[30] Nevertheless, despite the promising prospects, a thorough exploration of small molecule passivation for PSCs through the lens of ML remains notably absent.

In this study, we focus on developing ML model to identify key molecular traits responsible for effective PSC passivation and screen small molecules with remarkable passivation effects.

X. Zhang, Y. Wang, Y. Liu, G. Zhang, L. Zeng, L. Yang, C.-J. Li, G. Yang, B. Chen
State Key Laboratory for Mechanical Behavior of Materials
Xi'an Jiaotong University
Xi'an, Shaanxi 710049, P. R. China
E-mail: ygj@mail.xjtu.edu.cn; bochen@xjtu.edu.cn

B. Ding, M. K. Nazeeruddin
Institute of Chemical Sciences and Engineering
École Polytechnique Fedérale de Lausanne (EPFL)
Lausanne 1015, Switzerland
E-mail: mdkhaja.nazeeruddin@epfl.ch

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/adfm.202314529

**DOI: 10.1002/adfm.202314529**

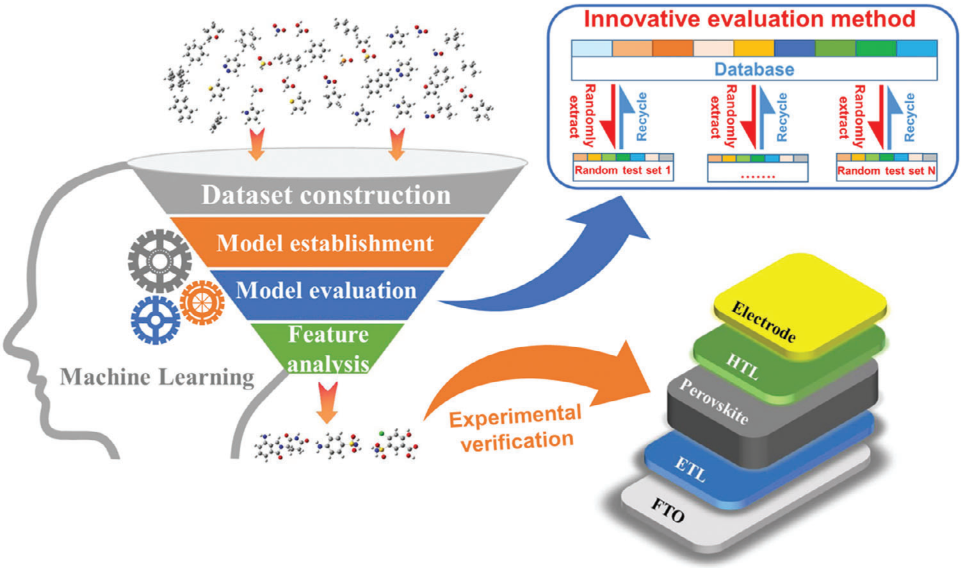


**Figure 1.** The machine learning process involves four steps: dataset construction, model establishment, model evaluation, and feature analysis. An innovative RE-RCV evaluation method is introduced, where test set is randomly extracted from the dataset and subsequently recycled after evaluation.

The dataset encompasses hundreds of small molecules that have been documented as effective passivation materials, thus providing a dataset as foundation for ML analysis. It is crucial to note that a limited dataset may potentially yield evaluation results with significant deviations.[27] To mitigate this issue, we proposed a random-extracted and recoverable cross-validation (RE-RCV) method to replace the commonly employed K-fold cross-validation method.[31] The establishment of high-accuracy ML model provides a quantitative assessment for the impact of all molecular traits on the passivation effects. This analysis pinpointed three pivotal molecular traits that could guide the selection of outstanding passivation molecules. The predictions generated by our model have led us to identify three representative passivation molecules—4-aminobenzenesulfonamide, 4-Chloro-2-hydroxy-5-sulfamoylbenzoic acid, and phenolsulfonphthalein—that can significantly increase absolute PCE values by more than 2%. Crucially, these predictions have been validated through experimental results, where the PCE of PSCs boosted from 22.40% to 24.99%, 25.41%, and 25.26%, respectively. This further emphasizes the effectiveness and potential of our ML-based approach in advancing perovskite photovoltaics.

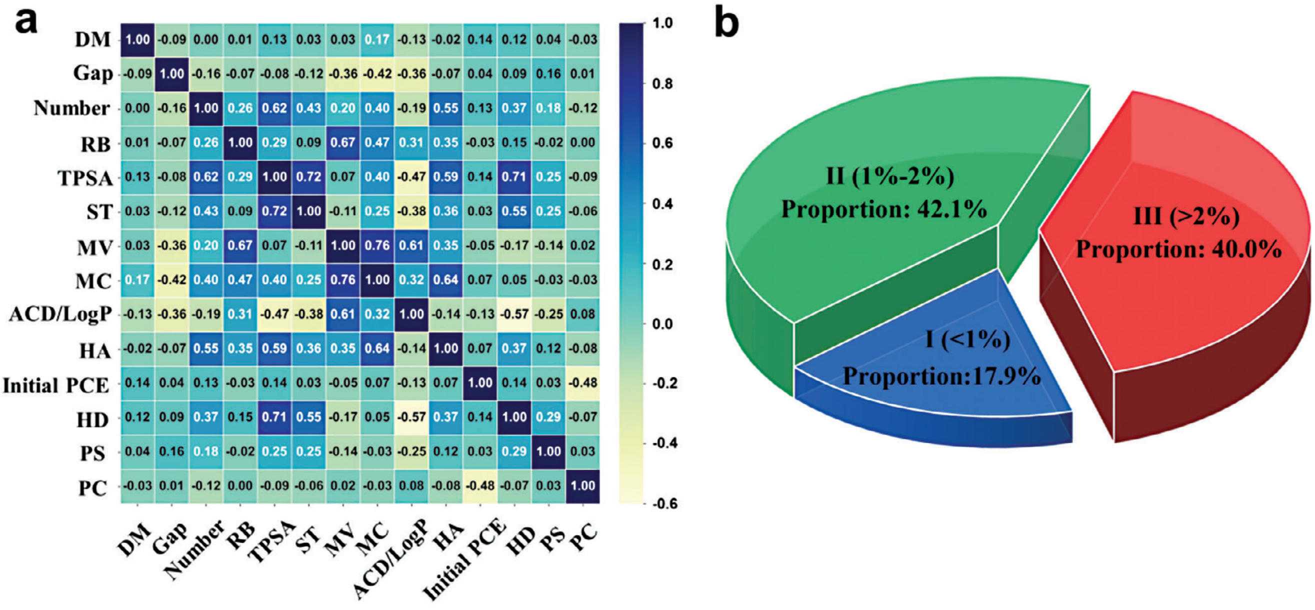
## 2. Results and Discussion

### 2.1. Machine Learning Workflow

**Figure** 1 outlines the four key steps in developing a machine learning model dedicated to select effective passivation molecules from an extensive molecular library for the application in PSCs. In the initial stage, a comprehensive dataset is constructed from a wealth of documented small molecules functioning as passivation materials. This dataset comprises several hundred small molecules, with their molecular traits transformed into a digital dataset. Transitioning to the second stage, ML models are established utilizing diverse algorithms, with training based on the dataset constructed in the first step. In the third step, model evaluation is performed with a goal of minimizing assessment deviation. Notably, the dataset size for reported passivation molecules in this context is relatively modest, in contrast to other domains with large datasets exceeding 10000 or even 100000 data points,[32] resulting in the conventional K-fold cross-validation method (K-Fold) which yields considerable deviation. To mitigate this, we introduce a novel evaluation technology—random-extracted and recoverable cross-validation (RE-RCV) method—tailored for limited datasets whereby enhancing precision in model assessment. The fourth stage centers on feature analysis, quantitatively gauging the influence of each molecular trait on passivation effects. This comprehensive insight enables the ML model to anticipate small molecules exhibiting remarkable passivation outcomes for PSCs.

### 2.2. Dataset Construction

The initial and pivotal step in machine learning is to construct the dataset. We gathered all literatures about small molecule as passivation material in PSCs from 2016 to April 2023. This dataset is divided into input (Input X) and output (Output Y) data to facilitate the subsequent training of ML model. The objective is to uncover correlations between molecular traits and passivation effects on perovskite films. Output Y represents the difference between PCE of perovskite solar cells after and before treatment with passivation molecules ($\Delta PCE = PCE_{\text{passivated}} - PCE_{\text{inital}}$). Input X consists of various features, including physicochemical traits of molecules (like dipole moment, HOMO-LUMO gap, molar volume, rotatable bond count, etc.), perovskite film properties (such as perovskite composition and perovskite film quality), as well as passivation style (bulk passivation, top surface passivation, and bottom surface passivation). There are different types of PSCs device architecture, such as n-i-p or p-i-n, planar or



| | DM | Gap | Number | RB | TPSA | ST | MV | MC | ACD/LogP | HA | Initial PCE | HD | PS | PC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DM | 1.00 | -0.09 | 0.00 | 0.01 | 0.13 | 0.03 | 0.03 | 0.17 | -0.13 | -0.02 | 0.14 | 0.12 | 0.04 | -0.03 |
| Gap | -0.09 | 1.00 | -0.16 | -0.07 | -0.08 | -0.12 | -0.36 | -0.42 | -0.36 | -0.07 | 0.04 | 0.09 | 0.16 | 0.01 |
| Number | 0.00 | -0.16 | 1.00 | 0.26 | 0.62 | 0.43 | 0.20 | 0.40 | -0.19 | 0.55 | 0.13 | 0.37 | 0.18 | -0.12 |
| RB | 0.01 | -0.07 | 0.26 | 1.00 | 0.29 | 0.09 | 0.67 | 0.47 | 0.31 | 0.35 | -0.03 | 0.15 | -0.02 | 0.00 |
| TPSA | 0.13 | -0.08 | 0.62 | 0.29 | 1.00 | 0.72 | 0.07 | 0.40 | -0.47 | 0.59 | 0.14 | 0.71 | 0.25 | -0.09 |
| ST | 0.03 | -0.12 | 0.43 | 0.09 | 0.72 | 1.00 | -0.11 | 0.25 | -0.38 | 0.36 | 0.03 | 0.55 | 0.25 | -0.06 |
| MV | 0.03 | -0.36 | 0.20 | 0.67 | 0.07 | -0.11 | 1.00 | 0.76 | 0.61 | 0.35 | -0.05 | -0.17 | -0.14 | 0.02 |
| MC | 0.17 | -0.42 | 0.40 | 0.47 | 0.40 | 0.25 | 0.76 | 1.00 | 0.32 | 0.64 | 0.07 | 0.05 | -0.03 | -0.03 |
| ACD/LogP | -0.13 | -0.36 | -0.19 | 0.31 | -0.47 | -0.38 | 0.61 | 0.32 | 1.00 | -0.14 | -0.13 | -0.57 | -0.25 | 0.08 |
| HA | -0.02 | -0.07 | 0.55 | 0.35 | 0.59 | 0.36 | 0.35 | 0.64 | -0.14 | 1.00 | 0.07 | 0.37 | 0.12 | -0.08 |
| Initial PCE | 0.14 | 0.04 | 0.13 | -0.03 | 0.14 | 0.03 | -0.05 | 0.07 | -0.13 | 0.07 | 1.00 | 0.14 | 0.03 | -0.48 |
| HD | 0.12 | 0.09 | 0.37 | 0.15 | 0.71 | 0.55 | -0.17 | 0.05 | -0.57 | 0.37 | 0.14 | 1.00 | 0.29 | -0.07 |
| PS | 0.04 | 0.16 | 0.18 | -0.02 | 0.25 | 0.25 | -0.14 | -0.03 | -0.25 | 0.12 | 0.03 | 0.29 | 1.00 | 0.03 |
| PC | -0.03 | 0.01 | -0.12 | 0.00 | -0.09 | -0.06 | 0.02 | -0.03 | 0.08 | -0.08 | -0.48 | -0.07 | 0.03 | 1.00 |

**Figure 2.** a) Correlation matrix depicting Pearson coefficients among the 14 screened features. The values within the square grid signify Pearson's correlation coefficients, in which positive and negative values represent positive and negative correlation between two features, respectively. b) Proportions of the three classification classes (I-class, II-class, and III-class) in the dataset.

mesoporous. Including all types of device architecture will increase the complexity of the dataset. Thus, we selected the planar n-i-p PSCs due to its largest proportion in published papers, allowing us to center our attention on the passivation effect. With the same rationale, we refined our focus to lead-based perovskites with $ABX_3$ structure, specifically noting changes in X-site anion (such as $I^-$ and $Br^-$) and A-site cation (such as $MA^+$ and $FA^+$) by an input feature of "Perovskite composition". Since different qualities of un-passivated perovskite films with different defect densities influence the passivation effect, thus we utilized an input feature of "Initial PCE" to represent the quality of un-passivated perovskite film. In recent years, computational chemistry has made significant advancements in the field of photovoltaics.[33–35] Thus we converted data derived from density functional theory (DFT) calculations and pre-existing molecular libraries (ChemSpider and PubChem) into digital format to incorporate molecular traits. The description and origin of each individual feature are elaborated upon in Table S1 (Supporting Information). To enhance molecular passivation features, a novel trait "Number" is introduced: the coordination group number of molecules. The criterion for the coordination group is whether the Lewis acid/base groups of small molecules bind with uncoordinated $Pb^{2+}$ or uncoordinated $I^-$ by DFT calculation (Figure S1, Supporting Information), which are the main forms of passivation bond for small molecules.[36,37] Ultimately, the dataset integrates molecular traits and technological parameters, comprising 330 data entries and 31 features (detailed in Table S1, Supporting Information).

Excessive feature dimensions, particularly with a limited dataset, can lead to overfitting.[28] Thus it is necessary to eliminate redundant features to avoid that. Calculating Pearson's correlation coefficient ($p$) between pairs of features allow us to depict feature correlations through correlation matrix, where the magnitude of $|p|$ indicates the strength of correlation.[38] Following the insights of previous studies, a high correlation between two features is observed when $|p|$ exceeds 0.8, allowing us to select one representative feature while excluding its correlated counterpart.[39,40] Based on this criterion and the corresponding correlation matrix of 31 features in Figure S2 (Supporting Information), we were able to effectively reduce the number of representative features to 14 by eliminating redundancy. Subsequently, **Figure 2a** visually illustrates the interrelation among the 14 screened features (detailed in **Table 1**), with all $|p|$ values remaining below 0.8.

Supervised ML algorithms are categorized as regression and classification, each tailored to distinct applications. Despite of that both algorithms are suited for addressing our tasks, the classification model outperformed the regression model in prediction capability. Therefore, our approach centered on constructing a classification ML model to establish a mapping between Input X and Output Y. Based on the enhancement in PCE resulting from defect passivation ($\Delta PCE$), the dataset was divided into three categories: I-class represents $\Delta PCE < 1\%$, II-class represents $1\% \leq \Delta PCE \leq 2\%$, and III-class represents $\Delta PCE > 2\%$. Figure 2b shows the proportion of these three categories within the dataset (I-class: 17.9%, II-class: 42.1%, and III-class: 40.0%).

### 2.3. Model Establishment and Evaluation

We established ML model using mainstream classification algorithms, including support vector machine (SVM),[41] neural network model (NNM),[42] random forest (RF),[43] k-nearest neighbor (KNN),[44] and naive Bayes (NB).[45] It is important to highlight that when dealing with limited datasets, relying solely on a single algorithm can lead to overfitting. To address this concern,

**Table 1.** Detailed definitions of the 14 features screened as Input X in machine learning model.

| Feature | Detailed definition | Feature | Detailed definition |
|---|---|---|---|
| DM | Dipole moment | Number | Number of coordination group |
| Gap | HOMO-LUMO gap | Initial PCE | PCE of un-passivated PSCs |
| RB | Rotatable bond count | ACD/LogP | Lipid-water partition coefficient used to describe hydrophobicity |
| HA | Hydrogen bond acceptor count | HD | Hydrogen bond donor count |
| MV | Molar volume | MC | Molecular complexity |
| TPSA | Topological polar surface area | ST | Surface tension |
| PS | Passivation style | PC | Perovskite composition |

we combined the strengths of these five classification algorithms to create complex ML models, thereby overcoming the limitations associated with individual algorithms and leveraging diverse data-processing architectures via stacked generalization.[46]

Model evaluation plays a crucial role in bridging the gap between the establishment of a ML model and its practical application, offering valuable feedback to select the optimal model among various candidates. For classification model, accuracy stands out as a pivotal criterion for model evaluation. The widely used K-Fold partitions the dataset into k segments with utilizing one as the test set and the others for training.[31] However, for the application scenario with limited dataset (<1000 samples), K-Fold often yields substantial accuracy deviations, undermining the reliability of model assessment.[27] To address this limitation, we introduce a novel model evaluation technique: the random-extracted and recoverable cross-validation method (RE-RCV). As illustrated in the inset of Figure 1, RE-RCV involves random extraction of 20% of test data from the dataset, with the remaining portion used as the training data, followed by calculating model accuracy. After the calculation, the extracted data are reintegrated into the established dataset. Subsequently, another 20% of the data is randomly selected as a test set, and this cycle is repeated 100 times, accomplished through Python code as shown in the "Machine-learning code" section (Supporting Information). The final model accuracy is then determined by averaging the accuracy values obtained from these 100 calculations. The detailed explanation of RE-RCV is provided in the "Model evaluation method" section in Supporting Information.

By applying RE-RCV as our model evaluation approach, we conducted a comprehensive assessment of five mainstream ML classification models. Our analysis revealed that SVM's accuracy is the highest, exceeding 80% (**Figure** 3a), for single algorithm scenario. For a more detailed overview, the specific hyperparameters used for these ML models are provided in Table S2 (Supporting Information). Among different complex models, the combination of SVM and RF models (denoted as S-R model) attains the highest accuracy of 83.7% (Figure 3b). The attempts to combine all five models yielded a similar accuracy of 83.3%, however, it required significantly more computational time than S-R model. As a result, we selected the S-R model for subsequent feature analysis and the screening of passivation molecules.

Moreover, we constructed a regression model with $\Delta PCE$ as the output Y. To assess the effectiveness of the regression models, we employed the coefficient of determination ($R^2$).[47,48] A higher $R^2$ value, closer to 1, indicates better prediction accuracy. However, the $R^2$ values of SVM, KNN, RF, MNN, and SVM+RF models calculated by RE-RCV method, as illustrated in Figure S3 (Supporting Information), we observed that all $R^2$ values were considerably below 1. This suggests that the prediction accuracy of these regression models did not meet our requirements, which is also the reason why we chose classification model to handle the data.

Subsequently, we conducted a rigorous comparison between the RE-RCV and K-Fold methods, using the S-R model as the basis. The results of model evaluation obtained from both methods are influenced by the distribution of the test data, thus we performed 50 iterations for each method by continuously altering the test set for each iteration. The accuracy and corresponding error for the S-R model with varying proportions of test data in RE-RCV and varying K values in K-Fold are illustrated in Figure 3c. It is evident that RE-RCV exhibits a narrower assessment error compared to K-Fold, indicating its superior precision in model evaluation. To quantify the dispersion of accuracy, we introduced the Coefficient of Variation (CV) as a metric:

$$s = \sqrt{\frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}{n}} \tag{1}$$

$$\mathrm{CV} = \frac{s}{\bar{x}} \times 100\% \tag{2}$$

where $s$ and $\bar{x}$ are standard deviation and average value, respectively. Figure 3d demonstrates that RE-RCV provides a lower CV value with reduced dispersion compared to K-Fold. This further underscores the suitability of the RE-RCV method for the precise evaluation of our S-R model. It is worth noting that this RE-RCV method also holds potential applicability to other datasets with limited samples.

In addition to evaluate model accuracy, which assesses the overall probability of accurately predicted data across all Output Y classes, it is equally crucial to understand the precise prediction probability within each individual Output Y class. This probability of predicted class matches with the actual class is denoted as the model precision (detailed in Methods). Calculating the confusion matrix of the S-R model offers insights into classification model precision (Figure 3e). Notably, the highest model precision of 88.6% is observed for the III-class, and model precision exceeding 80% are attained for the I-class and II-class as well. These results indicate that the model excels in accurately predicting molecules within the III-class, which are pivotal passivation

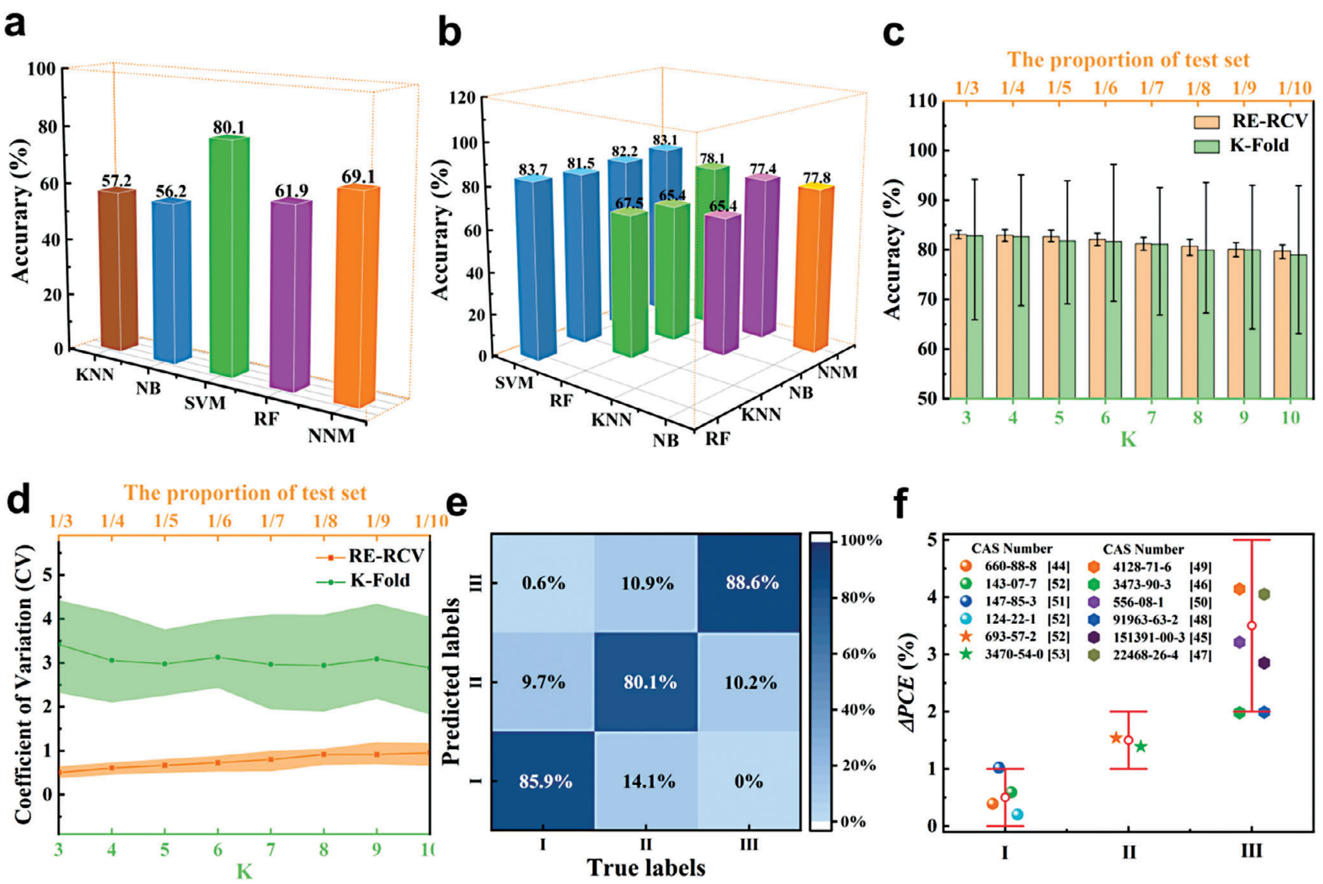


**Figure 3.** Accuracies of a) single ML model and b) complex ML models based on evaluation method of RE-RCV. c) Accuracy and corresponding error of S-R model obtained by RE-RCV and K-Fold. d) Comparison of CV values of S-R model accuracies based on RE-RCV and K-Fold. e) Confusion matrix of the S-R model on the test set. The values within gridding represent the model precision, with color bar from 0–100%. f) Comparison of the predicted PCE Enhancement (orange deviation bar) and reported PCE enhancement (colored dots data) of PSCs after small molecule passivation in the latest published papers.[49–58].

materials for enhancing the photovoltaic performance of PSCs in practical applications. This aligns with the original objective for constructing the ML model.

Furthermore, a third evaluation approach is introduced by comparing our model's predictions with data from the latest publications after April 2023 on defect passivation. As shown in Figure 3f, this comparison between the predicted outcomes from the S-R model and experimental results shows a strong agreement, affirming the reliable prediction ability of our ML model. This suggests that the ML model has the potential to expedite the screening of outstanding passivation materials, avoiding the need for extensive trial and error via experimental methods.

### 2.4. Feature Analysis

To quantitatively illustrate the contribution of each Input X, we employed SHapley Additive exPlanations (SHAP), developed based on generalized game theory for machine learning.[40] **Figure** 4a reveals the importance ranking of Input X through magnitudes of their associated mean absolute SHAP value for the combination of three Output Y classes, where larger absolute SHAP values correspond to a greater contribution to the probability of these molecules classifying into specific classes. Notably, DM, HA, Gap, and Initial PCE are the top four contributors among all Input X. Besides, we also presented the importance ranking of Input X for each Output Y (shown in Figure S4a–c, Supporting Information), wherein the top four ranked features agree with those depicted in Figure 4a. Considering the practical significance of the III-class, Figure 4b depicts the distribution relationship between feature values and their corresponding SHAP values. Here, the x-axis represents the SHAP values of each feature, where positive and negative SHAP values correspond to their positive and negative contribution to classifying molecules as the III-class, and the color bar on the right reflects the relative values of each feature, distinguished by color (transition from blue to red indicates a progression of feature values from low to high). It can be found that positive correlations between feature values and SHAP values are observed for DM and HA, while Gap and Initial PCE exhibit negative correlations. These insights shed light on how specific molecular traits influence passivation and their roles in classifying molecules into different passivation categories.

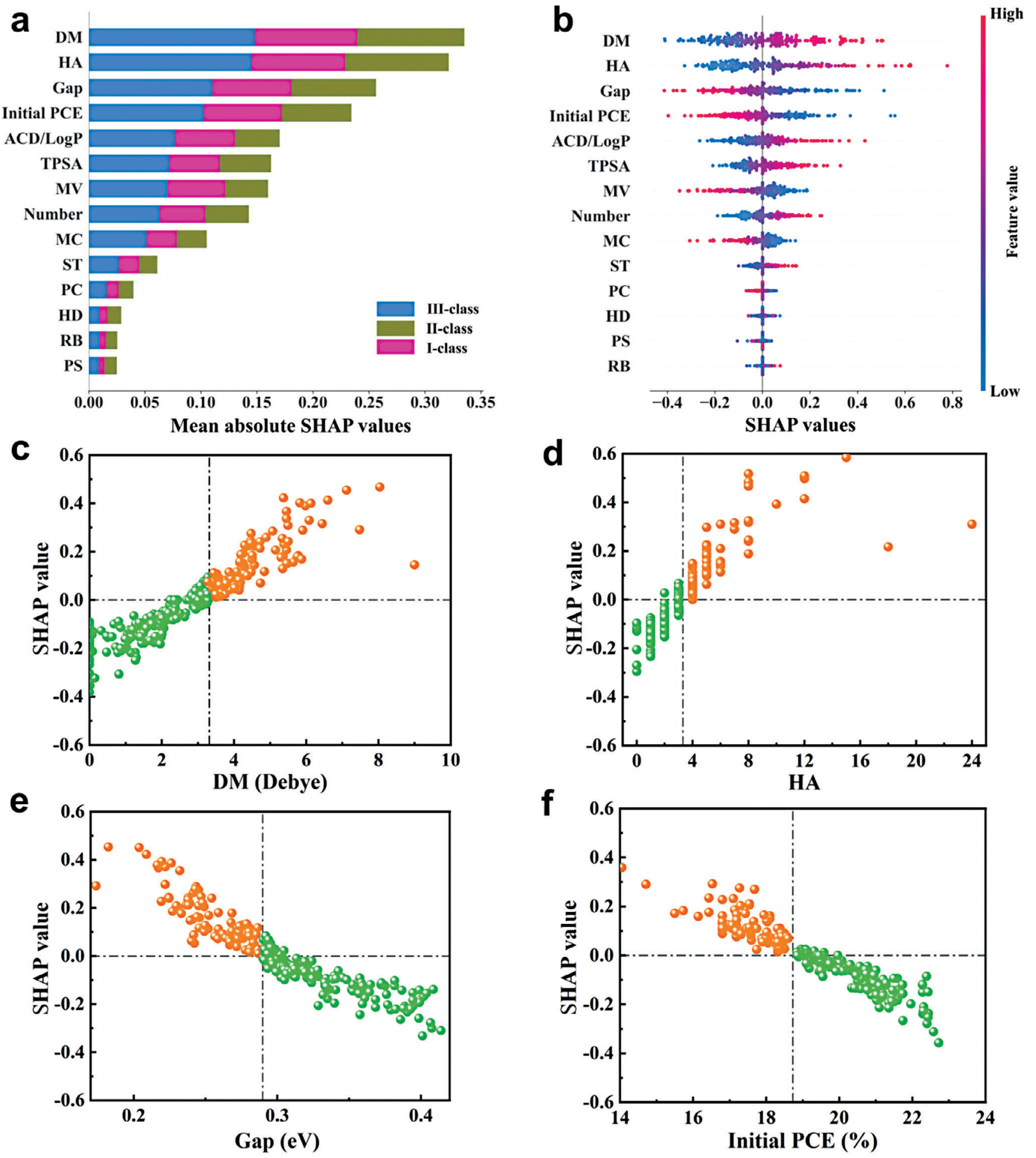


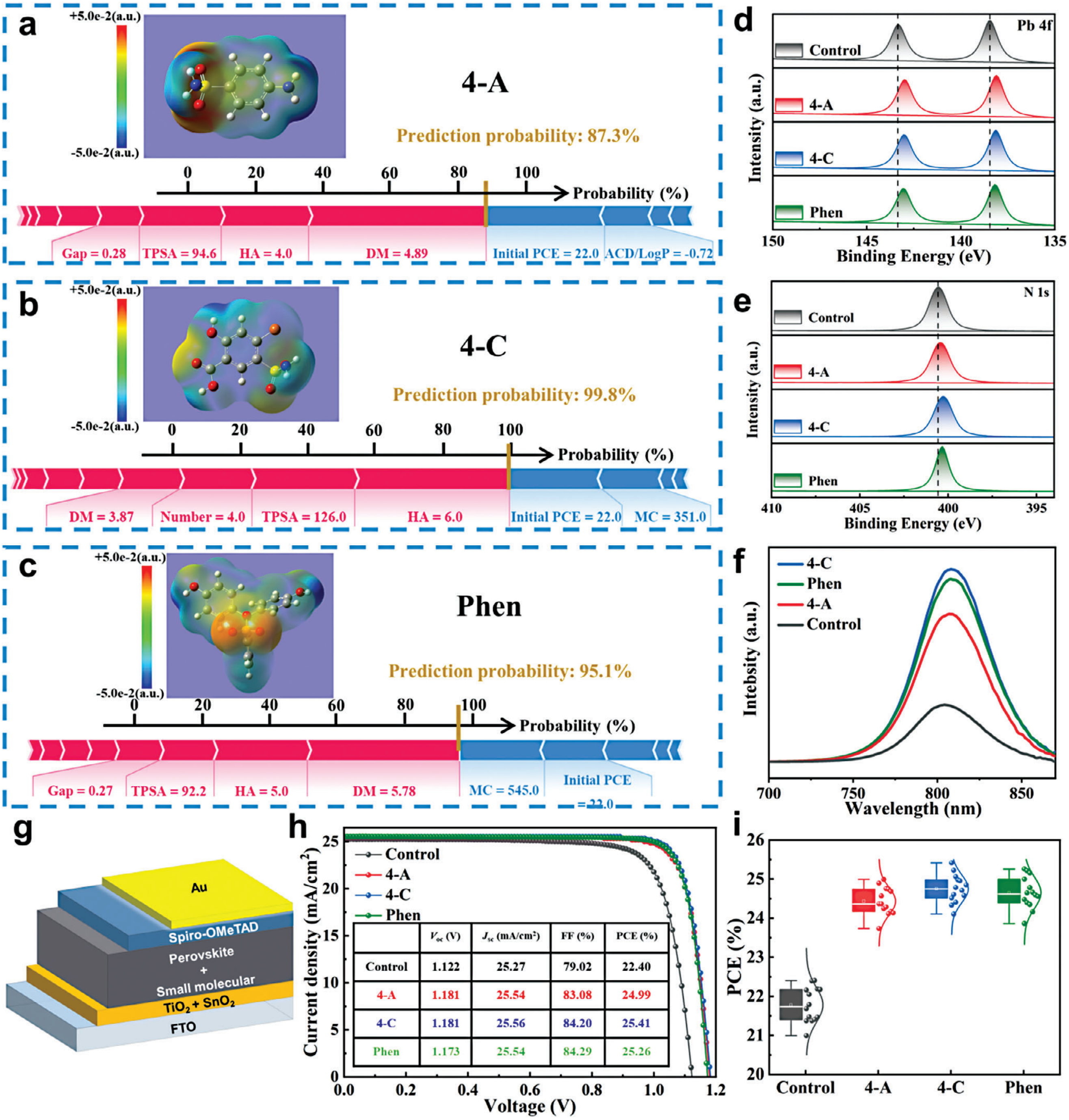
**Figure 4.** a) Importance ranking of Input X obtained from S-R model for the combination of I-class, II-class, and III-class. b) Importance ranking of Input X specifically for III-class, with different Input X arranged in descending order of importance. The red and blue of color bar represents high value and low value of each feature, respectively. Distribution diagrams illustrate the relationship between feature values and corresponding SHAP values for: c) DM, d) HA, e) Gap, and f) Initial PCE.

The analysis for III-class explores further into specific features. Figure 4c–f present distribution diagrams for feature values and their corresponding SHAP values, with a focus on key attributes: DM, HA, Gap, and Initial PCE. For dipole moment (DM) (Figure 4c), larger dipole moment within molecules enhance their polarity, leading to stronger electron-donating ability and increased binding energy with uncoordinated $Pb^{2+}$ by coordinate bond, and whereby improves the passivation effect.[24,59,60] Besides, Passivating molecules can adjust the energy levels of perovskite surfaces based on their own dipole moments to enhance charge carrier extraction and reduce recombination by effectively aligning energy levels at the interface.[50–53] A similar positive relationship applies to hydrogen bond acceptor count (HA) (Figure 4d). For hydrogen bond acceptor count (HA), this trait is related to the ability of a molecule to form hydrogen bonds with neighboring molecules. A higher HA provides more binding sites with the hydrogen bond donators ($FA^+$ or $MA^+$), aiding strong hydrogen bond, improving coverage and stability of the passivation layer and preventing aggregation and $FA^+$/ $MA^+$ migration.[10,61] HOMO-LUMO gap (Gap) of PSCs (Figure 4e) exhibits a negative relationship with SHAP values. Although the precise impact of Gap on passivation effect has not been extensively studied, we speculate that a narrow Gap facilitates carrier transmission through grain boundaries or interfaces between charge-transport-layer and perovskite, improving overall passivation effect. Some literatures reported a connection between high molecular conductivity and a narrow Gap,[62,63] which might cause the influence of Gap on passivation. Given the relatively unexplored nature of Gap in the field of perovskite solar cells, more comprehensive studies are needed to fully understand how Gap influences passivation effect. Furthermore, a lower Initial PCE corresponds to more defects in the perovskite layer, and consequently, an increase of PCE value becomes more pronounced by defect passivation (Figure 4f). Based on these findings in Figure 4c–f, we propose a universal criterion (DM> 3.3 HA> 3.3, Gap < 0.29) that can guide researchers in the selection of excellent passivation materials. Although DM and Gap require calculation using DFT, we have detailed the calculation process in the "DFT calculation" section of the Supporting Information. These parameters can be easily achieved and utilized to guide researchers in the selection of excellent passivation materials.

### 2.5. Screening Small Molecule and Experimental Verification

Our machine learning model holds the capability to predict the passivation effect of individual molecules when their molecular traits are provided as input. However, it is important to acknowledge the practical constraints posed by the vast number of small molecules that have been developed. With millions of such molecules in existence, attempting to predict the passivation effect for all molecules become unfeasible due to the substantial workload required to incorporate traits for every molecule. To efficiently screen for excellent passivation molecules, we utilize the aforementioned criterion (DM> 3.3, HA> 3.3, Gap < 0.29) for preliminary screening among thousand of molecules that is available from Sigma–Aldrich and contains amino ($-NH_2$), cyano group (–CN), carboxyl (–COOH), carbonyl (–C=O), or oxhydryl (–OH), which resulting in a smaller set of dozens of un-reported passivating molecules. Subsequently, the S-R model is employed to predict the passivation effects of those un-reported candidates, and we discover a dozen of these molecules can be classified as III-class (the feature data of screened molecules are displayed in supporting data). Considering factors such as accessibility and cost, we select three representative molecules—4-aminobenzenesulfonamide (4-A), 4-Chloro-2-hydroxy-5-sulfamoylbenzoic acid (4-C), and phenolsulfonphthalein (Phen)—among the predicted materials for subsequent experimental verification. Their molecular configurations and physicochemical parameters are depicted in **Figure 5a–c** and Table S3 (Supporting Information), in line with the pre-screening criterion. Furthermore, Figure 5a–c illustrates the prediction probabilities of three molecules being classified as III-class (4-A: 87.3%; 4-C: 99.8%; Phen: 95.1%). It's important to note that starting from the same initial PCE, a higher predicted probability of being classified as III-class indicates a more effective passivation effect. Even as the initial PCE increases, these molecules can still enhance the PCE for passivated samples, thanks to the consistent passivation mechanism, although there may be changes in the $\Delta PCE$ value. Additionally, the contribution degree related to SHAP values of dominated molecular traits is presented in Figure 5a–c, with more details available in Figure S5 (Supporting Information), showing their contributions to the probability of these molecules classifying into III-class.

As previously mentioned, high HA and DM values substantially contribute to the passivation effect, enhancing the interaction between molecules and perovskite. To validate this, we employed DFT to calculate the binding energies between the three selected molecules and perovskite. Figure S6 (Supporting Information) shows that the Lewis base groups ($-NH_2$, –COOH, –OH, etc.) and hydrogen bond acceptors (O, N, and Cl atoms) within these molecules form strong bonds with uncoordinated $Pb^{2+}$ and $MA^+$/$FA^+$ in perovskite, respectively. To explore the practical impact of these screened molecules on perovskite films, we incorporated them as passivation materials in CsFAMA-based precursor solutions to prepare the passivated perovskite films. Subsequently, X-ray photoelectron spectroscopy (XPS) characterizations were employed to examine the interaction between small molecules and perovskite film. As shown in Figure 5d, the Pb 4f main peaks for perovskite films without passivation treatment (control sample) are located at 138.45 and 143.34 eV. These peaks shift to lower binding energy after passivation with the three small molecules (4-A: 138.08 and 142.96 eV, 4-C: 138.13 and 143.02 eV, Phen: 138.10 and 143.99 eV, respectively), indicating strong coordination bonding formation between uncoordinated $Pb^{2+}$ and Lewis base groups of the three small molecules. A similar shift trend is observed for the N 1s peaks (Figure 5e), attributing to hydrogen bonding between the molecules and $MA^+$ or $FA^+$ ions within the perovskite film. Furthermore, we conducted $^1H$ NMR analysis, in which the peaks corresponding to NH in $FA^+$ split into two peaks when any of the three small molecules were introduced (Figure S7, Supporting Information), demonstrating all three molecules can form strong hydrogen bonds with $FA^+$. The results obtained from XPS and NMR are consistent with DFT calculations, thereby affirming the strong binding between these molecules and perovskite. Subsequently, we carried out steady-state photoluminescence (PL) measurements to explore the non-radiative carrier recombination behavior of the perovskite layer



| | $V_{oc}$ (V) | $J_{sc}$ (mA/cm$^2$) | FF (%) | PCE (%) |
|---|---|---|---|---|
| Control | 1.122 | 25.27 | 79.02 | 22.40 |
| 4-A | 1.181 | 25.54 | 83.08 | 24.99 |
| 4-C | 1.181 | 25.56 | 84.20 | 25.41 |
| Phen | 1.173 | 25.54 | 84.29 | 25.26 |

**Figure 5.** Configurations and predicted probabilities of screened small molecules by S-R ML model: a) 4-A, b) 4-C and c) Phen. The length of the color bar corresponds to the contribution degree related to SHAP values of dominant molecular traits to the probability of classification prediction, with red and blue denoting positive and negative roles, respectively. The intersection between red and blue zones corresponds to the prediction probability of classifying that molecule as III-class. Comparison of d) Pb 4f and e) N 1s XPS spectra of perovskite film before and after passivation treatment by 4-A, 4-C and Phen. f) steady-state PL spectra of perovskite films deposited on FTO substrate. g) Device configuration of the CsFAMA-based PSCs used for experimental verification. h) *J-V* curves and i) statistical PCE of CsFAMA-based PSCs with and without 4-A, 4-C and Phen treatment.

deposited on an FTO substrate. Compared to the unpassivated control sample, the PL intensities of the passivated perovskite films are significantly enhanced after adding any of the three small molecules (Figure 5f), affirming the reduction in defect density within the treated perovskite films.

The most direct approach to assess the efficacy of our ML model is through experimental verification of the photovoltaic performance of PSCs after passivation. Accordingly, we fabricated PSCs with device architecture of FTO/$TiO_2$/$SnO_2$/CsFAMA-based perovskite/spiro-OMeTAD/Au, as depicted in Figure 5g. Figure 5h displays the current density–voltage ($J$–$V$) curves of champion PSCs with and without addition of 4-A, 4-C, and Phen, respectively, under AM1.5G simulated solar illumination. The PCEs of passivated PSCs by different molecules were significantly improved compared to the control device with PCE of 22.40%. Specifically, the devices treated with 4-A, 4-C, and Phen achieved a champion PCE of 24.99%, 25.41%, and 25.26%, respectively. It is noted that 4-C treatment demonstrates an impressive PCE increasement of 3.01%. Hysteresis between forward and reverse scans was significantly mitigated by molecular passivation (Figure S8, Supporting Information). The incident photo-to-electron conversion efficiency (IPCE) measurements (Figure S9, Supporting Information) showed the integrated $J_{SC}$ of the PSCs with and without passivation treatment by molecules, which well matched the measured $J_{SC}$ from $J$-$V$ curves. Statistical diagrams of PCE, open-circuit voltage ($V_{OC}$), short-circuit current density ($J_{SC}$), and fill factor (FF) are shown in Figure 5i; Figure S10 (Supporting Information), and Table S4 (Supporting Information).

In addition, we selected other three molecules: 4-Amino-6-chlorobenzene-1 (4-A-6), 6-aminopyridine-3-carbonitrile (6-A-3), Carzenide (Carz) from the candidates to further validate the predictive capability of the model. The molecular structures and corresponding passivation effect are shown in Figure S11 (Supporting Information), and Table S4 (Supporting Information). It is noteworthy that the observed enhancement in PCE values of the PSCs after passivation treatment with the three screened molecules are all greater than 2%, in strong agreement with the outcomes predicted by our model. Besides, we have conducted a comprehensive comparison of the effectiveness and cost of the new passivation molecules with those newly published since 2023, as documented in Table S5 (Supporting Information). Our analysis revealed that the cost of these new passivation molecules screened by our model falls within the affordable range of $1-10.5/g. This cost-effectiveness, coupled with the promising PCE enhancements, positions these newly identified passivating molecules as valuable candidates for further exploration and application in the field.

## 3. Conclusion

In summary, we explored the relationship between molecular traits and passivation effect of PSCs, employing a ML model to identify superior molecular passivation materials. The dataset containing 330 data were established by collecting published literatures about small molecules as passivation material for PSCs. We constructed a complex S-R model with best accuracy of 83.6% by combining SVM and RF algorithms. To address the challenge of evaluating model with limited dataset and the associated potential for significant evaluation deviations, we developed a brand-new model assessment method (RE-RCV) with lower evaluation error compared to the conventional K-Fold method. Our investigation pinpointed three key molecular traits—dipole moment, hydrogen bond acceptor count, and HOMO-LUMO gap—as pivotal factors influencing the passivation effect. Building on these findings, we established a preliminary screening criterion to identify promising candidate materials for further exploration. Subsequently, we selected three previously unreported molecules—4-A, 4-C, and Phen—from this pool of candidates for experimental verification. Remarkably, the champion PSCs treated with these molecules exhibited impressive PCEs of 24.99%, 25.41%, and 25.26% respectively. Each of these results represented a $\Delta PCE$ increase of over 2% when compared to untreated device with PCE of 22.40%. This remarkable alignment between experimental outcomes and prediction results emphasizes the effectiveness of our proposed ML-based approach. Our study underscores the potential of ML to identify outstanding passivation materials, which could expedite the development and commercialization of PSCs.

## Supporting Information

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

## Acknowledgements

X.Z. and B.D. contributed equally to this work. This work was financially supported by the National Key R&D Program of China (grant no. 2019YFB1503200), the Valais Energy Demonstrators fund and the National Program for Support of Top-notch Young Professionals.

## Conflict of Interest

The authors declare no conflict of interest.

## Data Availability Statement

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

## Keywords

cross-validation, machine learning, passivation, screening, small molecule

Received: November 18, 2023
Revised: February 15, 2024
Published online: March 27, 2024

[1] J. Park, J. Kim, H.-S. Yun, M. J. Paik, E. Noh, H. J. Mun, M. G. Kim, T. J. Shin, S. I. Seok, *Nature.* **2023**, *616*, 724.
[2] G. E. Eperon, T. Leijtens, K. A. Bush, R. Prasanna, T. Green, J. T.-W. Wang, D. P. Mcmeekin, G. Volonakis, R. L. Milot, R. May, A. Palmstrom, D. J. Slotcavage, R. A. Belisle, J. B. Patel, E. S. Parrott, R. J. Sutton, W. Ma, F. Moghadam, B. Conings, A. Babayigit, H.-G. Boyen, S. Bent, F. Giustino, L. M. Herz, M. B. Johnston, M. D. Mcgehee, H. J. Snaith, *Science.* **2016**, *354*, 861.

[3] K. Wang, Z. Jin, L. Liang, H. Bian, D. Bai, H. Wang, J. Zhang, Q. Wang, S. Liu, *Nat. Commun.* **2018**, *9*, 4544.
[4] Z. Song, C. Li, L. Chen, Y. Yan, *Adv. Mater.* **2022**, *34*, 2106805.
[5] G. Nan, X. Zhang, M. Abdi-Jalebi, Z. Andaji-Garmaroudi, S. D. Stranks, G. Lu, D. Beljonne, *Adv. Energy Mater.* **2018**, *8*, 1702754.
[6] X. Li, X. Wu, B. Li, Z. Cen, Y. Shang, W. Lian, R. Cao, L. Jia, Z. Li, D. Gao, X. Jiang, T. Chen, Y. Lu, Z. Zhu, S. Yang, *Energy Environ. Sci.* **2022**, *15*, 4813.
[7] B. Turedi, M. N. Lintangpradipto, O. J. Sandberg, A. Yazmaciyan, G. J. Matt, A. Y. Alsalloum, K. Almasabi, K. Sakhatskyi, S. Yakunin, X. Zheng, R. Naphade, S. Nematulloev, V. Yeddu, D. Baran, A. Armin, M. I. Saidaminov, M. V. Kovalenko, O. F. Mohammed, O. M. Bakr, *Adv. Mater.* **2022**, *34*, 2202390.
[8] J. Wu, H. Cha, T. Du, Y. Dong, W. Xu, C. T. Lin, J. R. Durrant, *Adv. Mater.* **2022**, *34*, 2101833.
[9] B. Chen, P. N. Rudd, S. Yang, Y. Yuan, J. Huang, *Chem. Soc. Rev.* **2019**, *48*, 3842.
[10] H. Zhang, L. Pfeifer, S. M. Zakeeruddin, J. Chu, M. Grätzel, *Nat. Rev. Chem.* **2023**, *7*, 632.
[11] J. Deng, H. Zhang, K. Wei, Y. Xiao, C. Zhang, L. Yang, X. Zhang, D. Wu, Y. Yang, J. Zhang, *Adv. Funct. Mater.* **2022**, *32*, 2209516.
[12] Q. Zhou, D. He, Q. Zhuang, B. Liu, R. Li, H. Li, Z. Zhang, H. Yang, P. Zhao, Y. He, J. Chen, *Adv. Funct. Mater.* **2022**, *32*, 2205507.
[13] J. Guo, J. Sun, L. Hu, S. Fang, X. Ling, X. Zhang, Y. Wang, H. Huang, C. Han, C. Cazorla, Y. Yang, D. Chu, T. Wu, J. Yuan, W. Ma, *Adv. Energy Mater.* **2022**, *12*, 2200537.
[14] Y. Wang, H. Yang, K. Zhang, M. Tao, M. Li, Y. Song, *ACS Energy Lett.* **2022**, *7*, 3646.
[15] M. Wang, Y. Zhao, X. Jiang, Y. Yin, I. Yavuz, P. Zhu, A. Zhang, G. S. Han, H. S. Jung, Y. Zhou, W. Yang, J. Bian, S. Jin, J.-W. Lee, Y. Yang, *Joule.* **2022**, *6*, 1032.
[16] T.-H. Han, J.-W. Lee, C. Choi, S. Tan, C. Lee, Y. Zhao, Z. Dai, N. De Marco, S.-J. Lee, S.-H. Bae, Y. Yuan, H. M. Lee, Y. Huang, Y. Yang, *Nat. Commun.* **2019**, *10*, 520.
[17] J. Wu, M.-H. Li, J.-T. Fan, Z. Li, X.-H. Fan, D.-J. Xue, J.-S. Hu, *J. Am. Chem. Soc.* **2023**, *145*, 5872.
[18] Y. Zhu, P. Lv, M. Hu, S. R. Raga, H. Yin, Y. Zhang, Z. An, Q. Zhu, G. Luo, W. Li, F. Huang, M. Lira-Cantu, Y. Cheng, J. Lu, *Adv. Energy Mater.* **2023**, *13*, 2203681.
[19] S. M. Park, M. Wei, J. Xu, H. R. Atapattu, F. T. Eickemeyer, K. Darabi, L. Grater, Y. Yang, C. Liu, S. Teale, B. Chen, H. Chen, T. Wang, L. Zeng, A. Maxwell, Z. Wang, K. R. Rao, Z. Cai, S. M. Zakeeruddin, J. T. Pham, C. M. Risko, A. Amassian, M. G. Kanatzidis, K. R. Graham, M. Grätzel, E. H. Sargent, *Science.* **2023**, *381*, 209.
[20] C. Luo, G. Zheng, X. Wang, F. Gao, C. Zhan, X. Gao, Q. Zhao, *Energy Environ. Sci.* **2023**, *16*, 178.
[21] Y. Zhao, J. Zhang, Z. Xu, S. Sun, S. Langner, N. T. P. Hartono, T. Heumueller, Y. Hou, J. Elia, N. Li, G. J. Matt, X. Du, W. Meng, A. Osvet, K. Zhang, T. Stubhan, Y. Feng, J. Hauch, E. H. Sargent, T. Buonassisi, C. J. Brabec, *Nat. Commun.* **2021**, *12*, 2191.
[22] B. Lin, J. Jiang, X. C. Zeng, L. Li, *Nat. Commun.* **2023**, *14*, 4110.
[23] 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.
[24] Z. Liu, N. Rolston, A. C. Flick, T. W. Colburn, Z. Ren, R. H. Dauskardt, T. Buonassisi, *Joule.* **2022**, *6*, 834.
[25] N. T. P. Hartono, J. Thapa, A. Tiihonen, F. Oviedo, C. Batali, J. J. Yoo, Z. Liu, R. Li, D. F. Marrón, M. G. Bawendi, T. Buonassisi, S. Sun, *Nat. Commun.* **2020**, *11*, 4172.
[26] J. Xu, H. Chen, L. Grater, C. Liu, Y. Yang, S. Teale, A. Maxwell, S. Mahesh, H. Wan, Y. Chang, B. Chen, B. Rehl, S. M. Park, M. G. Kanatzidis, E. H. Sargent, *Nat. Mater.* **2023**, *22*, 1507.
[27] C. Zhi, S. Wang, S. Sun, C. Li, Z. Li, Z. Wan, H. Wang, Z. Li, Z. Liu, *ACS Energy Lett.* **2023**, *8*, 1424.
[28] W. Liu, N. Meng, X. Huo, Y. Lu, Y. Zhang, X. Huang, Z. Liang, S. Zhao, B. Qiao, Z. Liang, Z. Xu, D. Song, *J. Energy Chem.* **2023**, *83*, 128.
[29] W. Liu, Y. Lu, D. Wei, X. Huo, X. Huang, Y. Li, J. Meng, S. Zhao, B. Qiao, Z. Liang, Z. Xu, D. Song, *J. Mater. Chem. A.* **2022**, *10*, 17782.
[30] F. Gentile, J. C. Yaacoub, J. Gleave, M. Fernandez, A.-T. Ton, F. Ban, A. Stern, A. Cherkasov, *Nat. Protoc.* **2022**, *17*, 672.
[31] P. Burman, *Biometrika.* **1989**, *76*, 503.
[32] M. Callaghan, C.-F. Schleussner, S. Nath, Q. Lejeune, T. R. Knutson, M. Reichstein, G. Hansen, E. Theokritoff, M. Andrijevic, R. J. Brecha, *Nat. Clim. Change.* **2021**, *11*, 966.
[33] Y. Cui, P. Zhu, X. Liao, Y. Chen, *J. Mater. Chem. C.* **2020**, *8*, 15920.
[34] Q. Zhang, Y. J. Zheng, W. Sun, Z. Ou, O. Odunmbaku, M. Li, S. Chen, Y. Zhou, J. Li, B. Qin, K. Sun, *Adv. Sci.* **2022**, *9*, 2104742.
[35] Y. Wu, J. Guo, R. Sun, J. Min, *npj Comput. Mater.* **2020**, *6*, 120.
[36] Y. Xu, X. Guo, Z. Lin, Q. Wang, J. Su, J. Zhang, Y. Hao, K. Yang, J. Chang, *Angew. Chem., Int. Ed.* **2023**, *62*, e202306229.
[37] L. Zhu, X. Zhang, M. Li, X. Shang, K. Lei, B. Zhang, C. Chen, S. Zheng, H. Song, J. Chen, *Adv. Energy Mater.* **2021**, *11*, 2100529.
[38] S. M. Lundberg, S.-I. Lee, *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4768.
[39] A. Pujol, N. Brokhattingen, G. Matambisso, H. Mbeve, P. Cisteró, A. Escoda, S. Maculuve, B. Cuna, C. Melembe, N. Ndimande, H. Munguambe, J. Montaña, L. Nhamússua, W. Simone, K. K. A. Tetteh, C. Drakeley, B. Gamain, C. E. Chitnis, V. Chauhan, L. Quintó, A. Chidimatembue, H. Martí-Soler, B. Galatas, C. Guinovart, F. Saúte, P. Aide, E. Macete, A. Mayor, *Nat. Commun.* **2023**, *14*, 4004.
[40] S. Durand, Q. Lian, J. Jing, M. Ernst, M. Grelon, D. Zwicker, R. Mercier, *Nat. Commun.* **2022**, *13*, 5999.
[41] C. Chih-Chung, *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 1.
[42] I. A. Basheer, M. Hajmeer, *J. Microbiol. Methods.* **2000**, *43*, 3.
[43] T. K. Ho, *IEEE Trans. Pattern Anal. Mach. Intell.* **1998**, *20*, 832.
[44] T. Cover, P. Hart, *IEEE Trans. Inf. Theory.* **1967**, *13*, 21.
[45] H. Zhang, *Aa.* **2004**, *1*, 3.
[46] D. H. Wolpert, *Neural Networks.* **1992**, *5*, 241.
[47] N. Amigo, S. Palominos, F. J. Valencia, *Sci. Rep.* **2023**, *13*, 348.
[48] Z. Wu, N. Cui, D. Gong, F. Zhu, L. Xing, B. Zhu, X. Chen, S. Wen, Q. Liu, *J. Hydrol.* **2023**, *617*, 128947.
[49] Y. Li, P. J. Lohr, A. Segapeli, J. Baltram, D. Werner, A. Allred, K. Muralidharan, A. D. Printz, *ACS Appl. Mater. Interfaces.* **2023**, *15*, 24387.
[50] S. Jiang, S. Xiong, H. Wu, D. Zhao, X. You, Y. Xu, M. Jia, W. Bai, Z. Ma, X. Liu, *Adv. Energy Mater.* **2023**, *12*, 2300983.
[51] L. Yin, C. Ding, C. Liu, C. Zhao, W. Zha, I. Z. Mitrovic, E. G. Lim, Y. Han, X. Gao, L. Zhang, H. Wang, Y. Li, S. Wilken, R. Österbacka, H. Lin, C.-Q. Ma, C. Zhao, *Adv. Energy Mater.* **2023**, *13*, 2301161.
[52] F. Li, X. Huang, C. Ma, J. Xue, Y. Li, D. Kim, H.-S. Yang, Y. Zhang, B. R. Lee, J. Kim, B. Wu, S. H. Park, *Adv. Sci.* **2023**, *12*, 2301603.
[53] R. Wang, A. Altujjar, N. Zibouche, X. Wang, B. F. Spencer, Z. Jia, A. G. Thomas, M. Z. Mokhtar, R. Cai, S. J. Haigh, J. M. Saunders, M. S. Islam, B. R. Saunders, *Energy Environ. Sci.* **2023**, *16*, 2646.
[54] N. Gu, Y. Feng, L. Song, P. Zhang, P. Du, L. Ning, Z. Sun, H. Jiang, J. Xiong, *J. Mater. Chem. C.* **2023**, *11*, 8942.
[55] C. Chen, Y. Zhu, D. Gao, M. Li, Z. Zhang, H. Chen, Y. Feng, C. Wang, J. Sun, J. Chen, H. Tian, L. Ding, C. Chen, *Small.* **2023**, *19*, 2303200.
[56] B. Jiao, Z. Che, Z. Quan, W. Wu, K. Hu, X. Li, F. Liu, *Small.* **2023**, *19*, 2301630.
[57] H.-T. Hsu, Y.-M. Kung, S. Venkatesan, H. Teng, Y.-L. Lee, *Sol. RRL.* **2023**, *7*, 2300122.
[58] K. Wang, B. Yu, C. Lin, R. Yao, H. Yu, H. Wang, *Sol. RRL.* **2023**, *7*, 2300137.
[59] L. Liu, C. Zheng, Z. Xu, Y. Li, Y. Cao, T. Yang, H. Zhang, Q. Wang, Z. Liu, N. Yuan, J. Ding, D. Wang, S. F. Liu, *Adv. Energy Mater.* **2023**, *13*, 2300610.

[60] X. Jiang, B. Zhang, G. Yang, Z. Zhou, X. Guo, F. Zhang, S. Yu, S. Liu, S. Pang, *Angew. Chem., Int. Ed.* **2023**, *62*, 202302462.

[61] F. Li, X. Deng, Z. Shi, S. Wu, Z. Zeng, D. Wang, Y. Li, F. Qi, Z. Zhang, Z. Yang, S.-H. Jang, F. R. Lin, S. W. Tsang, X.-K. Chen, A. K. Y. Jen, *Nat. Photonics.* **2023**, *17*, 478.

[62] B. Q. Xu, X. L. Li, X. Y. Xiao, H. Sakaguchi, N. J. Tao, *Nano Lett.* **2005**, *5*, 1491.

[63] R. Casares, Á. Martínez-Pinel, S. Rodríguez-González, I. R. Márquez, L. Lezama, M. T. González, E. Leary, V. Blanco, J. G. Fallaque, C. Díaz, F. Martín, J. M. Cuerva, A. Millán, *J. Mater. Chem. C.* **2022**, *10*, 11775.