Materials Today Sustainability 31 (2025) 101176

Contents lists available at ScienceDirect

Materials Today Sustainability

journal homepage: www.journals.elsevier.com/materials-today-sustainability

Predicting high-performance perovskite solar cells using AI-based machine learning models

Shafidah Shafian [a,*], Mohd Nizam Husen [b,**], Lin Xie [c,d], Kyungkon Kim [e,***]

[a] Solar Energy Research Institute, Universiti Kebangsaan Malaysia, Bangi, 43600, Selangor, Malaysia
[b] Malaysian Institute of Information Technology, Universiti Kuala Lumpur, Kuala Lumpur, 50250, Malaysia
[c] School of Materials and Energy, Yunnan University, Kunming, 650091, China
[d] International Joint Research Center for Optoelectronic and Energy Materials, Yunnan University, Kunming, 650091, China
[e] Department of Chemistry and Nanoscience, Ewha Womans University, Seoul, 03760, Republic of Korea

ARTICLE INFO

Keywords:
Artificial intelligence
Machine learning
Model algorithms
Material discovery
Device optimization
Perovskite solar cell

ABSTRACT

Perovskite solar cells (PSCs) have garnered significant attention in the photovoltaic field due to their remarkable power conversion efficiencies (PCEs), with the certified PCE reaching 27.0% for single-junction cells and 30.1% for tandem perovskite/perovskite multijunction cells over the past decade. However, challenges remain including material instability, compositional inconsistency, and limited long-term performance. Machine learning (ML) has emerged as a transformative tool to address these challenges by accelerating material discovery, optimizing device design, and enabling data-driven insights from large and complex datasets. This review presents a comprehensive analysis of how ML is being applied to advance PSCs technologies. It begins with an overview of perovskite structures, device architectures, and performance parameters relevant to ML modelling. A structured ML workflow is introduced, covering data acquisition, feature selection, model development, performance evaluation, and model interpretability through explainable AI (XAI) techniques. Recent studies are examined across two major domains: material discovery and device performance optimization. Unlike previous reviews, this work emphasis on quantitative comparisons of ML algorithms by systematically assessing models reported in recent literature to identify the most effective predictors across various tasks. Furthermore, it discusses the strengths and limitations of current datasets and modelling strategies. The review concludes with insights into existing challenges and outlines future directions to support the efficient, interpretable, and scalable application of ML in PSCs research.

1. Introduction

The rapid development of renewable energy sources is crucial in addressing the global energy crisis and mitigating the adverse effects of climate change. Among the various renewable energy technologies, solar photovoltaics has gained significant attention due to its abundance and sustainability. In recent years, perovskite solar cells (PSCs) have emerged as a promising alternative to traditional silicon- and thin-film-based solar cells due to their exceptional power conversion efficiencies (PCE), ease of fabrication, and low production costs [1–5]. The unique properties of perovskite materials, such as high absorption coefficients, tunable bandgaps, and excellent charge transport characteristics, make them ideal candidates for next-generation photovoltaic devices. However, challenges such as long-term stability, scalability, and degradation under environmental stress remain significant barriers to the commercialization of solar cell technology [6,7]. As a result, substantial research and development efforts have been dedicated to addressing these issues [1,8–11].

Building upon these advancements, artificial intelligence (AI) through machine learning (ML) have gained increasing attention in the field of material science and photovoltaics, with researchers exploring their potential to accelerate the discovery and optimization of high-performance materials [12,13]. AI-driven ML models have proven to be effective tools in predicting the properties and performance of perovskite materials, offering the possibility of overcoming the time-consuming and costly processes typically associated with

* Corresponding author.
** Corresponding author.
*** Corresponding author.
E-mail addresses: norshafidah@ukm.edu.my (S. Shafian), mnizam@unikl.edu.my (M.N. Husen), kimkk@ewha.ac.kr (K. Kim).

https://doi.org/10.1016/j.mtsust.2025.101176
Received 7 April 2025; Received in revised form 28 June 2025; Accepted 2 July 2025
Available online 8 July 2025

experimental research. By leveraging large datasets, ML algorithms can identify complex relationships between material properties and device performance, enabling the design of more efficient and stable PSCs.

The use of AI-driven ML models in PSCs research can be broadly divided into two main areas: (1) material discovery and design, and (2) device performance and stability prediction. In the first area, ML algorithms analyse large datasets of material properties such as crystal structure, electronic characteristics, optical properties, and charge transport behavior. These models help identify new perovskite materials with better performance or suggest ways to modify existing materials to improve their properties. For example, ML can predict how changes in cation substitutions or halide combinations might impact key factors like bandgap and material stability. These predictions guide the synthesis of materials with desired characteristics, offering a more efficient alternative to traditional trial-and-error experimental methods. In the second area, ML models are used to predict the overall performance of PSCs by analysing the complex relationships between various device parameters, including open-circuit voltage (V_{OC}), short-circuit current (J_{SC}), fill factor (FF), and PCE. These models are particularly useful for optimizing the PSCs architecture, such as selecting the best electron and hole transport materials, choosing stable perovskite materials, adjusting the thickness of the perovskite layer, and refining deposition techniques. Additionally, ML is increasingly applied to predict PSCs stability, addressing ongoing challenges related to long-term performance, hysteresis, and reproducibility which is the key issues for advancing PSCs technology and ensuring its commercial viability.

This review offers a comprehensive and systematic overview of the integration of AI-driven ML models in PSCs research. We begin by examining the structure and composition of PSCs and the key performance parameters that are relevant for ML modelling. The manuscript then outlines a robust ML pipeline ranging from data acquisition, feature engineering, and model selection to interpretability and evaluation. Recent advances are critically reviewed across two primary domains: material property discovery and device performance optimization, with detailed comparative analyses of model types and predictive accuracy.

Unlike other reviews that mainly describe how ML is used in PSCs research, this work focuses on comparing models using actual performance numbers. It summarizes and contrasts key metrics from recent studies, helping readers see which ML models work best for predicting bandgaps, stability, and device performance. This clear, number-based comparison gives researchers practical guidance on choosing the right models for their own work.

Ultimately, we conclude with forward-looking perspectives that explore current limitations, propose best practices for reproducible research, and suggest future directions for expanding ML's role in PSCs. By consolidating state-of-the-art methodologies and benchmark studies, this review aims to guide both newcomers and experts toward the intelligent and sustainable development of next-generation PSCs.

2. Perovskite solar cells (PSCs)

PSCs represent one of the most promising advancements in photovoltaic technology due to their exceptional PCE and cost-effective manufacturing processes. Since their initial laboratory fabrication, PSCs have undergone rapid development, achieving PCE comparable to, and in some cases surpassing traditional silicon-based solar cells. Recent reports from the National Renewable Energy Laboratory (NREL) confirm that the highest certified PCE for single-junction PSCs has reached 27.0%, while tandem perovskite/perovskite multijunction solar cells have achieved a certified PCE of 30.1% [14–16]. Their unique characteristics, such as tunable bandgap, high absorption coefficients, and long carrier diffusion lengths, have garnered significant attention from researchers and industries alike [17]. In this context, ML has emerged as a powerful tool to optimize the design, predict the performance, and enhance the stability of PSCs by analysing their intrinsic and extrinsic properties. A key aspect of these properties lies in the unique structure and composition of the perovskite material itself, which fundamentally governs its photovoltaic behavior.

PSCs are named after their active layer, which adopts a crystal structure resembling the mineral perovskite of calcium titanate ($CaTiO_3$). This structure is characterized by a unique ABX_3 composition, where "A" represents a monovalent cation, "B" is a divalent metal cation, and "X" is a halide anion. The A-site cation can be materials like methylammonium, MA ($CH_3NH_3^+$), formamidinium, FA ($HC(NH_2)_2^+$), cesium (Cs^+) or rubidium (Rb^+). The B-site typically hosts lead (Pb^{2+}), through tin (Sn^{2+}), germanium (Ge^{2+}), zinc (Zn^{2+}), copper (Cu^{2+}) and magnesium (Mg^{2+}) are also explored as alternatives to reduce toxicity. The X-site consist of halide anion, such as iodide (I^-), bromide (Br^-), or chloride (Cl^-). Fig. 1 shows the perovskite cubic crystal structure and the periodic table showing the element corresponding to A, B and X materials. This versatile composition allows for precise tuning of material properties, enabling the optimization of key attributes like bandgap, stability, and charge transport, which are crucial for enhancing the efficiency of PSCs. The ability to vary the A, B, and X components makes perovskite materials highly adaptable, allowing researchers to tailor them for a wide range of photovoltaic applications. For instance, by modifying the A-site cation or the halide anion, the bandgap of the material can be adjusted to optimize light absorption for different solar spectrum conditions. Similarly, substituting lead with tin or germanium can improve the environmental friendliness of PSCs, though challenges with stability and performance remain.

PSCs generally consist of several key layers that work together to facilitate efficient charge transport and light absorption. These include the perovskite absorber layer, which is responsible for capturing sunlight and generating charge carriers. The electron transport layer (ETL) [19,20] and hole transport layer (HTL) [4,21] are critical for efficiently extracting electrons and holes from the photoactive layer and transporting them to the respective electrodes. The choice of materials for the ETL and HTL, as well as their configuration, can significantly impact the device's efficiency and stability by influencing the charge extraction and minimizing recombination losses. Finally, the electrodes, typically made of materials like gold, silver, or transparent conductive oxides, serve to collect and transfer the charge carriers to an external circuit. The interplay of these layers, combined with the tunable properties of the perovskite material itself, underscores the flexibility and potential of PSCs as a leading technology in the field of photovoltaics. Each layer's material selection and structural configuration directly impact the overall performance, including factors like efficiency, stability, and manufacturing scalability.

3. Machine learning (ML)

ML typically follows a structured workflow that includes four key stages: data acquisition, feature selection, model selection, and model evaluation as shown in Fig. 2. To ensure the practical utility and trustworthiness of ML predictions, a fifth stage which is model interpretability through explainable AI (XAI) techniques is increasingly recognized as essential [22].

3.1. Data acquisition

Data preparation is a fundamental and critical initial step in the development of ML models, drawing from a variety of sources, including existing literature, theoretical simulation tools, and open-access available datasets. The quality and comprehensiveness of these datasets are essential, as they provide the foundation for building accurate and reliable ML models. To ensure the validity and robustness of these models, thorough data validation processes are necessary to minimize errors and biases.

An increasingly important trend in recent PSCs research is the integration of ML with density functional theory (DFT) simulations and experimental validation. In such hybrid workflows, DFT is used to

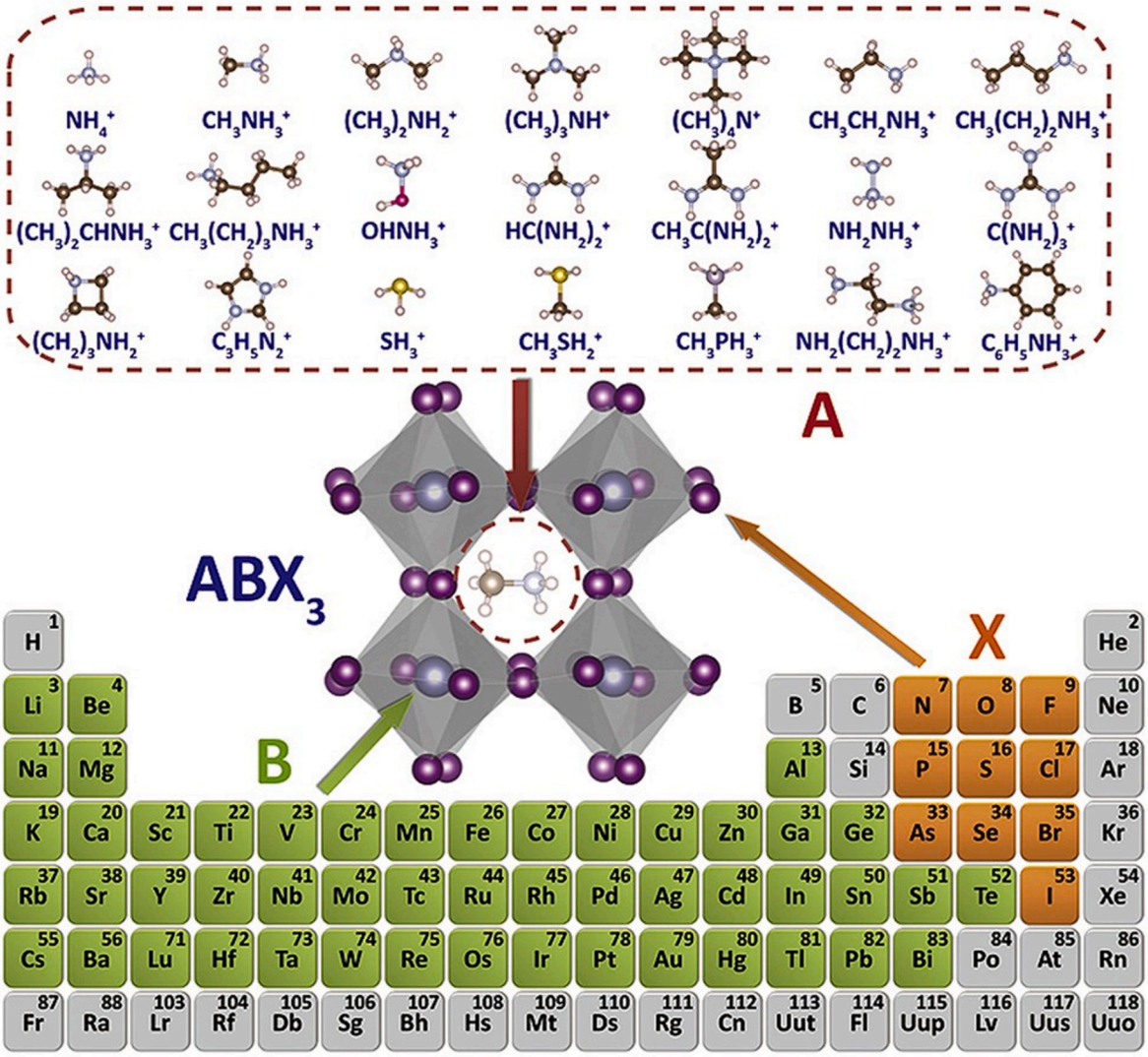

Fig. 1. Perovskite cubic crystal structure and the elements of the periodic table to form perovskites. Reproduced with permission from Ref. [18]. Copyright (2019) Elsevier.

generate computational data such as electronic structures and formation energies, which are then used to train ML models capable of exploring much larger chemical and composition spaces. These models guide the selection of promising candidates for synthesis and device fabrication, while experimental outcomes feed back into the model to further refine predictive accuracy. This integration significantly accelerates the discovery and optimization of PSCs materials by bridging theoretical insights with practical implementation, and it also underpins the development of hybrid datasets that link computational and experimental domains.

Table 1 presents a compilation of commonly used datasets in the context of perovskite materials and device fabrication for ML applications. As shown in Table 1, datasets can be categorized into three types: computational simulations, experimental data, and hybrid datasets that combine both computational and experimental approaches.

Calculation-based simulation databases provide a foundational backbone for ML applications in perovskite research due to several key advantages. These resources supply large volumes of high-quality data on structural, electronic, and thermodynamic properties typically derived from DFT. Their scale and consistency enable rapid screening of large and diverse chemical spaces, including hypothetical materials not yet synthesized and facilitate systematic trend analysis across composition and property domains. Furthermore, their open-access design and compatibility with application programming interfaces (APIs) make them highly accessible for automated data retrieval and model integration. However, there are several limitations. The coverage of PSCs relevant materials is often incomplete, particularly for organic–inorganic hybrid perovskites, low-dimensional structures, and lead-free compositions, which are underrepresented in many entries. In addition, these databases are subject to computational biases arising from fixed simulation parameters which may introduce systematic errors, especially for complex perovskites with soft lattices or dynamic structural behavior. Lastly, the absence of uncertainty quantification or experimental cross-validation in many entries may affect the transferability and interpretability of ML models trained on this data.

Experimental databases provide empirically validated data on crystal structures, physical properties, synthesis routes, and device performance metrics. These datasets are valuable for ML applications because they offer a direct reflection of real-world material behavior, enabling the development of models that are grounded in experimental reality. This type of data is crucial for improving the generalizability, reliability, and end-use relevance of ML models, especially when predicting synthesis feasibility, phase stability, or practical device efficiency. However, the completeness and consistency of experimental databases can vary significantly depending on material qualities, experimental methodology, geographic location, operator skills and reporting standards. In the

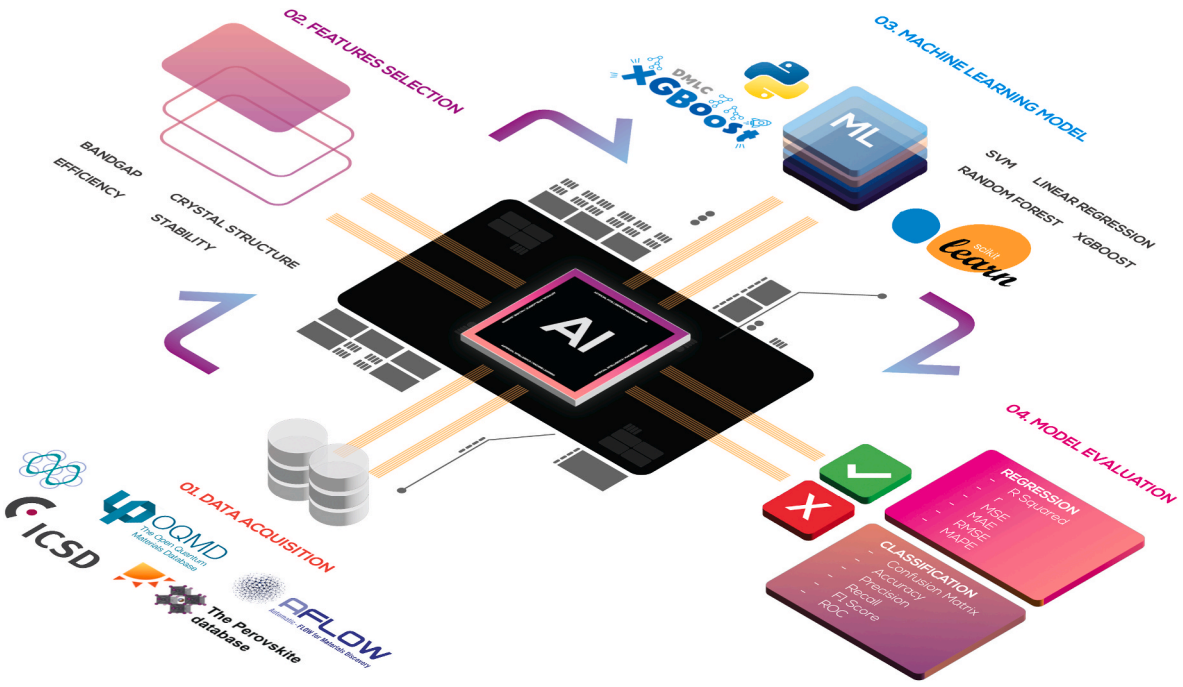

Fig. 2. Schematic workflow of ML in PSCs research.

context of PSCs, many databases tend to focus on well-characterized and stable compositions such as $MAPbI_3$ or its common variants while less attention is given to emerging materials, including lead-free, 2D layered, and more complex perovskites structure, are often underrepresented. This bias reflects broader trends in the literature, where failed experiments, unstable materials, or low-efficiency devices are less frequently reported, resulting in a publication and reporting bias that skews dataset distributions. Additionally, inconsistencies in metadata such as incomplete details on temperature, pressure, precursor purity, or fabrication techniques can introduce variability and noise into ML models, potentially reducing predictive accuracy.

Hybrid databases, which integrate both computational and experimental data, offer a balanced and often more application-specific resource. These databases are particularly valuable as they link structural and compositional information with processing parameters, device architectures, and photovoltaic performance metrics such as PCE and stability. By combining theoretical predictions with experimental results, hybrid datasets facilitate more holistic ML modelling and enable cross-validation between simulated and observed behaviour, thereby improving both accuracy and practical relevance. Nonetheless, hybrid databases also face a limitation. Their size is generally smaller compared to standalone computational databases, which may restrict the diversity of training data available for ML models. Moreover, they may exhibit curation bias, as the integration process often depends on published studies that emphasize high-performing or well-characterized materials, leading to underrepresentation of failed experiments or less-explored perovskite formulations. Variability in data quality, reporting standards, and metadata completeness across sources can also pose challenges for model consistency and generalization.

Each dataset type offers unique strengths that collectively support diverse ML application in PSCs research. However, a significant and growing concern is the lack of standardized benchmark datasets, which hampers fair comparison across studies. Differences in dataset structure, property definitions, and train-test splits often make it difficult to assess whether improvements in performance stem from better models or merely from favorable data conditions. Addressing these issues through community-driven standards, better data curation practices, and the development of shared benchmarks will be essential for advancing reliable and reproducible ML research in the PSCs domain.

3.2. Model selection

Selecting an appropriate ML model is a critical step in developing predictive tools for PSCs research. Most applications rely on supervised learning, where tasks are framed as either regression (predicting continuous outputs) or classification (categorizing discrete labels). Each ML category brings distinct advantages depending on the complexity and nature of the dataset. Linear models, such as Linear Regression (LR), Ridge Regression (RR), and Lasso, are widely used for their simplicity, computational efficiency, and ease of interpretation. These models assume a linear relationship between input features and output, making them suitable for straightforward problems with well-behaved data. Bayesian models, including Naïve Bayes and Gaussian Process Regression (GPR), incorporate uncertainty estimation and prior knowledge, which can be valuable for risk-sensitive applications or when data is limited. Instance-based models, like K-Nearest Neighbours (KNN), are non-parametric and rely on distance metrics to predict outcomes based on the similarity between samples. While intuitive, they can become computationally intensive on large datasets. Tree-based models, such as Decision Trees (DT), Random Forests (RF), and Extra Trees (ET), split data into hierarchical segments based on feature thresholds. These models are capable of capturing non-linear relationships and are robust to feature scaling. Support Vector Machines (SVM) aim to find the optimal hyperplane for classification and can be extended to regression tasks via Support Vector Regression (SVR), offering strong performance in high-dimensional spaces. Ensemble methods like Gradient Boosting Decision Trees (GBDT) combine multiple decision trees to create a strong predictive model. Notably, XGBoost and LightGBM are optimized implementations of GBDT that enhance training speed, regularization, and scalability. Lastly, deep learning models including Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), and Graph Neural Networks (GNN) excel in modelling complex, high-dimensional, or structured data such as images, graphs, or molecular representations. Table 2 summarizes these models by category, highlighting their typical use (regression or classification) and providing a brief description of each.

Table 1
Available databases used in perovskite materials.

Types	Databases	Materials	Properties	Entries	URL	Ref.
Calculation	MP - Materials Project	Molecules, inorganic compounds	Structural, electronic, magnetic, thermodynamic, etc.	>751,876	https://materialsproject.org/	[23]
	OQMD - Open Quantum Materials Database	Inorganic compounds	Physical properties such as material structure, energy, and stability calculated by DFT.	>1,226,781	https://oqmd.org/	[24]
	NOMAD	Inorganic compounds, organic molecules, metals, organometallic frameworks, etc.	DFT, other computational data	>12,235,230	https://nomad-lab.eu/nomad-lab/	[25]
	Materials Cloud	Molecules, inorganic compounds, organic compounds, crystals, etc.	Crystal structure, electronic properties, thermodynamic properties, etc.	>7,502,686	https://www.materialscloud.org/home	[26]
	CMR - Computational Materials Repository	Semiconductor, metal, polymer, etc.	Electronic, mechanical, structural properties, etc	>150,000	https://cmr.fysik.dtu.dk/	[27]
	AFLOW -Automatic-FLOW for Materials Discovery	Inorganic compounds, alloys, etc.	Crystal structure, electronic properties, etc.	>3,530,330	https://www.aflowlib.org/	[28]
Experiment	CSD - Cambridge Structural Database	Organic compounds, organometallic compounds, and some inorganic compounds.	Crystal structure data from X-ray.	>1,250,000	https://www.ccdc.cam.ac.uk	[29]
	COD - Crystallography Open Database	Organic and inorganic compounds, minerals, metals, etc	Crystal structure data from X-ray.	>210,000	https://icsd.nist.gov/	[30]
	MatDat – Materials Data Network	Engineering materials such as metals, composites, ceramics, polymers, etc.	Experimental mechanical performance characterization data	>1500	https://www.matdat.com/	[31]
	MatWeb	Polymers, metals, ceramics, etc.	Experimental physical, mechanical property, etc.	>170,000	https://www.matweb.com	[31]
	Materiae	Polymers, composites, metals, ceramics, etc.	Physical and mechanical properties, etc.	>39,519	https://cmpdc.iphy.ac.cn/materiae/#/	[32]
	ICSD -Inorganic Crystal Structure Database	Inorganic crystal structures.	Crystal structure data from X-ray.	>210,000	https://icsd.nist.gov/	[33]
Hybrid (Calculation + Experiment)	PDP -Perovskite Database Project	Perovskites	Fabrication processes, performances, etc.	>42,400	https://app.materials.zone/home	[34]
	HybriD^3 Materials Database	organic-inorganic hybrid semiconductors	Formula, Physical Properties	>10,000	https://hybrid3.duke.edu/	[35]
	MPDS -Materials Platform for Data Science	Organic and inorganic compounds, metals, alloys, etc.	Crystal structure, physical properties, thermodynamic data, etc.	>247,000	https://mpds.io/#modal/menu	[36]
	Springer Materials	Organic and inorganic compounds, metals, alloys, composites, etc.	Characterization data of physical chemistry and material properties	>1,530,930	https://materials.springer.com/	
	MatNavi	Inorganic compounds, metals, ceramics, composites, functional materials, etc.	Data from physical and chemical characterization to mechanical testing.	>29,000	https://mits.nims.go.jp/	[37]

3.3. Model evaluation

Once model validation and optimization methods are established, it is essential to evaluate the performance of the model using appropriate metrics. Different evaluation metrics are suited to various types of problems, ensuring an objective assessment of model effectiveness. This section focuses on the key metrics for evaluating ML models in both regression and classification contexts. As outlined in Table 3, common evaluation metrics for regression problems include Mean Square Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Coefficient of Determination (R^2), and Correlation Coefficient (r), while for classification problems, the metrics typically include the Confusion Matrix, Accuracy, Precision, Recall, F1 Score, and Receiver Operating Characteristic (ROC).

Metrics such as MSE, MAE, RMSE, and MAPE measure the generalization error of regression models, with lower values indicating better performance. R^2, widely used in regression tasks, reflects the proportion of variance explained by the model. R^2 ranges from 0 to 1, where 1 indicates perfect prediction and 0 indicates no predictive power. r quantifies the strength of the linear relationship between the predicted and true values, with values nearing 1 indicating a stronger correlation.

For classification tasks, the confusion matrix provides a detailed visualization of classification outcomes, as it provides a comprehensive overview of the model's performance, showing the true positives (TP) - the prediction and truth are positive, true negatives (TN) - The prediction and truth are negative, false positives (FP) - The prediction is positive, but truth is negative, and false negatives (FN) - The prediction is negative, but truth is positive. It forms the foundation for deriving other metrics. Once the confusion matrix is available, calculating accuracy comes next. Accuracy measures the proportion of correct predictions. Precision measures the proportion of positive predictions that are actually correct and Recall measures the proportion of actual positives that are correctly identified. The F1-Score combines Precision and Recall into a single metric, balancing their contributions. Lastly, the ROC curve, plotted with the False Positive Rate (FPR) on the x-axis and True Positive Rate (TPR) on the y-axis, illustrates the trade-offs between sensitivity and specificity, with better model performance represented by a curve.

3.4. Model interpretability and explainable AI (XAI) techniques

In materials science research, it is not sufficient to only assess how well a ML model performs, it is equally important to understand why the model makes specific predictions. This is especially crucial when ML guides material selection or experimental design. However, many high-performing ML models such as deep neural networks and ensemble-based methods are often considered "black boxes", as their internal decision-making processes are complex and not easily interpretable. To overcome this challenge, various XAI techniques have been developed. These tools help uncover the relationships between input features and model outputs, enhancing scientific understanding and trust in ML predictions. Table 4 provides a summary of widely used XAI techniques,

Table 2
Overview of ML models, including their categories, names, descriptions, and application types (regression and/or classification).

Categories	Models	Full Name	Description	Types	
				Regression	Classification
Linear	LR	Linear Regression	Foundational statistical method for modelling linear relationships	✓	
	RR	Ridge Regression	Introduced L2 regularization to address multicollinearity	✓	
	Lasso	Least Absolute Shrinkage and Selection Operator	Adds L1 penalty for feature selection and sparsity	✓	✓
	KRR	Kernel Ridge Regression	Combines ridge regression with kernel trick for non-linear problems	✓	
Bayesian	Naïve Bayes	Naïve Bayes	One of the earliest classifiers assumes feature independence		✓
	GPR	Gaussian Process Regression	Bayesian non-parametric regression using distributions over functions	✓	
Instance/Distance-Based	KNN	K-Nearest Neighbours	Predicts label/value based on majority/average of nearest neighbours	✓	✓
	NCC	Nearest Centroid Classifier	Assigns to class whose mean feature vector is closest		✓
	RN	Radius Neighbours	Variant of KNN that uses a fixed radius instead of k	✓	✓
Tree	DT	Decision Tree	Foundation of all tree-based methods	✓	✓
	RF	Random Forest	An ensemble of decision trees using bagging	✓	✓
	ET/ERT	Extra Tree/Extremely Randomized Trees	Similar to Random Forest, but splits are more randomized	✓	✓
	GPSR	Genetic Programming Symbolic Regression	Evolves mathematical expressions using genetic programming	✓	
Support Vector Machine (SVM)	SVR	Support Vector Regression	Fits data within a margin while minimizing prediction error	✓	
	SVC	Support Vector Classification	Classifies data by finding the optimal separating hyperplane		✓
Ensemble Learning	AdaBoost	Adaptive Boosting	First practical boosting method using adaptive weights	✓	✓
	GBDT	Gradient Boosting Decision Tree	General gradient boosting framework	✓	✓
	GBR	Gradient Boost Regression	Scikit-learn implementation of GBDT	✓	
	Stacking	Stacking Generalization	Combines multiple models using a meta-learner	✓	✓
	XGBoost	Extreme Gradient Boosting	Optimized GBDT with regularization and parallelization	✓	✓
	LightGBM	Light Gradient Boosting Machine	Optimized GBDT using histogram-based split finding	✓	✓
	CatBoost	Categorical Boosting	Optimized GBDT for categorical variables and imbalanced data	✓	✓
Deep Learning	ANN	Artificial Neural Network	The foundation of all neural networks, inspired by biological neurons.	✓	✓
	MLP	Multilayer Perceptron	First practical deep neural network with backpropagation	✓	✓
	RNN	Recurrent Neural Network	Designed to handle sequential data by retaining memory	✓	
	CNN	Convolution Neural Network	Designed for spatial data like images, widely used in vision	✓	✓
	AE	Autoencoder	Learns compressed representations of input data	✓	✓
	DNN	Deep Neural Network	Multi-layered ANN for complex pattern learning	✓	✓
	LSTM	Long Short-Term Memory	Improved RNNs for long-term sequence learning	✓	✓
	VAE	Variational Autoencoder	Probabilistic generative version of AE	✓	✓
	GNN	Graph Neural Networks	Designed for structured data like molecules and crystals	✓	✓
	BERT	Bidirectional Encoder Representations from Transformers	Transformer-based model for language and chemical sequences (e.g., SMILES)	✓	✓

categorized as model-agnostic or model-specific, along with their main applications, strengths, and limitations. This serves as a practical guide for selecting appropriate interpretability tools in the context of ML-driven PSCs research.

4. Characteristics of PSCs relevant to ML

4.1. Material structure

***Bandgap Energy*:** The bandgap of perovskite materials is one of the most critical characteristics in determining the performance of PSCs and plays a significant role in ML applications aimed at optimizing material properties. The bandgap directly influences the material's ability to absorb light, particularly within the visible spectrum, which is essential for efficient solar energy conversion. In ML, the bandgap serves as a key feature for predicting the optical properties and efficiency of perovskite materials. By incorporating bandgap data into ML models, researchers can identify new material compositions that maximize light absorption while minimizing energy losses due to recombination. Additionally, the ability to tune the bandgap through different cation and halide combinations allows for the design of perovskite materials that are better suited to specific applications, such as tandem solar cells or high-efficiency devices. ML algorithms use bandgap as a crucial input to guide the discovery of materials with optimal performance characteristics, making it an indispensable parameter in advancing PSCs technology.

***Crystal Structure*:** The crystal structure of perovskite materials is a crucial characteristic that significantly affects their electronic, optical, and mechanical properties, making it an essential feature in ML applications for PSCs optimization. The arrangement of atoms within the crystal lattice determines key factors such as charge carrier mobility, light absorption efficiency, and stability under environmental stressors like humidity and temperature. In ML models, the crystal structure is used to predict how different configurations can impact a material's photovoltaic performance, including its ability to efficiently convert photon to electron. By analysing large datasets of crystallographic information, ML algorithms can uncover patterns that link specific structural features to enhanced performance metrics, such as higher V_{OC} or reduced defect density. Moreover, understanding and predicting the influence of crystal structure on material properties allows researchers to design and discover new perovskite compositions with optimized structural stability and performance, thereby accelerating the development of next-generation solar materials.

Table 3
Common evaluation metrics of ML models for regression and calculation.

Types	Metrics	Formula
Regression	MSE	$MSE(y,\hat{y}) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$
	RMSE	$RMSE(y,\hat{y}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$
	MAE	$MAE(y,\hat{y}) = \frac{1}{N}\sum_{i=1}^{N}\lvert(y_i - \hat{y}_i)\rvert$
	MAPE	$MAPE(y,\hat{y}) = \frac{1}{N}\sum_{i=1}^{N}\frac{\lvert(y_i - \hat{y}_i)\rvert}{\lvert y_i\rvert}$
	R^2	$R^2(y,\hat{y}) = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$ $\hat{y} = \frac{1}{N}\sum_{i=1}^{N}\hat{y}_i$
	r	$r = \frac{\sum_{i=1}^{n}(y_i - \overline{y})(\hat{y}_i - \overline{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2\sum_{i=1}^{n}(\hat{y}_i - \overline{\hat{y}})^2}}$
Classification	Confusion Matrix	$Confusion\ Matrix = \begin{matrix} TP & FN \\ FP & TN \end{matrix}$
	Accuracy	$Accuracy = \frac{TP + TN}{TP + FN + FP + TN}$
	Precision	$Precision = \frac{TP}{TP + FP}$
	Recall	$Recall = \frac{TP}{TP + FN}$
	F1 Score	$F1 = \frac{2TP}{2TP + FP + FN}$
	ROC	$X : FPR = \frac{FP}{FP + TN}$
		$Y : TPR = \frac{TP}{TP + FN}$

N: number of samples; y_i: label value of the i_{sh} sample; $\hat{y}_i$: predicted value of the i_{sh} sample $\overline{y}$; $\overline{\hat{y}}$: mean value of label and predicted values.

4.2. Device performance

Efficiency: The PCE of PSCs is a key performance metric that reflects how effectively a solar cell converts photon to electron. It is determined by three essential parameters: V_{OC}, J_{SC}, and FF. V_{OC} represents the maximum voltage output when no current is flowing and is influenced by the bandgap of the perovskite material, which determines the energy required to excite electrons and generate charge carriers. A higher V_{OC} typically leads to better efficiency by enabling the cell to generate more voltage from the absorbed sunlight. J_{SC}, on the other hand, measures the amount of current generated under short-circuit conditions and is directly related to the material's ability to absorb light and produce charge carriers. Maximizing J_{SC} involves enhancing light absorption properties and reducing energy losses from recombination. FF, the ratio of the maximum power output to the theoretical maximum, reflects how efficiently the solar cell converts the generated current and voltage into useable power. A high FF is achieved by optimizing the charge transport properties and minimizing losses within the device, such as recombination.

ML plays a pivotal role in predicting and optimizing the efficiency of PSCs by analysing vast datasets of material properties and device parameters. Through ML, researchers can model the relationships between V_{OC}, J_{SC}, FF, PCE and various material. For example, ML algorithms can predict how changes in the perovskite material composition such as tuning the A-site, B-site, and X-site components that affect the key performance parameters of PSCs. Mixed-cation perovskites, which combine materials like MA^+ and Cs^+, can be optimized for both V_{OC} and J_{SC} by expanding the absorption spectrum and reducing defects, with ML models helping to identify the optimal combinations. Similarly, adjustments in halide compositions (I^-, Br^-, Cl^-) can be predicted by ML to improve the bandgap and enhance light absorption, further boosting efficiency. Additionally, ML can assist in optimizing device architecture, such as determining the ideal thickness of the perovskite layer or the most effective combination of ETL and HTL, ensuring the highest possible FF and reducing energy losses. By enabling precise control over these parameters, ML accelerates the discovery of materials and designs that maximize PSCs efficiency.

Furthermore, ML aids in surface passivation and contact engineering, where it can predict how additives or dopants will affect the crystallinity and morphology of the perovskite layer, minimize defects and improving charge carrier mobility. This process helps reduce recombination losses and enhances the overall device efficiency. As these parameters like V_{OC}, J_{SC}, and FF are optimized through the integration of ML models with experimental data, PSCs are positioned to achieve significant performance improvements.

Stability: Despite the remarkable PCE of PSCs, their stability remains a significant challenge for large-scale deployment and commercialization. One of the primary issues affecting the stability of PSCs is the susceptibility of the perovskite absorber layer to degradation when exposed to humidity, oxygen, and light. Temperature fluctuations can further exacerbate this issue, causing thermal stress that accelerates degradation. These environmental factors can trigger chemical reactions and phase transitions within the perovskite material, leading to a loss in PCE over time. The perovskite layer can also undergo structural instability, which results in the breakdown of the material's crystalline structure, creating defects that impair charge transport and reduce the overall performance of the device. In addition, the use of Pb^{2+} in many perovskite formulations introduces a toxicity concern, as lead can leach out of the material over time, potentially contaminating the surrounding environment and limiting the widespread adoption of PSCs.

ML plays a crucial role in addressing these stability challenges by enabling the prediction and optimization of perovskite materials and

Table 4
Summary of common XAI techniques.

Category	XAI Technique	Main Use	Strengths	Limitations
Model-Agnostic	SHAP (SHapley Additive exPlanations)	Quantifying feature impact on predictions	Theoretically sound, consistent, works globally and locally	Computationally intensive for large datasets
	LIME (Local Interpretable Model-Agnostic Explanations)	Explaining individual predictions with local surrogate models	Simple, model-agnostic, intuitive for local explanations	Can be unstable; sensitive to perturbations and sampling
	PDP (Partial Dependence Plot)	Visualizing average effect of a feature on predictions	Easy to visualize and interpret feature effects	Assumes feature independence; can mislead with correlated features
	ALE (Accumulated Local Effects)	Assessing feature effects while accounting for feature interactions	Handles correlated features better than PDPs	Requires more technical setup; less widely adopted
	Permutation Feature Importance	Ranking feature importance by impact on model error	Fast to compute and intuitive for global understanding	Does not account for feature interactions
Model-Specific	Integrated Gradients	Attributing output changes to input features in neural networks	Works well with complex deep models, mathematically grounded	Requires access to model internals; limited to deep models
	Saliency Maps	Highlighting input regions (e.g., pixels) most influencing output	Visualizes relevance in input space, good for CNNs	Limited to vision tasks; may miss abstract features
	LRP (Layer-wise Relevance Propagation)	Decomposing prediction layer-by-layer in neural networks	Layer-wise attribution, detailed explanation in deep nets	Requires neural network access and can be complex to implement
	DeepLIFT (Deep Learning Important FeaTures)	Comparing activations against reference input for feature contribution	Provides clearer baselines for feature contribution	Needs baseline reference; choice of reference affects results

devices for enhanced longevity and resilience. By analysing large datasets of environmental stressors and material properties, ML algorithms can identify the key factors that contribute to degradation, such as specific material compositions or structural features that are more prone to moisture, light, or thermal instability. For example, ML models can predict how the inclusion of certain cations or halides in the perovskite structure will affect its stability under different environmental conditions. This enables researchers to design more stable perovskite materials that are less sensitive to degradation, such as lead-free perovskites or mixed-cation and mixed-halide perovskites, which have shown increased resistance to humidity and temperature fluctuations. ML can also aid in the development of encapsulation techniques by predicting the most effective protective coatings or barrier layers to shield the solar cells from harmful environmental factors, thereby extending their operational lifespan. In addition, ML plays a vital role in improving interface stability between the perovskite absorber layer and the charge transport layers (ETL/HTL). By analysing large datasets on interfacial properties and device performance, ML models can suggest optimal material combinations and processing conditions that minimize recombination losses and enhance the durability of the entire device.

5. Application of ML in PSCs

ML has emerged as a transformative tool in advancing the development of PSCs, offering predictive capabilities across both material discovery and device optimization. By analysing datasets, ML models can predict key material properties such as bandgap and crystal structure, as well as device performance parameters including V_{OC}, J_{SC}, FF, PCE, and stability. This section presents detailed tabular summaries of recent studies (Tables 5–7), followed by a concise synthesis that discusses common trends, compares performance outcomes, and outlines considerations related to model selection and limitations.

5.1. Material discovery

The following Table 5 compiles recent studies that employed ML to analyse and predict material properties, particularly focusing on bandgap and crystal structure of perovskite compounds.

Li, J. et al. [65] predict the bandgaps of ABX_3-type perovskite materials with varying compositions using a dataset of 333 entries from 2000 papers. Among five algorithms, the ANN performed best, achieving the smallest RMSE (0.06 eV) and the highest r = 0.97. Based on this prediction, the authors synthesized three new materials, $Cs_xMA_{1-x}PbI_3$, $CsPb(I_xBr_{1-x})_3$, and $MAPb_{1-x}Sn_xI_3$. Experimental results showed that $MAPb_{1-x}$ Sn_xI_3 films had bandgaps below 1.4 eV, $CsPb(I_xBr_{1-x})_3$ films had bandgaps above 1.8 eV, and $Cs_xMA_{1-x}PbI_3$ films have bandgaps between 1.4 and 1.8 eV. The predicted and experimental bandgap values were closely aligned, with contour plots showing that Sn results in a smaller bandgap than Pb. (Fig. 3(a)).

Park, H. et al. [64] constructed ML models based on DL, GBM, GLM, RF and XGBoost to predict A-site cation in perovskite materials. Among these methods, DL, GBM, and XGboost demonstrated higher accuracy, particularly in predicting octahedral deformation-related quantities. The ensemble methods of GBM and XGBoost were more robust than RF, which struggled to learn from variable orientations and favored mean values. GLM, being a linear regression approach, had the lowest accuracy. Fig. 3(b) shows the predicted ΔHc using XGBoost, revealing Cs−MA−EA mixed perovskite as stable with low EA concentration.

Liu, H. et al. [63] used the MP database to screen thermodynamically stable ABO_3 perovskite. They employed a classification model to predict new ABO_3 materials and evaluated the model accuracy. GBDT model achieved the highest accuracy in predicting stable perovskite candidates.

Park, H. et al. [62] investigated the relationship between the bandgap and the chemical properties of the constituent elements, including ionic radii, affinities, and ionization energies. They also identified significant correlations with structural features, such as octahedral quadratic elongation, angle variance, and rotation. The MAE for bandgap prediction was notably low, with DL achieving 0.17 eV and XGBoost achieving 0.12 eV and R^2 values were 0.94 and 0.97, respectively.

Li, Y. et al. [56] developed a strategy to tune the bandgap of mixed halide perovskites by adjusting cations and halide ions, using an experimental dataset. Among the models, LR achieved the best performance with an RMSE of 0.032 eV, followed by NN and RF with 0.050 eV and 0.145 eV, respectively. All models showed low RMSE, indicating high accuracy in predicting bandgap values, with LR achieving the highest r = 0.997. The r-values were greater than 0.94 for all algorithms, indicating a strong correlation between predicted and experimental values. Fig. 3(c) illustrates the bandgap variation of $FA_{0.75}Cs_{0.25}Pb(Cl_{(1-x-y)}Br_xI_y)_3$ with halide ion ratios, showing that Cl^- doping increases the bandgap, while increasing Br doping has a smaller effect. The figure also shows how the bandgap of $Cs_aFA_bMA_{(1-a-b)}Pb(Cl_{0.05}Br_{0.15}I_{0.8})_3$ varies with cation ratios, demonstrating that high Cs concentrations result in a wider bandgap in I-dominated perovskites. To achieve wide bandgaps in I-dominated perovskites, increasing the doping ratios of Cl^- and Cs^+ is crucial.

Chenebuah, ET [60]. employed 12 ML models (Ada-Boost Regression (ABR), Bayesian Ridge Regression (BRR), DT, GBR, GPR, KNN, KRR, Multi-layer Perceptron (MLP), Passive Aggressive Regression (PAR), RF, Stochastic Gradient Descent (SGD) and SVR) to predict the electronic properties of perovskites, specifically on the formation energy and energy bandgap of two crystal configurations: ABX_3 and $A_2BB'X_6$. In Fig. 4 (a), provides a comparative analysis of the performance of 4 best ML (SVR, GBR, DT, RF) models in predicting the formation energy and energy bandgap. The results indicate that SVR is the most effective model for predicting formation energy due to its capability to capture complex data relationships, thus proving reliable for applications in materials science. Conversely, GBR and RFR demonstrated superior performance in predicting the energy bandgap, attributed to their ability to manage larger sample sizes and effectively model complex, non-linear data relationships. This study underscores the critical role of feature selection and dataset diversity in enhancing model performance within materials science applications.

Gao, Z. et al. [59] found that the XGBoost algorithm outperformed both the ANN and SVR algorithms in terms of accuracy (Fig. 4(b)), indicating that XGBoost offers a promising and faster approach for material discovery. Leveraging the high accuracy and efficiency of the ML model, they screened 5796 candidate inorganic double perovskites and identified two new lead-free options which are K_2NaInI_6 and Na_2MgMnI_6. These materials demonstrated desirable characteristics, including appropriate band gaps, high thermal stability, and excellent optical properties. With stable, direct band gaps close to $MAPI_3$, K_2NaInI_6 and Na_2MgMnI_6 are promising candidates for lead-free hybrid PSCs.

Li, C. et al. [58] developed an optimal feature set for predicting the band gap of ABO_3 perovskite materials by considering both chemical composition and structural characteristics. They introduced five structural descriptors that specifically capture the distortion of the B-O octahedron in perovskites. Incorporating these structural features significantly improved the prediction accuracy compared to models based solely on chemical composition, with R^2 values for SVR, bagging, RF, and GBR models increasing by 0.0891, 0.1076, 0.1078, and 0.134, respectively. These descriptors are easy to calculate and offer valuable insights into the structure-performance relationship of perovskite materials.

Li. L. et al. [57] used ML to improve the prediction of perovskite structure formation in ABX_3 and $A_2B'B''X_6$ compounds. They evaluated the performance of five models which is RF, DT, SVC, KNN, and LR using k-fold cross-validation and leave-one-out cross-validations (LOOCV) as shown in Fig. 4(c). The RF algorithm outperformed the others, making it the final choice for modelling. The RF classification model achieved an

Table 5
Summary of recent studies utilizing ML to analyse the bandgap and crystal structure of perovskite materials.

Year	ML Algorithm	Target	Outcome	Ref
Zhu et al., 2024	ANN, GBDT, KNN, RF, XGBoost	Predict bandgap of lead-free halide perovskite	**Best model: XGBoost** • RMSE: 0.20 eV **Model Performance Comparison:** • RMSE: XGBoost > RF	[38]
Moeini et al., 2024	SVR, RF, GBR, XGBoost	Predict bandgap in complex (low symmetry double and layered structure) perovskite	**Best model for double structure: SVR** • R^2: 96% **Model Performance Comparison:** • Direct bandgap R^2: SVR > XGBoost > GBR > RF • Indirect bandgap R^2: SVR > XGBoost = GBR > RF **Best model for layered structure: SVR** • R^2: 96% **Model Performance Comparison:** • Direct bandgap R^2: SVR > GBR > XGBoost > RF • Indirect bandgap R^2: SVR > GBR > XGBoost > RF	[39]
Feng et al., 2024	SVR, RF, XGBoost, Light-GBM (LGBM)	Predict bandgap of organic-inorganic hybrid perovskite	**Best model: XGBoost** • MAE: 0.0901 • MSE: 0.0173 • R^2: 0.99 **Model Performance Comparison:** • R^2: LGBM > LGBM > RF > SVR	[40]
Djeradi et al. 2024	RF, XGBoost, LGBM	Predict bandgap of double perovskite	**Best model: LGBM** • R^2: 0.934 **Model Performance Comparison:** • R^2: LGBM > RF > XGBoost	[41]
Obada et al., 2023	CatBoost, XGBoost, RF, LGBM, DT, CompoundNet	Predict bandgap of ABX_3 perovskite	**Best model for direct bandgap: CatBoost** • R^2: 0.887 **Model Performance Comparison:** • R^2: CatBoost > XGBoost > RF > CompoundNet > DT > LGBM **Best model for indirect bandgap: CatBoost** • R^2: 906 **Model Performance Comparison:** • R^2: CatBoost > XGBoost > RF > CompoundNet > LGBM > DT	[42]
Mishra et al., 2023	LR, SVR, LR, SVR, RF, KNN, XGBoost, AdaBoost	Predict bandgap for perovskite suitable for indoor application	**Best model: RF** • RMSE: 0.0503 eV • r: 0.988 **Model Performance Comparison:** • RMSE: RF > XGBoost > SVR > KNN ≫ KNN > LR > AdaBoost • r: RF > SVR > XGBoost > AdaBoost > KNN > LR	[43]
Zhao et al., 2022	Lasso, KRR, KNN, SVM, DT, Extreme Randomized Tree (ERT), AdaBoost, RF, GBDT, Light-GBDT (LGBT), XGDT, ANN	Predict formation energy & bandgap of ABX_3	**Best model for formation energy: XGDT** • RMSE: 0.269 **Model Performance Comparison:** • RMSE: XGDT > AdaBoost > LGDT > GBDT > KNN > RF > Ridge > Lasso > ERT > ANN > DT > SVM **Best model for bandgap: XGDT** • RMSE: 0.106	[44]

(*continued on next page*)

Table 5 (*continued*)

Year	ML Algorithm	Target	Outcome	Ref
			Model Performance Comparison: • RMSE: XGDT > GBDT > LGDT > RF > AdaBoost > KNN > Ridge > ERT > SVM > DT > Lasso > ANN	
Yilmaz et al., 2022	XGBoost, RF, ANN	Predict bandgap of 2D/3D perovskite	**Best model: XGBoost** • RMSE: 0.14	[45]
			Model Performance Comparison: • RMSE: XGBoost > RF > ANN	
Wang et al., 2022	GBDT, RF, DT, LR	Predict bandgap of double perovskite	**Best model: GBDT** • Accuracy: 92%	[46]
			Model Performance Comparison: • Accuracy: GBDT > DT > RF > LR	
Wang et al., 2022	Stacking, GBDT, XGBoost, RF, LGBM, SVR, Linear (LR, RR, Elastic-R)	Predict bandgap of inorganic compound for perovskite	**Best model: Stacking** • R^2: 0.92 • MAPE: 0.084 • RMSE: 0.634	[47]
			Model Performance Comparison: • R^2: Stacking > XGBoost > GBDT > RF > LGBM > SVR > Linear • MAPE & RMSE: Stacking > XGBoost > RF > GBDT > LGBM > SVR > Linear	
Rath et al., 2022	XGBoost, RF, LR, SVC, ANN, DT, Gradient Boost Classifier (GBC)	Predict bandgap of ABX_3 perovskite	**Best model: XGBoost** • Precision: 81%	[48]
			Model Performance Comparison: • Precision: XGBoost > RF > GBC > SVC > DT > LR > ANN	
Liu et al., 2022	XGBoost, RF, MLP, LR, SVR, KNN	Predict bandgap of ABX_3 perovskite	**Best model: XGBoost** • RMSE: 0.55 • r: 0.99	[49]
			Model Performance Comparison: • RMSE & r: XGBoost > RF > MLP > LR > SVR > KNN	
Kanakkithodi et al., 2022	GPR, NN, RF	Detect defect and impurities in $MAPbX_3$ (X = Cl, Br, I)	**Best model: GPR**	[50]
Hu et al., 2022	KNN, Kriging, RF, Rpart, SVM, XGBoost	Evaluate interaction between ions and 2D halide perovskite, focus on adsorption energy	**Best model: XGBoost** • R^2: 0.93 • r: 0.97	[51]
			Model Performance Comparison: R^2: XGBoost > Kriging	
Cai et al., 2022	LR, SVR, KNN, RF, GBR	Predict relationship between band gap and Sn-Pb composition in $MASn_xPb_{1-x}I_3$	**Best model: GBR** • R^2: 0.9172 • RMSE: 0.0386 • MAE: 0.0325	[52]
			Model Performance Comparison:	
Zhang et al., 2021	XGBoost, GBC, SVC, KNN	Predict formability of hybrid organic−inorganic perovskites (HOIP)	• Not mention **Best model: XGBoost** • Accuracy: 0.88	[53]
			Model Performance Comparison:	
Yang et al., 2021	SVC, LR, RFC, SVR, BPANN, RFR	Discover $A_2B'B''O_6$-type perovskites	• Not mention **Best model: SVC & SVR** • Accuracy: 0.968 • r: 0.919	[54]
			Model Performance Comparison:	
Priya et al., 2021	LR, SVM, NN, KNN, RF, XGBoost	Predict conductivity of ABO_3 perovskite	• Not mention **Best model: XGBoost** • RMSE: 0.24 • R^2: 0.987	[55]

(*continued on next page*)

Table 5 (*continued*)

Year	ML Algorithm	Target	Outcome	Ref
Li et al., 2021	LR, RF, NN	Predict bandgap of $Cs_aFA_bMA_{(1-a-b)}Pb(Cl_XBr_yI_{(1-x-y)})_3$	**Model Performance Comparison:** • R^2: XGBoost > RF > KNN > NN > SVM = LR **Best model: LR** • RMSE: 0.032 eV • r: 0.997	[56]
Li et al., 2021	RF, DT, SVC, KNN, LR	Predict formability ABX_3 and A_2 B′B″X6	**Model Performance Comparison:** • RMSE: LR > NN > RF • r: LR > NN > RF **Best model: RF** • Accuracy: 0.968	[57]
Li et al., 2021	SVR, RF, Bagging, GBR	Predict bandgap of ABO_3 perovskite based on chemical composition and physical structure	**Model Performance Comparison:** • Accuracy, F1, Precision: RF > DT > LR > SVC > KNN • Recall: RF > SVC > LR > DT > KNN **Best model: GBR** • R^2: 0.134	[58]
Gao et al., 2021	XGBoost, SVR, ANN	Predict bandgap and stable lead-free perovskite materials	**Model Performance Comparison:** • R^2: GBR > RF > Bagging > SVR **Best model: XGBoost** • MSE: 0.102 • R^2: 0.956 • r: 0.977	[59]
Chenebuah et al., 2021	ABR, BRR, DT, GBR, GPR, KNN, KRR, MLP, PAR, RF, SGD, SVR	Predict the electronic properties (formation energy and bandgap) of ABX_3 and $A_2BB'X_6$	**Model Performance Comparison:** • R^2: XGBoost > SVR > ANN **Best model for formation energy: SVR** • MAE: 0.055 eV/atom • RMSE: 0.096 eV/atom • R^2: 99%	[60]
Xie et al., 2020	SISSO, SVM, NN, DT, RF	Predict octahedral tilt in oxide perovskite	**Model Performance Comparison:** • MAE: SVR > GBR > RF > others (not mention in detail) **Best model for bandgap: GBR & RF** • Metrix does not mention **Best model: SISSO** • Accuracy: 86.7	[61]
Park et al., 2020	DL, XGBoost	Predict $A_XA'_{(1-X)}BX_3$ bandgap	**Model Performance Comparison:** • Accuracy: SISSO > NN > DT > RF > SVM **Best model: XGBoost** • MAE: 0.12 eV • R^2: 0.97	[62]
Liu et al., 2020	LR, SVM, RF, GBDT	Predict stable and metastable ABO_3 perovskite	**Model Performance Comparison:** • R^2: XGBoost > DL **Best model: GBDT** • Accuracy: 0.946 • F1 score: 0.980	[63]
Park et al., 2019	DL, GBM, GLM, RF, XGBoost	Predict influence of A-site cations on the phase stability perovskite	**Model Performance Comparison:** • Accuracy: GBDT > RF > SVM > LR **Best model: XGBoost** • R^2: 0.99 ΔHc, 0.88 σ^2, 0.87 λ	[64]
Li et al. 2019	LR, KNN, SVR, RF, ANN	Predict bandgap of ABX_3 from material composition synthesize new materials	**Model Performance Comparison:** • R^2: XGBoost > GBM > DL > RF > GLM **Best model: ANN** • RMSE: 0.06 eV • r: 0.97	[65]

(*continued on next page*)

Table 5 (*continued*)

Year	ML Algorithm	Target	Outcome	Ref
			Model Performance Comparison: • RMSE: ANN < RF = SVR < KNN < LR • r: ANN = SVR > RF = KNN = LR	

accuracy of 96.55% for ABX_3 compounds and 91.83% for $A_2B'B''X_6$ compounds. Using this model, 241 ABX_3 perovskites and 1131 $A_2B'B''X_6$ perovskites with high probabilities of formability were successfully identified from large datasets of candidate compounds. This approach provides valuable insights for accelerating the discovery of perovskite materials.

Yang, X. et al. [54] developed a ML-based strategy to accelerate the discovery of $A_2B'B''O_6$-type perovskites. Using a perovskite classification model, they found the SVC model outperformed others, achieving LOOCV and test set accuracies of 0.967 and 0.968, respectively. For bandgap prediction, the SVR model performed best, with R-values of 0.924 and 0.919. They screened 3098 perovskites from 6529 virtual samples and identified 60 oxide double perovskites with bandgaps between 1.00 and 1.60 eV. Notably, 19 perovskites, with bandgaps close to 1.34 eV, are promising for solar cell applications.

Cai, X. et al. [52] propose and realize a novel ML approach based on forward-reverse framework to establish the relationship between key parameters and photovoltaic performance in high-profile $MASn_xPb_{1-x}I_3$ perovskite materials. The task of forward analysis procedure is to establish the bandgap (Eg) and device-performance models using ML, and that of reverse engineering procedure is to predict the optimization parameters of mixed Sn-Pb perovskites and experimental realization. Predicted of bandgap by 5 ML algorithms LR, SVR, KNN, RF and GBR shows GBR give the best result with the corresponding values of R^2, RMSE and MAE of 0.9172, 0.0386, and 0.0325, respectively.

Liu, Y. et al. [49] applied ML techniques to predict the bandgaps of ABX_3-type perovskites. They compiled a dataset of 227 experimental bandgap values from 1254 recent publications on perovskites. To build their prediction models, they employed six ML algorithms: XGBoost, RF, MLP, LR, SVR, KNN. Among these models, XGBoost demonstrated the best performance, achieving a RMSE of 0.55 and r of up to 99%. By using SHAP, they identified that the composition of the B-site metal and X-site halogen ions were the most influential factors in determining the bandgap of ABX_3-type perovskites.

Rath, S. et al. [48] also utilized ML techniques to predict the bandgaps of ABX_3-type perovskites. They employed a dataset from the MP database, which contained 1528 ABX_3 compounds (where X = O, F, Cl, Br, I, S, Se, Te, N, or P), along with information on the nature of their direct and indirect bandgaps. Seven ML algorithms were used in their study: XGBoost, RF, LR, SVC, ANN, DT, and GBC. They found that XGBoost provided the best accuracy, achieving 72.8%. Through SHAP analysis, they discovered that the absence of transition metals and elements from groups IIIA to VIIIA with atomic numbers greater than 20 increased the likelihood of the perovskite exhibiting a direct bandgap.

Wang, T. et al. [47] explored ML for predicting semiconductor band gaps, highlighting the need for further optimization. explored ML for predicting semiconductor band gaps, highlighting the need for further optimization. They used a stacking approach (Fig. 5(a)), combining multiple baseline models to generate a new ensemble model to improve prediction accuracy. Ten baseline models were optimized, and their outputs fed into the stacking model. The research utilized two datasets, a benchmark with 3896 inorganic compounds and the E-AFLOW database with 21,534 compounds. The stacking model achieved high performance in predicting the bandgap (Fig. 5(b)), with R^2 values of 0.917, MAPE 0.084, RMSE of 0.634 and outperformed individual models by up to 17.54%. The model predicted band gaps of new materials with 78.57% accuracy within ±8%.

Based on Table 5, ensemble models, particularly XGBoost, often outperform other model categories such as linear and trees-based model in predicting material properties like band gap and crystal structure, due to their superior ability to balance bias and variance, capture complex nonlinear relationships, and generalize well on structured scientific datasets. Linear models, while simple and interpretable, assume a direct linear relationship between features and the target variable, which fails to capture the nonlinear and complex interactions inherent in materials data such as atomic interactions, lattice distortions, or dopant effects leading to underfitting and poor predictive accuracy. Similarly, trees-based model, although capable of handling nonlinearity, tend to overfit the training data and are highly sensitive to small variations, resulting in high variance and poor generalization on unseen samples. In contrast, ensemble methods like GBDT, and particularly their optimized implementations such as XGBoost and CatBoost, aggregate the predictions of multiple trees to reduce overfitting and enhance predictive robustness. XGBoost, as a highly optimized form of gradient boosting, goes further by incorporating regularization, efficient tree pruning, and parallel computation, making it more accurate, scalable, and interpretable. It consistently performs better than other methods by addressing both bias and variance more effectively, particularly when using engineered descriptors common in materials science, such as elemental features, bonding characteristics, and crystallographic parameters.

5.2. Device optimization

ML has also been widely applied to optimize device performance metrics of PSCs. This section is divided into two critical aspects: efficiency prediction and stability analysis.

5.2.1. Efficiency prediction

The studies summarized in Table 6 reflect diverse ML approaches employed to predict and optimize PSCs efficiency based on various input parameters, including material composition, interface engineering, and fabrication conditions.

Li, J. et al. [65] predict the V_{OC}, J_{SC}, FF and PCE of PSCs based on the bandgap (Eg), the energy difference between HOMO of HTL and perovskite material (ΔH), and LUMO energy difference between the ETL and perovskite materials (ΔL). ANN outperforms other algorithms with the smallest RMSE (3.27%) and highest correlation (r = 0.72). Due to its superior performance, ANN is used to optimize and predict PSCs PCE. Fig. 6(a) displays a 4D scatter plot of predicted PCE, showing that lower ΔH and ΔL values benefit low-bandgap materials, while higher values favour high-bandgap materials for higher PCE. The predicted results align closely with the Shockley–Queisser limit (Fig. 6(b)), suggesting that PSCs with bandgaps between 1.15 and 1.35 eV could achieve over 25% PCE under AM 1.5G illumination.

Cai, X. et al. [52] used the bandgap result with best GBR ML algorithm out of other 5 ML algorithms predict the PCE of n-i-p and p-i-n of PSCs using 6 ML algorithms, where 5 is same as bandgap prediction and add on NN algorithm. Based on the prediction, NN models behaves the best among competitive algorithms due to higher R^2 (0.9096 for PCE, 0.8378 for *Jsc*, 0.9024 for *Voc* and 0.9026 for FF and smaller RMSE and MAE.

Cueto, M.d. et al. [77] used KRR algorithm to investigate how hole-transporting material (HTM) impact PCE of PSCs. The HTM features considered included molecular fingerprints, structural characteristics, electronic properties, and types of additives. The model was trained using two datasets: a heterogeneous database containing 269

Table 6
Overview of recent studies using ML to analyse and optimize the performance of PSCs.

Year	ML Algorithm	Target	Outcome	Ref
Saqib et al., 2025	GBR, Bagging, RF, LR	Predict PCE based on HTM	**Best model: GBR** • R^2: 0.89 **Model Performance Comparison:** • R^2: GBR > RF > Bagging > LR	[66]
Zhu et al., 2024	ANN, GBDT, KNN, RF, XGBoost	Predict PCE of lead-free halide perovskite	**Best model: XGBoost** • RMSE: 1.77% **Model Performance Comparison:** • RMSE: XGBoost > RF	[38]
Zhou et al., 2024	SEResNet, CNN, RF, KNN, DT, SVR, LR	Predict PCE based on relationship between material properties, energy level, and device properties	**Best model: SEResMet** • RMSE: 0.833 • r: 0.98 **Model Performance Comparison:** • RMSE & r: SEResNet > CNN > RF	[67]
Robert et al., 2024	XGBoost, NN, GPR	Predict J_{SC} in metal-halide perovskite	**Best model: XGBoost** • RMSE: 3.58 • R^2: 0.35 **Model Performance Comparison:** • RMSE & R^2: XGBoost, GPR, NN	[68]
Malek et al., 2024	DT, RF, KNN, MLP	Predict V_{OC}, J_{SC}, FF, PCE values based on HTL and back contact	**Best model: MLP** • RMSE: 0.0005 • R^2: 0.99998 **Model Performance Comparison:** RMSE & R^2: MLP > RF > DT > KNN	[69]
Liu et al., 2024	SVM, KNN, RF, XGCoost, MLP	Predict V_{OC}, J_{SC}, FF, PCE, EQE curve based on bandgap of perovskite and mobility of transport layer	**Best model: XGBoost** • RMSE: 1.25% • r: 0.94 **Model Performance Comparison:** RMSE & r: XGBoost > RF > MLP > KNN > SVM	[70]
Lee et al., 2024	XGBoost, NN, GPR	Predict PCE based on materials utilized in each layer and fabrication process	**Best model: XGBoost** • MAE: 0.963 • R^2: 0.774 **Model Performance Comparison:** • MAE & R^2: XGBoost > GPR > NN	[71]
Zhi et al., 2023	MLP, SVM, GP, RF, GB, Ensemble	Predict PCE based on interface passivation materials to form 2D perovskite	**Best model: Ensemble** • MAE: 0.0484 **Model Performance Comparison:** MAE: Ensemble > GB > RF > GP > SVCM > MLP	[72]
Salah et al., 2023	RF, GBR, KNN, LR	Predict PCE based on perovskite and transport layer material, thickness, doping and defect	**Best model: RF** • RMSE: 1st data – 0.003%, 2nd data – 0.0817%, 3rd data – 0.1815% **Model Performance Comparison:** • RMSE: RF = GBRt > KNN > LR	[73]
Lu et al., 2023	LR, NN, RF, XGBoost, GBDT	Predict PCE based on fabrication parameter	**Best model: XGBoost** • RMSE: 1.28% • r: 0.768 **Model Performance Comparison:** RMSE & r: XGBoost > RF > GBDT > NN > LR	[74]
Yilmaz et al., 2022	XGBoost, RF, ANN	Predict PCE of 2D/3D perovskite based on n-i-p and p-i-n structure	**Best model for n-i-p: RF** n-i-p structure • RMSE: 3.81 p-i-n structure • RMSE: 3.943 **Model Performance Comparison:** • RMSE: RF > XGboost > ANN	[45]
Liu et al., 2022	LR, RF, XGBoost, NN	Predict PCE and interface passivation material at the atomic level	**Best model: RF** • RMSE: 0.64% • r: 0.90 **Model Performance Comparison:** • RMSE & r: RF > NN > XGBoost > LR	[75]
Hu et al., 2022	RR, SVR, GPR, Polynomial least-squares regression (PLSR)	Predict V_{OC}, J_{SC}, FF, PCE, stability based on bandgap, trap density, grain size, carrier lifetimes, and surface roughness	**Best model: SVR** • R^2: 0.94 **Model Performance Comparison:** • R^2: SVR > GPR > PLSR > RR	[76]
Cueto et al., 2022	KRR	Predict PCE based on HTM	**Best model: KRR** Homogeneous database • RMSE: 2.7% • r: 0.57	[77]

(*continued on next page*)

Table 6 (*continued*)

Year	ML Algorithm	Target	Outcome	Ref
			Heterojunctions database • RMSE: 3% • r: 0.72 **Model Performance Comparison:** • Not mention	
Cai et al., 2022	LR, SVR, KNN, RF, GBR, NN	Predict V_{OC}, J_{SC}, FF, PCE based on predictive bandgap	**Best model: NN** • RMSE: 1.9626 • MAE: 1.5439 • R^2: 0.9026 **Model Performance Comparison:** • RMSE, MAE, R^2: NN > GBR > RF > KNN > SVR > LR	[52]
She et al., 2021	DT, ET, RF, Adaboost, GBoost, XGBoost, MLP, SVM	Predict PCE based on doped ETL materials	**Best model: RF** • Accuracy: >66.48% **Model Performance Comparison:** • Accuracy: RF > DT > AdaBoost > SVM > XGBoost = ET > GBoost > MLP	[78]
Corre et al., 2021	DT, RF	Predict dominant recombination process by different light intensity, doping and mobilities	**Best model: RF** • Accuracy: >82% **Model Performance Comparison:** Accuracy: RF > DT	[79]
Li et al., 2019	LR, KNN, SVR, RF, ANN	Predict V_{OC}, J_{SC}, FF, PCE based on predictive bandgap, HOMO HTL-perovskite (ΔH), LUMO ETL-perovskite (ΔL)	**Best model: ANN** • RMSE: 3.27% • r: 0.72 **Model Performance Comparison:** • RMSE: ANN > RF > SVR > KNN > LR • r: ANN > SVR > RF > KNN	[65]

entries and a homogeneous subset of this data. By integrating multiple HTM features, the model's ability to predict PCE was enhanced. The study revealed that certain molecular components in HTMs, such as arylamine and thiophene groups, showed a strong correlation with the average PCE.

Liu, W. et al. [75] applied ML to analyse the relationship between PCE and interface passivation materials at the atomic level, aiding efficient material screening. They studied ~100 perovskite/HTL interface materials using LR, RF, XGBoost, and NN models. Among them, the RF model performed best on the test set, achieving an RMSE of 0.7%. In contrast, XGBoost, though strong on the training set, overfitted and failed on the test set due to the small dataset and numerous features, which hyperparameter tuning could not fix. Their high-throughput predictions, supported by DFT calculations, revealed that materials with high binding energy to the $[PbI_4]^{2-}$ surface and small organic

Table 7
Overview of recent studies using ML to analyse and enhance the stability of PSCs.

Year	ML Algorithm	Target	Outcome	Ref
Chen et al., 2025	LR, RF, GBDT, XGBoost	Predict stability of ABX_3, focus on T_{80}, T_{90} and aging curves	**Best model: XGBoost** n-i-p structure • R^2: 0.9984 • RMSE: 33.73 h p-i-n structure • R^2: 0.9882 • RMSE: 118.18 h **Model Performance Comparison:** • R^2: XGBoost > GBDT > RF > LR	[80]
Dunlap-Shohl et al., 2024	LR with GFS, Lasso, RR	Predict stability on T_{80} of $CH_3NH_3PbI_3$ at varying temperature and relative humidity	**Best model: Lasso** • R^2: 0.7014 **Model Performance Comparison:** • R^2: Lasso > RR > LR(GFS)	[81]
Mammeri et al. 2023	DT, ET	Predict stability by trained device sample with different materials, deposition methods and storage condition	Does not compare different ML models but identified optimized conventional and inverted structure materials	[82]
Graniero et al. 2023	RF, Elastic Net (eNet)	Predict stability on T_{80} from Perovskite Database and in-house database	Does not compare different ML models but underscores the importance of data quality in improving model performance for analysing PSCs stability	[83]
Liang et al., 2022	RF, Bagging, ET, DT, AdaBoost, XGBoost	Predict thermodynamic phase stability of lead-free halide double perovskite	**Best model: XGBoost** • Accuracy: 0.882 • Precision: 0.884 • Recall: 0.8883 • F1 Score: 0.884 **Model Performance Comparison:** • Accuracy, Precision, Recall, & F1 Score: XGBoost > ET > RF > Bagging > AdaBoost > DT	[84]
Zhao et al. 2021	GBT	Predict stability on cation engineering in PSCs structure at different temperature	**Best model: NMF** **Model Performance Comparison:** • Not mention	[85]
Kim et al., 2021	Non-negative Matrix	Predict stability of charge transport	**Best model: NMF** **Model**	[86]

(*continued on next page*)

Table 7 (*continued*)

Year	ML Algorithm	Target	Outcome	Ref
	Factorization (NMF)	dynamics under 1 sun and environment stimuli	**Performance Comparison:** • Not mention	
Guo et al., 2021	RF, RR, SVR, XGBoost	Predict stability of lead-free halide double perovskite materials	**Best model: XGBoost** • R2: 0.9935 • MAE: 0.0126 **Model Performance Comparison:** • R^2 & MAE: XGBoost > RR > RF > SVM	[87]
Higgins et al., 2020	NMF, GPR	Predict long-term stability of metal halide perovskite in ambient conditions	**Best model: Both Model Performance Comparison:** • Not mention	[88]
Hartono et al., 2020	RF, GBR	Predict stability on effect of capping layer materials	**Best model: Both Model Performance Comparison:** • Not mention	[89]
Hartono et al., 2020	LR, KNN, RF, GBR-DT, NN, SVR	Predict stability on effect of capping layer materials	**Best model: RF** • RMSE: 104 min **Model Performance Comparison:** • RMSE: RF > GBR > NN > KNN > LR > SVR	[90]
Li et al. 2019	KRR, KNN, SVR	Predict perovskite stability with compositional ionic radii.	**Best model: KRR** • RMSE: 25.5 meV/atom • R2: 0.92 **Model Performance Comparison:** • RMSE: KRR > KNN > SVR • R^2: KRR > KNN=SVR	[91]
Li et al. 2018	ET, ANN, SVM, DT, Logistic Regression (LR1), Linear Regression (LR2) KRR	Predict thermodynamic phase stability.	**Best model fpr classification: ET** • Accuracy: 0.93 • Precision: 0.89 • Recall: 0.87 • F1 score: 0.88 **Model Performance Comparison:** • Accuracy: ET = ANN=SVM > DT > LR1 • Precision: ET = SVM > ANN > DT > LR1 • Recall: ET = LR1>SVM = ANN > DT • F1 score: ET = SVM > ANN > DT > LR1	[92]

Table 7 (*continued*)

Year	ML Algorithm	Target	Outcome	Ref
			Best model fpr regression: KRR • RMSE: 28.5 meV/atom • MAE: 16.7 meV/atom • R^2: 0.894 **Model Performance Comparison:** • RMSE & R^2: KRR > ET > ANN > DT > LR2 • MAE: ET > KRRT > ANN > DT > LR2	

cations with NH_3^+ terminals exhibit strong passivation effects. Experimental validation with MAI and phenethylammonium iodide confirmed the predictions. This study enables rapid screening and design of interface materials for efficient PSCs and establishes general screening rules at the atomic level.

Salah M. M. et al. [73] employed a systematic ML approach to predict and optimize the PCE of PSCs. They developed three datasets of varying sizes and complexities for $CsPbX_3$ (X = I, Br, or Cl) PSCs and evaluated four ML regression models: RF, GBR, KNN, and LR. The first dataset linked material properties such as porosity and ETL layer thickness to PCE, the second focused on ETL thickness, doping, and defects, while the third included nine inputs related to HTL, ETL, and perovskite thin films. The study revealed that doping in the ETL had a significant impact on PCE, with its importance varying across datasets.

Zhao, S. et al. [67] propose a prediction model to analyse the relationship between perovskite material properties, energy level alignment, device structure, and PCE. They generate data for the chemical formula $MA_{1-x}FA_xPb_{1-y}Sn_yI_{3(1-z)}Br_{3z}$ (where x, y, and z range from 0 to 1 in increments of 0.05) to study how different features affect PCE. Fig. 6 (c) shows that PCE peaks at x = 1, y = 0, and z = 0 ($FAPbI_3$). In Fig. 6(d), the model's predicted PCE values are compared with experimental results, demonstrating a strong correlation with a RMSE of 0.762% and a r of 0.973. These findings confirm the model's accuracy in predicting PSCs performance trends.

Liu, Y. et al. [70] segregated and clean the data from Perovskite Database (2016–2022) to achieve three goals: accurate prediction of device performance for PSCs, ML analysis of influencing factors, and high-throughput PSCs design. They used ML models (SVM, KNN, RF, XGBoost, MLP-NN) to assess device performance. The GPSR algorithm and SHAP technique were employed to link macroscale performance with microscale factors. The XGBoost model outperformed RF model with the best performance, achieving a RMSE of 1.25% PCE, 1.43 mA/cm^2 J_{SC}, 0.049V V_{OC}, and 4.06% EQE (Fig. 7(a)). Fig. 7(b) shows SHAP analysis revealed that key features influencing performance include the A-site element (FA^+ content) for PCE, the X-site element (Br^- content) for J_{SC}, and the B-site element (Pb content) for V_{OC} and FF.

Saqib et al. [66] use four ML model (GBR, Bagging, RFR and LR) for designing and screening of HTM for PSCs applications. ML algorithms were proficient to forecast reorganization energy with the GBR model emerging as the best, achieving R^2 values of 0.89.

In predicting efficiency-related parameters based on Table 6, ensemble models once again demonstrate superior performance, with tree-based models also showing strong results. Among them, the ensemble model XGBoost and the tree-based model RF stand out due to their effective handling of nonlinear relationships. The correlation between PSCs material features such as band gap, ionic radius,

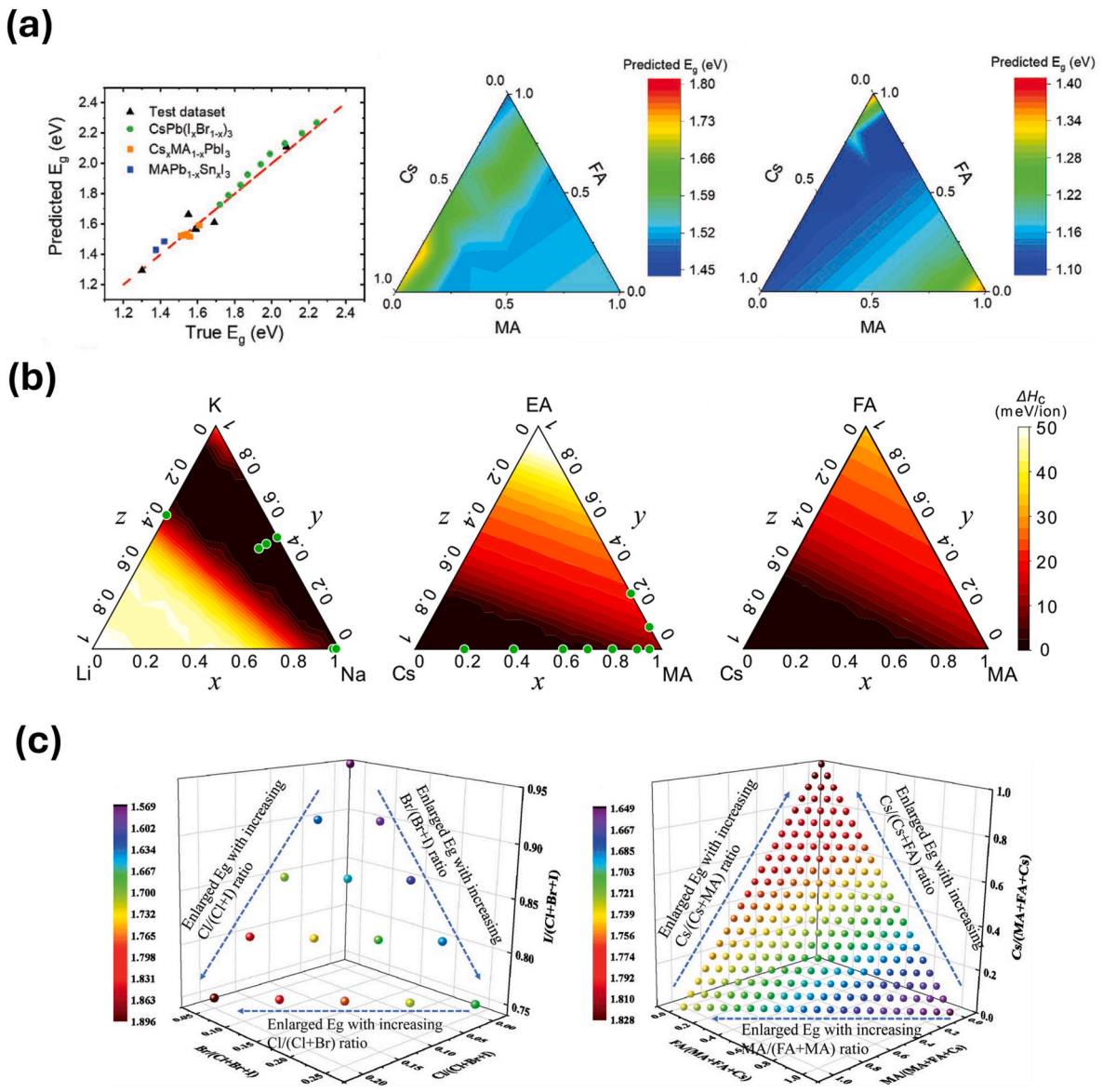

Fig. 3. (a) Performance of the PSCs material design model by the ANN algorithm showing the correlation between the true data and predicted results and ternary contour plots of Cs/MA/FAPbI$_3$ and Cs/MA/FASnI$_3$. Reproduced with permission from Ref. [65]. Copyright (2019) Wiley. (b) XGBoost model in ΔHc prediction of Li$_x$Na$_y$K$_z$NbO$_3$ (left), Cs$_x$MA$_y$EA$_z$PbI$_3$ (middle), and Cs$_x$MA$_y$FA$_z$PbI$_3$(right). Reproduced with permission from Ref. [64]. Copyright (2019) American Chemical Society. (c) 4D plots of the predicted bandgaps by change the ratio of halide ions, while the cations ratios of FA, MA, and Cs are fixed to be 0.75, 0, and 0.25 (left) change the ratio of cations, while the halide ratios of Cl, Br, and I are fixed to be 0.05, 0.15, and 0.8 (right). Reproduced with permission from Ref. [56]. Copyright (2021) Royal Society of Chemistry.

electronegativity and PCE is inherently complex and nonlinear. Both RF and XGBoost, as tree-based ensemble methods, are well-suited for capturing these interactions without requiring prior feature transformation. Additionally, both models exhibit strong resistance to overfitting, albeit through different mechanisms. RF achieves this by averaging predictions from multiple uncorrelated trees (bagging), while XGBoost uses boosting with regularization, pruning, and iterative error correction. These characteristics make both models highly effective for PSCs research, where datasets are often of medium size and contain moderate noise. Furthermore, their ability to generalize well with limited data makes them particularly advantageous compared to deep learning approaches, which typically require larger datasets to perform reliably. In contrast, linear regression consistently underperforms due to its assumption of a linear relationship between inputs and output, which fails to capture the inherently nonlinear nature of material–property interactions in PSCs, resulting in underfitting and poor predictive accuracy.

5.2.2. *Stability analysis*

ML has been employed to optimize the stability of PSCs, with various case studies and applications compiled in Table 7.

Li, W. et al. [92] developed ML models to predict the thermodynamic stability of perovskite oxides. Using 5 models' classification (LR, SVR, DT, ET, ANN) and 5 models for regression (LR, KRR, DT, ET, ANN) they trained models on a dataset of over 1900 DFT-calculated perovskite oxide energies. Fig. 8(a) shows the models achieved high accuracy of 0.93 and F1 score of 0.88 for classification and minimum RMSE of 28.5 meV/atom, 16.7 meV/atom of MAE and highest R^2 of 0.894 for regression.

Hartono, N. T. P. et al. [90] studied the stability of 21 organic halide salts used as capping layers on MAPbI$_3$ films. They sourced these materials from the PubChem database and aged them under accelerated

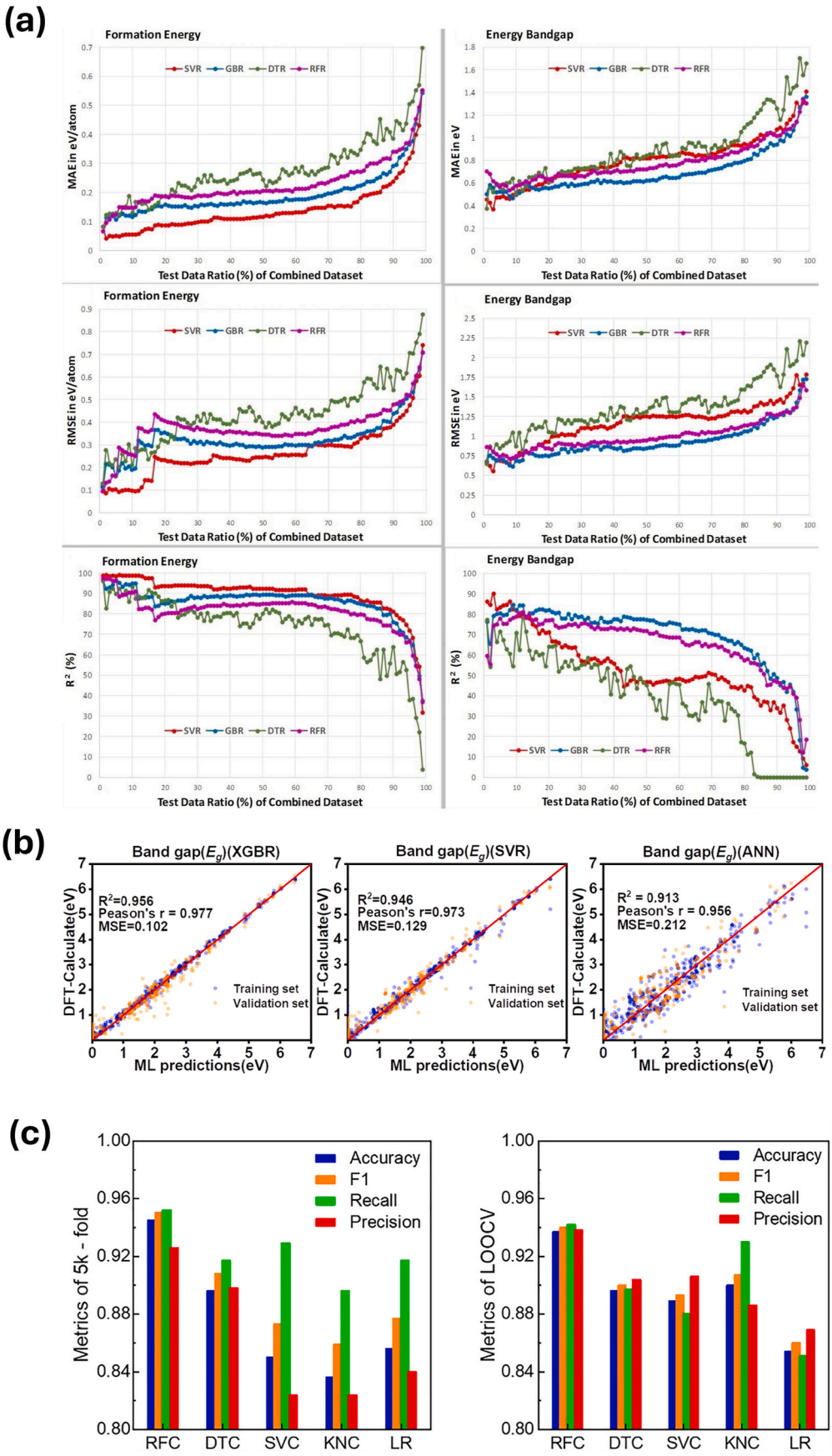

Fig. 4. (a) Comparison of 4 ML algorithm (SVR, GBR, DT, RF). Reproduced with permission from Ref. [60]. Copyright (2021) Elsevier. (b) Comparison of 3 ML algorithm (XGBoost, SVR, ANN). Reproduced with permission from Ref. [59]. Copyright (2021) Elsevier. (c) Comparison of 5 ML algorithm (RF, DT, SVC, KNN, LR). Reproduced with permission from Ref. [57]. Copyright (2021) Elsevier.

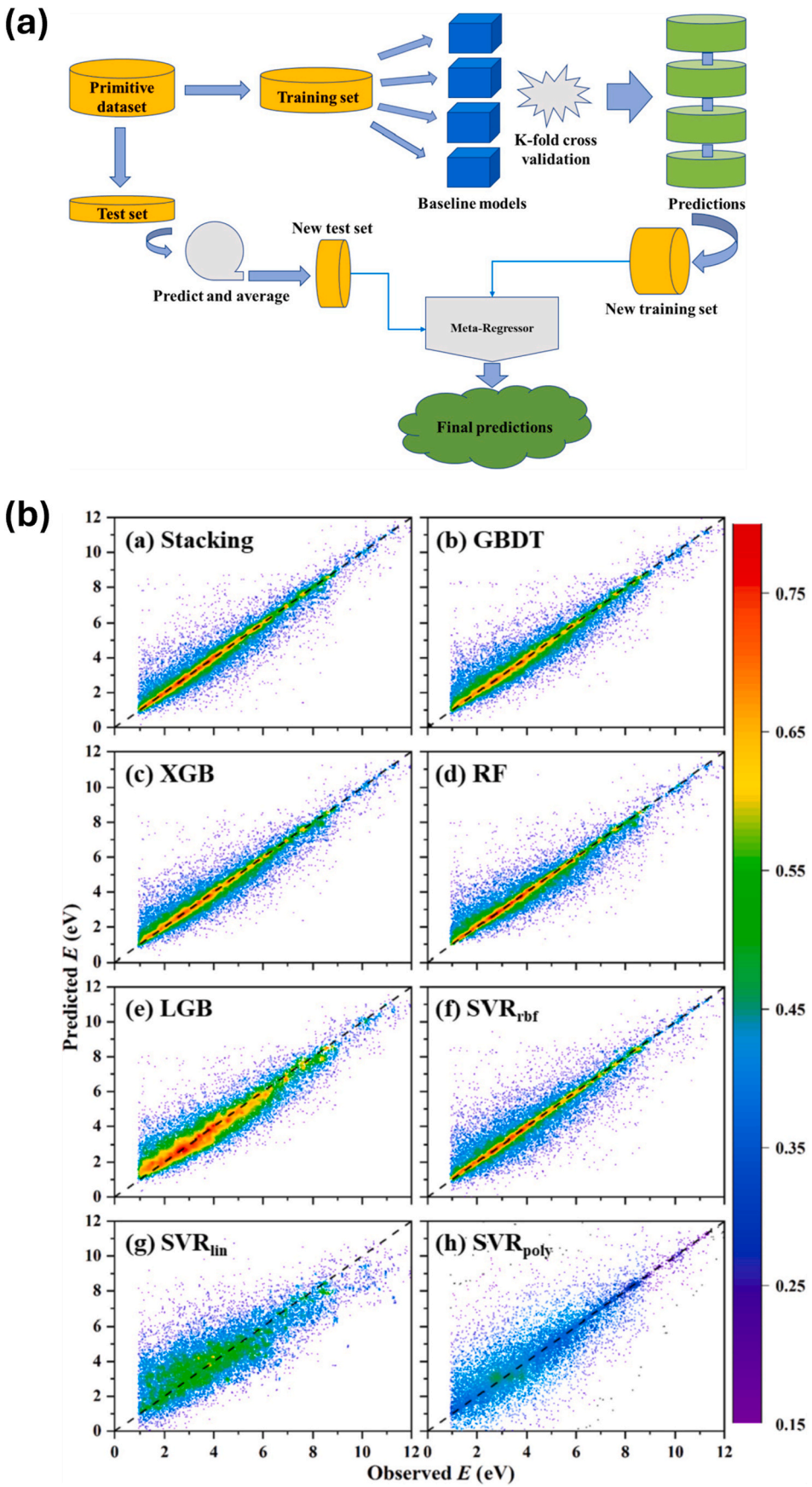

Fig. 5. (a) Flowchart illustrating the ML regression using a stacking approach. (b) Comparison of observed versus predicted bandgap values for stacking, GBDT, XGB, RF, LGB, and SVR. The colormap indicates data density, with the ideal correlation between observed and predicted bandgap represented by the 45° dashed line. Reproduced with permission from Ref. [47]. Copyright (2022) Elsevier.

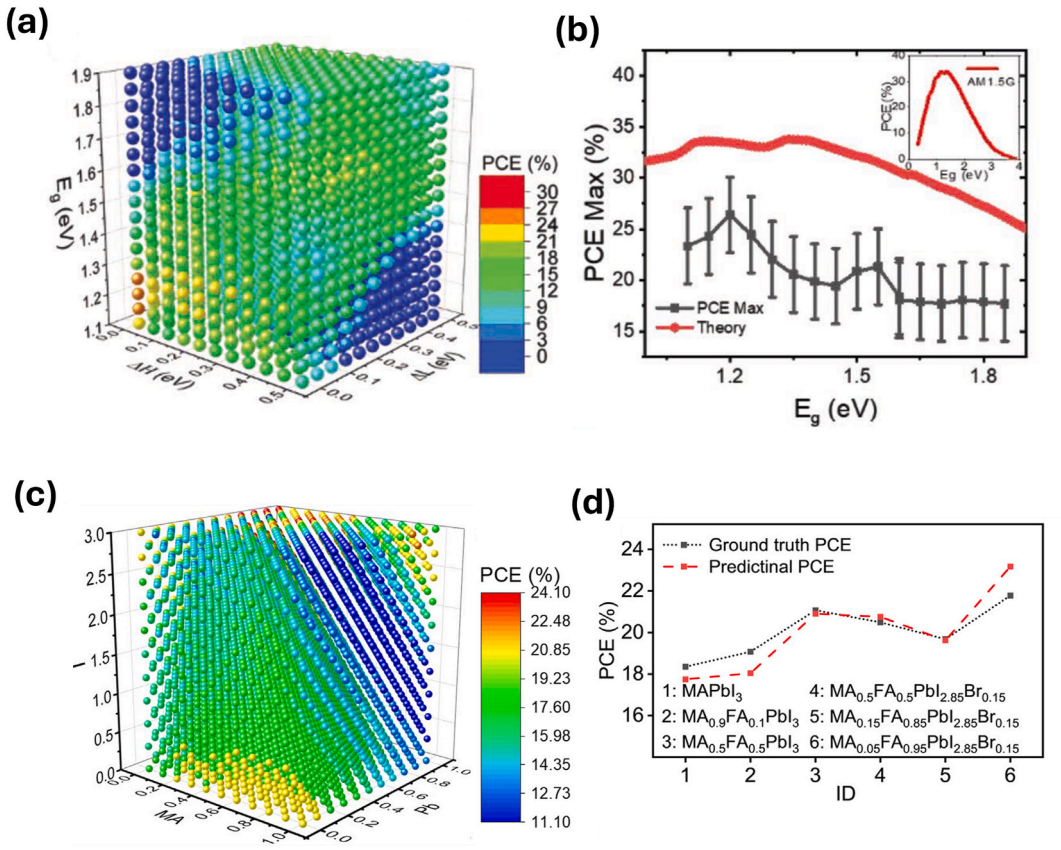

Fig. 6. (a) 4D plot of PCE vs. Eg, ΔH, and ΔL, showing the highest PCE values occur between 1.2 and 1.3 eV. (b) The black line shows the maximum PCE from (a) as a function of Eg, while the red line represents the Shockley–Queisser limit. Reproduced with permission from Ref. [65]. Copyright (2019) Wiley. (c) 4D plot of PCE with respect to Pb^{2+}, I^{-}, and MA^{+}. (d) Comparison graph of experimental and predicted data (ID represents number of 6 different sets of PSCs samples. Reproduced with permission from Ref. [67]. Copyright (2019) Wiley.

conditions, varying annealing temperature and precursor solution concentration. Stability was assessed using six ML algorithms and SHAP values. The degradation was monitored by observing color changes in images, with the transition from black to yellow being a key stability indicator. Fig. 8(b) upper-left shows 21 materials were grouped into four categories: organic materials with fewer than five carbon atoms, long-chain organics, phenyl-based organics, and branched organics. Phenyl-based and branched organics were found to improve stability. Fig. 8(b) upper-right shows out of 6 ML algorithm to calculate RMSE, RF gives the lowest value of 104 min followed by GBR, NN, KNN, LR and SVR. Fig. 8(b) bottom-left demonstrates the randomly-split 20%:80% test:train set observation and prediction results of RF regression and LR, which shows that the RF model is the best-performing model for our data, without the evidence of overfitting demonstrated that phenyltriethylammonium iodide (PTEAI), as the best-performing capping material. Fig. 8(b) bottom-right analyse the RF results using SHAP. The yellow color corresponds to high value of the features consisting of molecular properties and processing conditions, whereas the purple colour corresponds to low value of the features. If SHAP values (x-axis) is positive, the degradation onset increases hence the film lasts longer, and vice versa. y-axis listing all the 14features, are ranked based on their contribution to the degradation onset.

In separate study by Hartono, N.T.P et al., [89] the topological polar surface area and the number of hydrogen-bond donors were identified as the most significant features for capping layer stability, as determined through the RF and GBR algorithms. While a direct comparison was not provided, the RF algorithm reveal that capping materials based on tetrapheptyl-ammonium iodide (TPAI) and PTEAI in $MAPbI_3$ demonstrated degradation onsets times of 353 ± 66 min and 462 ± 115 min, respectively.

Higgins, K. et al. [88] employed a combination of robotic synthesis and ML to explore the stability of multicomponent lead halide perovskites. They used NMF and GPR to analyse the photoluminescent properties of 368 compositions, revealing nonuniform stability across the composition space.

Guo, Z. et al. [87] presents a ML framework to assess the thermodynamic stability of lead-free halide double perovskites with high speed and precision. The framework evaluates feature importance and guides the discovery of potential lead-free perovskites. Four ML algorithms (RF, RR, SVR, and XGBoost) are used. Among them, XGBoost achieves the highest predictive performance, with an R^2 of 0.9935 and MAE of 0.0126, as shown in Fig. 9(a). XGBoost excels due to its efficient handling of sparse data and weighted quantile sketch for instance weighting. Stability predictions are evaluated using MSE, R^2, and MAE (Fig. 9(b)).

Kim, D. et al. [86] investigated the transport behavior of hybrid PSCs under different environmental conditions using ML analysis of impedance spectroscopy. They employed NMF to analyse the distribution of relaxation times from experimental impedance data. The results revealed distinct relaxation time constants corresponding to different environmental conditions, providing insights into the changes in ionic and electronic transport mechanisms. Quantitative results included a 280% increase in charge carrier concentration under oxygen, whereas a substantial decrease of approximately 300% was observed under ambient air conditions. This study significantly contributes to understanding the environmental effects on PSCs and offers effective methods for interpreting impedance responses.

Zhao Y.C. et al. [85] employed GBT with SHAP to analyse perovskite stability across various aging temperatures. Their findings revealed a temperature-dependent impact of organic cations, showing that MA^{+}

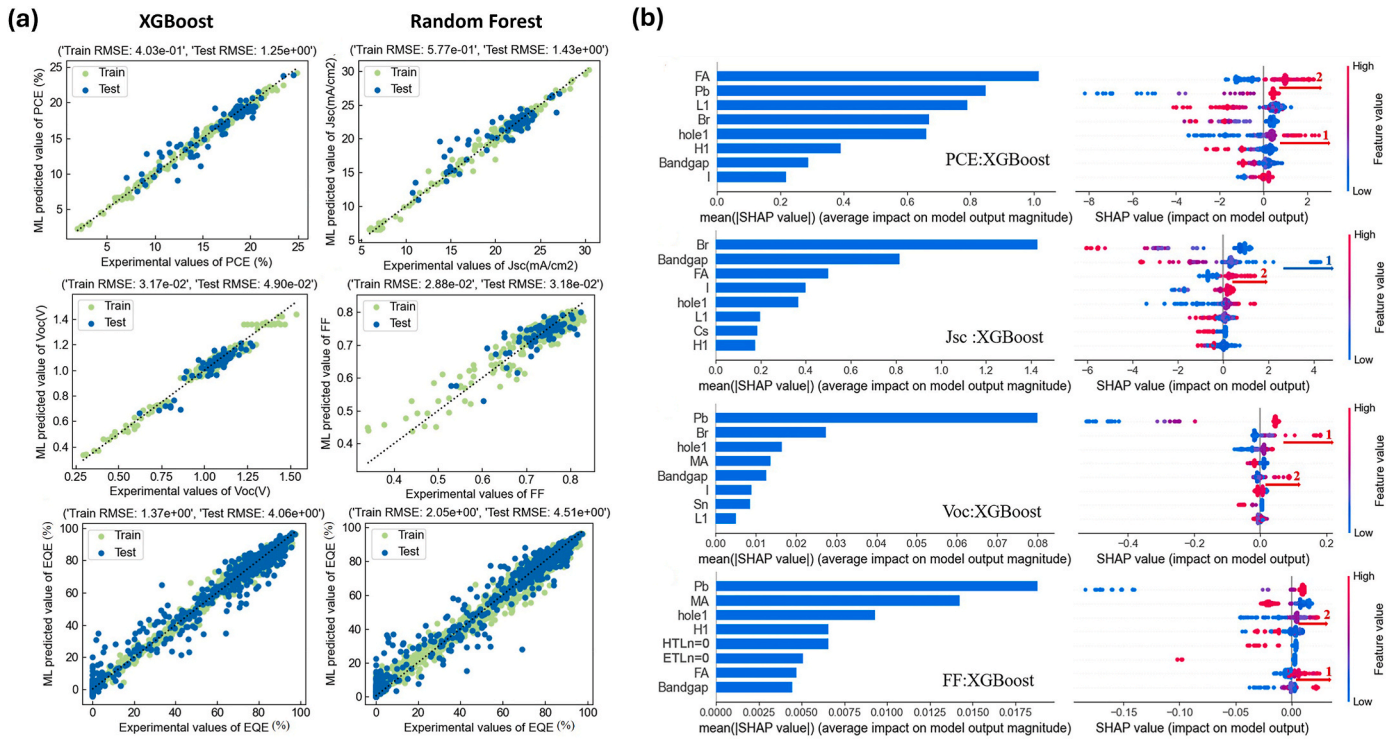

Fig. 7. (a) The score of predictive models in between XGBoost (left) and RF (right). X-axis is the experimental values recorded in the publication and the Y-axis is the predicted values obtained by the predictive model. (b) SHAP full feature importance and contribution analysis was conducted for the photovoltaic parameter prediction model based on the XGBoost algorithm. Reproduced with permission from Ref. [70]. Copyright (2024) Elsevier.

enhances stability, while Cs^+ and Rb^+ reduce it below 100 °C. They also demonstrated long-term device stability with negligible efficiency loss after 1800 h of operation at 30 °C This discovery highlights a stability reversal phenomenon, contrasting with high-temperature behaviors where Cs^+/Rb^+ typically improve stability.

Liang, G-Q. et al. [84] employed ML to develop an optimal model for predicting the thermodynamic phase stability of lead-free halide double perovskites. Their study utilized a dataset of 469 $A_2B'BX_6$ double perovskites with DFT-calculated energy above the convex hull (E_{hull}) values and 24 primary features derived from the periodic table (Fig. 10 (a)). Six ML algorithms, Adaboost, Bagging, DT, ET, RF, and XGBoost were implemented to calculate both classification and regression models. Through performance comparisons using cross-validation, the authors found that XGBoost outperformed the other algorithms in both classification and regression models. As shown in Fig. 10(b), all models achieved accuracy, precision, recall, and F1 scores above 0.85, with XGBoost achieving the highest values: accuracy (0.882 ± 0.030), precision (0.884 ± 0.032), recall (0.883 ± 0.030), and F1 score (0.884 ± 0.032). For regression, XGBoost demonstrated the best performance, with R^2 of 0.84 (±0.05), RMSE of 17.48 (±2.45) meV/atom, and MAE of 12.05 (±1.33) meV/atom (Fig. 10(c)).

Graniero, P. et al. [83] utilized a comprehensive dataset of 42,400 entries from the Perovskite Database and over 1000 in-house aging experiments to analyse the stability of PSCs by using two ML algorithms which is Elastic Net (eNet) for linear model and RF for non-linear model. Fig. 11(a) presents a scatter plot of operational lifetimes devices to degrade to 80% of their initial efficiency (T_{80}) against four other stability figures of merit, displaying the correlation coefficient (r) on the upper edge of every sub-figure. The figure highlights that even figures of merit with similar definitions can yield significantly different values for PSCs' aging curves. Moreover, these metrics do not consistently agree on which devices are more stable, underscoring the challenge of identifying a universal figure of merit for stability due to the diverse aging behaviors of PSCs.

Mammeri, M. et al. [82] employed DT constituent of the ET to analyse the stability of PSCs. They constructed a dataset of 1050 samples, covering various materials and method. They highlight the relative influence of various manufacturing processes and materials on the stability of PSCs with the highest influence is based on HTL. The study identified that TiO_2/m-TiO_2 as ETL, 2D-3D perovskite as active layer, and carbon as back contact with DMF + DMSO as precursor solution and Chlorobenzene as an anti-solvent solvent enhance regular cell stability, while BCP, PCBM as ETL, and $MAPbI_{3-x}Cl_x$, NiO and DEA as HTL and Al as back contact improve inverted cell stability.

Dunlap-Shohl et al. [81] developed ML models to predict T_{80} of $CH_3NH_3PbI_3$ PSCs under varying temperature and relative humidity. The significant average reduction in J_{SC} at T_{80} indicates that perovskite decomposition is a process of major importance in device failure. In Fig. 11(b) show dark field images of pristine devices from initially flat and featureless, indicating uniform device layers largely free of macroscopic defects towards degradation progresses under 50% relative humidity conditions and 1 sun illumination in air at 25 °C material transformation is at first most noticeable at the device edges and at intermediate times, isolated defects in the interior by the appearance of large numbers of bright spots in the dark field images throughout the device. Photooxidation of the absorber can occur in areas not covered by the top Ag contact or at the edges, where moisture and oxygen can enter, while fabrication defects like thin spots in the contact caused by dust may also allow oxidizing substances in, and iodine can react with Ag to form AgI. Additionally, electric field gradients at the edges can push dipolar species like MA^+, damaging the perovskite absorber, with defects in the Ag contact further contributing to this degradation. They utilized LR with greedy feature selection (GFS), LASSO, and RR techniques to analyse data from 45 accelerated degradation experiments. The models incorporated features such as initial rates of change in device parameters, ambient conditions, and a physiochemical feature derived from chemical reaction kinetics. Based on Fig. 11(c), models trained using GFS, LASSO, and RR are relatively consistent with one

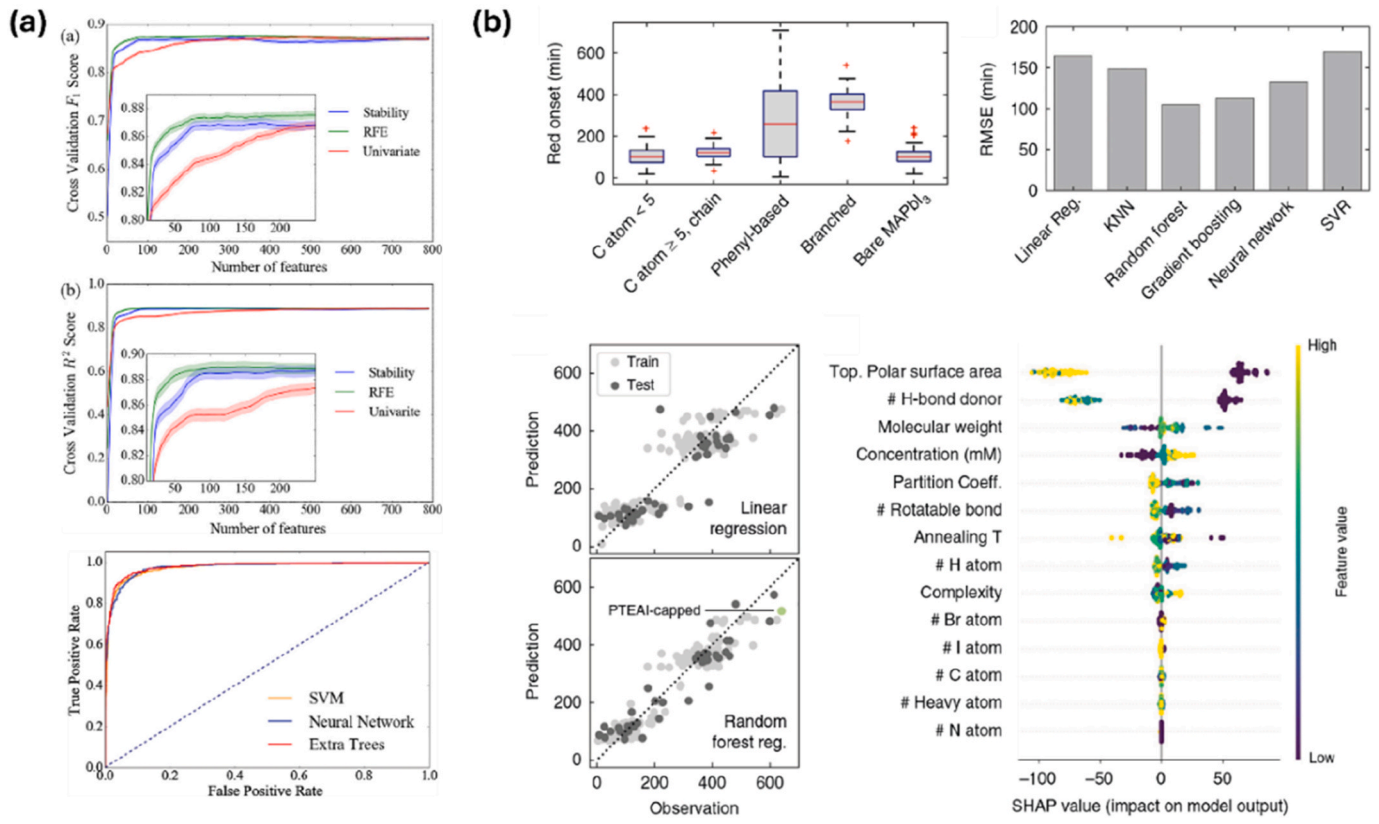

Fig. 8. (a) Upper-left: F1 scores from cross-validation are shown for three feature selection methods in classification: stability selection, recursive feature elimination (RFE), and univariate selection based on mutual information, Bottom-left: R^2 scores from cross-validation are shown for the same feature selection methods in regression, Middle-right: ROC curve of SVM, NN and ET. Reproduced with permission from Ref. [92]. Copyright (2018) Elsevier. (b) Upper-left: Extracted red onset points in minutes and their distributions from time-dependent camera data of capping-layer groups and bare $MAPbI_3$ controls, synthesized using 12 different processing conditions. Upper-right: The cross-validated RMSE of various ML models. Bottom-left: The comparison of red onset prediction and observation output between LR and RF models. Bottom-right: The feature importance ranking obtained from the RF algorithm and SHAP library. Reproduced with permission from Ref. [90]. Copyright (2020) The Authors.

another. Parity plots of test set predictions show that average prediction accuracy for all models lies in the range of ~35–45%. The R^2 values for the test set predictions are also similar at 0.66–0.7, showing that the models can explain about 2/3 of the variance in unseen samples. The results showed that these models could predict T_{80} with an accuracy of about 40%, highlighting the significant role of perovskite decomposition in performance loss.

Chen, J. et al. [80] also investigate the T_{80} in n-i-p and p-i-n configurations, using four algorithms: LR, RF, GBDT, and XGBoost. Data from the Perovskite Database were analysed. Based on the distribution of T_{80} and T_{90} (time to degrade to 90% efficiency) over time in Fig. 11 (d), XGBoost outperforms the other models, with the lowest RMSE, indicating strong predictive accuracy. R^2 values of 0.9984 and 0.9882 further confirm the model's excellent fit. Prior to 2020, the average T_{80} was below 500 h, while from 2021 to 2024, the average T_{90} consistently exceeded 1000 h, demonstrating improved stability in PSCs. The XGBoost algorithm was then used to classify T_{90} data into two categories, above and below 1000 h, identifying highly stable PSCs. The classification model, based on XGBoost, achieved an accuracy of 85% in predicting PSCs with high stability (T_{90} > 1000 h). In Fig. 11(e), the authors used SHAP analysis to evaluate ML data and found that a wider perovskite bandgap improves the stability of PSCs, while a narrower bandgap reduces their stability. They also discovered that humidity should not exceed 20%, as higher humidity negatively affects stability, and elevated temperatures lead to decreased stability. In terms of structure, maintaining the FA content above 80% and the Cs content below 10% is crucial. This is because FA, with its larger ionic radius, enhances thermal stability when it predominates as the A-site cation in the perovskite. Additionally, incorporating Cs^+ helps improve the structural stability of the perovskite, reduces trap density, and ultimately boosts the stability of PSCs.

Based on Table 7, XGBoost, RF, and NMF have proven to be effective in predicting the stability of PSCs due to their ability to handle complex, nonlinear relationships and diverse data types. XGBoost excels through its gradient boosting framework, regularization, and feature interaction capabilities, making it highly accurate and robust, even with noisy or limited datasets. RF similarly performs well by averaging over multiple decision trees, reducing variance and overfitting, while offering interpretability through feature importance. NMF, though unsupervised, is valuable for extracting latent degradation patterns from time-series or spectroscopic data, providing mechanistic insights into environmental effects like temperature and humidity. In contrast, SVR and LR consistently underperform in this domain due to their limited capacity to model nonlinear and complex interactions inherent in stability-related features. SVR is highly sensitive to kernel choice and prone to overfitting in noisy datasets, while LR oversimplifies the problem by assuming linearity, resulting in underfitting and poor generalization.

6. Limitation and challenges

While ML offers powerful capabilities in accelerating PSCs research, several challenges must be addressed to ensure robust, interpretable, and reproducible outcomes. These challenges relate to the quality and structure of available datasets, the transparency of ML models, and the reproducibility of results across studies. This section outlines five key limitations affecting the development and deployment of ML in PSCs

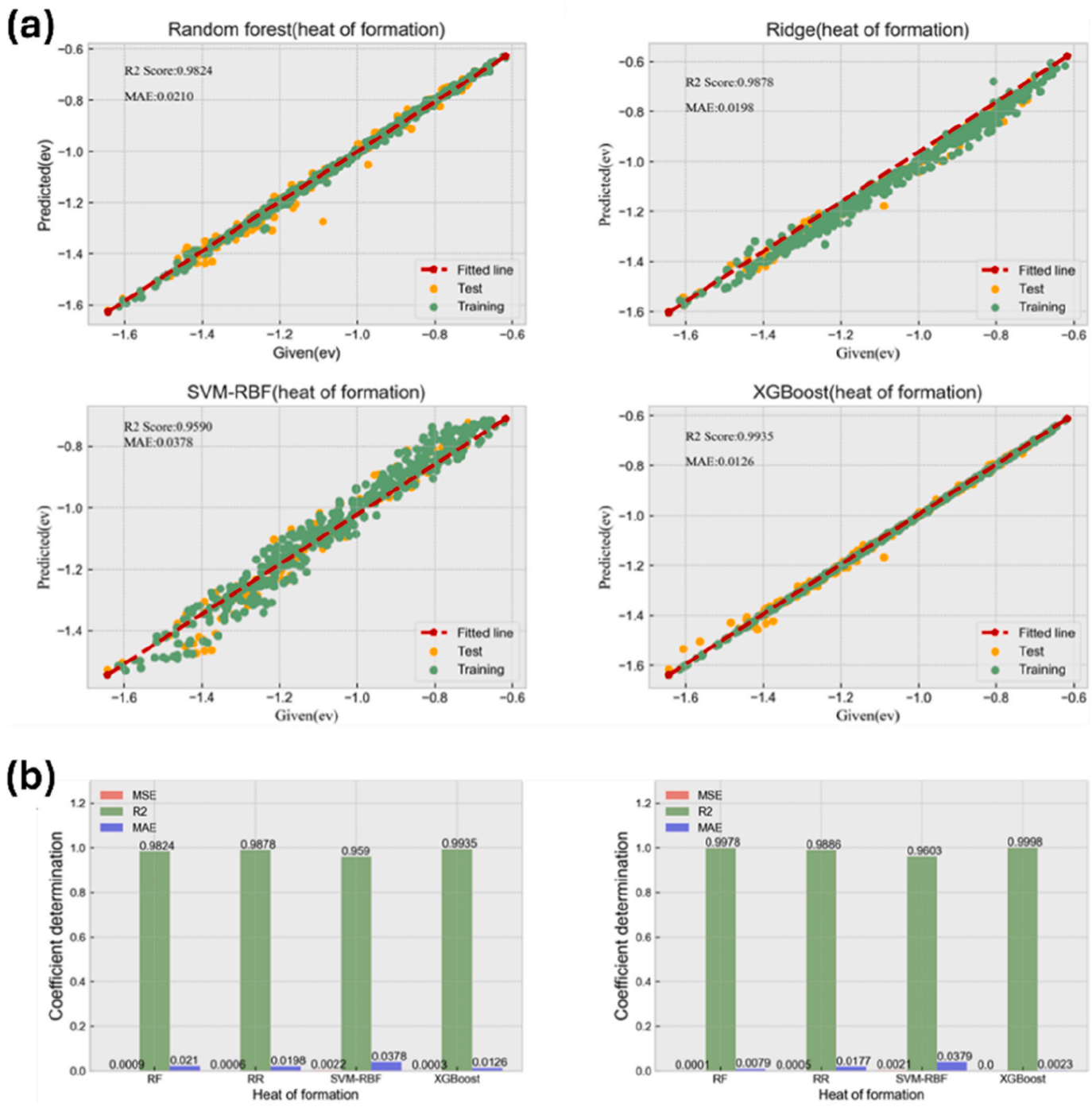

Fig. 9. (a) 4 model comparisons of stability prediction for halide double perovskites. (b) Prediction of heat of formation on the test set. Reproduced with permission from Ref. [87]. Copyright (2021) Elsevier.

research.

6.1. Lack of standardized benchmark datasets

A major limitation in ML for PSCs is the lack of standardized benchmark datasets. Existing datasets differ in structure, features, and metadata quality, making fair model comparisons and replication difficult. Many emphasize high-performing materials, while others lack key details like synthesis or environmental conditions. Without benchmarks, it remains unclear whether performance gains from model improvements or differences in dataset preparation. Developing well-curated, open-access benchmarks will be critical for enabling reproducible and comparable ML research.

6.2. Data bias and imbalance

A common limitation in PSCs datasets is the prevalence of publication and selection bias, where data is skewed toward well-performing materials, high-performance device, and good-stability data, while failed experiments, unstable phases, and low-performing results are rarely reported or publicly shared. This publication bias results in skewed datasets that can limit the generalizability of ML models to novel or exploratory materials. Additionally, class imbalance such as over-representation of high PCE values may cause models to perform poorly

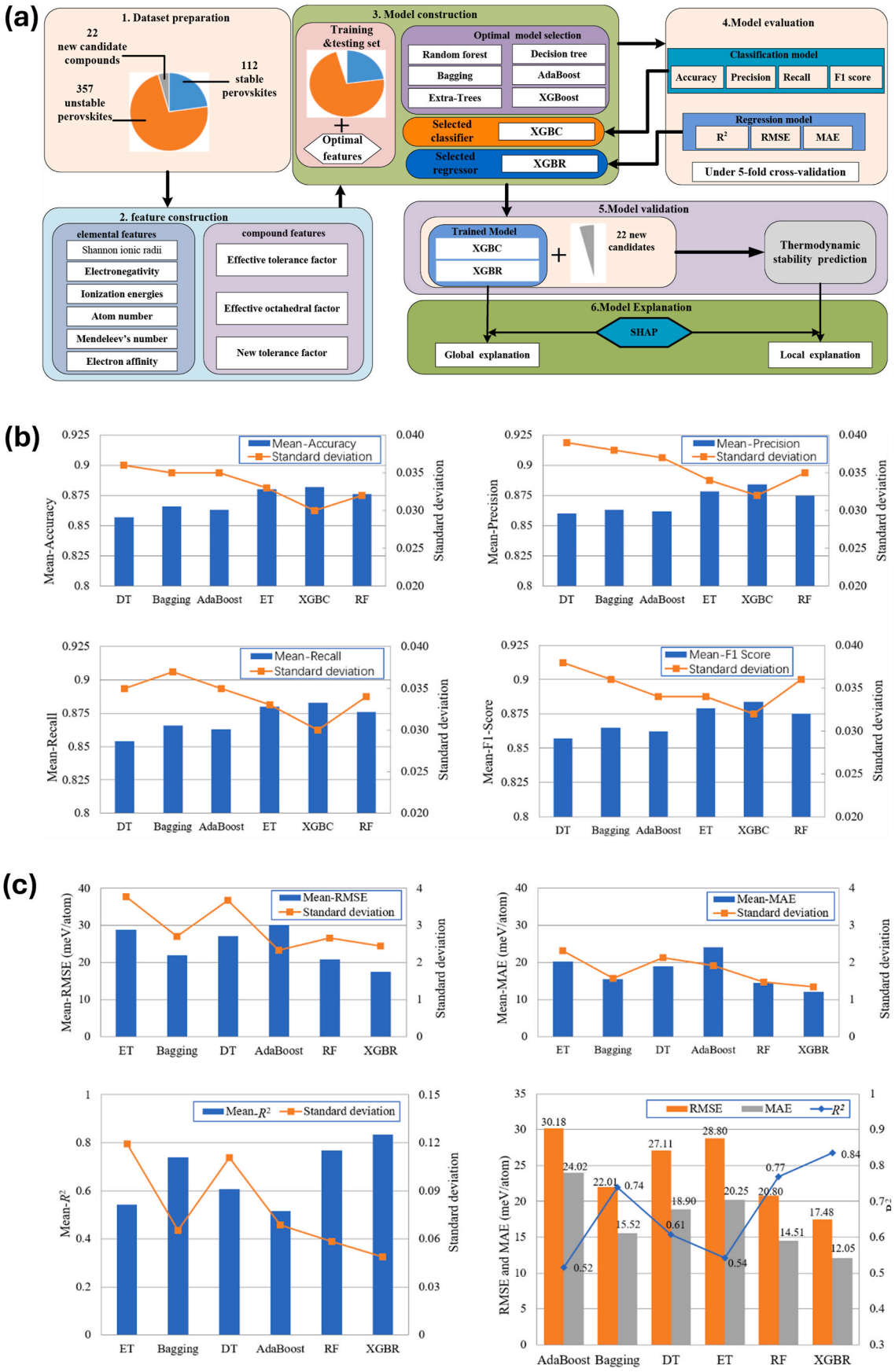

Fig. 10. (a) ML workflow for stable halide double perovskite. (b) Performance of different classification models by the average value of accuracy, precision, F1 score, and recall. (c) Performance of different regression model by the mean value of RMSE, MAE and R^2. Reproduced with permission from Ref. [84]. Copyright (2022) Elsevier.

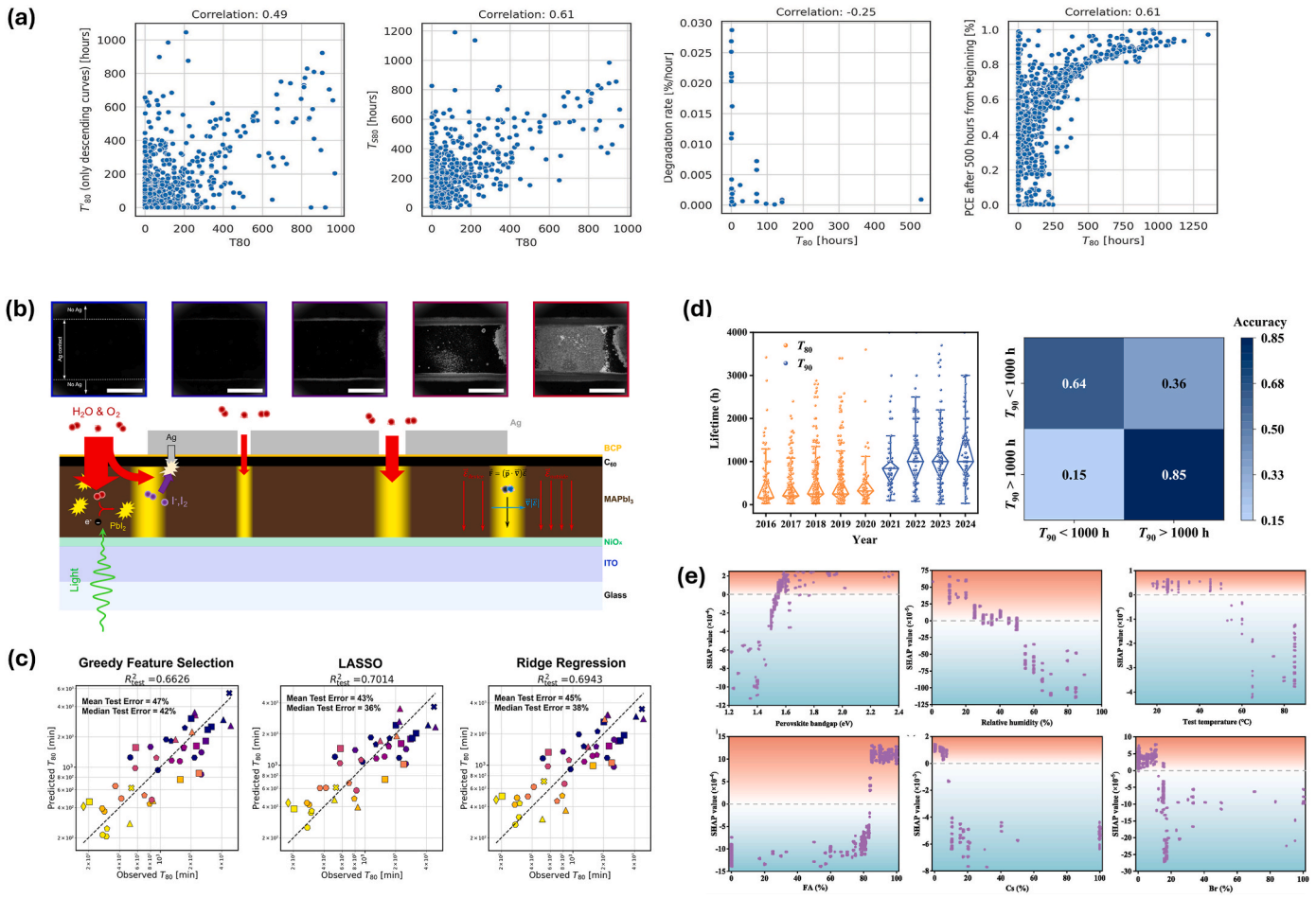

Fig. 11. (c) Scatter plot of T80 against four other stability figures of merit. Reproduced with permission from Ref. [83]. Copyright (2023) The Authors. (b) Dark field image of degradation progress under 50% relative humidity and 1 sun illumination in air at 25 °C. (c) ML result of GFS, Lasso, and RR. Reproduced with permission from Ref. [81]. Copyright (2025) Royal Society of Chemistry. (d) Data distributions of T_{80} and T_{90} over the years and confusion matrix of the optimal T_{90} classification model based on the XGBoost algorithm. (e) SHAP value analysis of key parameters (perovskite bandgap, relative humidity, test temperature, FA ratio, Cs ratio, Br ratio) affecting the stability of PSCs in the T_{90} model. Reproduced with permission from Ref. [80]. Copyright (2025) Elsevier.

on underrepresented but scientifically important subsets. Addressing this requires more inclusive and transparent data reporting, as well as careful data curation and data cleaning.

6.3. *Reproducibility and reporting practices*

Reproducibility remains a major concern in ML-based PSCs research due to inconsistent reporting of data preprocessing steps, feature selection methods, hyperparameters, and training protocols undermines the reproducibility of many ML studies. Without clear documentation and access to code and datasets, it is challenging for other researchers to validate or build upon published work. Promoting open-source practices and standardized reporting formats will be essential for improving transparency and reliability in ML-driven PSCs research.

6.4. *Overfitting and model robustness*

Many ML models reported in PSCs research achieve high accuracy on training datasets but struggle to generalize to unseen data. This is often a result of overfitting, where models capture noise or dataset-specific patterns rather than underlying physical relationships. The risk of overfitting is especially high when using small, biased, or imbalanced datasets. Without rigorous cross-validation, external test sets, or regularization strategies, reported performance metrics may overstate real-world applicability. Ensuring model robustness requires careful model selection, transparent validation procedures, and ideally, evaluation on independent datasets.

6.5. *Model interpretability (black-box nature)*

Many ML models, particularly complex ones such as deep learning and ensemble methods often function as "black box" with limited transparency into how predictions are made. A black box refers to a model whose internal logic or decision-making process is not easily interpretable by humans, even when the input and output are observable. This lack of interpretability reduces trust in the model and makes it difficult to extract scientific insights. Incorporating XAI methods, such as SHAP, LIME, and PDPs can help identify key features driving predictions and enable researchers to validate model outputs against domain knowledge. While these tools do not make the model inherently interpretable, they help bridge the gap between complex predictions and human understanding.

6.6. *Limited transferability across material domains*

Another challenge is the limited transferability of ML models when applied to new classes of materials. Models trained on well-studied lead-based perovskites, for instance, often perform poorly when applied to lead-free, 2D, or mixed-dimensional systems due to underlying structural and chemical differences. This constraint hampers the exploration

of novel or emerging materials using existing models. Improving transferability may involve incorporating domain adaptation techniques, expanding training datasets to include broader chemical diversity, or leveraging transfer learning strategies to adapt models across material domains.

7. Future perspective

AI has rapidly evolved into a central tool for accelerating the discovery, optimization, and deployment of PSCs. From atomic-scale understanding to industrial-scale fabrication, AI offers diverse solutions across the innovation pipeline. However, realizing this potential requires addressing key challenges ranging from limited data and interpretability to environmental durability and scalability. This section synthesizes key future directions by merging related themes, identifying critical barriers, and proposing strategic AI-driven solutions.

7.1. AI for understanding and designing materials from atom to device

Accurately modelling the fundamental behaviour of perovskite materials remains a challenge in advancing PSCs technology. Understanding charge transport, defect states, and excitonic interactions at the quantum level is critical but computationally intensive and often disconnected from device-level outcomes [93,94].

Challenge: The limited availability of quantum-scale data and the difficulty of integrating information across scales such as structural, morphological, and electronic hinder comprehensive understanding and material optimization.

Solution: Physics-informed ML, quantum-ML, and transfer learning techniques can be applied to DFT-derived datasets to model charge dynamics and defect tolerance efficiently. In parallel, integrating multimodal and multi-scale datasets ranging from atomic-level simulations to optical/electrical measurements using deep learning architectures can enable comprehensive, explainable predictions of PSCs behaviour. Embedding XAI techniques such as SHAP, LIME into these models can further provide interpretability and scientific insights, guiding material refinement and hypothesis-driven experimentation.

7.2. Emerging AI paradigms: AutoML, generative models, and reinforcement learning

The increasing complexity of PSCs materials and devices calls for advanced AI strategies that can streamline discovery and design with minimal manual intervention. Recent developments in Automated ML (AutoML), generative models, and reinforcement learning (RL) offer new paradigms for accelerating innovation, enabling autonomous decision-making, and expanding the materials design space.

Challenge: Traditional ML workflows in PSCs research often require expert knowledge for model selection, hyperparameter tuning, or feature engineering. Moreover, the vast design space of possible material compositions, device architectures, and fabrication pathways remains underexplored due to limitations in human-driven iteration and data sparsity.

Solution: AutoML can streamline the ML process by automatically selecting models, tuning parameters, and building pipelines, reducing the need for manual intervention and enabling broader adoption. Generative models like Generative Adversarial Network (GAN) and Variational Autoencoder (VAE) can suggest new material compositions or device structures by learning from high-performing examples, helping explore unexplored design spaces. RL enables closed-loop optimization by learning from trial-and-error interactions, making it useful for optimizing synthesis conditions, device configurations, or experimental workflows.

7.3. AI for autonomous laboratory automation

Experimental validation remains a bottleneck in materials discovery due to the manual nature of synthesis, processing, and testing. AI-driven laboratory automation holds promise for accelerating this process through real-time feedback, robotic systems, and closed-loop optimization [95].

Challenge: Hardware-software integration remains a barrier to implementing fully autonomous AI-experiment cycles in PSCs research.

Solution: Modular robotic platforms combined with Bayesian optimization or RL can guide intelligent experiment selection, synthesis parameter adjustment, and real-time decision-making. Standardizing data formats and adopting Findable, Accessible, Interoperable, Reusable (FAIR) data principles will further support reproducibility and cross-laboratory collaboration.

7.4. AI for tandem solar cell optimization

Tandem solar cells such as perovskite/perovskite, perovskite/silicon, perovskite/CIGS or perovskite/organic combinations offer a promising path to surpass single-junction efficiency limits [96–98]. However, device design is complex due to current matching, recombination and interface engineering challenges.

Challenge: Multi-variable optimization of layer thickness, band alignment, and charge transport pathways presents a prohibitively complex design.

Solution: AI tools such as surrogate modelling and evolutionary algorithms can rapidly explore and identify high-performing material and structural configurations. RL can further refine device designs based on performance feedback, while XAI techniques help clarify which features have the greatest impact on efficiency.

7.5. AI for real-world performance and environmental resilience

Most PSCs models are developed under laboratory conditions and do not account for environmental stressors such as humidity, UV exposure, and airborne pollutants that affect long-term performance [99]. Additionally, lifecycle sustainability including toxicity and recyclability is gaining attention as PSCs near commercialization [100–102].

Challenge: Limited long-term outdoor datasets and a lack of integrated environmental modelling constrain robust AI predictions and sustainability assessments.

Solution: AI models can be adapted to real-world conditions using domain adaptation and transfer learning, enabling reliable predictions from limited, application-specific data. Coupling real-time environmental monitoring with AI-driven degradation modelling allows for early detection of failure mechanisms and supports the development of more durable materials and encapsulation strategies. At the same time, integrating ML with life-cycle assessment (LCA) offers a practical pathway to identify greener materials and cleaner fabrication processes. Taken together, these approaches not only improve long-term performance but also support the broader goal of making PSCs technologies more sustainable and environmentally responsible.

7.6. AI for aesthetic and transparent solar technologies

Colorful and semitransparent PSCs are increasingly attractive for applications in building-integrated photovoltaics (BIPV), where architectural design and energy production must coexist [103–105]. These devices offer unique aesthetic and functional advantages but often suffer from reduced efficiency due to lower light absorption [106–109].

Challenge: Achieving an optimal balance between visual appeal (color, transparency) and photovoltaic performance is technically challenging.

Solution: AI models trained on datasets that include optical absorption spectra, bandgap properties, and human-perceived color

parameters can predict formulations that strike a balance between aesthetics and efficiency. Generative models like GANs can propose novel chromatic perovskite compositions, while AI-based control systems can dynamically adjust transparency in smart window applications based on ambient conditions and energy needs. These tools support the design of visually appealing, high-performance solar solutions for modern architecture.

7.7. AI for industrial-scale fabrication

Translating lab-scale PSCs performance into scalable, stable, and cost-effective large-area modules is a fundamental hurdle on the path to commercialization [110,111]. Fabrication techniques such as roll-to-roll printing or vacuum deposition often introduce variability that negatively impacts yield and device uniformity.

Challenge: Maintaining consistent film quality, layer thickness, and interface engineering across large surfaces under dynamic environmental conditions is difficult using conventional control systems.

Solution: AI-enabled digital twins of manufacturing lines can simulate and optimize process parameters in real-time, adjusting for fluctuations in temperature, humidity, or material feedstock. Machine vision systems powered by deep learning can monitor defects, while ML models can fine-tune fabrication settings, increasing throughput and reducing production costs.

8. Conclusion

This review underscores the transformative role of ML in advancing PSCs research, particularly in material discovery and device performance optimization. Drawing from a comprehensive workflow of spanning data acquisition, model selection, evaluation metrics, and explainable AI techniques. We critically analysed recent studies that apply ML to predict bandgaps, stability, and photovoltaic parameters. Quantitative comparisons reveal that ensemble models like XGBoost consistently achieve superior predictive accuracy, while hybrid datasets linking computational and experimental data enhance model generalizability. This review also outlines key limitations and challenges, along with future outlooks that are critical for guiding ongoing research. Addressing these aspects will be essential to fully harness AI-driven approaches for the efficient, reliable, and scalable development of PSCs technologies.

CRediT authorship contribution statement

Shafidah Shafian: Writing – review & editing, Writing – original draft, Visualization, Validation, Project administration, Investigation, Formal analysis, Conceptualization. **Mohd Nizam Husen:** Validation, Funding acquisition. **Lin Xie:** Writing – review & editing, Conceptualization. **Kyungkon Kim:** Writing – review & editing, Validation, Conceptualization.

Funding sources

This work was supported by the Universiti Kuala Lumpur, Malaysia.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

S.S acknowledges support from the Geran Galakan Penyelidik Muda (GGPM), grant number GGPM-2023-048, funded by Universiti Kebangsaan Malaysia, Malaysia. L.X. acknowledges support from the National Natural Science Foundation of China (22209144), the Project of Natural Science Foundation of Yunnan, and the Yunnan Revitalization Talent Support Program (202201AU070030 and 202201AT070114). The authors also acknowledge the financial support provided by Universiti Kuala Lumpur, Malaysia.

Data availability

Data will be made available on request.

References

[1] S. Qi, C. Ge, P. Wang, B. Wu, Y. Zhao, R. Zhao, S. Shafian, Y. Hua, L. Xie, ACS Appl. Mater. Interfaces 16 (2024) 51037–51045.
[2] Y. Zheng, Y. Li, R. Zhuang, X. Wu, C. Tian, A. Sun, C. Chen, Y. Guo, Y. Hua, K. Meng, Energy Environ. Sci. 17 (2024) 1153–1162.
[3] X. Zhang, L. Wang, S. Shafian, P. Wang, Y. Zhao, P. Wang, B. Wu, J. Zhai, J. Chen, L. Sun, Y. Hua, L. Xie, Small 21 (2024) 2408362.
[4] P. Wang, S. Shafian, F. Qiu, X. Zhang, Y. Zhao, B. Wu, K. Kim, Y. Hua, L. Xie, J. Energy Chem. 102 (2025) 151–160.
[5] S. Lee, Y.S. Yoon, S. Shafian, J.Y. Kim, K. Kim, Small Methods (2025) 2500104.
[6] R. Wang, M. Mujahid, Y. Duan, Z.K. Wang, J. Xue, Y. Yang, Adv. Funct. Mater. 29 (2019) 1808843.
[7] P. Cheng, X. Zhan, Chem. Soc. Rev. 45 (2016) 2544–2582.
[8] M. Hong, J. Youn, K.Y. Ryu, S. Shafian, K. Kim, ACS Appl. Mater. Interfaces 15 (2023) 20151–20158.
[9] H. Kim, Y.-J. Kong, W.-S. Kim, S. Shafian, K. Kim, ACS Appl. Polym. Mater. 6 (2024) 5814–5821.
[10] H. Kim, Y. Heo, Y. Na, S. Shafian, B. Kim, K. Kim, ACS Appl. Mater. Interfaces 16 (2024) 55873–55880.
[11] N.G. Park, Adv. Energy Mater. 10 (2020) 1903106.
[12] S. Shafian, F.N. Mohd Salehin, S. Lee, A. Ismail, S. Mohamed Shuhidan, L. Xie, K. Kim, ACS Appl. Energy Mater. 8 (2025) 699–722.
[13] B. Jo, W. Chen, H.S. Jung, J. Energy Chem. 101 (2025) 298–323.
[14] Best Research-Cell Efficiency Chart | Photovoltaic Research | NREL, https://www.nrel.gov/pv/cell-efficiency.html.
[15] M.A. Green, E.D. Dunlop, M. Yoshira, N. Kopidakis, K. Bothe, G. Siefer, D. Hinken, M. Rauer, J. Hohl-Ebinger, X. Hao, Prog. Photovoltaics Res. Appl. 32 (2024) 425–441.
[16] M.A. Green, E.D. Dunlop, M. Yoshira, N. Kopidakis, K. Bothe, G. Siefer, X. Hao, J. Y. Jiang, Prog. Photovoltaics Res. Appl. 33 (2024) 3–15.
[17] W. Zulhafizhazuan, K. Sobayel, S. Shafian, S. Sepeai, M.A. Ibrahim, Interactions 246, 2025, p. 29.
[18] T. Wu, J. Wang, Nano Energy 66 (2019) 104070.
[19] K.Y. Ryu, S. Shafian, J. Shin, Y.J. Lee, M. Lee, K. Kim, J. Power Sources 542 (2022) 231772.
[20] S. Shin, S. Shafian, K.Y. Ryu, Y.K. Jeon, W.S. Kim, K. Kim, Adv. Mater. Interfac. 9 (2022) 2200118.
[21] S. Shafian, H. Hwang, K. Kim, Opt. Express 24 (2016) 25308–25316.
[22] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J.M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, F. Herrera, Inf. Fusion 99 (2023) 101805.
[23] A. Jain, S.P. Ong, G. Hautier, W. Chen, W.D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, APL Mater. 1 (2013) 011002.
[24] S. Kirklin, J.E. Saal, B. Meredig, A. Thompson, J.W. Doak, M. Aykol, S. Rühl, C. Wolverton, npj Comput. Mater. 1 (2015) 1–15.
[25] C. Draxl, M. Scheffler, J. Phys.: Materials 2 (2019) 036001.
[26] L. Talirz, S. Kumbhar, E. Passaro, A.V. Yakutovich, V. Granata, F. Gargiulo, M. Borelli, M. Uhrin, S.P. Huber, S. Zoupanos, Sci. Data 7 (2020) 299.
[27] D.D. Landis, J.S. Hummelshøj, S. Nestorov, J. Greeley, M. Dułak, T. Bligaard, J. K. Nørskov, K.W. Jacobsen, Comput. Sci. Eng. 14 (2012) 51–57.
[28] S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R.H. Taylor, L.J. Nelson, G. L. Hart, S. Sanvito, M. Buongiorno-Nardelli, Comput. Mater. Sci. 58 (2012) 227–235.
[29] C.R. Groom, I.J. Bruno, M.P. Lightfoot, S.C. Ward, Structural Science 72 (2016) 171–179.
[30] S. Gražulis, A. Daškevič, A. Merkys, D. Chateigner, L. Lutterotti, M. Quiros, N. R. Serebryanaya, P. Moeck, R.T. Downs, A. Le Bail, Nucleic acids research 40 (2012) D420–D427.
[31] L. Rumor, A. Andrade-Campos, Adv. Mech. Eng. 14 (2022) 16878132221130575.
[32] T. Zhang, Y. Jiang, Z. Song, H. Huang, Y. He, Z. Fang, H. Weng, C. Fang, Nature 566 (2019) 475–479.
[33] M. Hellenbrandt, Crystallogr. Rev. 10 (2004) 17–22.
[34] T.J. Jacobsson, A. Hultqvist, A. García-Fernández, A. Anand, A. Al-Ashouri, A. Hagfeldt, A. Crovetto, A. Abate, A.G. Ricciardulli, A. Vijayan, A. Kulkarni, A. Y. Anderson, B.P. Darwich, B. Yang, B.L. Coles, C.A.R. Perini, Nat. Energy 7 (2022) 107–115.
[35] X. Qin, X. Du, S. Bhattacharya, C. Clayton, J. Hu, M. Jana, S. Janke, R. Lau, R. Laasner, A. Levin. HybriD 3: materials property database for hybrid organic-inorganic perovskitesAPS March Meeting Abstracts, 2021. X54. 007.
[36] E. Blokhin, P. Villars, Handbook of Materials Modeling: Methods: Theory and Modeling, 2020, pp. 1837–1861.

[37] M. Yamazaki, Y. Xu, M. Murata, H. Tanaka, K. Kamihira, K. Kimura, Baltica VII (2007) 193.
[38] C. Zhu, Y. Liu, D. Wang, Z. Zhu, P. Zhou, Y. Tu, G. Yang, H. Chen, Y. Zang, J. Du, W. Yan, Cell Rep. Phys. Sci. 5 (2024) 102321.
[39] A. Sabagh Moeini, F. Shariatmadar Tehrani, A. Naeimi-Sadigh, Sci. Rep. 14 (2024) 26736.
[40] S. Feng, J. Wang, Molecules 29 (2024) 499.
[41] S. Djeradi, T. Dahame, M.A. Fadla, B. Bentria, M.B. Kanoun, S. Goumri-Said, Machine Learning and Knowledge Extraction 6 (2024) 435–447.
[42] D.O. Obada, E. Okafor, S.A. Abolade, A.M. Ukpong, D. Dodoo-Arhin, A. Akande, Mater. Sci. Semicond. Process. 161 (2023) 107427.
[43] S. Mishra, B. Boro, N.K. Bansal, T. Singh, Mater. Today Commun. 35 (2023) 106376.
[44] R. Zhao, B. Xing, H. Mu, Y. Fu, L. Zhang, Chin. Phys. B 31 (2022) 056302.
[45] B. Yılmaz, Ç. Odabaşı, R. Yıldırım, Energy Technol. 10 (2022) 2100948.
[46] Z. Wang, Y. Han, X. Lin, J. Cai, S. Wu, J. Li, ACS Appl. Mater. Interfaces 14 (2021) 717–725.
[47] T. Wang, K. Zhang, J. Thé, H. Yu, Comput. Mater. Sci. 201 (2022) 110899.
[48] S. Rath, G.S. Priyanga, N. Nagappan, T. Thomas, Comput. Mater. Sci. 210 (2022) 111476.
[49] Y. Liu, W. Yan, H. Zhu, Y. Tu, L. Guan, X. Tan, Org. Electron. 101 (2022) 106426.
[50] A. Mannodi-Kanakkithodi, M.K. Chan, J. Mater. Sci. 57 (2022) 10736–10754.
[51] W. Hu, L. Zhang, Z. Pan, ACS Appl. Mater. Interfaces 14 (2022) 21596–21604.
[52] X. Cai, F. Liu, A. Yu, J. Qin, M. Hatamvand, I. Ahmed, J. Luo, Y. Zhang, H. Zhang, Y. Zhan, Light Sci. Appl. 11 (2022) 234.
[53] S. Zhang, T. Lu, P. Xu, Q. Tao, M. Li, W. Lu, J. Phys. Chem. Lett. 12 (2021) 7423–7430.
[54] X. Yang, L. Li, Q. Tao, W. Lu, M. Li, Comput. Mater. Sci. 196 (2021) 110528.
[55] P. Priya, N. Aluru, npj Comput. Mater. 7 (2021) 90.
[56] Y. Li, Y. Lu, X. Huo, D. Wei, J. Meng, J. Dong, B. Qiao, S. Zhao, Z. Xu, D. Song, RSC Adv. 11 (2021) 15688–15694.
[57] L. Li, Q. Tao, P. Xu, X. Yang, W. Lu, M. Li, Comput. Mater. Sci. 199 (2021) 110712.
[58] C. Li, H. Hao, B. Xu, Z. Shen, E. Zhou, D. Jiang, H. Liu, Comput. Mater. Sci. 198 (2021) 110714.
[59] Z. Gao, H. Zhang, G. Mao, J. Ren, Z. Chen, C. Wu, I.D. Gates, W. Yang, X. Ding, J. Yao, Appl. Surf. Sci. 568 (2021) 150916.
[60] E.T. Chenebuah, M. Nganbe, A.B. Tchagang, Mater. Today Commun. 27 (2021) 102462.
[61] S.R. Xie, P. Kotlarz, R.G. Hennig, J.C. Nino, Comput. Mater. Sci. 180 (2020) 109690.
[62] H. Park, R. Mall, A. Ali, S. Sanvito, H. Bensmail, F. El-Mellouhi, Comput. Mater. Sci. 184 (2020) 109858.
[63] H. Liu, J. Cheng, H. Dong, J. Feng, B. Pang, Z. Tian, S. Ma, F. Xia, C. Zhang, L. Dong, Comput. Mater. Sci. 177 (2020) 109614.
[64] H. Park, R. Mall, F.H. Alharbi, S. Sanvito, N. Tabet, H. Bensmail, F. El-Mellouhi, J. Phys. Chem. 123 (2019) 7323–7334.
[65] J. Li, B. Pradhan, S. Gaur, J. Thomas, Adv. Energy Mater. 9 (2019) 1901891.
[66] M. Saqib, U. Shoukat, M.M. Soliman, S. Bashir, M.H. Tahir, H.K. Thabet, M. Kallel, Chem. Phys. 589 (2025) 112515.
[67] S. Zhao, J. Wang, Z. Guo, H. Luo, L. Lu, Y. Tian, Z. Jiang, J. Zhang, M. Chen, L. Li, J. Energy Chem. 94 (2024) 441–448.
[68] N. Roberts, D. Jones, A. Schuy, S.C. Hsu, L.Y. Lin, Advanced Physics Research 3 (2024) 2400060.
[69] I.I. Malek, H. Imtiaz, S. Subrina, Sol. Energy 278 (2024) 112737.
[70] Y. Liu, X. Tan, P. Xiang, Y. Tu, T. Shao, Y. Zang, X. Li, W. Yan, Materials Today Physics 42 (2024) 101359.
[71] S. Lee, Y.J. Park, J. Kim, J. Im, S. Yoon, S.I. Seok, Curr. Appl. Phys. 68 (2024) 98–107.
[72] C. Zhi, S. Wang, S. Sun, C. Li, Z. Li, Z. Wan, H. Wang, Z. Li, Z. Liu, ACS Energy Lett. 8 (2023) 1424–1433.
[73] M.M. Salah, Z. Ismail, S. Abdellatif, Mater. Renew. Sustain. Energy 12 (2023) 187–198.
[74] Y. Lu, D. Wei, W. Liu, J. Meng, X. Huo, Y. Zhang, Z. Liang, B. Qiao, S. Zhao, D. Song, J. Energy Chem. 77 (2023) 200–208.
[75] W. Liu, Y. Lu, D. Wei, X. Huo, X. Huang, Y. Li, J. Meng, S. Zhao, B. Qiao, Z. Liang, J. Mater. Chem. A 10 (2022) 17782–17789.
[76] Y. Hu, X. Hu, L. Zhang, T. Zheng, J. You, B. Jia, Y. Ma, X. Du, L. Zhang, J. Wang, Adv. Energy Mater. 12 (2022) 2201463.
[77] M. Del Cueto, C. Rawski-Furman, J. Arago, E. Orti, A. Troisi, J. Phys. Chem. C 126 (2022) 13053–13061.
[78] C. She, Q. Huang, C. Chen, Y. Jiang, Z. Fan, J. Gao, J. Mater. Chem. A 9 (2021) 25168–25177.
[79] V.M. Le Corre, T.S. Sherkar, M. Koopmans, L.J.A. Koster, Cell Rep. Phys. Sci. 2 (2021).
[80] J. Chen, Y. Zhan, Z. Yang, Y. Zang, W. Yan, X. Li, Mater. Today Energy 48 (2025) 101769.
[81] W.A. Dunlap-Shohl, Y. Meng, P.P. Sunkari, D.A. Beck, M. Meilă, H.W. Hillhouse, J. Mater. Chem. A 12 (2024) 9730–9746.
[82] M. Mammeri, L. Dehimi, H. Bencherif, F. Pezzimenti, Sol. Energy 249 (2023) 651–660.
[83] P. Graniero, M. Khenkin, H. Köbler, N.T.P. Hartono, R. Schlatmann, A. Abate, E. Unger, T.J. Jacobsson, C. Ulbrich, Front. Energy Res. 11 (2023) 1118654.
[84] G.-Q. Liang, J. Zhang, Comput. Mater. Sci. 204 (2022) 111172.
[85] Y. Zhao, J. Zhang, Z. Xu, S. Sun, S. Langner, N.T.P. Hartono, T. Heumueller, Y. Hou, J. Elia, N. Li, Nat. Commun. 12 (2021) 2191.
[86] D. Kim, E.S. Muckley, N. Creange, T.H. Wan, M.H. Ann, E. Quattrocchi, R. K. Vasudevan, J.H. Kim, F. Ciucci, I.N. Ivanov, Adv. Sci. 8 (2021) 2002510.
[87] Z. Guo, B. Lin, Sol. Energy 228 (2021) 689–699.
[88] K. Higgins, S.M. Valleti, M. Ziatdinov, S.V. Kalinin, M. Ahmadi, ACS Energy Lett. 5 (2020) 3426–3436.
[89] N.T.P. Hartono, J. Thapa, A. Tiihonen, F. Oviedo, C. Batali, J.J. Yoo, Z. Liu, R. Li, D.F. Marrón, M.G. Bawendi, Capping layers design Guidelines for stable perovskite solar cells via machine learning. 2020 47th IEEE Photovoltaic Specialists Conference (PVSC), 2020, pp. 693–695. IEEE.
[90] N.T.P. Hartono, J. Thapa, A. Tiihonen, F. Oviedo, C. Batali, J.J. Yoo, Z. Liu, R. Li, D.F. Marrón, M.G. Bawendi, Nat. Commun. 11 (2020) 4172.
[91] Z. Li, Q. Xu, Q. Sun, Z. Hou, W.J. Yin, Adv. Funct. Mater. 29 (2019) 1807280.
[92] W. Li, R. Jacobs, D. Morgan, Comput. Mater. Sci. 150 (2018) 454–463.
[93] Q. Tao, P. Xu, M. Li, W. Lu, npj Comput. Mater. 7 (2021) 23.
[94] C. Chen, A. Maqsood, T.J. Jacobsson, J. Alloys Compd. 960 (2023) 170824.
[95] N.J. Szymanski, B. Rendy, Y. Fei, R.E. Kumar, T. He, D. Milsted, M.J. McDermott, M. Gallant, E.D. Cubuk, A. Merchant, Nature 624 (2023) 86–91.
[96] Y. Cheng, L. Ding, SusMat 1 (2021) 324–344.
[97] K.O. Brinkmann, P. Wang, F. Lang, W. Li, X. Guo, F. Zimmermann, S. Olthof, D. Neher, Y. Hou, M. Stolterfoht, T. Wang, A.B. Djurišić, T. Riedl, Nat. Rev. Mater. 9 (2024) 202–217.
[98] J. Lim, N.-G. Park, S.I. Seok, M. Saliba, Energy Environ. Sci. 17 (2024) 4390–4425.
[99] W. Tress, K. Domanski, B. Carlsen, A. Agarwalla, E.A. Alharbi, M. Graetzel, A. Hagfeldt, Nat. Energy 4 (2019) 568–574.
[100] J. Gong, S.B. Darling, F. You, Energy Environ. Sci. 8 (2015) 1953–1968.
[101] J. Zhang, X. Gao, Y. Deng, Y. Zha, C. Yuan, Sol. Energy Mater. Sol. Cell. 166 (2017) 9–17.
[102] X. Tian, S.D. Stranks, F. You, Nat. Sustain. 4 (2021) 821–829.
[103] F.N.M. Salehin, P. Chelvanathan, A.A. Goje, N.A. Ludin, M.A. Ibrahim, S. Shafian, Results Phys. 70 (2025) 108172.
[104] S. Shafian, Future Energy and Environment Letters 2 (2025) 13–32.
[105] Y.-J. You, M.A. Saeed, S. Shafian, J. Kim, S.H. Kim, S.H. Kim, K. Kim, J.W. Shim, Nanotechnology 32 (2021) 465401.
[106] S. Park, S. Shafian, J. Lee, S. Jo, S. Jeon, S. Lee, D. Shangxian, H. Ahn, K. Kim, D. Y. Ryu, Adv. Opt. Mater. 11 (2023) 2301357.
[107] B.Y. Kim, S. Shafian, K. Kim, Adv. Mater. Interfac. 10 (2023) 2300421.
[108] S. Shafian, J. Son, Y. Kim, J.K. Hyun, K. Kim, ACS Appl. Mater. Interfaces 11 (2019) 18887–18895.
[109] Y. Kim, J. Son, S. Shafian, K. Kim, J.K. Hyun, Adv. Opt. Mater. 6 (2018) 1800051.
[110] L. Meng, J. You, Y. Yang, Nat. Commun. 9 (2018) 5265.
[111] Y. Rong, Y. Hu, A. Mei, H. Tan, M.I. Saidaminov, S.I. Seok, M.D. McGehee, E. H. Sargent, H. Han, Science 361 (2018) eaat8235.