

## ORIGINAL RESEARCH ARTICLE

# Data-driven identification of functional additives and solution parameters in mixed Sn-Pb perovskite solar cells via $\beta$ -VAE augmentation

Behzad Iranipour<sup>1</sup> , Mohammadreza Sadeghian<sup>2</sup> , and Ezeddin Mohajerani<sup>1\*</sup> 

<sup>1</sup>Laboratory of Photonics of Organic Materials and Polymers (POMP), Laser and Plasma Research Institute, Shahid Beheshti University, Tehran, Iran

<sup>2</sup>Faculty of Physics, Shahid Beheshti University, Tehran, Iran

## Abstract

Optimizing perovskite solar cells (PSCs) requires precise control of solution chemistry and functional additives. However, limited experimental data hinder systematic discovery. Here, we integrate 1,540 carefully selected experimental device records with 4,000 synthetic data points generated by a beta-variational autoencoder to investigate solution parameters and organic additives governing device performance. A residual neural network trained on this hybrid dataset achieves strong predictive accuracy with an  $R^2$  of 0.87 for power conversion efficiency. Even when trained solely on synthetic data, the model attains an  $R^2$  of 0.785. Within this framework, 733 organic additives with diverse functional groups were evaluated to identify molecular features that enhance absorber quality. High-efficiency PSCs are associated with solution concentrations above 1.3 molar and elevated formamidinium iodide (FAI) ratios, in combination with additives containing benzene rings, methylene, and amine groups. Notably, a composition comprising FAI (1.05), cesium iodide (0.03), methylammonium chloride (0.3), lead(II) iodide (1.5), and a molybdenum trioxide interlayer, combined with 1,3-dihydro-1-[1-(phenylmethyl)-4-piperidinyl]-2H-benzimidazol-2-one as an additive, yields a PCE of 25.66%. This additive was absent from the training data, demonstrating the capability of the proposed framework to discover novel and effective organic additives for PSC optimization.

### \*Corresponding author:

Ezeddin Mohajerani  
(e-mohajerani@sbu.ac.ir)

**Citation:** Iranipour B, Sadeghian M, Mohajerani E. Data-driven identification of functional additives and solution parameters in mixed Sn-Pb perovskite solar cells via  $\beta$ -VAE augmentation. *Int J AI Mater Design*. 2026;3(1):69-93.  
doi: 10.36922/IJAMD025480051

**Received:** November 30, 2025

**Revised:** January 31, 2026

**Accepted:** February 9, 2026

**Published online:** March 26, 2026

**Copyright:** © 2026 Author(s). This is an Open-Access article distributed under the terms of the Creative Commons Attribution License, permitting distribution, and reproduction in any medium, provided the original work is properly cited.

**Publisher's Note:** AccScience Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Keywords:** Perovskite solar cell; Experimental data; Synthetic data; Additive

## 1. Introduction

The development and discovery of optimal and efficient materials and structures in materials science and chemistry require substantial investment and extended timeframes. Furthermore, the interplay of multiple factors influencing outcomes, as well as the interdependence of various parameters, underscores the necessity for additional experimentation. In recent decades, considerable efforts have been devoted to enhancing efficiency and sustainability while minimizing the use of toxic elements, leading to significant advancements in the field. However, researchers continue to face numerous challenges before these devices can be commercialized. For instance, although organic additives have emerged as a crucial factor in enhancing stability and final efficiency—demonstrating significant potential for further stability improvements and Power

Conversion Efficiency (PCE)—a comprehensive study examining the various types of these additives and their effects on the absorber layer has yet to be conducted. A practical solution is to leverage past experimental experiences and their outcomes to avoid unnecessary repetitions. Recent advancements in artificial intelligence (AI) and innovations in neural network architecture and machine learning algorithms have inspired researchers across scientific disciplines to adopt these emerging technologies. Notably, promising developments have been reported in applying AI within materials science. Nevertheless, more rigorous and innovative efforts are required to streamline experiments, accelerate progress, and develop new optimal compounds. In the domain of perovskite solar cells (PSCs), there is substantial potential for applying AI, owing to the abundance of laboratory data and the collective experience of researchers worldwide.<sup>1–3</sup> While the effectiveness of machine learning-based algorithms in evaluating factors influencing cell performance is well established, their application in predicting photovoltaic parameters remains in its nascent stages.<sup>4–6</sup> This limitation arises primarily from several challenges. First, the complex and numerous relationships affecting final device performance complicate the analysis of these interactions.<sup>7–11</sup> Second, limitations in existing datasets necessitate the collection of new data, which is often time-consuming and costly.<sup>12</sup> For instance, in the case of a mixed-cation, mixed-metal perovskite containing formamidinium, cesium, tin, and lead with iodide anions, the lack of experimental data led A.D. Kapim Kenfack and colleagues to generate non-experimental data using the solar cell capacitance simulator in one dimension (SCAPS-1D).<sup>13</sup> However, the multitude of parameters and diverse manufacturing methods necessitate a substantial amount of data to effectively train machine learning models and neural networks.<sup>14–18</sup> Additionally, the use of laboratory simulators has proven insufficient due to the numerous assumptions introduced to simplify calculations.<sup>19,20</sup> For instance, efficiency values and photovoltaic parameters calculated using SCAPS-1D software exhibit significant discrepancies when compared with laboratory measurements.<sup>21,22</sup> Furthermore, critical factors influencing final efficiency, such as the precursor molar ratio, the antisolvent employed, and other laboratory parameters, are not incorporated into the software.<sup>23,24</sup> These factors, combined with existing simplifications, contribute to the observed discrepancies between laboratory and simulation results.<sup>25</sup> Third, perovskites incorporating newer compounds and lead-substituted cations have not received sufficient attention, further impeding data collection in this domain.<sup>26</sup> In addition, the quality of available data varies considerably, leading to

potential overlap or redundancy.<sup>27,28</sup> Similar to machine learning algorithms, multilayer neural networks require large datasets for effective learning, with the required dataset size increasing alongside the number of features and model complexity.<sup>29–31</sup> The perovskite database, developed by Jacobson and colleagues, serves as a valuable resource, encompassing over 43,000 PSC performance entries spanning from 2009 to 2024. This database supports machine learning-based research and reduces the need for extensive experimental testing.<sup>32</sup> Leveraging this database, F.J. Kusuma *et al.* achieved a coefficient of determination ( $R^2$ ) of 0.751 for PCE prediction.<sup>33</sup> However, the practical utility of this dataset remains limited due to incomplete data and the lack of consideration for laboratory characteristics, such as precursor quantities, among other factors. Additionally, to address data scarcity, Sh. Zhao *et al.* enhanced training samples through feature transformation with the feature mask method, achieving a root mean square error (RMSE) of 0.833% and a Pearson correlation coefficient of 0.980.<sup>34</sup> However, the model was tested using only 10 samples, indicating a need for additional data to further evaluate its robustness. Therefore, the adoption of innovative methods for predicting the performance of PSCs is essential for advancing the development and practical application of these technologies. In our previous work, we improved model accuracy in predicting final efficiency by generating only 100 synthetic data points using neural networks such as autoencoders and a conditional generative adversarial network (CGAN), along with architectural modifications.<sup>35</sup> In this study, we have successfully generated a synthetic dataset comprising 4,000 data points using models based on deep learning and neural networks. This was achieved by employing a beta-variational autoencoder ( $\beta$ -VAE) architecture, which was adapted for dense layers. The training data utilized for the network were sourced from reliable journals, including *Nature*, *the American Chemical Society*, *Science*, *Elsevier*, *the Royal Society of Chemistry*, and *Wiley Online Library*. Following a rigorous data separation and cleaning process to eliminate outliers, 1,540 structures were obtained. The synthetic dataset was generated incrementally, allowing the network to acquire the necessary values for effective training. Initially, 200 data points were generated sequentially. At each stage, the newly generated data were integrated into the synthetic dataset, resulting in a total of 4,000 data points. All data were sourced from the optimized model and subjected to multiple quality assessments to ensure integrity. By refining this process and optimizing the model, additional data can be generated to further enhance network training and predictive performance. For evaluation, both classical machine learning models and modern deep neural

networks were employed, utilizing the original dataset and a composite dataset consisting of both synthetic and original data. Among classical models, Gradient Boosting (GBoost), Extreme Gradient Boosting (XGBoost), Extra Tree (ETree), and Random Forest (RF) demonstrated the highest accuracy when tested on 151 new data points, achieving coefficient of determination ( $R^2$ ) values of 0.822, 0.822, 0.812, and 0.798, respectively. Conversely, after training on the composite dataset, the residual neural network (ResNet), XGBoost, GBoost, and ETree attained  $R^2$  values of 0.870, 0.867, 0.857, and 0.856, respectively, on the test set. Overall, this comparison highlights a substantial improvement in accuracy for modern neural networks, underscoring the superior capacity of deep learning models. The incorporation of synthetic data enables deep models to leverage additional information, thereby enhancing their generalizability through effective feature learning. On the other hand, classical models typically do not require large datasets to achieve their optimal learning potential. Their accuracy improvements are driven more by data quality and diversity than by dataset size. This suggests that the model has effectively learned the underlying data distribution and that the generated data meet the required quality standards. In the subsequent phase, we analyzed 733 structures containing various organic additives and constructed a dataset based on their functional groups. This led to the development of a trained model with a mean absolute error (MAE) of 0.3982, designed to identify organic molecules that effectively enhance the photovoltaic properties of the absorber layer, thereby improving PCE and overall stability. Additionally, 20 new candidate molecules are proposed for experimental consideration. In conclusion, to enable researchers in this field to leverage models trained on artificial and large-scale datasets while minimizing the need for extensive experimentation, a Python-based system that learns from laboratory and synthetic data has been developed. This system predicts the photovoltaic parameters of structures and recommends optimal material configurations, thereby reducing the reliance on costly experimental trials. By extending this methodology across various laboratory disciplines, researchers can avoid repetitive experimentation in the pursuit of optimal material structures.

## 2. Methodology

This research is structured as follows: (i) research and testing to evaluate the most important features affecting the performance of the perovskite cell; (ii) collecting targeted, high-quality data while eliminating structures with incomplete features and unreasonable PCE. This process removed redundant features and enhanced model

performance by leveraging the most relevant information; (iii) implementing various generative neural network models, such as generative adversarial network (GAN), CGAN,  $\beta$ -VAE, and U-shaped convolutional neural network (U-Net), while modifying the architecture and implementation on the existing dataset. Table S1 lists all the abbreviations used. Among these models, the  $\beta$ -VAE architecture was selected as the most efficient model for generating synthetic data to enhance the primary dataset and improve prediction accuracy; (iv) application of classical machine learning models and deep neural network-based models for training, validation, and evaluation using both the main and concatenated datasets. Finally, the features influencing PCE were identified, and the enhanced models were utilized to predict optimal structures and to assess the effects of new organic additives on these structures. The overall enhanced learning workflow is shown in Figure 1.

### 2.1. Construction of the dataset

To address the scarcity of laboratory data in materials-related fields, particularly in perovskite research, several steps were undertaken to develop a more accurate and robust model. First, a dataset comprising 1,540 high-quality data points was compiled from 310 publications between 2013 and 2024. This collection includes articles from esteemed journals such as *Nature*, *the American Chemical Society*, *Science*, *Elsevier*, *the Royal Society of Chemistry*, and *Wiley Online Library*. The dataset contains values ranging from 0.12 to 24.33. Based on prior knowledge, intuition, and testing of significant parameters, a total of 24 parameters were identified as influential features. These include the molar amounts of precursors (methylammonium iodide, formamidinium iodide (FAI), cesium iodide (CsI), cesium bromide, methylammonium bromide, methylammonium chloride, lead(II) iodide [ $\text{PbI}_2$ ], lead(II) bromide, lead(II) chloride, tin(II) iodide [ $\text{SnI}_2$ ], tin(II) bromide, tin(II) chloride, lead(II) thiocyanate, and tin(II) fluoride), solvents (N, N-Dimethylformamide, dimethyl sulfoxide, and gamma-butyrolactone), antisolvent polarity index values, perovskite annealing temperature, electron and hole transport layers (ETL and HTL), back contact, and interlayer or insulating layer (Table 1). Feature interrelationships were analyzed using Spearman's rank correlation matrix (Figure S1). All 24 features were retained, as each contributes distinct information to the PSC structures, and no redundant variables were identified for removal. To reduce unnecessary complexity, low-impact parameters such as the thickness of the electron and hole transport layers were excluded. Distribution plots for all features are shown in Figure S2. To further enhance the relevance and applicability of the research, only structures fabricated using the one-step perovskite

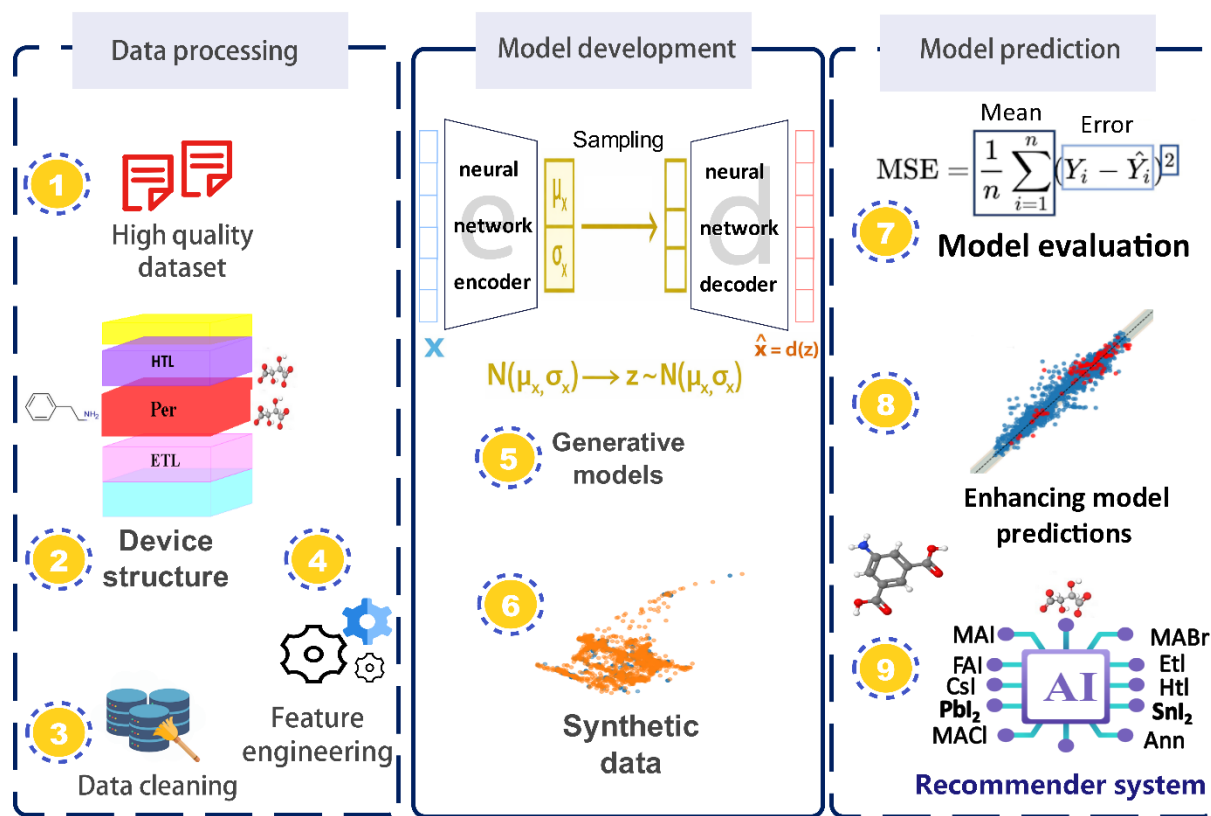
layer deposition and spin coating methods were included. Structures prepared using alternative methods, such as dip coating, chemical vapor deposition, Vienna ab initio simulation package, and spray, were excluded. The final cell efficiency was also considered a target variable in the reverse prediction. The PSC structures used in this study are presented in Table S2. To enhance model usability and enable accurate predictions prior to the layer synthesis and fabrication, specific molar amounts of precursors and solvent ratios were utilized, rather than relying solely on perovskite type or cation–anion nomenclature. Furthermore, to broaden the applicability of this research and to examine the effects of various additives, data were compiled based on the presence of distinct functional groups associated with different organic additives. These functional groups include hydrazine, ester, hydroxyl, nitrile, imide, amine, amide, ether, methyl, methylene, ketone, halides (iodine, bromine, and chlorine), alkene, ammonium, trifluoromethyl, benzene, thiocyanate, aryl halide, tert-butyl, pyridine, thiophene, thioketone, sulfonate, and imine. In this context, the difference between the final PCE of the reference structure and that of the structures containing the additive is regarded as the target variable.

For an initial preview of the data and for the statistical

and intuitive analysis, it is essential to present the dataset and the relationships among variables using various formats, including statistical charts and correlation analyses. A comprehensive understanding of both quantitative and qualitative relationships within the data significantly aids researchers in data collection and in discerning the relationships among different parameters and their relevance. Recognizing the importance of different features helps identify critical ones and eliminate those that are unnecessary. In this context, a bar chart, derived from both quantitative and qualitative analyses, is used to illustrate the distribution and range of power conversion efficiencies (Figure S3), thereby providing an overview of the dataset. Descriptive statistics, including the mean, median, standard deviation, and graphical representations, are employed to convey the data clearly and effectively. According to the diagram, the PCE values predominantly fall within the ranges of 4–10% and 14–20%, which together contain the highest data point values. This relatively normal distribution is expected to improve model efficacy and enhance learning outcomes. Additionally, an analysis of data density reveals that the values are more concentrated in the mid-range. The decline in data density for PCE values exceeding 20% highlights the need for additional data in this region. Considering

**Table 1. Descriptions of the 24 features and the one target feature for DL**

Category	Input feature name	Feature description
Perovskite composition	MAI, FAI, CsI, CsBr, MABr, MACl, PbI <sub>2</sub> , PbBr <sub>2</sub> , PbCl <sub>2</sub> , SnI <sub>2</sub> , SnCl <sub>2</sub> , SnBr <sub>2</sub> , SnF <sub>2</sub> , Pb (SCN) <sub>2</sub>	The ratios of various components of lead and tin halide perovskite.
Processing parameters	Solvent system (DME, DMSO, GBL)	Ratios of solvents used (e.g., N, N-dimethylformamide: dimethyl sulfoxide, N, N-Dimethylformamide: gamma-butyrolactone).
	Antisolvent (Polarity index)	Type of anti-solvent used (e.g., toluene, chlorobenzene).
	Electron transport layer (ETL)	Material used as ETL (e.g., TiO <sub>2</sub> , SnO <sub>2</sub> , PCBM).
	Hole transport layer (HTL)	Materials used as HTL (e.g., Spiro-OMeTAD, PTAA, PEDOT: PSS).
	Annealing	The temperature applied to the perovskite film after deposition.
	Back contact	Materials used as Back contact (e.g., Au, Ag).
	Interlayer	Materials used as Interlayer (e.g., LiF, Al <sub>2</sub> O <sub>3</sub> ).
Additives/dopants	Additive type	Specific additives (e.g., SnF <sub>2</sub> , Pb(SCN <sub>2</sub> )).
Output target	Best power conversion efficiency (%)	The champion power conversion efficiency.



**Figure 1.** Flowchart of the enhanced learning workflow designed to enhance the prediction accuracy of classical models and deep networks while identifying efficient structures  
Abbreviations: AI: Artificial intelligence; MSE: Mean Squared Error.

the limited data available, there is an increased need for alternative data generation methods, such as synthetic data generation.

## 2.2. Model framework

Overall, to address the different objectives of device optimization and additive discovery, two independent predictive models were developed within the proposed framework, each tailored to a distinct prediction task operating at different scales. First, both classical machine learning models and neural network-based models were employed to predict the absolute PCE of PSCs using fabrication parameters and solution chemistry descriptors as input features. To mitigate data scarcity and improve generalization, the experimental dataset comprising 1,540 samples was augmented with 4,000 synthetic samples generated using a  $\beta$ -VAE, resulting in a hybrid training set with increased diversity. Second, a molecular-level additive screening model was developed to evaluate the effect of organic additives on device efficiency. A separate dataset comprising 733 additives was compiled, in which each molecule was represented using molecular

descriptors and functional group features. The target variable was defined as the change in PCE induced by each additive. Independent predictive models based on XGBoost and deep neural networks were trained to estimate this performance variation, with the best-performing model achieving a MAE of 0.3982. This model was subsequently employed to screen candidate molecules and to propose 20 new organic additives with potential performance improvements. Because these two tasks involve different datasets, feature spaces, and prediction targets—namely absolute PCE prediction versus additive-induced performance variation—the corresponding models were designed and trained independently, forming two complementary components of the overall predictive framework. The classical models evaluated included decision tree, ETree, RF, Adaptive Boosting, GBoost, XGBoost, Multilayer Perceptron, and Support Vector Machine. Additionally, the neural network models utilized included ResNet, one-dimensional convolutional neural network (1D CNN), deep neural network (Deep NN), wide and deep neural network, attention neural network, simple neural network (Simple NN), and TabTransformer.

In contrast to machine learning models, which primarily emphasize feature selection for final predictions, deep learning methods focus on developing suitable prediction models for the dataset. Currently, popular deep learning models include convolutional neural networks, deep neural networks, and recurrent neural networks. To enhance prediction accuracy, structures yielding unreasonable returns were removed, thereby improving the reliability of the predicted results relative to real and experimental outcomes. The training dataset comprised 1,390 structures, while the test set consisted of 150 data points that were entirely different from the main dataset. To enhance the reliability of the results, the main dataset was not randomly split into training and testing subsets. Given the continuous range of final efficiency, model training was formulated as a regression problem. To further address the data limitations inherent in deep networks and to improve the accuracy of the models, synthetic data were generated using a generative model such as  $\beta$ -VAE. This process was conducted in multiple stages. Using Shapley Additive Explanations (SHAP) analysis, we identified the features most strongly associated with final efficiency. To further enhance the applicability of the framework and identify effective organic additives, organic additives were collected and categorized according to the functional groups present in their molecular structures. SHAP analysis was also employed to determine the functional groups that contribute most significantly to the stability and passivation of the absorber layer.

### 2.3. Development of the model

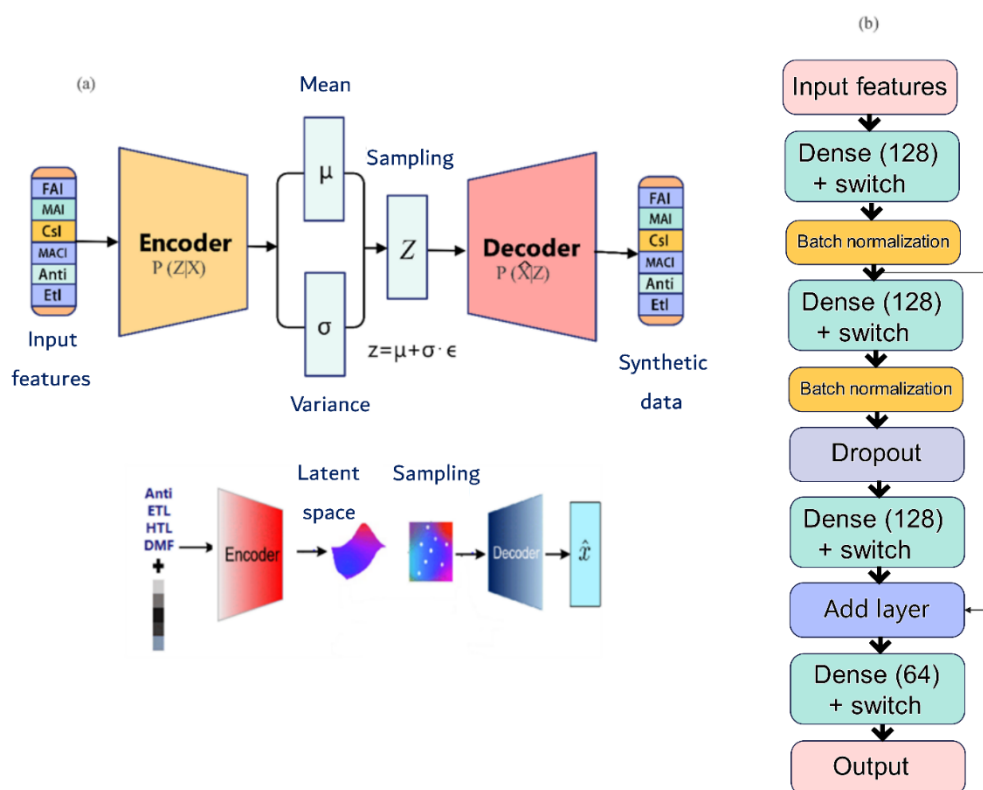
To enhance the original dataset and address the data scarcity in deep learning models, generative models were employed to generate synthetic data. To identify the most effective network for generating high-quality synthetic data, both unsupervised learning and supervised learning methods were employed. The architectures utilized in this research include the CGAN,  $\beta$ -VAE, and U-Net. Based on the nature of the data and the quality of the network outputs, the  $\beta$ -VAE architecture was selected as the primary model for this study. As reported in previous studies, unsupervised learning-based networks, while capable of enhancing model accuracy, are often limited in their ability to generate substantial volumes of high-quality synthetic data. The network architecture employed in this research is depicted in Figure 2, which provides a comprehensive overview of the integrated deep learning framework used in this study. Figure 2a illustrates the  $\beta$ -VAE architecture, which effectively learns compact and semantically meaningful latent representations of the input data through probabilistic encoding and decoding. This architecture enables accurate reconstruction of input

samples but supports the generation of new data and sampling from the learned distribution.  $\beta$ -VAE combines neural networks with statistical inference to learn the data distribution  $p(x)$ . However, direct computation of the marginal likelihood  $\int p(x|z) p(z) dz$  is intractable due to the integral over the latent variable  $z$ . To address this, Variational Autoencoders (VAEs) introduce an approximate posterior distribution  $q(z|x)$  and optimize a lower bound on the data log-likelihood, known as the Evidence Lower Bound (ELBO):

$$\log p(x) \geq E_{q\phi(z|x)} [\log p\theta(x|z)] - \hat{\alpha} DK(q\phi(z|x) \| p(z)) \quad (1)$$

The reconstruction term ( $E_{q\phi(z|x)} [\log p\theta(x|z)]$ ) encourages the decoder, parameterized by  $\theta$ , to accurately reconstruct the input data  $x$  from the sampled latent variable  $z$ . The regularization term ( $DK(q\phi(z|x) \| p(z))$ ) ensures that the learned latent distribution  $q\phi(z|x)$  remains close to the prior distribution  $p(z)$ , which is typically chosen as a standard normal distribution. Unlike generative models that rely on data retention, such as traditional autoencoders, VAEs learn a probability distribution of the data in a low-dimensional latent space by optimizing the ELBO. This methodology enables the generation of new samples while preserving the critical characteristics of the training dataset. However,  $\beta$ -VAE is highly sensitive to the selection of hyperparameters, feature distribution, and the architecture of both the encoder and decoder layers in maintaining the integrity of the data distribution. Consequently, optimizing these parameters can be challenging due to the complexity of the data and its feature distribution. Furthermore, the ResNet architecture, shown in Figure 2b, enhances the learning capacity of deep features, particularly when faced with large and complex datasets, and effectively extracts significant new features. The incorporation of residual blocks in ResNet mitigates common challenges in training deep networks, such as the vanishing gradient problem, thereby improving model convergence and enhancing the stability of the training process. Given the critical need for producing precise synthetic data and achieving optimal output values, the models were trained using regression algorithms. Regression serves as a vital data analysis technique, enabling the exploration of relationships between a dependent variable and one or more independent variables. We employed label encoding methods for feature encoding, removed missing values to enhance model accuracy and data quality, and conducted feature screening. Label encoding is a technique that converts strings into numerical values ranging from 0 to a specified maximum, making the data more interpretable for neural networks and machine learning models. To





**Figure 2.** Schematic representation of (a) the variational autoencoder and (b) the superior model based on integrated data (residual neural network)

evaluate the performance of the models and assess their accuracy and error rates, we employed metrics such as mean squared error (MSE), MAE, RMSE, and  $R^2$ . These metrics provide insights into the average squared errors, the mean absolute differences between actual and predicted values, the RMSE, and the proportion of variance explained by the model. MAE is less affected by large errors, whereas large errors carry more weight in the MSE criterion, as demonstrated in Equations 2 and 3.

$$MAE = \frac{1}{N} \sum_{i=1}^N |true_i - predicted_i| \quad (2)$$

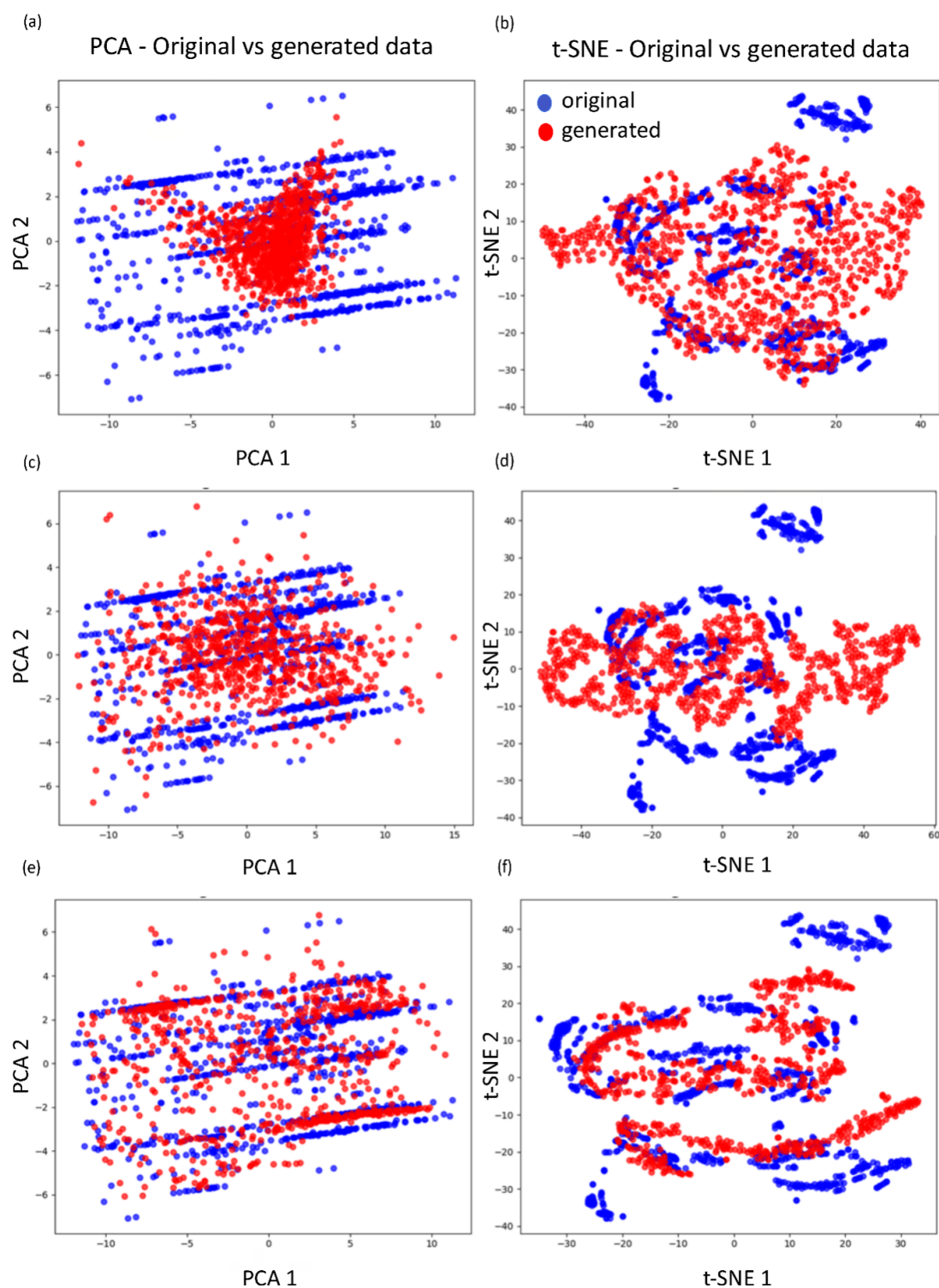
$$MSE = \frac{1}{N} \sum_{i=1}^N (true_i - predicted_i)^2 \quad (3)$$

### 3. Results and discussion

#### 3.1. Model evaluation

Figure 3 shows the process of  $\beta$ -VAE model updating and model learning enhancement through hyperparameter tuning and layer optimization. To illustrate the learning improvement process and evaluate the data distribution effectively, the dataset features were reduced to two

dimensions using Principal Component Analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE) (Section S2.2). PCA identifies the directions (principal components) along which the data exhibit the greatest variance,<sup>36</sup> then projects the high-dimensional data onto these directions to reduce dimensionality while retaining as much information as possible.<sup>37</sup> As shown in Figure 3a, the output data are more significantly concentrated and less dispersed than the original dataset. In contrast, the original data exhibits greater scattering and distinct distribution bands, whereas the output data is more focused around a central cluster. This over-concentration in the output data often signifies reduced diversity or incomplete modeling, suggesting that the data may not adequately represent the entire feature space. Consequently, the model may reproduce only a portion of the data distribution rather than accurately capturing the original distribution, leading to under-diversified and incomplete output data. In contrast, Figure 3e demonstrates that the generated data (red dots) closely overlap with the original data (blue dots), with the dispersion of the generated data closely mirroring that of the original dataset. This overlap and similarity in dispersion indicate that the generative model has effectively preserved the overall diversity and distribution of the original data. Furthermore, both datasets display



**Figure 3.** The process of enhancing data pattern learning through the Variational Autoencoder model by adjusting and optimizing network hyperparameters. (a, b) Initial output: Over-concentrated generated data with insufficient diversity. (c, d) Intermediate improvement: Partial overlap but persistent cluster separation. (e, f) Optimized output: Synthetic data (red) exhibits complete alignment with original data (blue) in both global distribution (PCA) and local structure (t-SNE), demonstrating successful hyperparameter tuning in preserving the original data manifold. Abbreviations: PCA: Principal Component Analysis; t-SNE: t-distributed stochastic neighbor embedding.

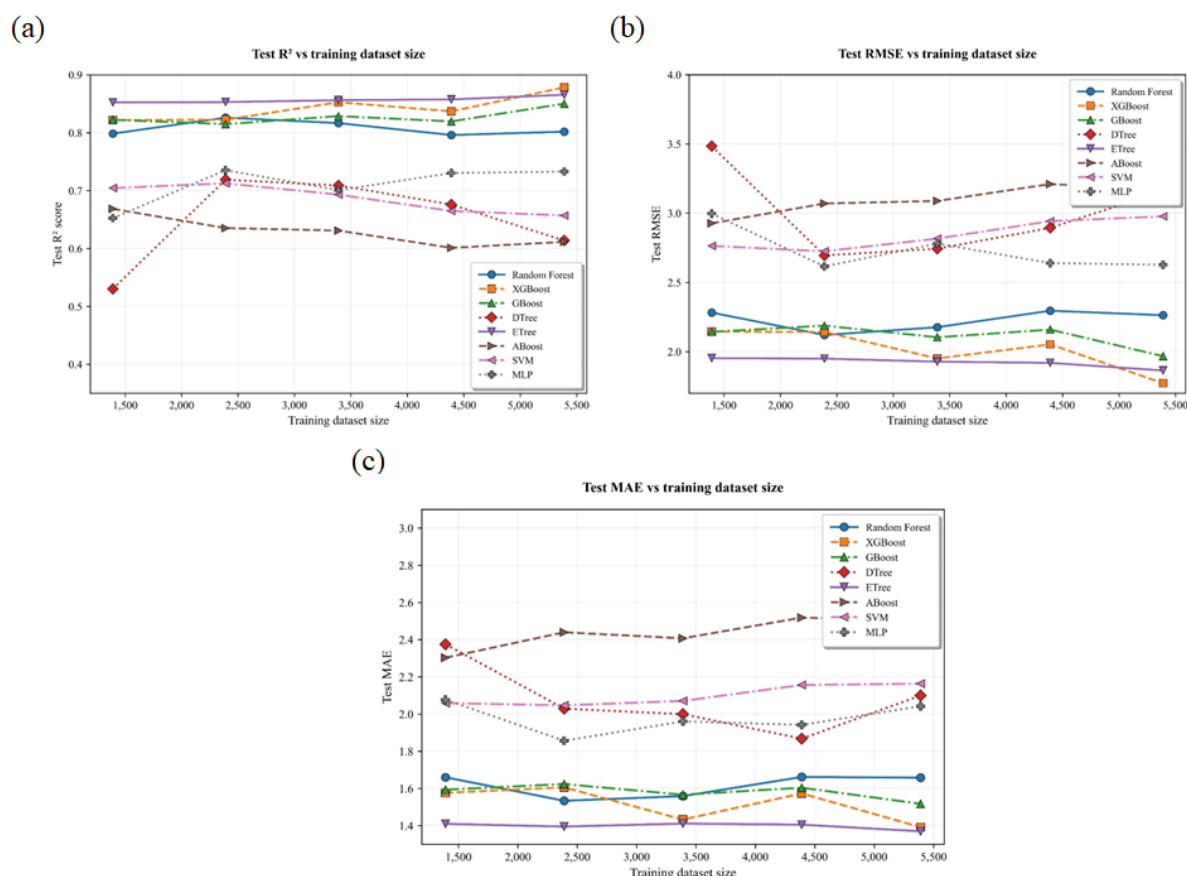


a multi-cluster or linearly scattered pattern, suggesting that the overall structures in the generated data have been maintained. t-SNE is a nonlinear dimensionality reduction algorithm that primarily focuses on preserving local structures within the data.<sup>38</sup> Unlike PCA, which aims to maintain overall variance, t-SNE attempts to preserve proximity relationships, keeping points that are close in the original space similarly close in the lower-dimensional representation.<sup>39</sup> This characteristic makes t-SNE particularly suitable for visualizing complex and nonlinear clusters and structures within the data. [Figure 3d](#) reveals a much clearer and stronger separation between the original and generated datasets. The original data form several distinct clusters, while the generated data appear as compressed, isolated bands within the space. This notable separation suggests that the generative model may struggle to replicate the intricate local structures of the original data, resulting in substantial differences between the generated samples and the actual samples in terms of neighborhood and clustering. In [Figure 3f](#), the original data points and generated data exhibit partial overlap within certain clusters; this overlap indicates that the generative model has successfully reproduced some local structures of the original data. Optimization of hyperparameters to enhance the understanding of the data distribution is depicted in [Figure S4](#). However, a clear separation persists in some clusters, highlighting subtle differences in local neighborhoods, suggesting that challenges remain in accurately reproducing all local complexities and patterns, particularly for complex, highly nonlinear data distribution. In this study, the key hyperparameters include the learning rate, hidden layers, and latent dimension.<sup>40–42</sup> Increasing the number of layers and neurons in the network can lead to overfitting.<sup>43</sup> However, implementing dropout mitigates this by preventing the network from memorizing the training data.<sup>44</sup> In this study, six layers were employed in the encoder, with a dropout rate of 0.1 and a learning rate of 0.001. Considering the input dimension of 25 features, the latent dimension was set to 40 after conducting multiple tests. This choice accommodates the highly nonlinear relationships among the features and facilitates the effective learning of the hidden data distribution.

In the subsequent stage, 200 synthetic data points were added to the synthetic dataset. Stage 3 similarly involved adding 200 synthetic data points, and the original dataset with synthetic data was incrementally augmented in a stepwise manner. The Reconstruction Loss parameter quantifies the discrepancy between the original input and the network-reconstructed output,<sup>45</sup> serving as an indicator of the model's effectiveness in data reconstruction. Given the continuous nature of the data, MSE was employed as the primary reconstruction metric. Additionally, the

Kullback–Leibler loss measures how closely the latent space approximates a normal distribution.<sup>46</sup> Selected examples from the synthetic dataset generated by the VAE network are presented in [Table S3](#). These data effectively illustrate the learning rate of the network, with structural parameters, including those of precursors and solutions, showing a significant correlation with the final efficiency. To evaluate the robustness of the synthetic data generation process and its effect on model performance, we systematically varied the volume of synthetic samples from 1,390 to 5,390 in five increments, evaluating classical and neural network-based models independently on a fixed, unseen test set. For ensemble-based classical models, such as RF and XGBoost, increasing the number of synthetic data points led to a monotonic improvement in generalization performance, indicated by a progressive increase in test  $R^2$  and decreases in RMSE and MAE. This trend reflects a systematic reduction in model bias. High-variance classical models, such as decision trees, exhibited unstable performance with limited synthetic data; however, their variability decreased significantly with larger synthetic volumes, demonstrating improved generalization and enhanced variance control. Notably, performance gains for most classical models reached a point of diminishing returns beyond approximately 4,000 synthetic samples, suggesting that the core structure of the data distribution is adequately captured at this scale. The results of this analysis for classical models are summarized in [Figure 4a–c](#), which visualizes the  $R^2$ , RMSE, and MAE trends across synthetic data volumes. A comparable visualization for neural network-based models is provided in [Figure 5](#).

In contrast, neural network-based models exhibit an even stronger dependence on synthetic data volume. Across seven architectures evaluated over the same range of data volumes, neural models demonstrated larger performance gains and a steeper improvement trend than classical models, consistent with their higher representational capacity and flexibility. Importantly, no performance degradation was observed for neural models at higher synthetic volumes; instead, improvements continued beyond the saturation point observed in classical models. This indicates that synthetic data is particularly effective in reducing both bias and variance in neural architectures. Quantitatively, notable improvements were observed across architectures. For instance,  $R^2$  scores increased from 0.4209 to 0.7823 for Simple NN, from 0.6954 to 0.8555 for the Wide and Deep model, and from 0.6622 to 0.8117 for the 1D CNN. Correspondingly, error reduction was also pronounced, with RMSE decreasing most significantly for the Wide and Deep model (from 2.8051 to 1.9320) and the Simple NN (from 3.8681 to 2.3717). More complex



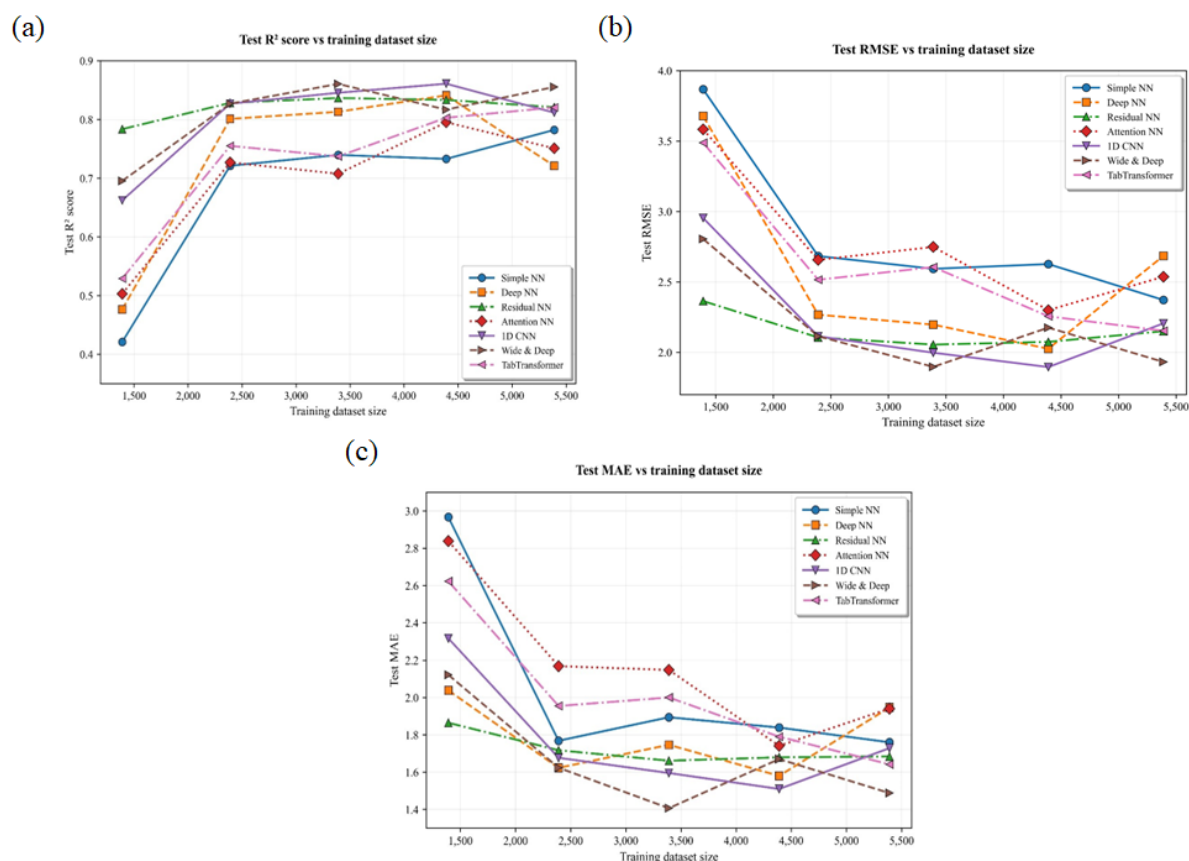
**Figure 4.** Performance of classical machine learning models across varying volumes of synthetic data. (a)  $R^2$  score, (b) RMSE, and (c) MAE. The trends demonstrate bias reduction and variance stabilization with increasing data volume, with performance gains saturating beyond approximately 4000 synthetic samples for most models.

Abbreviations: ABoost: Adaptive Boosting; DTree: Decision Tree; ETree: Extra Trees; GBoost: Gradient Boosting; MAE: Mean absolute error; MLP: Multilayer Perceptron;  $R^2$ : coefficient of determination; RMSE: Root mean square error; SVM: Support Vector Machine; XGBoost: Extreme Gradient Boosting.

architectures, such as Wide and Deep and 1D CNN, benefited most from increased data volume, reflecting their enhanced capacity to capture complex patterns. With increased data volume, performance variance decreased across all neural models, resulting in more stable predictions. Architectures such as Residual NN maintained consistently high performance across all data volumes, suggesting a strong suitability for the problem domain. In contrast, some models, including the Deep NN, exhibited diminishing returns beyond a certain volume, suggesting potential capacity limitations or the need for further hyperparameter adjustment. Finally, the performance gap between different architectures narrowed with increasing data volume, implying that sufficient synthetic data can partially compensate for architectural differences.

After generating synthetic data using generative neural networks, we assessed the quality of the data and

its impact on the prediction accuracy of the models. Eight classical networks and seven neural networks were used for evaluation. In the first stage, the model was trained using the original dataset. Figure 6 presents the  $R^2$  values for the models applied to this training set in relation to the PCE value. Additionally, Figure 6b illustrates the correlation between the predicted values by these models and the actual PCE values. The specific values of the tuned hyperparameters are provided in Table S4. At this stage, classical and ensemble models have demonstrated superior performance compared to deep network-based models. The  $R^2$  values for the GBoost, XGBoost, ETree, and RF models are 0.968, 0.961, 0.933, and 0.946, respectively, for the training data, and 0.822, 0.822, 0.812, and 0.798, respectively, for the test data. These results indicate that the models effectively learn while avoiding overfitting. The coefficient of determination quantifies the extent to which the model explains data variation, with higher



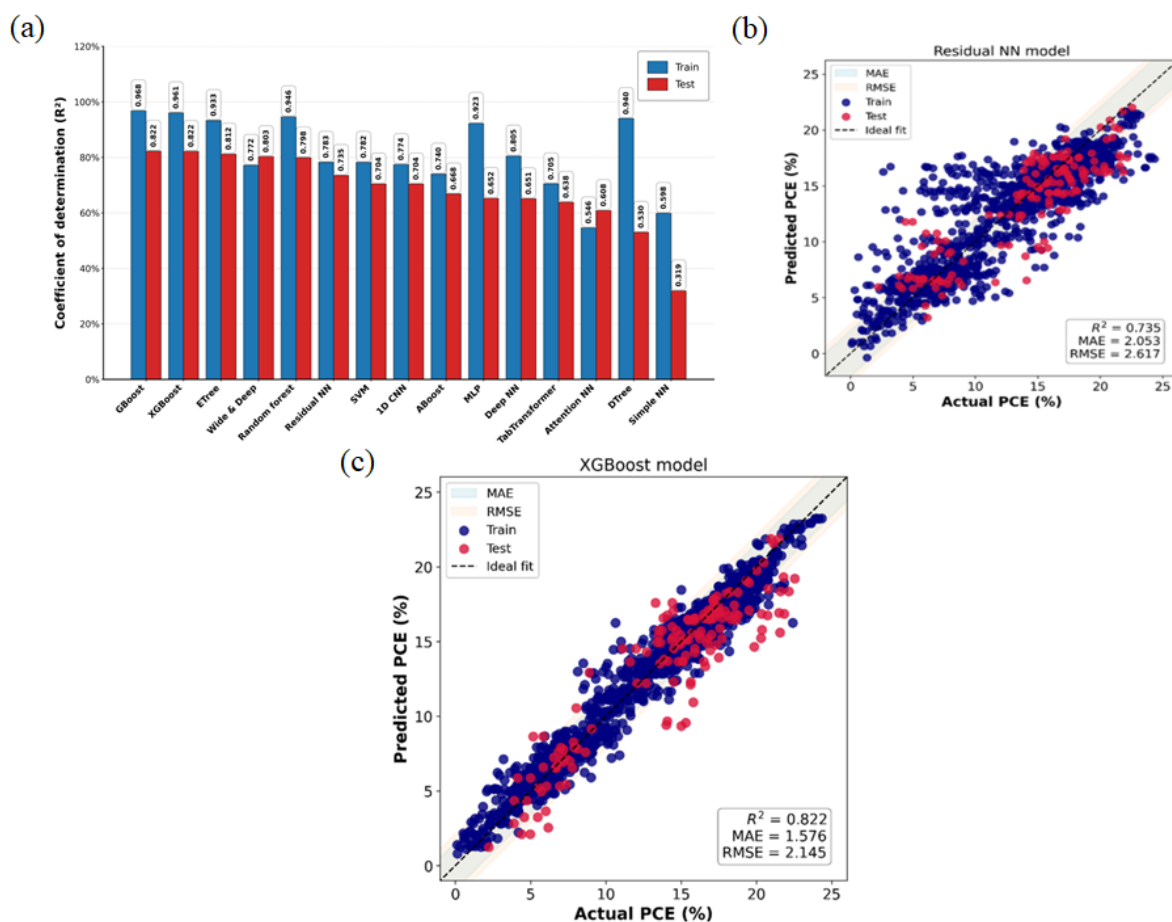
**Figure 5.** Performance of neural network-based models across varying volumes of synthetic data. (a)  $R^2$  score, (b) RMSE, and (c) MAE. In contrast to classical models, neural architectures show continued improvement without saturation in this range, highlighting their stronger dependence on and benefit from synthetic data augmentation.

Abbreviations: 1D CNN: One-Dimensional Convolutional Neural Network; Attention NN: Attention Neural Network; Deep NN: Deep Neural Network; MAE: Mean absolute error;  $R^2$ : coefficient of determination; Residual NN: Residual Neural Network; RMSE: Root mean square error; Simple NN: Simple Neural Network.

values reflecting greater learning capacity. The fitting graph (Figure 6b) illustrates the dispersion of predictions for the selected ideal model, XGBoost, from the ideal line. The blue and pink bars surrounding the ideal line represent the MAE and RMSE, respectively. The MAE for XGBoost is 1.576, indicating that, on average, the predictions are 1.576% away from the true PCE value. The RMSE value of 2.145 suggests the presence of significant errors, particularly for PCE values exceeding 10, highlighting the need for additional data in this range for improved training. In contrast, models based on deep networks exhibited lower performance at this stage, with the ResNet model achieving an  $R^2$  value of only 0.735 for the test data (Figure 6c). Although this performance is acceptable, further improvements of the model are needed to ensure high confidence and validity in the results.

In the subsequent step, the models were trained on a

concatenated dataset, resulting in a training set comprising 4,000 synthetic data points and 1,390 experimental data points. Figure 7 illustrates the  $R^2$  values for the various models on both the training and test datasets, highlighting the correlation between predicted and actual PCE values. Notably, the ResNet model demonstrates significantly reduced dispersion compared to previous analyses, indicating a significant improvement in evaluation metrics. ResNet achieves an  $R^2$  value of 0.87, an RMSE of 1.834, and an MAE of 1.496, outperforming the classical and ensemble-based models. Table S5 presents numeric results for the main dataset, while Table S6 shows results for the integrated synthesis dataset. These results underscore the potential of synthetic data to improve the learning capabilities of deep neural networks and suggest a higher degree of generalization in deep networks relative to classical and ensemble models. This implies that the model effectively captures the fundamental patterns within the



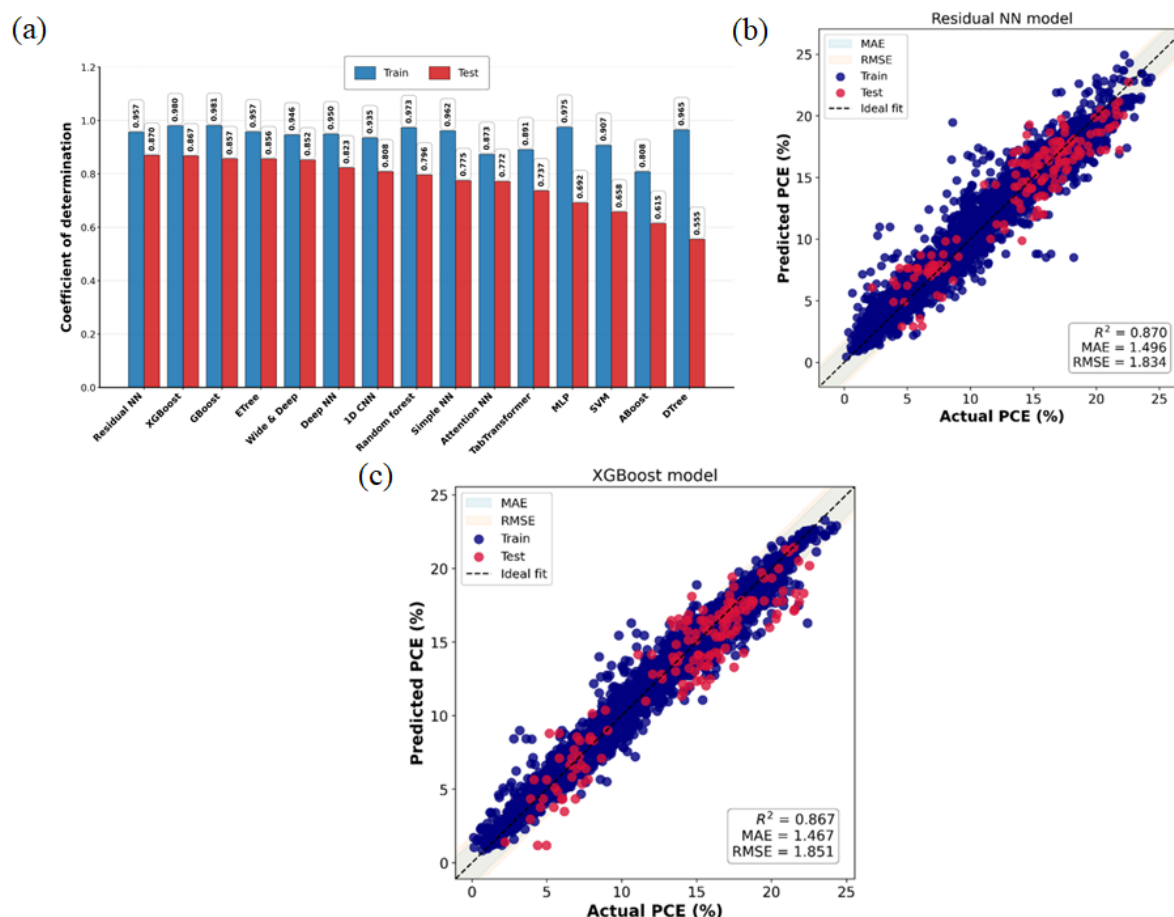
**Figure 6.** Bar chart and fitting graph of testing and training results based on the main dataset using 15 distinct models for PCE prediction. The blue bars represent the training, while the red bars represent the test results: (a) Bar chart illustrating the performance of various models. (b) Fitting plot for the XGBoost model. (c) Fitting plot for the ResNet model.

Abbreviations: 1D CNN: One-Dimensional Convolutional Neural Network; ABoost: Adaptive Boosting; Attention NN: Attention Neural Network; Deep NN: Deep Neural Network; DTree: Decision Tree; ETree: Extra Trees; GBoost: Gradient Boosting; MAE: Mean absolute error; MLP: Multilayer Perceptron; PCE: Power Conversion Efficiency;  $R^2$ : coefficient of determination; Residual NN/ResNet: Residual Neural Network; RMSE: Root mean square error; Simple NN: Simple Neural Network; SVM: Support Vector Machine; XGBoost: Extreme Gradient Boosting.

data rather than relying on specific details. This further suggests that, when generated effectively, synthetic data can introduce latent relationships or valuable noise into the model. Deep models are better equipped to recognize these complex relationships, whereas classical models may only respond to linear or simpler features. Conversely, while enhancing neural network-based models due to their significant potential, we also recognize the importance of strengthening the learning of classical and ensemble models. This approach not only increases the volume of data but also improves its quality, reflecting the model's effective training on the data distribution.

Using SHAP provides an effective framework for identifying significant factors influencing the final target variable,<sup>47</sup> particularly in cases where conventional physics

cannot fully elucidate the complex relationships among factors governing PCE. SHAP is grounded in Game Theory, specifically Shapley values, and is widely used to interpret the outputs of machine learning models by assigning an importance value to each feature.<sup>48</sup> In the SHAP summary plots, the x-axis represents the SHAP values for each feature, and positive and negative values indicate positive and negative contributions to the final PCE. The color bar on the right denotes feature magnitude, with values transitioning from blue (low) to red (high). Figure 8a shows the key features that significantly influence the final efficiency as well as the output of SHAP analysis based on the XGBoost model trained exclusively on the original dataset. These features include the  $\text{PbI}_2$  content, the amount of additive concentration in the absorber layer, the polarity index of the antisolvent used, the electron



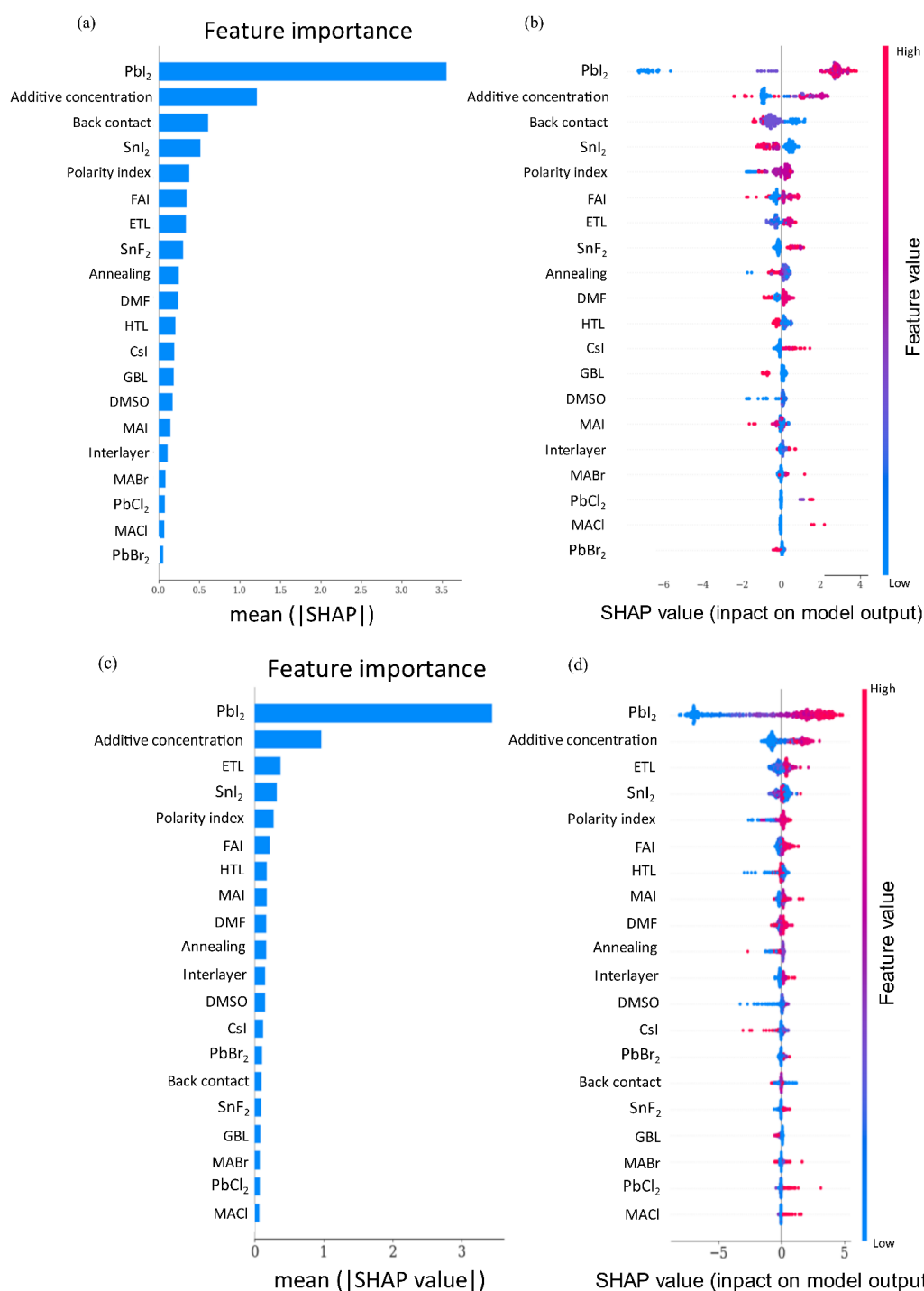
**Figure 7.** The bar chart and fitting graph of testing and training results based on the integrated main data and data generated by the  $\beta$ -VAE model using 15 distinct models for PCE prediction. The blue bars represent the training, while the red bars represent the test results: (a) bar chart illustrating the performance of various models, (b) fitting plot for the XGBoost model, (c) fitting plot for the ResNet model.

Abbreviations: 1D CNN: One-Dimensional Convolutional Neural Network; ABoost: Adaptive Boosting; Attention NN: Attention Neural Network; Deep NN: Deep Neural Network; DTrees: Decision Tree; ETrees: Extra Trees; GBoost: Gradient Boosting; MAE: Mean absolute error; MLP: Multilayer Perceptron; PCE: Power Conversion Efficiency;  $R^2$ : coefficient of determination; Residual NN/ResNet: Residual Neural Network; RMSE: Root mean square error; Simple NN: Simple Neural Network; SVM: Support Vector Machine; XGBoost: Extreme Gradient Boosting.

transport layer, the FAI value, and the hole transport layer. However, Figure 8a does not explicitly convey the specific impact of each feature on the final efficiency. In contrast, Figure 8b clarifies how variations in individual features influence device efficiency. Among the A-site cations in the absorber layer, FAI exhibits the most significant impact on PCE. A decrease in FAI content results in a slight decline in the PCE of PSCs, whereas increasing FAI content leads to a significant improvement in PCE, underscoring its critical role in achieving high PCE in PSCs. We analyzed the influence of various features on PCE by augmenting the original dataset with synthetic data and reanalyzed using the same XGBoost model for comparison with prior results. As illustrated in Figure 8b, the five most influential features identified were  $\text{PbI}_2$ , additive concentration,

polarity index, electron transport layer, and FAI value, respectively. Consistent with earlier observations, FAI remains the most influential A cations of the perovskite layer. Additionally, the importance of the  $\text{SnI}_2$  becomes more pronounced after data augmentation. Specifically, an increase in  $\text{SnI}_2$  content coupled with a decrease in  $\text{PbI}_2$  content significantly reduces the PCE of PSCs. This phenomenon can be attributed to the rapid oxidation of  $\text{Sn}^{2+}$  ions, which enhances non-radiative recombination,<sup>49</sup> thereby reducing the PCE of PSCs. Consequently, achieving high PCE in PSCs requires increasing FAI content and selecting perovskites with optimal  $\text{PbI}_2$  and  $\text{SnI}_2$  values. Other features also exert varying degrees of influence on PCE. For instance, the SHAP analysis indicates that higher CsI content correlates with a reduction in PCE, consistent





**Figure 8.** SHAP analysis identifying key features influencing PCE in perovskite solar cells. (a, c) A bar chart illustrating the average absolute SHAP values. (b, d) Swarm plots displaying the distribution of SHAP values for each feature. Abbreviation: SHAP: Shapley Additive Explanations.

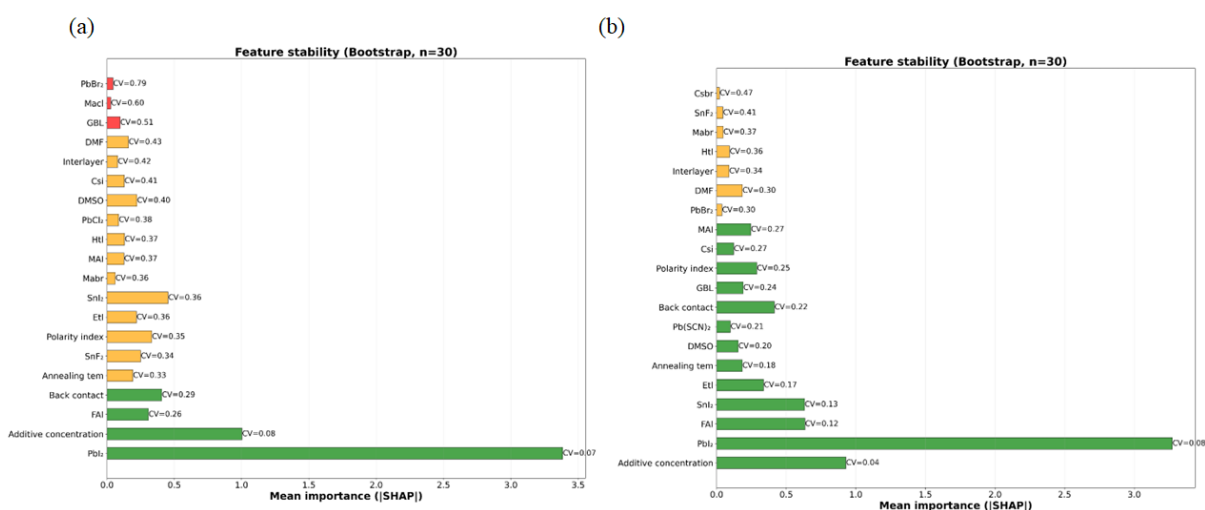
with the widening of the band gap and diminished photon absorption,<sup>50</sup> an effect that was not clearly evident in Figure 8b. Furthermore, reductions in SnI<sub>2</sub> content and back-contact materials (coded as 0 and 1 for gold and silver, respectively, and up to 4 for carbon) exert a more pronounced impact on PCE compared with earlier analyses. Finally, the electron transfer layer is shown to be more influential than the hole transfer layer, in agreement with previous studies highlighting the importance of the electron transfer layer in comparison to the hole transfer layer.<sup>51–54</sup>

To assess the robustness of SHAP-based feature importance, a bootstrap stability analysis was conducted across 30 iterations. For each iteration, a new bootstrap sample was drawn from the training data, an XGBoost model was retrained, and SHAP values were computed using a permutation explainer. Feature importance was calculated as the mean absolute SHAP value per feature. On the original dataset, the mean coefficient of variation (CV) across features was 0.3739, with 20% of features (4/20) classified as stable (CV < 0.3) and 7% (3/20) as unstable (CV > 0.5). The top predictive features exhibited lower CV values, indicating greater inherent stability. When the same analysis was applied to the combined dataset (original and synthetic data), stability improved markedly. The mean CV decreased to 0.2459, the number of stable features increased to 14 out of 20, and no features were classified as unstable. This demonstrates that synthetic data augmentation significantly enhances the consistency and reliability of feature importance interpretations. For example, the CV of the SnI<sub>2</sub>-related feature decreased from 0.36 (moderately unstable) to 0.13 (highly stable),

confirming its role as a genuinely important predictor rather than an artifact of data sparsity. Overall, these results indicate that interpretability is not only preserved but enhanced with synthetic data, yielding more stable and trustworthy explanations across different training conditions. Figure 9a,b displays the results of the bootstrap analysis for the original dataset and the combined dataset (original and synthetic data), respectively.

### 3.2. The impact of organic additives on cell performance

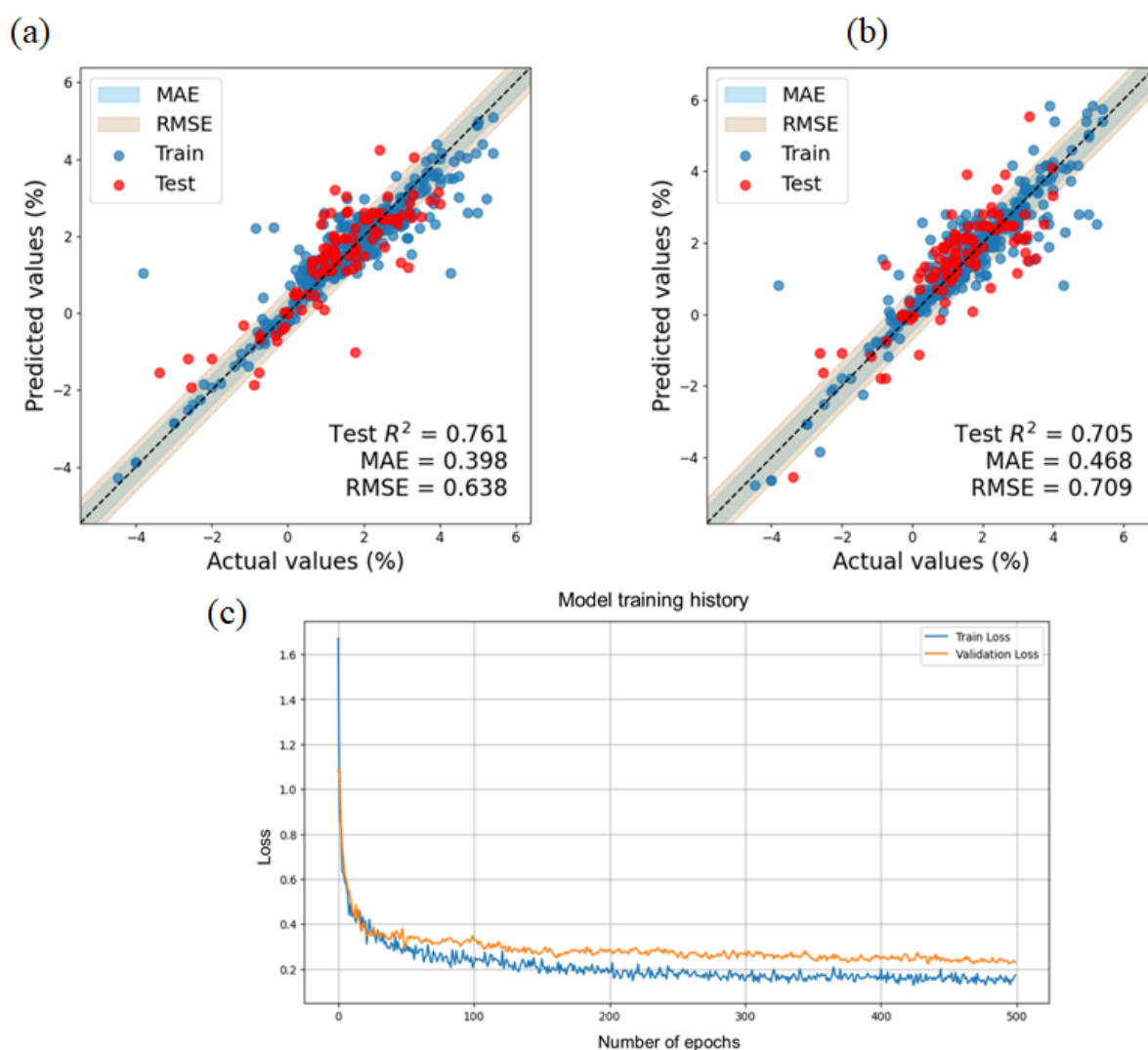
The application of organic additives to enhance the stability and overall efficiency of perovskite solar cells has increased significantly in recent years.<sup>55–59</sup> Perovskite fabrication processes, particularly solution-based methods, consistently introduce detrimental defects within the material layer.<sup>60</sup> Furthermore, the aggregation of multiple crystals exacerbates this issue.<sup>61</sup> The presence of Pb<sup>2+</sup> ions and uncoordinated halogens contributes to ion migration and undesirable recombination within the structure,<sup>62</sup> thereby adversely affecting the cell stability and performance. One of the most effective strategies for controlling grain growth and minimizing structural defects is the engineering and selection of appropriate additives within the cell architecture.<sup>63,64</sup> Different functional groups exert distinct effects on the absorber layer and its interface with adjacent layers.<sup>65–67</sup> For instance, hydroxyl (–OH) groups can effectively adsorb oxygen by donating hydrogen atoms and electrons, thereby providing robust protection against oxidation in tin-containing perovskites.<sup>68</sup> Here, we collected a dataset focusing on organic additives to investigate their effects on the final performance and



**Figure 9.** Bootstrap stability analysis of Shapley Additive Explanations feature importance. (a) Distribution of the coefficient of variation (CV) for each feature using only the original dataset. (b) Distribution of CV for each feature using the combined dataset (original and synthetic data).

efficiency of solar cell structures. The objective is to identify new additives that enhance both the efficiency and stability of these structures. To mitigate the potential for misleading results, we considered only the change in efficiency—either an increase or decrease—resulting from the incorporation of specific additives in the absorber layer. Furthermore, drawing inspiration from the SoftMax activation function, we differentiated the impact of additives based on their effects on PCEs, placing greater emphasis on improvements exceeding 10% compared to those below this threshold. After identifying the important features using the XGBoost regression model and evaluating the performance of various models, we developed a predictive model incorporating classical approaches, the XGBoost

model, and a deep neural network to predict the PCE increases across different structures. Figure 10 illustrates the plots of predicted PCE increases versus actual PCEs for both the training and test sets. Key regression metrics, including RMSE,  $R^2$ , and MAE, were employed to evaluate model predictions. RMSE measures the deviation between predicted and actual values, with lower RMSE signifying smaller discrepancies. Conversely,  $R^2$  represents the proportion of variance in the dependent variable explained by the independent variable(s) and serves as a critical indicator of model fit, reflecting the predictive capability of the dependent variable based on the independent variables. In this context, the dependent variable is the predicted PCE change, while the independent variable is the actual



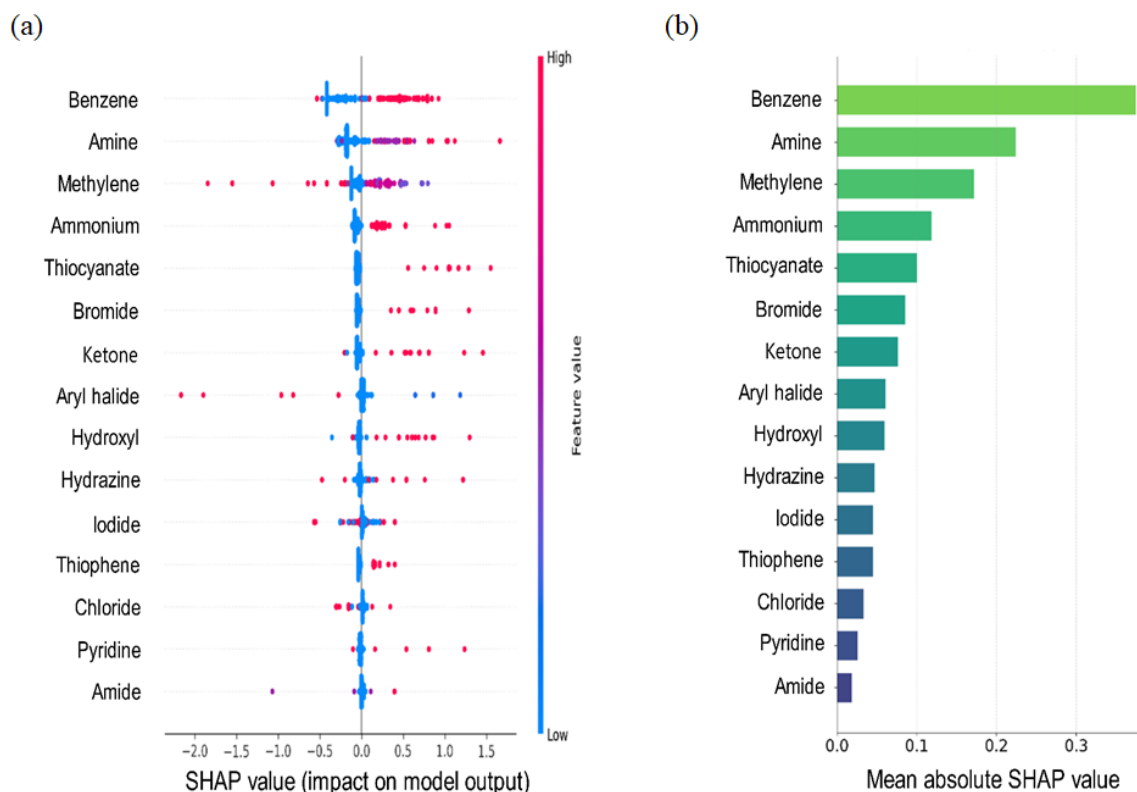
**Figure 10.** Performance comparison of XGBoost and deep neural network models on raised PCE. Training and testing fit curves for (a) the XGBoost model, and (b) the neural network model. (c) Training and validation loss curves for the neural network model.

Abbreviations: MAE: Mean absolute error;  $R^2$ : coefficient of determination; RMSE: Root mean square error; XGBoost: Extreme Gradient Boosting.

PCE change. The  $R^2$  value ranges from 0 to 1, with values greater than 0.6 indicating substantial explanatory power. As presented in Figure 10a,b, the RMSE values for the XGBoost and deep neural network models on the test set are 0.638 and 0.709, respectively, with corresponding  $R^2$  values of 0.761 and 0.705. Additionally, the MAE values are 0.3982 and 0.468, respectively. These results demonstrate the strong reliability of these models.

The  $R^2$  values for the training dataset are 0.864 and 0.866, respectively. These results further indicate that the extreme gradient boosting model provides a good fit for our data, demonstrating accurate predictions without overfitting. SHAP analysis was employed to assess the significance of 26 functional groups as inputs for enhancing the quality of the absorber layer, thereby improving PCE. Figure 11a shows the relationship between feature values and their corresponding SHAP values derived from extreme gradient boosting. The x-axis denotes the SHAP values of each feature, with positive and negative SHAP values reflecting their positive and negative contributions, respectively. The color bar on the right indicates the relative values of each feature, with a

gradient from blue to red signifying an increase in feature values from low to high. The five most influential features identified are benzene, amine, methylene, ammonium, and thiocyanate. Notably, positive correlations are observed between feature values and SHAP values for benzene, amine, ammonium, and thiocyanate, whereas methylene exhibits a negative correlation. Molecules with long carbon chains may inhibit electron exchange between effective active functional groups and the cations and anions within the absorber layer.<sup>69</sup> The impact of specific features on PCE was also investigated. Molecules containing a benzene ring tend to enhance thermal and moisture stability due to the high electron density of aromatic structures, which can engage in  $\pi$ - $\pi$  interactions or weak coordination with lead ions ( $\text{Pb}^{2+}$ ) or halides (iodide [ $\text{I}^-$ ]/bromide [ $\text{Br}^-$ ]).<sup>70</sup> The amine group serves as a ligand for lead cations ( $\text{Pb}^{2+}$ ) and mitigates surface defects,<sup>71</sup> while also improving adhesion between the perovskite layers and the charge transport layer.<sup>72</sup> In perovskite, ammonium-containing groups are crucial for maintaining structural stability and influencing the energy gap. Ammonium can form hydrogen bonds with halide anions ( $\text{I}^-/\text{Br}^-$ ), thereby



**Figure 11.** SHAP analysis identifying the impact of additives on the efficiency of perovskite solar cells. (a) A bar chart illustrating the average absolute SHAP values. (b) Swarm plots displaying the distribution of SHAP values for each feature.

Abbreviation: SHAP: Shapley Additive Explanations.

stabilizing the structure.<sup>73</sup> Thiocyanate further enhances the morphology of the perovskite layer and reduces surface defects.<sup>74</sup> The XGBoost model thus generates guidelines for identifying additives that can improve the performance of methylammonium lead iodide perovskite systems. To enhance device performance, researchers should prioritize key characteristics such as benzene, amine, ammonium, and thiocyanate when selecting additives. Additionally, the model was utilized to predict the PCE improvement rates for 20 new candidate molecules.

Table 2 shows the predictive performance of the XGBoost regression model on a set of organic additives sourced from recent, reputable literature that were not included in the initial training dataset. This evaluation assesses the model's generalizability beyond the compounds it was originally trained on. The predicted values for the PCE increase show strong agreement with the experimental results for all five samples. For instance, the predicted PCE increase for

terephthalic acid (sample 4) is 3.208, closely matching the experimental value of 3.203, with a minimal error of 0.005. Similar accuracy is observed for the other samples, with all predictions remaining within a reasonable deviation range. This demonstrates that the XGBoost model effectively captures the structural and chemical characteristics correlating with PCE increases. Consequently, the model can be confidently used to screen new molecular additives prior to experimental validation, reducing the time and costs of experimental testing. The average relative error between the predicted and experimental values for the five unseen organic additives is approximately 11.5%, indicating that the XGBoost regression model achieves an effective accuracy of about 89% on external data. This level of performance demonstrates the robust generalizability of the model in predicting PCE enhancements from new molecular additives and highlights its practical utility in guiding experiments.

**Table 2. Organic additives extracted from the literature (not included in the primary dataset), evaluated using an Extreme Gradient Boosting regressor to predict the power conversion efficiency improvement**

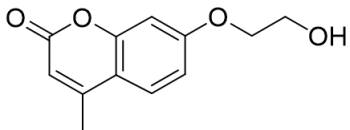
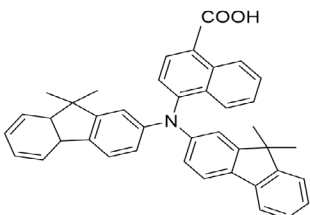
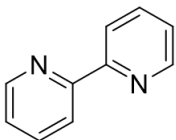
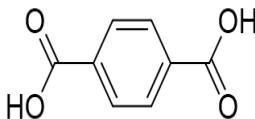
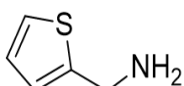
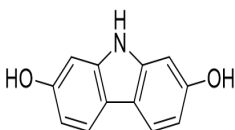
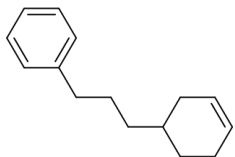
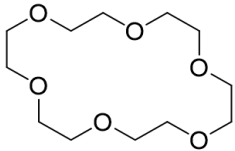
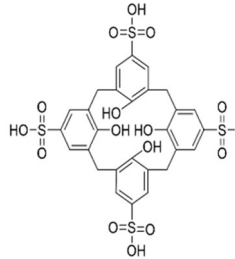
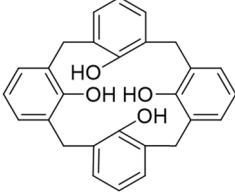
Sample number	Molecular name	Predicted	Experimental	Chemical structure	Reference
1	Coumarin hydroxyethyl	2.62	2.29		2
2	4-(bis (9,9-dimethyl-9H-fluoren-2-yl) amino)-1-naphthoic acid	2.59	2.75		3
3	2,2'-bipyridine	1.847	1.57		4
4	Terephthalic acid	3.208	3.203		5
5	2-thienylmethylamine hydrochloride	1.85	1.575		6



Table 3 presents examples of new candidates sourced from PubChem, selected for their diverse functional groups to support further research. These candidates were evaluated using a pre-trained model to predict their target values. The 20 most effective candidates are detailed in Table S7. The influence of various functional groups on the absorber layer is clearly illustrated. For instance, the molecule 4-sulfocalix[4]arene, which features four additional sulfonate functional groups, is more complex than calix[4]arene-25,26,27,28-tetrol but demonstrates

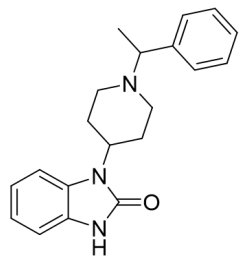
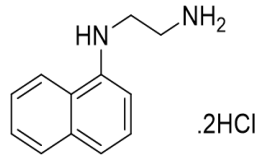
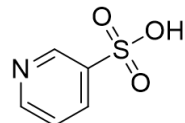
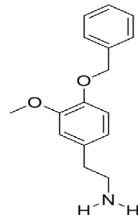
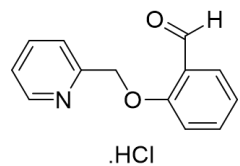
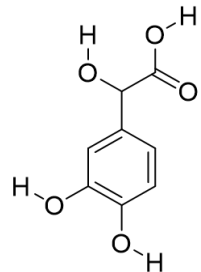
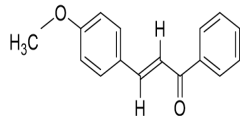
inferior performance. This may be due to the combined effects of hydroxyl and sulfonate groups on the solution, which increased acidity (lower pH), leading to more defects and reduced surface passivation. Conversely, samples 6 and 10 exhibit superior performance due to their effective incorporation of functional groups such as amine, ketone, ether, methylene, and phenyl. Phenyl has demonstrated superior ability to neutralize defects compared to iodo and cyano groups when coordinated with amide, supporting perovskite grain growth, reducing defects, and enhancing

**Table 3. New organic additives candidates with potentially high power conversion efficiency (PCE) utilizing the Extreme Gradient Boosting regression model and the amount of raised PCE as the target variable**

Sample number	Candidate	Classification	Target	Chemical structure
1	9H-Carbazole-2,7-diol	Aromatic alcohol	2.69	
2	[3-(Cyclohex-3-en-1-yl)propyl]benzene	Aromatic hydrocarbon	2.55	
3	18-Crown-6-	Polyether	2.17	
4	4-Sulfocalix[4]arene	Calixarene	1.89	
5	Calix[4]arene-25,26,27,28-tetrol	Calixarene	2.81	

(Cont'd...)

Table 3. Continued

Sample number	Candidate	Classification	Target	Chemical structure
6	1,3-dihydro-1-[1-(1-phenylethyl)-4-piperidinyl]-2H-benzimidazol-2-one	Heterocyclic compound	4.26	
7	N-1-Naphthylethylenediamine dihydrochloride	Aromatic amine salt	2.44	
8	Pyridine-3-sulfonic acid	Heterocyclic aromatic	1.76	
9	Phenethylamine, 4-(benzyloxy)-3-methoxy-	Phenethylamines	2.16	
10	2-(4-pyridinylmethoxy)benzaldehyde hydrochloride	Aromatic aldehyde	4.15	
11	2-hydroxy-2-(4-hydroxy-3-methoxyphenyl)acetic acid	Aromatic hydroxy acid	2.95	
12	4'-Methoxychalcone	Flavonoid/aromatic ketone	2.3	

charge transport more effectively than other end groups through  $\pi$ - $\pi$  conjugation and hydrophobicity.<sup>75</sup> Amine additives can effectively modulate crystal growth, film formation, and crystallization, improving perovskite film quality in terms of crystallinity, crystal orientation, smoothness, and uniformity. Acting as a Lewis base via their amino groups, these additives form coordination bonds with perovskites by sharing the N lone electron pair with the empty 6p orbital of  $\text{Pb}^{2+}$ .<sup>76,77</sup> This coordination slows the reaction between the metal halide and organic halide salt, reducing crystallization rate. As a result, highly uniform and compact perovskite films with preferential crystal growth orientation and improved crystallinity can be achieved. The coordination bonds formed by the ketone group with  $\text{Pb}^{2+}$  and tin(IV)/tin(II) ions, combined with the hydrogen bonds from the -OH group with  $\text{I}^-$  in perovskite, facilitate the passivation process. This multifunctional interface interaction effectively removes residual  $\text{PbI}_2$  from the grain boundaries, preventing perovskite degradation. Optimizing energy level alignment between perovskite and  $\text{SnO}_2$  quantum dots further reduces interface barriers, facilitating the establishment of an electron bridge for rapid electron extraction and transfer. Consequently, the D-fructose-based PSC achieved a champion efficiency of 24.91%, with negligible J-V hysteresis and excellent stability.<sup>78</sup>

## 4. Conclusion

In summary, we developed high-accuracy models for predicting and identifying optimal structures in the field of PSCs using a dataset of 1,540 experimental data points and 4,000 synthetic data points. The data were carefully selected and cleaned to enhance quality, enabling robust applicability to new experiments and providing a reliable recommendation system. We employed a generative neural network, specifically the  $\beta$ -VAE model, to generate synthetic data and learned the relationships between features and the target variable. For evaluation, we implemented various machine learning models, including classical models, ensemble methods, and deep networks. The results showed a significant enhancement in the learning capabilities of the models. ResNet achieved the highest  $R^2$  value of 0.87 among all models, reflecting a 15.5% improvement, while the  $R^2$  values for the XGBoost and GBoost models increased from 0.822 and 0.821 to 0.867 and 0.857, respectively. These findings demonstrate that synthetic data can effectively address data scarcity in laboratory sciences while enhancing predictive accuracy, ultimately reducing experimental costs and time. Furthermore, to increase the capabilities of the proposed system, we collected 733 data points on organic additives, analyzed their functional groups and quantities within the

additives used in the perovskite solution, and identified effective functional groups, enabling the proposal of new molecules for future research.

## Acknowledgements

The authors would like to thank Laser Plasma Research Institute for its support.

## Funding

None.

## Conflict of interest

The authors declare that they have no competing interests.

## Author contributions

*Conceptualization:* All authors

*Data curation:* Behzad Iranipour

*Formal analysis:* Behzad Iranipour, Mohammadreza Sadeghian

*Funding acquisition:* Ezeddin Mohajerani

*Investigation:* All authors

*Methodology:* All authors

*Project administration:* Ezeddin Mohajerani

*Software:* Behzad Iranipour, Mohammadreza Sadeghian

*Supervision:* Ezeddin Mohajerani

*Validation:* All authors

*Visualization:* Behzad Iranipour, Mohammadreza Sadeghian

*Writing-original draft:* Behzad Iranipour

*Writing-review & editing:* All authors

## Ethics approval and consent to participate

Not applicable.

## Consent for publication

Not applicable.

## Availability of data

The code and dataset supporting the findings of this study are publicly available in the Beta-Variational-Autoencoder repository on GitHub at <https://github.com/BehzadIranipour/Beta-Variational-Autoencoder>.

## References

1. Gómez-Peralta JI, Bokhimi X. Discovering new perovskites with artificial intelligence. *J Solid State Chem.* 2020;285:121253.  
doi: 10.1016/j.jssc.2020.121253
2. Liu M, Cao Z, Wang X, *et al.* Perovskite material-based memristors for applications in information processing and

- artificial intelligence. *J Mater Chem C*. 2023;11(39):13167-13188.  
doi: 10.1039/D3TC02309E
3. Tao Q, Xu P, Li M, Lu W. Machine learning for perovskite materials design and discovery. *Npj Comput Mater*. 2021;7(1):23.  
doi: 10.1038/s41524-021-00495-8
4. Liu Y, Yan W, Han S, *et al*. How machine learning predicts and explains the performance of perovskite solar cells. *Solar RRL*. 2022;6(6):2101100.  
doi: 10.1002/solr.202101100
5. Lu Y, Wei D, Liu W, *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*. 2023;77:200-208.  
doi: 10.1016/j.jechem.2022.10.024
6. Khan A, Kandel J, Tayara H, Chong KT. Predicting the bandgap and efficiency of perovskite solar cells using machine learning methods. *Mol Inform*. 2024 ;43(2):e202300217.  
doi: 10.1002/minf.202300217
7. Huang F, Pascoe AR, Wu WQ, *et al*. Effect of the microstructure of the functional layers on the efficiency of perovskite solar cells. *Adv Mater*. 2017;29(20):1601715.  
doi: 10.1002/adma.201601715
8. Lakhdar N, Hima A. Electron transport material effect on performance of perovskite solar cells based on CH<sub>3</sub>NH<sub>3</sub>GeI<sub>3</sub>. *Opt Mater*. 2020;99:109517.  
doi: 10.1016/j.optmat.2019.109517
9. Shao S, Loi MA. The role of the interfaces in perovskite solar cells. *Adv Mater Interfaces*. 2020;7(1):1901469.  
doi: 10.1002/admi.201901469
10. Bag A, Radhakrishnan R, Nekovei R, Jeyakumar R. Effect of absorber layer, hole transport layer thicknesses, and its doping density on the performance of perovskite solar cells by device simulation. *Sol Energy* 2020;196:177-182.  
doi: 10.1016/j.solener.2019.12.014
11. Li Z, Xiao C, Yang Y, *et al*. Extrinsic ion migration in perovskite solar cells. *Energy Environ Sci*. 2017;10(5):1234-1242.  
doi: 10.1039/C7EE00358G
12. Valencia A, Liu F, Zhang X, *et al*. Auto-generating a database on the fabrication details of perovskite solar devices. *Sci Data* 2025;12(1):270.  
doi: 10.1038/s41597-025-04566-z
13. Kenfack AK, Mashamba DR, Thantsha NM, Msimanga M. Prediction of band gap and optimum electrical parameters of a thin homojunction perovskite solar cell based on FA1-xCsxSnyPb1-yI3 through a combination of SCAPS-1D and machine learning based modelling. *Mater Today Commun*. 2023;37:107318.  
doi: 10.1016/j.mtcomm.2023.107318
14. Mammeri M, Dehimi L, Bencherif H, Pezzimenti FJ. Paths towards high perovskite solar cells stability using machine learning techniques. *Sol Energy*. 2023;249:651-660.  
doi: 10.1016/j.solener.2022.12.002
15. Li W, Hu J, Chen Z, *et al*. Performance prediction and optimization of perovskite solar cells based on the Bayesian approach. *Sol Energy*. 2023;262:111853.  
doi: 10.1016/j.solener.2023.111853
16. Yan W, Liu Y, Zang Y, *et al*. Machine learning enabled development of unexplored perovskite solar cells with high efficiency. *Nano Energy*. 2022;99:107394.  
doi: 10.1016/j.nanoen.2022.107394
17. Calvo ME. Materials chemistry approaches to the control of the optical features of perovskite solar cells. *J Mater. Chem. A*. 2017;5(39):20561-20578.  
doi: 10.1039/C7TA05666D
18. Yang B, Suo J, Di Giacomo F, *et al*. Interfacial passivation engineering of perovskite solar cells with fill factor over 82% and outstanding operational stability on nip architecture. *ACS Energy Lett*. 2021;6(11):3916-3923.  
doi: 10.1021/acsenenergylett.1c01811
19. Foster JM, Snaith HJ, Leijtens T, Richardson G. A model for the operation of perovskite based hybrid solar cells: Formulation, analysis, and comparison to experiment. *SIAM J Appl Math*. 2014;74(6):1935-1966.  
doi: 10.1137/130934258
20. Hernández-Balaguera E, Arredondo B, del Pozo G, Romero B. Exploring the impact of fractional-order capacitive behavior on the hysteresis effects of perovskite solar cells: A theoretical perspective. *Commun Nonlinear Sci Numer Simul*. 2020;90:105371.  
doi: 10.1016/j.cnsns.2020.105371
21. Hunde BR, Woldeyohannes AD. Performance analysis and optimization of perovskite solar cell using SCAPS-1D and genetic algorithm. *Mater Today Commun*. 2023;34:105420.  
doi: 10.1016/j.mtcomm.2023.105420
22. Danladi E, Gyuk PM, Tasie NN, *et al*. Impact of hole transport material on perovskite solar cells with different metal electrode: a SCAPS-1D simulation insight. *Heliyon*. 2023;9(6).  
doi: 10.1016/j.heliyon.2023.e16838
23. Wang Y, Wu J, Zhang P, *et al*. Stitching triple cation perovskite by a mixed anti-solvent process for high performance perovskite solar cells. *Nano Energy*. 2017;39:616-625.

- doi: 10.1016/j.nanoen.2017.07.046
24. Slimi B, Mollar M, Assaker IB, *et al.* Perovskite FA1-xMAxPbI3 for solar cells: films formation and properties. *Energy Procedia*. 2016;102:87-95.  
doi: 10.1016/j.egypro.2016.11.322
25. Abedini-Ahangarkola H, Soleimani-Amiri S, Rudi SG. Modeling and numerical simulation of high efficiency perovskite solar cell with three active layers. *Sol Energy*. 2022;236:724-732.  
doi: 10.1016/j.solener.2022.03.055
26. Xiao Z, Song Z, Yan Y. From lead halide perovskites to lead-free metal halide perovskites and perovskite derivatives. *Adv Mater*. 2019;31(47):1803792.  
doi: 10.1002/adma.201803792
27. Tang Y, Li Z, Nellikkal MA, *et al.* Improving data and prediction quality of high-throughput perovskite synthesis with model fusion. *J Chem Inf Model*. 2021;61(4):1593-1602.  
doi: 10.1021/acs.jcim.0c01307
28. Marchenko EI, Fateev SA, Petrov AA, *et al.* Database of two-dimensional hybrid perovskite materials: open-access collection of crystal structures, band gaps, and atomic partial charges predicted by machine learning. *Chem Mater*. 2020;32(17):7383-7388.  
doi: 10.1021/acs.chemmater.0c02290
29. Zhou L, Pan S, Wang J, Vasilakos AV. Machine learning on big data: Opportunities and challenges. *Neurocomputing*. 2017;237:350-361.  
doi: 10.1016/j.neucom.2017.01.026
30. Tufail S, Riggs H, Tariq M, Sarwat AI. Advancements and challenges in machine learning: A comprehensive review of models, libraries, applications, and algorithms. *Electronics*. 2023;12(8):1789.  
doi: 10.3390/electronics12081789
31. Jacobs R, Liu J, Abernathy H, Morgan D. Machine learning design of perovskite catalytic properties. *Adv Energy Mater*. 2024;14(12):2303684.  
doi: 10.1002/aenm.202303684
32. Jacobsson TJ, Hultqvist A, García-Fernández A, *et al.* An open-access database and analysis tool for perovskite solar cells based on the FAIR data principles. *Nat Energy*. 2022;7(1):107-115.  
doi: 10.1038/s41560-021-00941-3
33. Kusuma FJ, Widiyanto E, Santoso I, *et al.* Optimizing novel device configurations for perovskite solar cells: Enhancing stability and efficiency through machine learning on a large dataset. *Renew Energy*. 2025;247:122947.  
doi: 10.1016/j.renene.2025.122947
34. Zhao S, Wang J, Guo Z, *et al.* Exploring device physics of Perovskite solar cell via machine learning with limited samples. *J Energy Chem*. 2024;94:441-448.  
doi: 10.1016/j.jechem.2024.03.003
35. Iranipour B, Sadeghian M, Mohajerani E. Artificial data generation: A strategy to improve efficiency predictions in mixed Sn-Pb perovskite solar cells. *Mater Today Commun*. 2025;43:111625.  
doi: 10.1016/j.mtcomm.2025.111625
36. Boubchir M, Boubchir R, Aourag H. The Principal Component Analysis as a tool for predicting the mechanical properties of Perovskites and Inverse Perovskites. *Chem Phys Lett*. 2022;798:139615.  
doi: 10.1016/j.cplett.2022.139615
37. Li Y, Scheel KR, Clevenger RG, *et al.* Highly efficient and stable perovskite solar cells using a dopant-free inexpensive small molecule as the hole-transporting material. *Adv Energy Mater*. 2018;8(23):1801248.  
doi: 10.1002/aenm.201801248
38. Fukasawa R, Asahi T, Taniguchi T. Effectiveness and limitation of the performance prediction of perovskite solar cells by process informatics. *Energy Adv*. 2024;3(4):812-820.  
doi: 10.1039/D3YA00617D
39. Anowar F, Sadaoui S, Selim B. Conceptual and empirical comparison of dimensionality reduction algorithms (pca, kpca, lda, mds, svd, lle, isomap, le, ica, t-sne). *Comput Sci Rev*. 2021;40:100378.  
doi: 10.1016/j.cosrev.2021.100378
40. Srivastava M, Howard JM, Gong T, *et al.* Machine learning roadmap for perovskite photovoltaics. *J Phys Chem Lett*. 2021;12(32):7866-7877.  
doi: 10.1021/acs.jpcllett.1c01961
41. Li X, Dan Y, Dong R, *et al.* Computational screening of new perovskite materials using transfer learning and deep learning. *Appl Sci*. 2019;9(24):5510.  
doi: 10.3390/app9245510
42. Zhang R, Motes B, Tan S, *et al.* Machine Learning Prediction of Organic-Inorganic Halide Perovskite Solar Cell Performance from Optical Properties. *ACS Energy Lett*. 2025;10(4):1714-1724.  
doi: 10.1021/acsenergylett.4c03592
43. Tetko IV, Livingstone DJ, Luik AI. Neural network studies. 1. Comparison of overfitting and overtraining. *J Chem Inf Comput Sci*. 1995;35(5):826-833.  
doi: 10.1021/ci00027a006
44. Salehin I, Kang DK. A review on dropout regularization approaches for deep neural networks within the scholarly domain. *Electronics*. 2023;12(14):3106.  
doi: 10.3390/electronics12143106



45. Dias Da Cruz S, Taetz B, Stifter T, Stricker D. Autoencoder and partially impossible reconstruction losses. *Sensors*. 2022;22(13):4862.  
doi: 10.3390/s22134862
46. Asperti A, Trentin M. Balancing reconstruction error and kullback-leibler divergence in variational autoencoders. *IEEE Access*. 2020;8:199440-199448.  
doi: 10.1109/ACCESS.2020.3034828
47. Chen J, Zhan Y, Yang Z, *et al.* Predicting and analyzing stability in perovskite solar cells: Insights from machine learning models and SHAP analysis. *Mater Today Energy*. 2025;48:101769.  
doi: 10.1016/j.mtener.2024.101769
48. Priyanga GS, Sampath S, Shravan PV, *et al.* Advanced prediction of perovskite stability for solar energy using machine learning. *Sol Energy*. 2024;278:112782.  
doi: 10.1016/j.solener.2024.112782
49. Kumari N, Patel SR, Gohel JV. Enhanced stability and efficiency of Sn containing perovskite solar cell with SnCl<sub>2</sub> and SnI<sub>2</sub> precursors. *J MaterSci Mater Electron*. 2018;29(21):18144-18150.  
doi: 10.1007/s10854-018-9926-y
50. Wang L, Shahiduzzaman M, Muslih EY, *et al.* Double-layer CsI intercalation into an MAPbI<sub>3</sub> framework for efficient and stable perovskite solar cells. *Nano Energy*. 2021;86:106135.  
doi: 10.1016/j.nanoen.2021.106135
51. Foo S, Thambidurai M, Senthil Kumar P, *et al.* Recent review on electron transport layers in perovskite solar cells. *Int J Energy Res*. 2022;46(15):21441-21451.  
doi: 10.1002/er.7958
52. Noh MF, Teh CH, Daik R, *et al.* The architecture of the electron transport layer for a perovskite solar cell. *J Mater Chem C*. 2018;6(4):682-712.  
doi: 10.1039/C7TC04649A
53. Sinha NK, Ghosh DS, Khare A. Role of built-in potential over ETL/perovskite interface on the performance of HTL-free perovskite solar cells. *Opt Mater*. 2022;129:112517.  
doi: 10.1016/j.optmat.2022.112517
54. Szymkowski J, Galagan Y, Glowienka D. Exploring the interfacial effects at the ETL/perovskite boundary in the semitransparent perovskite solar cells. *Sol Energy*. 2023;266:112176.  
doi: 10.1016/j.solener.2023.112176
55. Zhang F, Zhu K. Additive engineering for efficient and stable perovskite solar cells. *Adv Energy Mater*. 2020;10(13):1902579.  
doi: 10.1002/aenm.201902579
56. Zhang Y, Li Y, Zhang L, *et al.* Propylammonium chloride additive for efficient and stable FAPbI<sub>3</sub> perovskite solar cells. *Adv Energy Mater*. 2021;11(47):2102538.  
doi: 10.1002/aenm.202102538
57. Zheng Z, Xia M, Chen X, *et al.* Enhancing the performance of Fa-based printable mesoscopic perovskite solar cells via the polymer additive. *Adv Energy Mater*. 2023;13(23):2204335.  
doi: 10.1002/aenm.202204335
58. Gao Y, Wu Y, Lu H, *et al.* CsPbBr<sub>3</sub> perovskite nanoparticles as additive for environmentally stable perovskite solar cells with 20.46% efficiency. *Nano Energy*. 2019;59:517-526.  
doi: 10.1016/j.nanoen.2019.02.070
59. Gong X, Li M, Shi XB, *et al.* Controllable perovskite crystallization by water additive for high-performance solar cells. *Adv Funct Mater*. 2015;25(42):6671-6678.  
doi: 10.1002/adfm.201503559
60. Wang J, Bi L, Fu Q, Jen AK. Methods for passivating defects of perovskite for inverted perovskite solar cells and modules. *Adv Energy Mater*. 2024;14(35):2401414.  
doi: 10.1002/aenm.202401414
61. Saidaminov MI, Williams K, Wei M, *et al.* Multi-cation perovskites prevent carrier reflection from grain surfaces. *Nat Mater*. 2020;19(4):412-418.  
doi: 10.1038/s41563-019-0602-2
62. Le Z, Liu A, Reo Y, *et al.* Ion migration in tin-halide perovskites. *ACS Energy Lett*. 2024;9(4):1639-1644.  
doi: 10.1021/acsenenergylett.4c00198
63. Afroz MA, Garai R, Gupta RK, Iyer PK. Additive-assisted defect passivation for minimization of open-circuit voltage loss and improved perovskite solar cell performance. *ACS Appl Energy Mater*. 2021;4(10):10468-10476.  
doi: 10.1021/acsaem.1c01205
64. Azam M, Liu K, Sun Y, *et al.* Recent advances in defect passivation of perovskite active layer via additive engineering: A review. *J Phys D Appl Phys*. 2020;53(18):183002.  
doi: 10.1088/1361-6463/ab6f8d
65. Liu S, Guan Y, Sheng Y, *et al.* A review on additives for halide perovskite solar cells. *Adv Energy Mater*. 2020;10(13):1902492.  
doi: 10.1002/aenm.201902492
66. Mahapatra A, Prochowicz D, Tavakoli MM, *et al.* A review of aspects of additive engineering in perovskite solar cells. *J Mater Chem A*. 2020;8(1):27-54.  
doi: 10.1039/C9TA07657C
67. Han G, Hadi HD, Bruno A, *et al.* Additive selection strategy for high performance perovskite photovoltaics. *J Phys Chem C*. 2018;122(25):13884-13893.

- doi: 10.1021/acs.jpcc.8b00980
68. Bai Y, Xing D, Luo H, *et al.* Facilitating the formation of SnO<sub>2</sub> film via hydroxyl groups for efficient perovskite solar cells. *Appl Surf Sci.* 2021;552:149459.  
doi: 10.1016/j.apsusc.2021.149459
  69. Fu C, Gu Z, Tang Y, *et al.* From structural design to functional construction: amine molecules in high-performance formamidinium-based perovskite solar cells. *Angew Chem Int Ed.* 2022;61(19):e202117067.  
doi: 10.1002/anie.202117067
  70. Lewinska G, Kanak J, Danel KS, *et al.* Effect of benzene-based dyes on optothermal properties of active layers for ternary organic solar cells. *Appl Surf Sci.* 2023;641:158535.  
doi: 10.1016/j.apsusc.2023.158535
  71. Wang Z, Ma T, Wang J, *et al.* Surface passivation for efficient and stable perovskite solar cells in ambient air: The structural effect of amine molecules. *Ceram Int.* 2024;50(5):7528-7537.  
doi: 10.1016/j.ceramint.2023.12.058
  72. Rasool S, Khan N, Jahankhan M, *et al.* Amine-based interfacial engineering in solution-processed organic and perovskite solar cells. *ACS Appl Mater Interfaces* 2019;11(18):16785-16794.  
doi: 10.1021/acsami.9b03298
  73. Akin S, Dong B, Pfeifer L, *et al.* Organic ammonium halide modulators as effective strategy for enhanced perovskite photovoltaic performance. *Adv Sci.* 2021;8(10):2004593.  
doi: 10.1002/advs.202004593
  74. Yu Y, Wang C, Grice CR, *et al.* Improving the performance of formamidinium and cesium lead triiodide perovskite solar cells using lead thiocyanate additives. *ChemSusChem.* 2016;9(23):3288-3297.  
doi: 10.1002/cssc.201601027
  75. Zhang Y, Xie J, Tao L, *et al.* Passivation strategies of Perovskite film defects for solar cells by bifunctional amides with various molecular structures. *Org Electron.* 2022;108:106597.  
doi: 10.1016/j.orgel.2022.106597
  76. Yang S, Wang Y, Liu P, *et al.* Functionalization of perovskite thin films with moisture-tolerant molecules. *Nat Energy.* 2016;1(2):1-7.  
doi: 10.1038/nenergy.2015.16
  77. Feng W, Tan Y, Yang M, *et al.* Small amines bring big benefits to perovskite-based solar cells and light-emitting diodes. *Chem.* 2022;8(2):351-383.  
doi: 10.1016/j.chempr.2021.11.010
  78. Wu R, Ding B, Xiao S, *et al.* Eco-friendly small molecule with polyhydroxyl ketone as buried interface chelator for enhanced carrier dynamics toward high-performance perovskite solar cells. *Sci China Mater.* 2025;68(4):1249-1258.  
doi: 10.1007/s40843-024-3228-1