Machine learning for perovskite optoelectronics: a review

Feiyue Lu,[a] Yanyan Liang,[a] Nana Wang,[a] Lin Zhu,[a,*] and Jianpu Wang[a,b,*]
[a]Nanjing Tech University, Institute of Advanced Materials and School of Flexible Electronics (Future Technologies), Key Laboratory of Flexible Electronics, Nanjing, China
[b]Changzhou University, School of Materials Science and Engineering and School of Microelectronics and Control Engineering, Changzhou, China

Abstract. Metal halide perovskite materials have rapidly advanced in the perovskite solar cells and light-emitting diodes due to their superior optoelectronic properties. The structure of perovskite optoelectronic devices includes the perovskite active layer, electron transport layer, and hole transport layer. This indicates that the optimization process unfolds as a complex interplay between intricate chemical crystallization processes and sophisticated physical mechanisms. Traditional research in perovskite optoelectronics has mainly depended on trial-and-error experimentation, a less efficient approach. Recently, the emergence of machine learning (ML) has drastically streamlined the optimization process. Due to its powerful data processing capabilities, ML has significant advantages in uncovering potential patterns and making predictions. More importantly, ML can reveal underlying patterns in data and elucidate complex device mechanisms, playing a pivotal role in enhancing device performance. We present the latest advancements in applying ML to perovskite optoelectronic devices, covering perovskite active layers, transport layers, interface engineering, and mechanisms. In addition, it offers a prospective outlook on future developments. We believe that the deep integration of ML will significantly expedite the comprehensive enhancement of perovskite optoelectronic device performance.

Keywords: perovskite; machine learning; optoelectronics.

1 Introduction

Metal halide perovskite materials are considered one of the most promising materials for the next generation of optoelectronic devices, attributed to their high light absorption coefficient, large carrier mobility, and simplicity in synthesis methods.[1–5] These characteristics have significantly increased the focus on perovskite materials within the scientific community, marking them a subject of growing research and development interest. Over the past few years, perovskite optoelectronic devices have witnessed remarkable advancements and widespread application across various fields.[6–10] Notably, perovskite solar cells (PSCs) have attained power conversion efficiencies exceeding 25%,[9,11–14] and perovskite light-emitting diodes (PeLEDs) have broken through the 20% external quantum efficiency (EQE) barrier.[15–26] Furthermore, significant progress has been made in the development of perovskite detectors and lasers.[27–32] However, traditional approaches to material design and device fabrication often rely on trial-and-error processes. These methods are time-consuming and inefficient, hindering further performance enhancements in these devices.

Machine learning (ML) is a subfield of artificial intelligence that utilizes computer science, statistics, and mathematics to develop models and algorithms capable of automatically learning from data and making predictions based on acquired knowledge. Its data analysis capabilities, automatic decision-making processes, and adaptive model adaptation make it a powerful tool for optimizing research processes and minimizing experimental waste.[33] In the perovskite materials, researchers are

*Address all correspondence to Lin Zhu, iamlzhu@njtech.edu.cn; Jianpu Wang, iamjpwang@njtech.edu.cn

increasingly leveraging ML techniques to streamline the research process, notably by reducing the consumption of experimental materials and saving time.

Metal halide perovskites can generally be classified into three-dimensional (3D) and low-dimensional structures. 3D perovskites are described by the general chemical formula ABX_3, where A signifies organic or inorganic cations, such as cesium (Cs^+), formamidinium (FA^+), and methylammonium (MA^+), B denotes divalent metal cations, such as lead (Pb^{2+}) and tin (Sn^{2+}), and X represents halide ions including iodide (I^-), bromide (Br^-), and chloride (Cl^-). When A is a large organic cation that exceeds the size restrictions imposed by the tolerance factor for 3D perovskites, such as phenylethylammonium (PEA^+) or 1-naphthylmethylammonium (NMA^+), the perovskite structure changes, resulting in the formation of a low-dimensional perovskite and subsequently altering its optoelectronic properties.[34,35] ML has been utilized to explore perovskite materials and predict their bandgap and optical properties, offering crucial technical support for the development of perovskite materials.[36] In addition, ML has played a pivotal role in investigating the stability of lead halide perovskites. This has enhanced our understanding of the various factors influencing the stability of perovskite materials, thereby facilitating the development of stable devices.[37] Furthermore, ML has been employed to optimize the crystallization process of perovskites. By fine-tuning synthesis parameters, it enables faster and more controlled crystallization, which is vital for the fabrication of high-quality perovskite crystals.[38]

Over the last 2 years, ML has significantly influenced the development of perovskite optoelectronics, showcasing a trend toward multifaceted collaborations.[39–44] ML is being strategically leveraged to tackle the practical challenges within perovskite optoelectronic devices, including the optimization of perovskite active layers, the selection of transport layers, and the elucidation of underlying functional mechanisms. Interpretable ML methodologies, such as SHapley Additive exPlanations (SHAP) values, are increasingly being applied to analyze and comprehend the factors affecting the performance of perovskite devices. The reliability of ML insights is often validated by harmonizing them with empirical device data and corroborative density functional theory (DFT) calculations. This review highlights the critical role of ML in boosting the efficiency of perovskite optoelectronic devices, encompassing developments, current challenges, and future prospects. It succinctly outlines the essential contribution of ML to the progression of perovskite technologies, with a special focus on solar cells and LEDs. The aim is to encourage interdisciplinary collaboration between perovskite optoelectronics and artificial intelligence, highlighting the significant potential for revolutionary advancements in optoelectronic technologies through this synergistic partnership.

2 Fundamentals of Machine Learning

2.1 Machine Learning Workflows

ML workflows usually include data collection, feature engineering, model selection, and performance evaluation (Fig. 1).[42,45,46] Data collection is the foundation of ML, where the key lies in selecting appropriate data sources and ensuring the accuracy of the data. In the perovskite optoelectronic devices, data typically originate from laboratory measurements, as well as results from
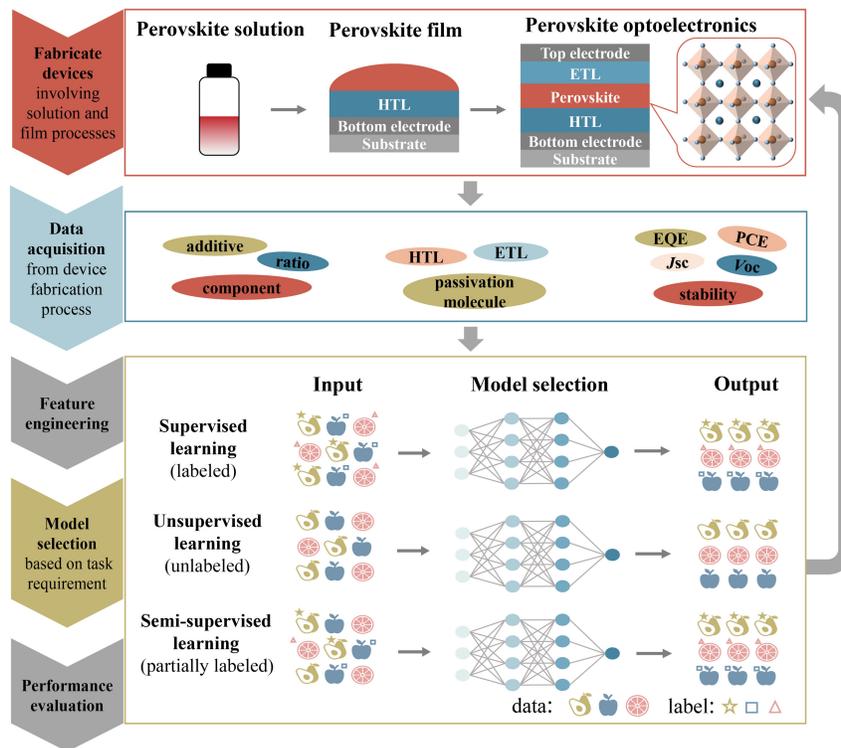
Fig. 1 Flowchart of ML-assisted preparation of perovskite optoelectronic devices. This includes the fabrication of devices, data acquisition, feature engineering, selection of appropriate models, and performance evaluation. The models encompass a range of learning algorithms: supervised, unsupervised, and semi-supervised.

computational simulations. In addition, it is necessary to label data samples according to the type of algorithm selected for the task.

Following data collection, feature engineering is undertaken with the objective of converting raw material data into a format that ML algorithms can interpret and utilize effectively. This process is crucial for uncovering underlying correlations and physicochemical properties of the materials in question. Feature engineering encompasses several key parts, including feature extraction, feature selection, and feature construction. Feature extraction is the process of distilling valuable features from the raw data. Feature selection focuses on identifying the most relevant or representative features, which helps in reducing the dimensionality of the data and enhancing the model's generalization capability. The Pearson correlation coefficient and variance threshold are commonly used methods. The Pearson correlation coefficient measures the linear relationship between pairs of features, filtering out highly correlated ones.[47] The variance threshold method sets a threshold value, retaining only features with variance exceeding this threshold.[48] Lastly, feature construction is about generating novel features through the combination, transformation, or derivation of existing ones, thereby bolstering the model's ability to capture and express complex patterns and relationships.

Subsequently, the selection of ML models is contingent upon factors, such as the complexity of the task at hand and the specific characteristics of the dataset. Tasks are generally classified as either classification or regression types, with data often encompassing a range of formats including numerical values, text, and images. The most suitable ML model is selected tailored to the particular requirements of the task and the data's attributes. Popular models frequently employed in such scenarios include random forests (RFs),[49] support vector machines (SVMs),[50] and diverse variants of neural networks.[51]

Performance evaluation is the process of assessing an ML model's predictive capability and accuracy. By evaluating the performance of different models on a given dataset, the most suitable model can be selected for subsequent training and prediction. The evaluation methods for models vary depending on the task. Common evaluation metrics in classification tasks include precision and accuracy, while in regression tasks, metrics, such as mean squared error (MSE) and mean absolute error (MAE), are frequently used. MSE is the average of the squared differences between predicted and actual values. A smaller MSE indicates lower error and thus more accurate predictions. MAE is the average of the absolute differences between predicted and actual values. Similarly, a smaller MAE signifies lower error, reflecting more accurate predictions.

With the advancement of ML applications, there is a heightened emphasis on deciphering the reasoning behind model decisions. This trend is highlighted by the increasing adoption of SHAP values, a sophisticated method for interpreting ML model predictions.[52] At the heart of SHAP is the concept of each feature's marginal contribution to the model's prediction, where "marginal contribution" signifies the impact of an individual feature on the predictive result of the model. SHAP endows the capability for both overarching (global) and specific (local) explanations of model behaviors. Global explanations shed light on the significance of each feature across the entire model, whereas local explanations delve into the rationale behind the model's specific predictions for individual instances. In practical scenarios, the explanatory insights of SHAP are graphically represented, predominantly through SHAP plots. These plots intuitively exhibit the positive or negative influence of each feature on a particular instance. By analyzing SHAP plots, one can gain insights into which features are pivotal in driving specific predictive outcomes.

2.2 Machine Learning Algorithms

ML algorithms cover a spectrum of approaches including supervised, unsupervised, and semi-supervised learning techniques (Fig. 1). Supervised learning is an ML method that involves training a model using labeled data sample.[53–55] The objective of this approach is for the model to learn the mapping between input features and output labels, enabling it to predict or classify new input features. In classification tasks, the model's output is categorical, effectively segregating the data into defined groups.[56–59] Conversely, in regression tasks, the output is continuous, yielding a numerical prediction for the input data.[60] Interestingly, classification and regression can be amalgamated to tackle certain types of problems.[61] Among the common algorithms employed in supervised learning is the decision tree algorithm. This algorithm creates a tree-like model by systematically dividing the data based on its features, with each terminal node (or leaf) representing a specific category or value.[62] Another notable algorithm is the SVM, which strives to maximize the margin between different categories. It achieves this by locating an optimal hyperplane within the feature space. Lastly, gradient boosting stands out as an ensemble learning method. It enhances the performance of a model by iteratively developing a "weak" classifier and continuously refining its predictive accuracy. This refinement process involves minimizing the gradient of the loss function, thereby progressively building a more robust classifier.[63]

Unlike supervised learning, unsupervised learning is an ML method that focuses on uncovering patterns, structures, or relationships in unlabeled data.[64–67] This method operates without predefined labels or output values, focusing solely on the input data. Unsupervised learning in the field of material science mainly performs tasks, such as data clustering, dimensionality reduction, or using natural language processing to retrieve materials from literature by discovering the internal structure and pattern of data.[68,69] Prominent algorithms in unsupervised learning include Gaussian mixture models (GMMs).[70] These models are used for data clustering and modeling, based on the premise that the data consist of multiple Gaussian distributions. GMMs work by estimating parameters to find the most accurate fit to these distributions. Another notable algorithm is self-organizing maps (SOMs), a type of neural network algorithm.[71] SOMs organize similar input samples into adjacent neurons in a self-regulating manner, proving useful for data visualization and clustering. This approach is particularly effective for mapping high-dimensional data into a more manageable, low-dimensional space.

Semi-supervised learning is an ML method that combines supervised learning and unsupervised learning. It leverages a combination of labeled and unlabeled data during the training process.[72–75] By utilizing unlabeled data alongside labeled data, semi-supervised learning aims to enhance the generalization ability and performance of the model.[75] A notable example of semi-supervised learning is the self-training method.[76] This technique begins by training an initial model using labeled data. Once trained, this model is then applied to make predictions on

unlabeled data. Samples for which the model has high prediction confidence are added to the labeled dataset. The model is then retrained with this augmented data, and this iterative process continues until the model reaches a point of convergence. Another approach within semi-supervised learning is co-training.[77] This method involves training multiple classifiers simultaneously, with each classifier based on the assumption that different perspectives or feature sets exist within the data, and each provides independent and valuable information. In co-training, each classifier is trained using features from one specific view. The classifiers then use the predictions from one view to aid in the filtering and labeling of the unlabeled data from another view.

3 Machine Learning in Perovskite Optoelectronic Devices

Fabrication of high-efficiency perovskite optoelectronic devices relies on various factors, including the composition and structure of the perovskite materials, device architecture, and fabrication methodologies. Conventional empirical (trial-and-error) methods have demonstrated limited efficacy in optimizing these dimensions. In recent years, ML has emerged as a powerful tool for leveraging existing data to uncover intrinsic patterns, thereby facilitating the realization of high-performance perovskite optoelectronics. In this context, we primarily focus on the application-centric research of ML within perovskite photovoltaics and PeLEDs.

3.1 Perovskite Solar Cells

PSCs represent a promising photovoltaic technology due to their superior optical absorption properties and charge transport capabilities. These PSCs feature distinct layers, including a light-absorbing layer, electron transport layer (ETL), hole transport layer (HTL), anode, and cathode. The light-absorbing perovskite material plays a pivotal role in efficiently converting light energy into electrical energy. To enhance the stability and power conversion efficiency (PCE) of PSCs, ML has already been applied to various aspects, such as the device structure, perovskite layer, transport layer, and interface engineering. This innovative application offers critical insights, expedites the optimization processes, and broadens the comprehension of PSC mechanisms. It holds substantial potential for further enhancing both the stability and efficiency of PSCs, marking a significant advancement in renewable energy technologies.

3.1.1 Optimizing surface capping layer

The degradation of perovskite films directly impacts the lifetime of perovskite photovoltaic devices. As a result, surface treatment of perovskites to form a low-dimensional perovskite capping layer has emerged as an effective strategy to enhance the stability of these photovoltaic devices. Nevertheless, a significant challenge persists in identifying appropriate materials for surface modification. Recently, ML has garnered significant attention in the development of passivation materials for perovskite photovoltaics.[78–81] The primary emphasis has been on the prediction and design of materials for passivation. This involves compiling a database of passivation materials from literature and then employing ML for predictions, with the use of SHAP to interpret the importance of features. In addition, there is a significant trend in comparing the results of ML predictions with actual experimental outcomes in devices, as well as with

insights from DFT calculations. This comparison serves to further substantiate the reliability of ML applications.

Hartono et al. developed an ML method to select overlay materials to suppress perovskite degradation.[78] They collected 21 organic salts from the PubChem database as overlay materials, using the structure and chemical characteristics of organic molecules as well as device fabrication conditions as model inputs, while using the initial degradation values and degradation rates of perovskites as model outputs [Fig. 2(a)]. Regression models [Fig. 2(b)] and SHAP value analysis [Fig. 2(c)] were employed to analyze the relationship between features and stability. It was found that a smaller number of hydrogen bond donors and a smaller topological polar surface area in organic molecules are associated with increased stability of the $MAPbI_3$ films. The best-performing organic halide, phenethyltriethylammonium iodide (PTEAI), successfully extended the stability lifetime of $MAPbI_3$ by 4 ± 2 times compared to control $MAPbI_3$. Moreover, characterization techniques (X-ray photoelectron spectroscopy and Fourier-transform infrared spectroscopy) have also revealed that PTEAI effectively forms a capping layer by modifying the surface structure and chemistry. This alteration correlates with the suppression of methylammonium loss and the formation of both PbI_2 and oxygen-containing compounds at the surface of the perovskite, ultimately achieving the goal of protecting the perovskite from degradation.[82]

Recently, Zhi et al. collected a dataset of 46 experimental data points documenting the improvement in power conversion efficiency after interface passivation with 19 different ammonium salts at various concentrations.[79] These ammonium salts encompassed 14 from the initial dataset and five additional data selected via Latin hypercube sampling (LHS). LHS achieves diversity and uniformity sampling by dividing the range of values for each parameter into equally spaced subintervals and ensuring that there is only one sample point within each subinterval. The model inputs were 12 molecular descriptor features and the concentration of the ammonium salt precursor solution. The model output was the rate of improvement in device PCE. It is noteworthy that, to enhance precision, they employed an ensemble learning method that integrates five distinct ML regression models with various architectures to train and make predictions on the dataset. Similar to the previous work, SHAP analysis was also utilized to investigate the impact of molecular features on device efficiency, identifying hydrogen bond donors, hydrogen atoms, and the molecular lipid−water partition coefficient (molecular LogP or MolLogP) as the three most critical molecular features for the passivating ammonium salts. Ultimately, they employed the trained ML model to screen 10 suitable ammonium salts from 112 candidates in the PubChem database and experimentally validated six different ammonium salts. Among them, 2-phenylprop-1-ammonium iodide achieved PCEs of 22.36% and 24.47% for FAMACs and FAMA-based PSCs, respectively.

3.1.2 Degradation mechanism

Despite numerous strategies having been employed to enhance the stability and power conversion efficiency of perovskite photovoltaics, the factors affecting stability and PCE remain elusive. Hu et al. constructed an ML regression model to investigate the key factors influencing the efficiency and stability of PSCs (Fig. 3).[83] They selected 102 sets of PSC device data from published literature. Through feature engineering, they retained key parameters: Eg (bandgap), size (the grain size of perovskite films, inversely representing grain boundary density), Rn (surface
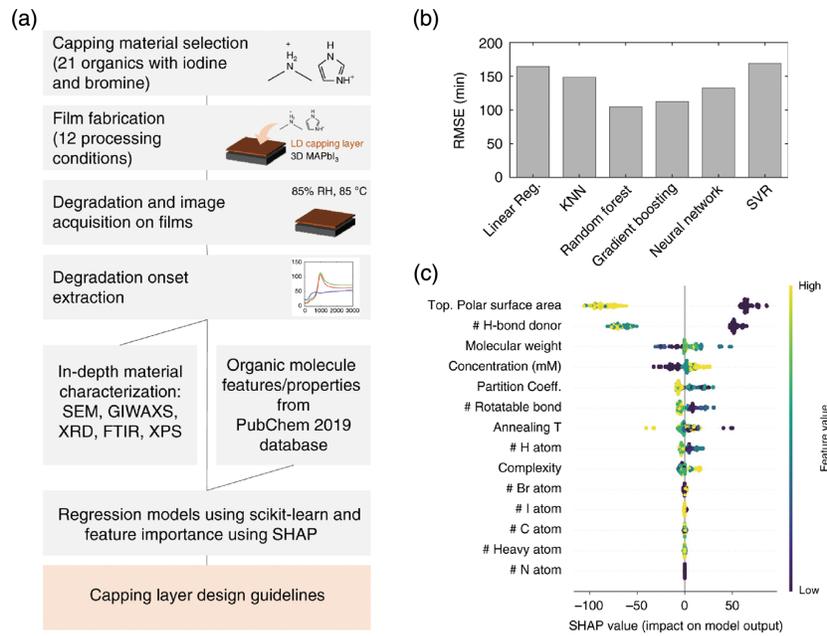
Fig. 2 (a) Schematic illustration of design rules for capping layer of PSCs. (b) The cross-validated root mean square error of various machine-learning models. (c) The feature importance ranking obtained from the RF regression algorithm and SHAP library, showing the chemical properties and processing conditions in descending order of importance (rank). The yellow and purple color indicates high and low values of a given feature, respectively. Reproduced with permission from Ref. 78 (CC-BY).
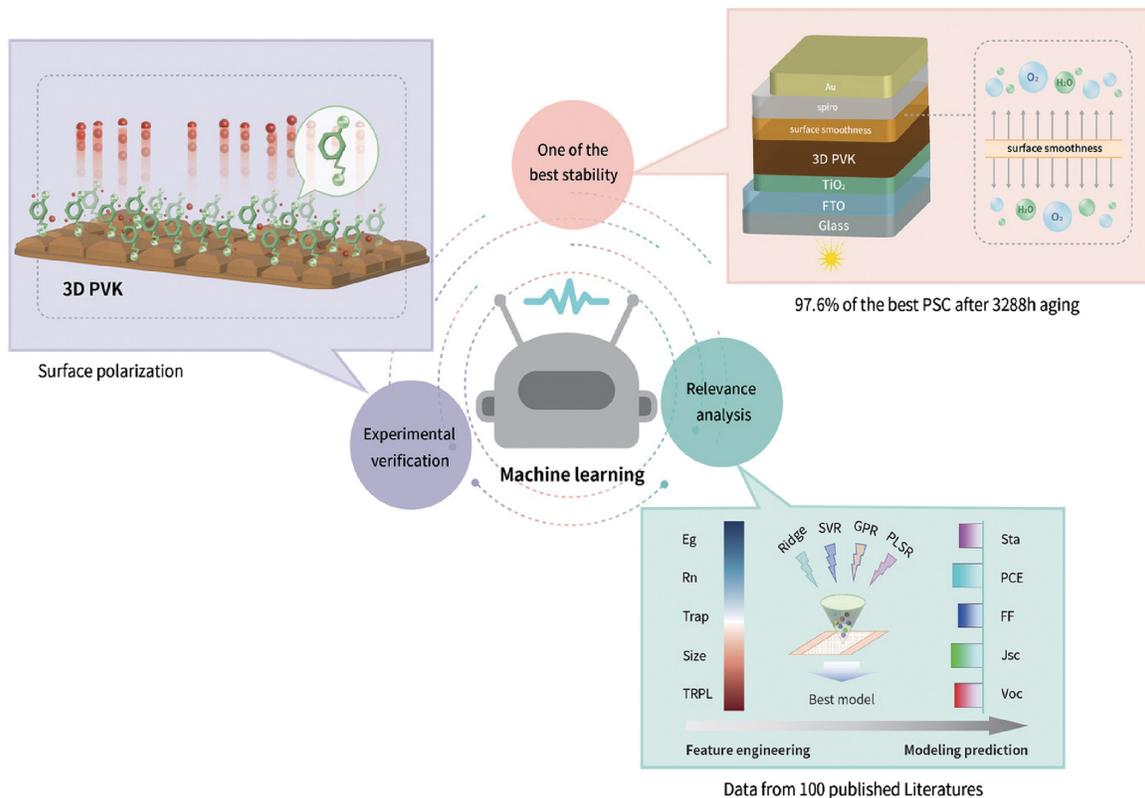
Fig. 3 Working principle of ML: relevance analysis—experimental verification—one with the best properties. Complex multivariable analysis by ML. PVK is perovskite. Reproduced with permission from Ref. 83 © 2022 Wiley-VCH.

roughness at the interface between the perovskite film and the HTL), Trap (trap density), and TRPL (average fluorescence lifetime from time-resolved photoluminescence measurements) as model inputs. These were utilized with the aim of predicting two critical outcomes: power conversion efficiency and stability. Several ML regression models containing support vector regression (SVR) and ridge regression were employed to discern the relationships between model inputs and outputs. Among these, SVR demonstrated the best performance. Moreover, it was found that the bandgap had the greatest impact on efficiency, while surface roughness and grain size predominantly influenced stability. Finally, guided by predictive insights, various annealing temperatures were strategically employed to modify grain sizes, and three unique organic molecules were utilized to change surface roughness. This approach yielded PSCs with an impressive 23.4% efficiency and exceptional long-term stability, maintaining 97.6% of their initial efficiency after an extended period of 3288 h under ambient conditions. It is noteworthy that the results from DFT calculations also show a remarkable correlation with the predictions of the ML model.

3.1.3 Optimizing electron transport layer

An ETL that matches the energy levels of the perovskite layer and exhibits high electron mobility is fundamental for ensuring efficient electron injection and extraction. In addition, it can impact the density of defect states at interfaces between the perovskite and the ETL, thereby preventing detrimental interfacial recombination. As a result, an appropriate ETL is essential for the performance of perovskite photovoltaic devices. This indicates the need not only to choose suitable ETL materials but also to perform surface modification or doping of the selected layer, to ensure efficient electron transport, and to facilitate perovskite crystal growth.

She et al. utilized a two-step ML approach, involving classification and regression, to study suitable ETLs (Fig. 4).[80] Initially, they extracted 1820 performance data points of PSCs from 795 articles published between 2013 and 2020, forming the first dataset. The data were then categorized based on the materials of the ETLs, dividing them into groups, such as TiO_2-based, SnO_2-based, and other ETL-based groups. In this dataset, the ETLs were undoped. To prevent sample imbalance, the samples were further classified into three categories based on their efficiency: category A for efficiency below 9%, category B for efficiency ranging from 9% to 18%, and category C for efficiency above 18%. By selecting nine features that influence the power conversion efficiency through feature engineering, they used these as model inputs with power conversion efficiency as the output, training different ML
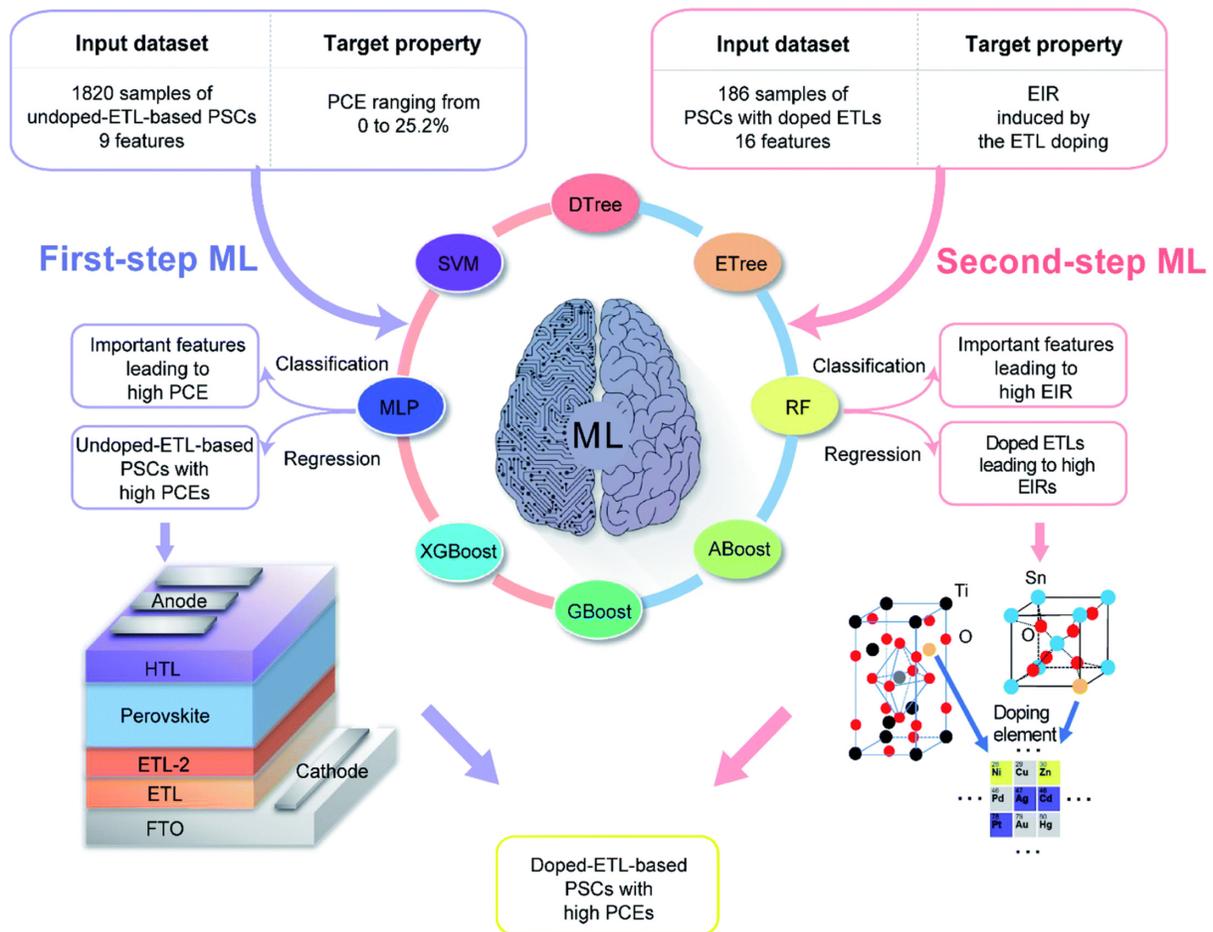
Fig. 4 Flow chart of the two-step ML in this work, which includes two steps, first-step ML (left side) and second-step ML (right side). Reproduced with permission from Ref. 80 © 2021 Royal Society of Chemistry.

classification models on this dataset. Ultimately, they further investigated feature importance based on the classification model, finding that TiO_2 or SnO_2 as ETL materials could lead to high-performance PSCs.

Subsequently, to further improve the PCE, they focused on the impact of doping ETLs. A second dataset was created, comprising 90 device data points for PSCs based on doped SnO_2 and 96 for PSCs based on doped TiO_2. The physical and chemical properties and concentrations of the doping elements were used as model input features. They calculated the efficiency improvement ratio (EIR), which is the PCE of PSCs with doped ETL divided by the PCE of corresponding PSCs with undoped ETL, and used EIR as the model output. Using the RF model and genetic algorithms (GAs) for data training, they predicted doped ETLs that might lead to high EIR. Finally, based on predictions, they achieved PCEs of up to 30.47% and 28.54% for Cs-doped TiO_2 ETL in CsFAMA-based PSCs and S-doped SnO_2 ETL in FAMA-based PSCs, respectively.

3.1.4 Optimizing lead-free perovskite

In order to develop environmentally friendly lead-free PSCs, ML is utilized to optimize the device architecture and components of Sn-based PSCs. Bak et al. initially collected real experimental data from 122 different photovoltaic devices with the perovskite structure $ASnI_3$ from 49 published journal articles. The data include information on perovskite chemical structure, metal electrodes, transparent electrodes, HTL, and ETL as input features, while the corresponding short-circuit current density (J_{SC}), open-circuit voltage (V_{OC}), and fill factor (FF) values are considered as output features. A deep neural network (DNN) algorithm was then developed, and data augmentation was performed using K-fold cross-validation to randomly automate hyperparameter optimization of the DNN model to obtain

the model's pre-trained weights. The performance of the model was assessed by comparing the experimental data's PCE values with the DNN's predicted values for each layer (i.e., perovskite layer, metal electrode, transparent electrode, HTL, and ETL). Finally, by random combination of a large number of input features, predictions were made through the optimized DNN model (Fig. 5). ML recommended a new Sn-based PSC device structure FTO/PEDOT:PSS/$EDA_{0.01}PEA_{0.07}Cs_{0.03}FA_{0.51}MA_{0.38}$ SnI_3/indene-C60 bisadduct (ICBA)/2,9-dimethyl-4,7-diphenyl-1,10-phenanthrolin (BCP)/Ag. Obviously, the composition of the perovskite is still extremely complex. Furthermore, experimental verification has confirmed that the PCE of the ML-optimized Sn-based perovskite photovoltaics reached 5.57%, which is a significant improvement over the traditional testing results (PCE of 1.72%).[84]

3.1.5 Vapor-deposited perovskite solar cells

Although the performance of solution-processed perovskite photovoltaics has rapidly improved in laboratory settings, scaling up to large-scale production remains a significant challenge. Vapor-deposited PSCs, which need to be compatible with the existing large-scale electronic industry, have substantially lagged behind. Wang et al. extracted information on 220 devices from 150 published papers on thermal evaporation PSCs between 2015 and 2022. Initially, the data were cleansed, reducing the sample size to 180 groups, and missing values in the dataset were filled with the mode of each column. The dataset was then expanded using Gaussian noise, tripling the sample size. Twenty features were chosen as input variables, encompassing aspects such as the material composition of the perovskite layer and device fabrication methods, with the device's J_{SC}, V_{OC}, FF, and PCE as output variables. Training was conducted using 10 different ML models, with the RF model scoring highest on the test
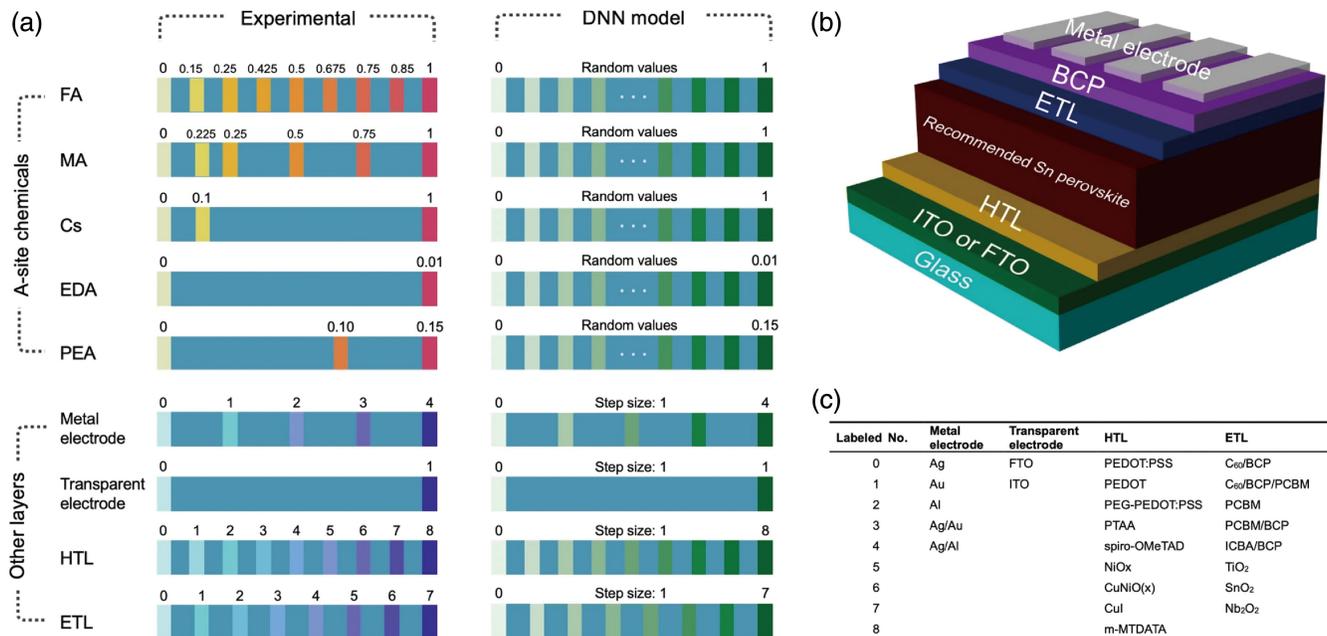
Fig. 5 Schematics of Sn-PSC design ML recommendation. (a) Parameter setting of each feature for the recommendation process. (b) Schematic of recommended Sn PSCs. (c) Labeled number and corresponding materials used as other layers in (a). Reproduced with permission from Ref. 84 (CC-BY).

set. To enhance the generalization ability of the ML model, it was combined with a GA, increasing the prediction accuracy of the GA-enhanced RF model on the test set from 60.2% to 67.4%. Subsequently, the importance of features influencing PCE was analyzed through the mean Gini index of feature contributions in each decision tree of the RF model. The study found that the five features contributing most to the PCE were the proportions of MA, FA, Br, annealing temperature, and pressure (Fig. 6). Finally, the optimal device structure proposed was indium tin oxide (ITO)/SnO_2/BCP/$FA_{0.25}MA_{0.25}Cs_{0.5}$ $Pb(I_{0.75}Br_{0.25})_3$/MeO-2PACz/Ag, achieving the highest PCE of 26.1% under conditions of 90°C to 105°C and an evaporation pressure of 3×10^{-5} to 4×10^{-5} mbar. This work not only identified the optimal conditions for vapor-deposited perovskite but also, more crucially, delineated the direction for vapor-deposited perovskite.[85]

3.1.6 Machine vision for large-area slot-die coated films

In the commercialization of perovskite photovoltaics, the controllable fabrication of large-area homogeneous perovskite films represents a critical technological challenge. Traditionally, the assessment of uniformity in large-area films has relied primarily on visual inspection, which is imprecise. Taherimakhsousi et al. developed a rapid, reliable, and non-destructive machine vision approach to quantify the uniformity of large-area perovskite films, thereby aiding process optimization.[86] Initially, white light photography was employed to sample slot-die coated perovskite films. Subsequently, a convolutional neural network based on the VGG16 architecture successfully segmented the original images into fully covered,[87] partially covered, and uncovered regions without the need for manual intervention. Optimized for pixel resolution, this approach spatially quantified multiple attributes of perovskite films (such as substrate coverage, film thickness, and defect density) from 25 cm^2 samples and established correlations between slot-die coating process parameters (wet film thickness and gas knife speed) and film properties (Fig. 7). This strategy facilitated multi-parameter, multi-objective optimization of the gas-knife assisted slot-die coating process, enabling the identification of optimal process conditions that simultaneously maximize coating throughput, film quality, and predicted device current density while reducing the characterization effort required for each process condition. This work demonstrates the significant potential of machine vision in optimizing large-area photovoltaic films.

3.2 Perovskite Light-Emitting Diodes

Perovskite LEDs are semiconductor devices that transform electrical energy into light, featuring an optically active layer constituted by perovskite materials. Their primary performance indicators include EQE, calculated as the ratio of photons emitted from the LED to the electrons circulating in the external circuit, and half-life (T_{50}), representing the duration for brightness (or efficiency) to decrease by half.

3.2.1 Predicting additive molecules

The quality of the perovskite active layer is crucial for the performance of perovskite LEDs. Currently, enhancing the quality of perovskite films involves additive engineering, which includes introducing organic molecules into the perovskite
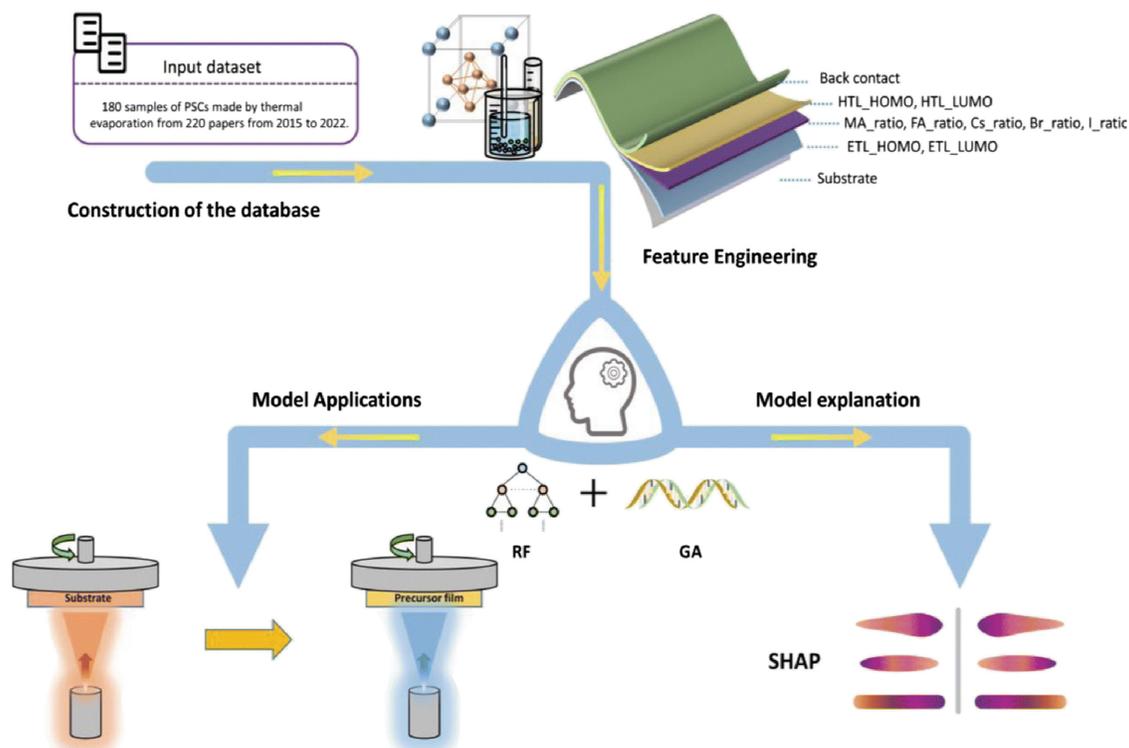
Fig. 6 ML flowchart for studying the variables that influence the performance of vapor-deposited PSCs, consisting of four main parts: building the dataset, feature engineering, SHAP interpretation model, and model prediction. Reproduced with permission from Ref. 85 © 2023 Royal Society of Chemistry.
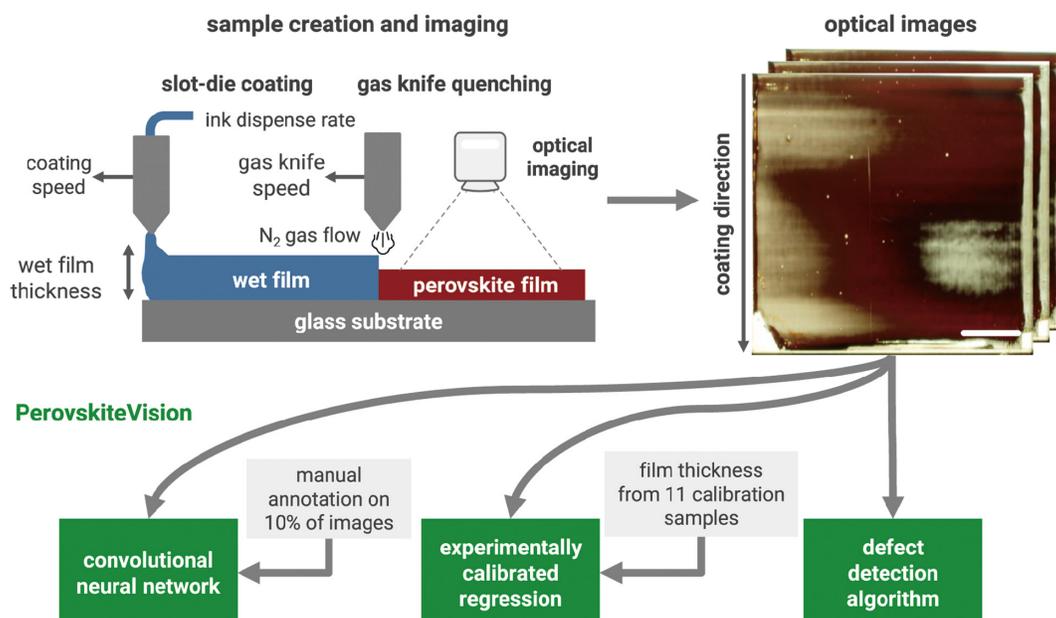
Fig. 7 Machine vision workflow for quantifying large-area perovskite film morphology from optical images using the PerovskiteVision tool. Reproduced with permission from Ref. 86 (CC-BY).

precursor solution. These molecules, while not integrating into the perovskite lattice, improve film quality by tuning the crystallization pathway. Yet, not every organic molecule is a suitable additive, and unsuitable choices can impair LED performance. Traditional additive selection based on trial-and-error is inefficient and challenged by the small size of perovskite LED additive databases for conventional ML. To address this, the enhanced molecular information model (EMIM) was developed, incorporating both qualitative molecular fingerprints and quantitative descriptors to enhance molecular information and improve predictive accuracy (Fig. 8).[39] It is noteworthy that in this work, the additive molecular database used was validated

through experimental lab device tests. By utilizing EQE as the output feature and additive molecular descriptors as inputs, along with a 10.3% EQE threshold for classification, the EMIM model attained an impressive 98% prediction accuracy, significantly outperforming traditional ML models. In addition, among the 12 chosen additives, only one showed inconsistent results between EMIM predictions and device fabrication outcomes, demonstrating EMIM's reliability. One of the additives even elevated the EQE of perovskite LEDs to 22.7%, making it one of the most efficient near-infrared LEDs. This work is pioneering in using ML for analyzing perovskite LEDs, especially for small data additives.
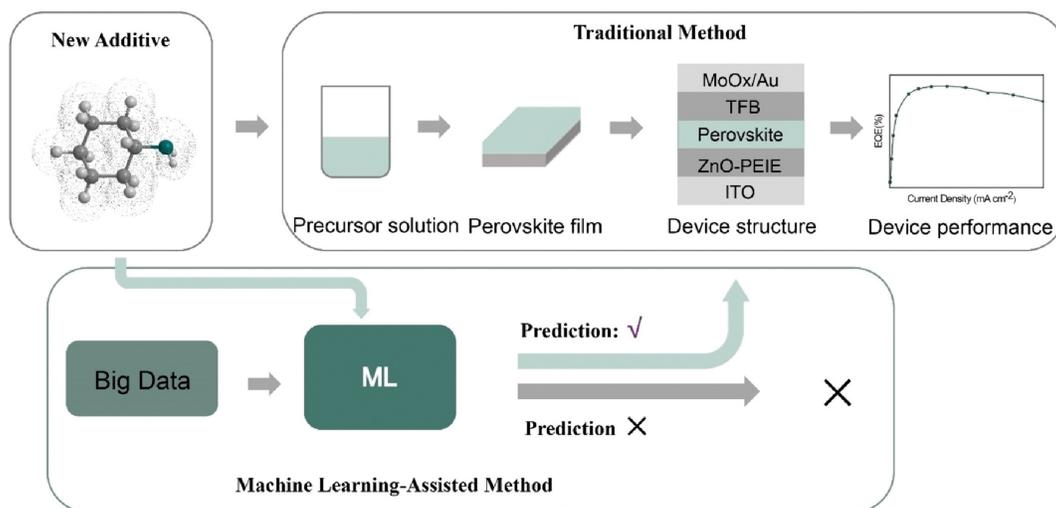
Fig. 8 Traditional method versus ML-assisted method. The traditional method is to verify the effectiveness of the additive for PeLEDs, while the ML-assisted method is to fabricate the device with the "good" additive predicted by ML. The additive is added into the perovskite precursor solution to form the perovskite film. Reproduced with permission from Ref. 39 © 2022 Wiley-VCH.
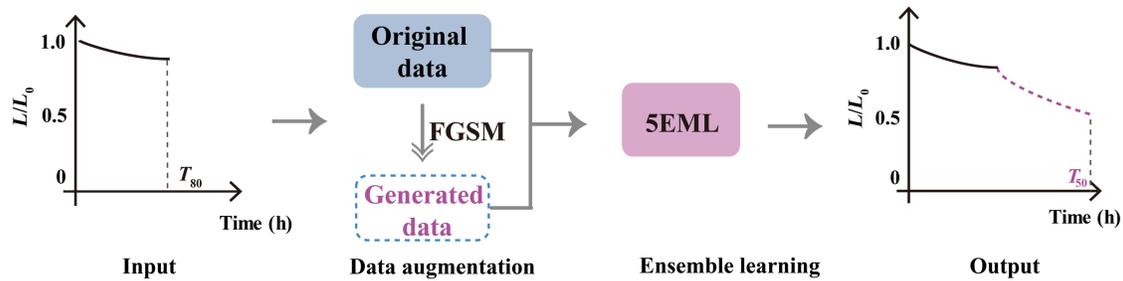
Fig. 9 Illustration of lifetime prediction by 5EML. Reproduced with permission from Ref. 40 © 2024 Wiley-VCH.

3.2.2 Half-lifetime prediction

Stability is a key objective in the development of perovskite LEDs. Stability measurements, unlike EQE measurements, are time-consuming, necessitating hundreds of hours for the most stable perovskite LEDs. Therefore, developing strategies to predict the half-life of perovskite LEDs is crucial. Recently, we found that traditional exponential fitting methods, such as monoexponential, biexponential, and stretched exponential functions, are inadequate for predicting half-life due to complex degradation mechanisms involving ion migration, Joule heating, and interfacial effects.[40] To address this, an ensemble learning model was developed to predict the T_{50} lifetime (time to 50% of initial performance) based on shorter tests T_{80} (time to 80% of initial performance) (Fig. 9). A database of 210 near-infrared and red perovskite LEDs was initially compiled from experimental data. To augment this dataset, a DNN combined with the fast gradient sign method was employed, effectively doubling the dataset to 420 samples. Comparative analysis of five different ML models, such as least absolute shrinkage and selection operator, SVR, gradient boosting regressor, Gaussian process regressor, and elastic net (EN), showed improved accuracy over the original data. An ensemble learning model (5EML) based on these five algorithms was then used for better fitting the degradation curves of perovskite LEDs, achieving a high score of 0.995. This model also applies to quantum dot LEDs, demonstrating its universality.

4 Conclusion and Perspectives

4.1 Conclusion

It is evident that the application of ML in the perovskite field has undergone a significant transformation. Initially focused primarily on predicting new perovskite materials, the emphasis has now shifted toward perovskite optoelectronic devices, including device design, crystallization regulation, stability analysis, and manufacturing processes. More importantly, in the fields of perovskite photovoltaics and perovskite LEDs, the focus of ML research differs. Specifically, in perovskite photovoltaics, ML is primarily used for interface engineering of the perovskite layer and optimization of the transport layers, whereas in perovskite LEDs, it mainly focuses on additive screening and stability prediction. This difference essentially arises from the distinct optimization strategies for these two types of devices.[88] Perovskite LEDs are extremely sensitive to defects, so the key is to suppress defects and enhance photoluminescence quantum efficiency. In contrast, perovskite photovoltaics are less sensitive to defects, with the primary focus being on charge carrier transport.

Certainly, the application of ML in perovskite optoelectronics has the following characteristics. First, the primary source of data for perovskite optoelectronic devices is derived from experimental results, ensuring their authenticity and dependability. Second, this research area strongly emphasizes the veracity of ML inferences. ML outcomes are typically corroborated through actual device fabrication or supplemented with DFT calculations. Third, the SHAP method is widely employed to investigate the attributes of passivation materials, offering critical insights into feature importance. This method serves as a theoretical compass for material design. Fourth, current algorithmic approaches are largely focused on supervised learning, including classification and regression models.

4.2 Perspectives and Challenges

Despite the notable progress and significant research potential demonstrated by ML in perovskite optoelectronics, the overall volume of research remains limited, and there is a substantial need for further in-depth exploration. The development of perovskite optoelectronic devices is in a relatively nascent stage, and a considerable amount of experimental data is yet to be made publicly available, which poses challenges for the application of ML techniques. Moreover, the optimization process of perovskite optoelectronic devices is exceedingly complex and involves a multitude of details. Slight variances during the device fabrication process can lead to significantly divergent experimental outcomes, making it difficult to assess the reliability of data samples. Therefore, regarding the intersection of ML and perovskite optoelectronic devices, we propose the following outlooks and suggestions.

Increasing the open-sourcing of perovskite data. The open-sourcing of algorithms has become increasingly common, enabling researchers in the perovskite field to effectively utilize these models. However, the open-sourcing of perovskite data remains relatively limited. Therefore, it is recommended to establish a platform for perovskite data sharing. In addition, an abundance of data samples can help address concerns about the reliability of experimental data. It is important to note that not all collected experimental data are used for ML. Instead, data samples that have undergone preprocessing and feature engineering are utilized. In other words, the robust data processing capabilities of ML can mitigate minor biases introduced during experimental data collection.

To tackle the challenge of limited datasets, researchers are encouraged to delve into data augmentation methodologies that

utilize the physical and chemical properties of perovskite molecules as integral feature descriptors in ML models. This strategy facilitates the extraction of enhanced molecular insights while preserving the intricate structural details. Recent breakthroughs in graph neural networks, specifically in molecular generation, have demonstrated significant potential in broadening the scope of datasets and elevating their quality.[89]

Strengthen the depth of interdisciplinary research between ML and perovskites. First, the interpretability of ML models is especially critical, as it reveals the decision criteria used by the models, thereby guiding the understanding and optimization strategies for perovskite optoelectronic devices. Second, enhance the application of machine image processing in analyzing perovskite film morphology. For LEDs, a discrete morphology is desirable as it helps improve light extraction efficiency. In contrast, the photovoltaic often prefers large and dense grains to enhance carrier transport. Currently, the study of perovskite film morphology primarily relies on empirical deductions, which fail to accurately analyze and extract underlying morphological patterns.

Code and Data Availability

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

References

1. G. Xing et al., "Long-range balanced electron-and hole-transport lengths in organic-inorganic $CH_3NH_3PbI_3$," *Science* **342**(6156), 344–347 (2013).
2. H. Tsai et al., "High-efficiency two-dimensional Ruddlesden–Popper perovskite solar cells," *Nature* **536**(7616), 312–316 (2016).
3. N. T. P. Hartono et al., "The effect of structural dimensionality on carrier mobility in lead-halide perovskites," *J. Mater. Chem. A* **7**(41), 23949–23957 (2019).
4. S. D. Stranks et al., "Electron-hole diffusion lengths exceeding 1 micrometer in an organometal trihalide perovskite absorber," *Science* **342**(6156), 341–344 (2013).
5. M. Liu, M. B. Johnston, and H. J. Snaith, "Efficient planar heterojunction perovskite solar cells by vapour deposition," *Nature* **501**(7467), 395–398 (2013).
6. M. M. Lee et al., "Efficient hybrid solar cells based on meso-super-structured organometal halide perovskites," *Science* **338**(6107), 643–647 (2012).
7. H.-S. Kim et al., "Lead iodide perovskite sensitized all-solid-state submicron thin film mesoscopic solar cell with efficiency exceeding 9%," *Sci. Rep.* **2**(1), 591 (2012).
8. Z.-K. Tan et al., "Bright light-emitting diodes based on organometal halide perovskite," *Nat. Nanotechnol.* **9**(9), 687–692 (2014).
9. L. C. Schmidt et al., "Nontemplate synthesis of $CH_3NH_3PbBr_3$ perovskite nanoparticles," *J. Am. Chem. Soc.* **136**(3), 850–853 (2014).
10. C. Motta et al., "Revealing the role of organic cations in hybrid halide perovskite $CH_3NH_3PbI_3$," *Nat. Commun.* **6**(1), 7026 (2015).
11. Z. Huang et al., "Anion–π interactions suppress phase impurities in $FAPbI_3$ solar cells," *Nature* **623**(7987), 531–537 (2023).
12. M. A. Green et al., "Solar cell efficiency tables (version 63)," *Prog. Photovoltaics Res. Appl.* **32**(1), 3–13 (2024).
13. H. Chen et al., "Improved charge extraction in inverted perovskite solar cells with dual-site-binding ligands," *Science* **384**(6692), 189–193 (2024).
14. J. Zhou et al., "Highly efficient and stable perovskite solar cells via a multifunctional hole transporting material," *Joule* **8**(6), 1691–1706 (2024).
15. M. Li et al., "Acceleration of radiative recombination for efficient perovskite LEDs," *Nature* **630**(8017), 631–635 (2024).
16. Y. Cao et al., "Perovskite light-emitting diodes based on spontaneously formed submicrometre-scale structures," *Nature* **562**(7726), 249–253 (2018).
17. Y. He et al., "$CsPbBr_3$ perovskite detectors with 1.4% energy resolution for high-energy γ-rays," *Nat. Photonics* **15**(1), 36–42 (2021).
18. J. Jiang et al., "Red perovskite light-emitting diodes with efficiency exceeding 25% realized by co-spacer cations," *Adv. Mater.* **34**(36), 2204460 (2022).
19. J. S. Kim et al., "Ultra-bright, efficient and stable perovskite light-emitting diodes," *Nature* **611**(7937), 688–694 (2022).
20. Y. Jiang et al., "Synthesis-on-substrate of quantum dot solids," *Nature* **612**(7941), 679–684 (2022).
21. Y. Sun et al., "Bright and stable perovskite light-emitting diodes in the near-infrared range," *Nature* **615**(7954), 830–835 (2023).
22. W. Bai et al., "Perovskite light-emitting diodes with an external quantum efficiency exceeding 30%," *Adv. Mater.* **35**(39), 2302283 (2023).
23. S. Yuan et al., "Efficient blue electroluminescence from reduced-dimensional perovskites," *Nat. Photonics* **18**(5), 425–431 (2024).
24. L. Kong et al., "Fabrication of red-emitting perovskite LEDs by stabilizing their octahedral structure," *Nature* **631**, 73–79 (2024).
25. Y. Feng et al., "Nucleophilic reaction-enabled chloride modification on $CsPbI_3$ quantum dots for pure red light-emitting diodes with efficiency exceeding 26%," *Angew. Chem. Int. Ed.* **63**(11), e202318777 (2024).
26. J. Jiang et al., "Efficient pure-red perovskite light-emitting diodes with strong passivation via ultrasmall-sized molecules," *Sci. Adv.* **10**(18), eadn5683 (2024).
27. H. Zhu et al., "Lead halide perovskite nanowire lasers with low lasing thresholds and high quality factors," *Nat. Mater.* **14**(6), 636–642 (2015).
28. C. Bao et al., "Bidirectional optical signal transmission between two identical devices using perovskite diodes," *Nat. Electron.* **3**(3), 156–164 (2020).
29. K. Sakhatskyi et al., "Stable perovskite single-crystal X-ray imaging detectors with single-photon sensitivity," *Nat. Photonics* **17**(6), 510–517 (2023).
30. K. Wen et al., "Continuous-wave lasing in perovskite LEDs with an integrated distributed feedback resonator," *Adv. Mater.* **35**(44), 2303144 (2023).
31. L. Gu et al., "High Q-factor and low threshold continuous-wave near-infrared lasing with quasi-2D perovskites," *Adv. Funct. Mater.* **33**(44), 2303900 (2023).
32. C. Qin et al., "Stable room-temperature continuous-wave lasing in quasi-2D perovskite films," *Nature* **585**(7823), 53–57 (2020).
33. Q. Xu et al., "Rationalizing perovskite data for machine learning and materials design," *J. Phys. Chem. Lett.* **9**(24), 6948–6954 (2018).
34. N. Wang et al., "Perovskite light-emitting diodes based on solution-processed self-organized multiple quantum wells," *Nat. Photonics* **10**(11), 699–704 (2016).
35. L. Zhu et al., "Large organic cations in quasi-2D perovskites for high-performance light-emitting diodes," *J. Phys. Chem. Lett.* **11**(20), 8502–8510 (2020).
36. V. Gladkikh et al., "Machine learning for predicting the band gaps of ABX_3 perovskites from elemental properties," *J. Phys. Chem. C* **124**(16), 8905–8918 (2020).

37. K. Higgins et al., "Chemical robotics enabled exploration of stability in multicomponent lead halide perovskites via machine learning," *ACS Energy Lett.* **5**(11), 3426–3436 (2020).

38. J. Kirman et al., "Machine-learning-accelerated perovskite crystallization," *Matter* **2**(4), 938–947 (2020).

39. L. Zhang et al., "Deep learning for additive screening in perovskite light-emitting diodes," *Angew. Chem. Int. Ed.* **61**(37), e202209337 (2022).

40. L. Zhang et al., "Prediction of operational lifetime of perovskite light emitting diodes by machine learning," *Adv. Intell. Syst.* **6**(6), 2300772 (2024).

41. N. K. Bansal et al., "Machine learning in perovskite solar cells: recent developments and future perspectives," *Energy Tech.* **11**(12), 2300735 (2023).

42. Y. Liu et al., "Machine learning for perovskite solar cells and component materials: key technologies and prospects," *Adv. Funct. Mater.* **33**(17), 2214271 (2023).

43. T. Liu et al., "Machine-learning accelerating the development of perovskite photovoltaics," *Sol. RRL* **7**(23), 2300650 (2023).

44. T. J. Jacobsson et al., "An open-access database and analysis tool for perovskite solar cells based on the FAIR data principles," *Nat. Energy* **7**(1), 107–115 (2022).

45. Q. Tao et al., "Machine learning for perovskite materials design and discovery," *NPJ Comput. Mater.* **7**(1), 23 (2021).

46. M. Srivastava et al., "Machine learning roadmap for perovskite photovoltaics," *J. Phys. Chem. Lett.* **12**(32), 7866–7877 (2021).

47. J. Benesty et al., "Pearson correlation coefficient," in *Noise Reduction in Speech Processing*, J. Benesty and W. Kellermann, Eds., Vol. **2**, pp. 1–4, Springer Berlin Heidelberg, Berlin, Heidelberg (2009).

48. N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst. Man Cybern.* **9**(1), 62–66 (1979).

49. L. Breiman, "Random forests," *Mach. Learn.* **45**(1), 5–32 (2001).

50. C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.* **20**(3), 273–297 (1995).

51. S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," in *Int. Conf. Eng. Technol.*, IEEE, Antalya, Turkey, pp. 1–6 (2017).

52. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," arXiv:1705.07874 (2017).

53. R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," in *Int. Conf. Mach. Learn.*, Association for Computing Machinery, New York, pp. 161–168 (2006).

54. P. C. Sen, M. Hajra, and M. Ghosh, "Supervised classification algorithms in machine learning: a survey and review," in *Emerging Technology in Modelling and Graphics*, J. K. Mandal, Ed., pp. 99–111, Springer, Singapore (2020).

55. P. Cunningham, M. Cord, and S. J. Delany, "Supervised learning," in *Machine Learning Techniques for Multimedia: Case Studies on Organization and Retrieval*, M. Cord, Ed., pp. 21–49, Springer, Berlin, Heidelberg (2008).

56. T. Hastie, J. Friedman, and R. Tibshirani, *The Elements of Statistical Learning*, Springer New York, New York (2001).

57. S. B. Kotsiantis, I. D. Zaharakis, and P. E. Pintelas, "Machine learning: a review of classification and combining techniques," *Artif. Intell. Rev.* **26**(3), 159–190 (2006).

58. I. G. Maglogiannis, Ed., *Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in EHealth, HCI, Information Retrieval and Pervasive Technologies*, IOS Press, Amsterdam (2007).

59. A. E. Maxwell, T. A. Warner, and F. Fang, "Implementation of machine-learning classification in remote sensing: an applied review," *Int. J. Remote Sens.* **39**(9), 2784–2817 (2018).

60. K. P. Murphy, *Machine Learning: A Probabilistic Perspective*, MIT Press (2012).

61. V. Stanev et al., "Machine learning modeling of superconducting critical temperature," *NPJ Comput. Mater.* **4**(1), 29 (2018).

62. A. J. Myles et al., "An introduction to decision tree modeling," *J. Chemom.* **18**(6), 275–285 (2004).

63. J. H. Friedman, "Stochastic gradient boosting," *Comput. Stat. Data Anal.* **38**(4), 367–378 (2002).

64. T. Hastie, R. Tibshirani, and J. Friedman, "Unsupervised learning," in *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, T. Hastie, Ed., pp. 485–585, Springer, New York (2009).

65. G. Hinton and T. J. Sejnowski, *Unsupervised Learning: Foundations of Neural Computation*, MIT Press, Cambridge, Masschusetts (1999).

66. Z. Ghahramani, *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2-14, 2003, Tübingen, Germany, August 4-16, 2003, Revised Lectures*, Springer, Berlin, Heidelberg (2004).

67. D. Greene, P. Cunningham, and R. Mayer, "Unsupervised learning and clustering," in *Machine Learning Techniques for Multimedia: Case Studies on Organization and Retrieval*, M. Cord, Ed., pp. 51–90, Springer, Berlin, Heidelberg (2008).

68. K. R. Chowdhary, "Natural language processing," in *Fundamentals of Artificial Intelligence*, K. R. Chowdhary, Ed., pp. 603–649, Springer India, New Delhi (2020).

69. A. Gardin et al., "Classifying soft self-assembled materials via unsupervised machine learning of defects," *Commun. Chem.* **5**(1), 82 (2022).

70. D. A. Reynolds, "Gaussian mixture models," *Encyclopedia Biometr.* **741**, 659–663 (2009).

71. T. Kohonen, "The self-organizing map," *Proc. IEEE* **78**(9), 1464–1480 (1990).

72. X. Zhu and A. B. Goldberg, *Introduction to Semi-Supervised Learning*, Springer Nature (2022).

73. X. Zhai et al., "S4L: self-supervised semi-supervised learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, IEEE, Seoul, South Korea, pp. 1476–1485 (2019).

74. O. Chapelle, B. Scholkopf, and A. Zien, Eds., "Semi-supervised learning (Chapelle, O. et al., Ed.; 2006) [Book reviews]," *IEEE Trans. Neural Netw.* **20**(3), 542 (2009).

75. X. Zhu, *Semi-Supervised Learning Literature Survey*, University of Wisconsin-Madison, Department of Computer Sciences, CS Technical Report TR1530 (2005).

76. M.-R. Amini et al., "Self-training: a survey," arXiv:2202.12040 (2023).

77. M.-F. Balcan, A. Blum, and K. Yang, "Co-training and expansion: towards bridging theory and practice," in *Adv. Neural Inf. Process. Syst.*, Vol. 17, pp. 89–96 (2004).

78. N. T. P. Hartono et al., "How machine learning can help select capping layers to suppress perovskite degradation," *Nat. Commun.* **11**(1), 4172 (2020).

79. C. Zhi et al., "Machine-learning-assisted screening of interface passivation materials for perovskite solar cells," *ACS Energy Lett.* **8**(3), 1424–1433 (2023).

80. C. She et al., "Machine learning-guided search for high-efficiency perovskite solar cells with doped electron transport layers," *J. Mater. Chem. A* **9**(44), 25168–25177 (2021).

81. W. Liu et al., "Machine learning enables intelligent screening of interface materials towards minimizing voltage losses for p-i-n type perovskite solar cells," *J. Energy Chem.* **83**, 128–137 (2023).

82. S. Kim et al., "PubChem substance and compound databases," *Nucl. Acids Res.* **44**(D1), D1202–D1213 (2016).

83. Y. Hu et al., "Machine-learning modeling for ultra-stable high-efficiency perovskite solar cells," *Adv. Energy Mater.* **12**(41), 2201463 (2022).

84. T. Bak et al., "Accelerated design of high-efficiency lead-free tin perovskite solar cells via machine learning," *Int. J. Precis. Eng. Manuf. Green Technol.* **10**(1), 109–121 (2023).

85. J. Wang et al., "Advancing vapor-deposited perovskite solar cells via machine learning," *J. Mater. Chem. A* **11**(25), 13201–13208 (2023).

86. N. Taherimakhsousi et al., "A machine vision tool for facilitating the optimization of large-area perovskite photovoltaics," *NPJ Comput. Mater.* **7**(1), 190 (2021).
87. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv:1409.1556 (2015).
88. S. Wang et al., "Carrier dynamics determines the optimization strategies of perovskite LEDs and PVs," *Research* **6**, 0112 (2023).
89. F. Scarselli et al., "The graph neural network model," *IEEE Trans. Neural Netw.* **20**(1), 61–80 (2009).

Feiyue Lu received his master's degree from the Institute of Advanced Materials and School of Flexible Electronics (Future Technologies), Nanjing Tech University (NanjingTech) in 2024. He currently focuses on the interdisciplinary research of machine learning and perovskite optoelectronic devices.

Lin Zhu received her PhD in chemistry from Nanjing University. She subsequently joined Nanjing Tech University, where her current research focuses on perovskite LEDs and machine learning.

Jianpu Wang received his PhD in organic semiconductor/inorganic nanocrystal devices from the Cavendish Laboratory at the University of Cambridge. He subsequently worked as a postdoctoral research associate at the Cavendish Laboratory from 2009 to 2013. His current research interests include organic and perovskite semiconductor devices, device physics, and artificial intelligence.

Biographies of the other authors are not available.