EES Solar

ROYAL SOCIETY OF CHEMISTRY

REVIEW

View Article Online
View Journal | View Issue

Check for updates

Cite this: *EES Sol.*, 2025, **1**, 927

Machine learning for perovskite solar cells: a comprehensive review on opportunities and challenges for materials scientists

Víctor de la Asunción-Nadal,*[a] Christopher Iliffe Sprague,[b] Bertha Guijarro-Berdiñas,[c] Ute B. Cappel[de] and Alberto García-Fernández*[df]

Perovskite solar cells (PSCs) have emerged as a promising technology due to their tunable optoelectronic properties, low-cost fabrication and high efficiency. Despite this progress, key challenges such as long-term stability, large-scale manufacturability, and recyclability remain unsolved. Moreover, traditional methods for discovering new materials and optimizing device architectures rely on trial-and-error experiments. Machine learning (ML) offers powerful tools to address these bottlenecks by uncovering hidden patterns in data, accelerating discovery, and guiding rational design. However, the growing number of ML-driven studies in PSC research can be difficult to navigate, particularly for experimentalists and scientists without a computational background. To address this gap, this review is written with accessibility in mind, and it is structured to serve as a bridge between ML experts and the broader materials science community. We provide an overview of how ML can be applied to PSCs, from databases and data preprocessing to model training, evaluation, and interpretability. Advantages and limitations of different approaches are critically assessed, with emphasis on how dataset choice, algorithms, and metrics affect reliability. We conclude by outlining current challenges and open questions, as well as potential directions where the integration of ML with experimental and theoretical research could further advance the development of perovskite solar cells.

Received 26th March 2025
Accepted 21st September 2025

DOI: 10.1039/d5el00041f

rsc.li/EESSolar

Open Access Article. Published on 04 1447. Downloaded on 25/11/47 10:58:40 .
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Broader context

The integration of machine learning (ML) into the research and development of perovskite solar cells (PSCs) represents a significant advancement in the field of photovoltaics. PSCs have already demonstrated remarkable potential with their high power conversion efficiencies and cost-effective fabrication processes. However, the challenges of stability, material degradation, and upscaling to large area have hindered their widespread adoption. In this regard, ML offers a new approach by leveraging large datasets to identify patterns, predict material properties, and optimize device architectures more efficiently. This synergy between ML and PSC research can accelerate the discovery of new materials and enhance the performance and stability of solar cells. This comprehensive review discuss the progress in ML-driven PSC research and highlights the importance of interdisciplinary approaches in advancing materials science. It also discuss potential future applications of ML in solar energy technologies.

1. Introduction

Perovskite solar cells (PSCs) have rapidly emerged as one of the most promising candidates for next-generation photovoltaic technologies, owing to their exceptional high absorption coefficients, long carrier diffusion lengths, and defect tolerance.[1–3] These properties enable PSCs to achieve high power conversion efficiencies (PCEs) while maintaining relatively low production costs.[4,5] Since their introduction in 2009, the efficiency of PSCs has increased from 3.8% to >26%,[6] making them one of the fastest-advancing solar technologies to date. Their solution-based fabrication methods allow for cost-effective and scalable manufacturing,[7–9] leading to the development of new perovskite-based technologies like flexible, lightweight, and tandem solar cells.[10–12]

Despite these advantages, material stability, scalability, and long-term performance remain major challenges. Environmental factors such as moisture, heat, and UV exposure significantly degrade PSCs over time, limiting their commercial

[a]Aiiso Yufeng Li Family Department of Chemical and Nano Engineering, University of California San Diego, La Jolla, California 92093, USA. E-mail: vdelaasuncionnadal@ucsd.edu
[b]The Alan Turing Institute, British Library, 96 Euston Road, London NW1 2DB, UK
[c]Universidade da Coruña, CITIC, Campus de Elviña s/n, A Coruña, 15071, Spain
[d]Division of X-ray Photon Science, Department of Physics and Astronomy, Uppsala University, Box 516, Uppsala, SE-751 20, Sweden
[e]Wallenberg Initiative Materials Science for Sustainability, Department of Physics and Astronomy, Uppsala University, 751 20 Uppsala, Sweden
[f]Universidade da Coruña, CICA (Interdisciplinary Center for Chemistry and Biology), Department of Chemistry, Faculty of Science, As Carballeiras, s/n, Campus de Elviña 15071 A Coruña, Spain. E-mail: alberto.garcia.fernandez@udc.es

© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 04 1447. Downloaded on 25/11/47 10:58:40 .
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

viability.[13,14] To address these limitations, extensive research is being conducted to enhance material durability, improve encapsulation techniques, and optimize fabrication processes.

In this context, machine learning (ML) has emerged as a transformative tool for accelerating perovskite research. By uncovering complex patterns in large datasets and enabling predictive modeling beyond the limits of traditional trial-and-error experimentation, ML offers unique opportunities to guide materials discovery, optimize device performance, and even design new fabrication strategies. The convergence of PSC research with data-driven methodologies is part of a broader paradigm shift in materials science, where artificial intelligence and computational techniques complement experimental efforts to shorten innovation cycles and reduce costs.[15,16]

Over the past few years, a growing number of studies have applied ML methods to various aspects of perovskite research, including bandgap engineering, defect analysis, stability prediction, processing optimization, and device performance forecasting. The field now encompasses diverse approaches, from simple regression models to advanced ensemble algorithms and neural networks, reflecting the increasing maturity and adoption of ML in photovoltaic research. At the same time, the availability of open-source datasets, high-throughput simulations, and automated laboratories is further fueling this integration.[17–20]

This review provides a comprehensive overview of machine learning applications in perovskite solar cells, with the dual aim of clarifying fundamental principles and highlighting key advancements and future opportunities. We begin by introducing the fundamentals of PSCs, including their composition, properties, and current technological limitations. Next, we present a general ML workflow in materials science, which serves as the framework for a detailed discussion of PSC-focused applications. Within this structure, we examine each stage of the workflow: data extraction, including current open-source databases; data exploration and preprocessing, with emphasis on computational tools and libraries; model selection, where we analyze the most widely used algorithms in PSC research and illustrate their strengths and limitations through representative examples; we continue by discussing model training, validation, performance assessment, and results interpretation, showing how these steps contribute to reliable and impactful predictions. Finally, we outline the current challenges in applying ML to PSCs, such as data quality, model interpretability, and scalability, and discuss future research directions where ML could play a decisive role in accelerating materials discovery, device optimization, and commercialization of perovskite technologies.

2. Perovskite solar cells

The objective of this section is to identify the parameter-space and challenges in perovskite solar cell research, from the preparation of the materials and devices to the evaluation of important metrics such as efficiency, stability, recyclability, and scalability. Indeed, the preparation of solar cells is a multi-parametric process that needs optimization from the laboratory-scale to industrial manufacturing, operational conditions and post-operational recycling and waste management. In this regard, ML offers unprecedented advantages compared to classical research and optimization that will be highlighted in the next sections of this review.

As previously introduced, perovskite solar cells (PSCs) are based on a family of materials known as hybrid perovskites. This class of materials combines organic and inorganic components and is characterized by a perovskite crystal structure, usually represented by the formula ABX_3. In this structure, shown in Fig. 1 (left), 'A' is typically an organic cation (*e.g.*, methylammonium (MA), formamidinium (FA)), but it can also be an inorganic cation such as cesium (Cs) or a mixture of organic and inorganic cations. 'B' is a metal cation (*e.g.*, Pb^{2+}, Sn^{2+}), and 'X' is a halide anion (*e.g.*, iodine, bromine, chlorine).[21–23]

PSCs, shown in Fig. 1 (right), typically consist of multiple layers (substrate|electrode|transport material|perovskite|transport material|electrode) usually processed using co-evaporation or solution-based methods, such as spin-coating, blade coating, or inkjet printing.[24,25]

The substrate can be rigid, like glass, or flexible, such as polymer films or even another solar cell for tandem device integration.[26,27] The electrodes collect and transport the generated charges. The transparent conductive electrode is usually indium tin oxide (ITO) or fluorine-doped tin oxide (FTO), while the back electrode is often gold, silver, carbon, or aluminum. The electron transport layer (ETL) facilitates electron extraction while preventing hole recombination. Popular ETL materials include TiO_2, SnO_2, ZnO, and fullerene derivatives such as PCBM (phenyl-C_{61}-butyric acid methyl ester). The hole transport layer (HTL) selectively extracts holes and blocks electrons. Common HTL materials include organic molecules (*e.g.*, spiro-OMeTAD), polymer-based materials (*e.g.*, PEDOT:PSS), and inorganic alternatives such as NiO_x and CuSCN.[28]

The choice of transport layers and electrodes plays a crucial role for the solar cell's efficiency and stability. For instance, inorganic transport layers (*e.g.*, NiO_x, SnO_2) improve thermal and moisture stability, while organic transport layers (*e.g.*, spiro-OMeTAD) offer high efficiency but require additives that can degrade over time.[29–32] Similarly, Ag/Au electrodes provide excellent conductivity but are expensive and prone to diffusion into the perovskite layer,[33,34] whereas carbon electrodes offer a cost-effective and stable alternative.[35,36] Engineering interfaces between these layers is essential to reduce charge recombination, improve carrier mobility, and enhance overall device performance.[37,38]

The perovskite absorber layer, responsible for capturing sunlight and generating charge carriers, offers a vast range of possibilities. The high tunability of these materials enables precise adjustments in bandgap, stability, and optoelectronic properties, making them highly versatile.[39] The study of perovskite absorber materials has evolved far beyond the traditional methylammonium lead iodide ($MAPbI_3$) and formamidinium lead iodide ($FAPbI_3$) systems. Research has expanded into mixed-halide and mixed-cation perovskites, which offer improved stability and efficiency.[40–42] Additionally, alternative material systems, such as lead-free perovskites,[43–45] and low-

© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 04 1447. Downloaded on 25/11/47 10:58:40 .
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

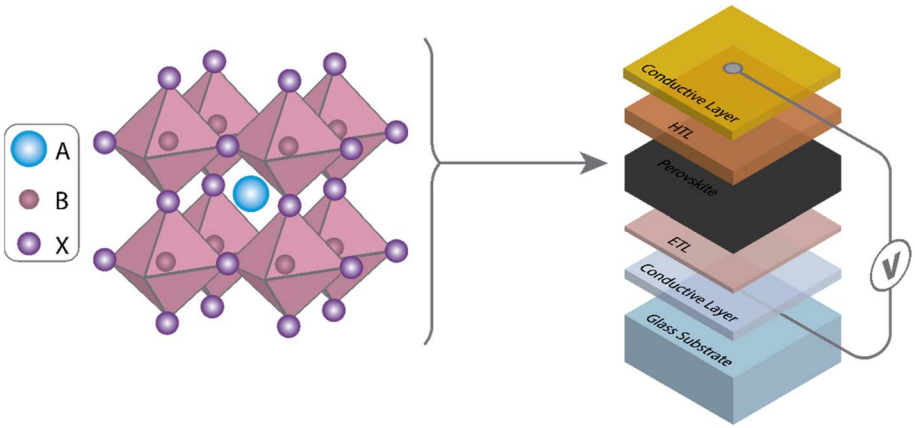

Fig. 1 Example of a perovskite structure (left) and general scheme of a perovskite solar cell (right).

dimensional lead halide perovskites incorporating larger cations like dimethylammonium (DMA),[46–48] guanidinium (GA),[49,50] ethylammonium (EA)[51,52] are being explored to address stability and toxicity concerns. These larger cations can influence structural stability, enhance moisture resistance, and tune optoelectronic properties, broadening the potential applications of perovskite materials in solar cells.[53,54]

To evaluate the performance of PSCs, it is crucial to measure their efficiency and assess their stability, providing key insights into their long-term reliability and practical applications. The power conversion efficiency (PCE) of PSCs is commonly evaluated by measuring current–voltage curves under 1 sun illumination, where the maximum power is determined by three main parameters: open-circuit voltage (V_{OC}), short-circuit current density (J_{SC}), and fill factor (FF). The rapid efficiency improvements in PSCs have made them one of the most competitive photovoltaic technologies, with lab-scale devices exceeding 26% efficiency.[6] Another critical metric is their stability, which is typically assessed by measuring operational lifetime.[55] To ensure real-world applicability, PSCs must be evaluated under standard test conditions (STC), which include 1-sun illumination (AM1.5G spectrum), 25 °C temperature, and controlled humidity levels.[56]

Stability is often compromised by moisture, oxygen, light exposure, and thermal stress, causing gradual degradation over time. These environmental factors can trigger structural decomposition, phase segregation, and ion migration, ultimately reducing the long-term performance of the device. Surfaces and interfaces play a critical role in the stability of PSCs, as they are key sites for defect formation, ion migration, and degradation processes under environmental stressors.[57–60]

Another challenge is toxicity, particularly due to the presence of lead (Pb) in most high-efficiency PSCs. Lead-based perovskites present environmental and regulatory concerns, driving research toward lead-free alternatives, and encapsulation strategies to prevent lead toxicity.[61,62]

Additionally, scalability and reproducibility remain key barriers, as many high-performance PSCs are fabricated using lab-scale techniques that may not be easily translated to industrial-scale manufacturing. Defect formation, particularly at grain boundaries and interfaces, also affects charge transport, recombination, and hysteresis, requiring improvements in film quality and interface engineering.[63,64] Finally, fabrication process control is crucial, as slight variations in humidity, annealing temperature, or precursor solution quality can lead to significant differences in device performance.[65] Overcoming these challenges requires a combination of advanced material design, improved processing techniques, and data-driven optimization methods.[66]

3. Machine learning principles for materials science: insights from perovskite solar cells

Machine Learning (ML), a branch of Artificial Intelligence (AI), is increasingly transforming the landscape of materials science by enabling data-driven discovery, optimization, and prediction.[67,68] In the context of perovskite solar cells (PSCs), ML offers powerful tools to deal with the compositional flexibility of hybrid organic–inorganic materials and the multiple options for device fabrication.[20,69] By learning patterns from experimental and computational datasets, ML models can predict material properties, identify degradation pathways, and recommend optimal synthesis and processing conditions, thereby reducing dependance on costly and time-consuming trial-and-error experimentation.[70,71]

The application of ML solutions for perovskite research is a thriving field as highlighted by the interest of the scientific community on the combination of perovskite materials with ML techniques. In this regard, the evolution of the number of reports shows rapid growth, especially in the last 5 years (Fig. 2). Approximately half of the reports in the Web of Science platform on perovskites linked to ML techniques are related to solar energy applications. This helps to understand the impact of these novel technologies in the research and development of new promising materials to help tackle the new environmental and energy challenges that may arise from the current socio-economic landscape.

ML has proven particularly valuable in addressing two of the most pressing challenges in PSCs: performance optimization

Open Access Article. Published on 04 1447. Downloaded on 25/11/47 10:58:40 .
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

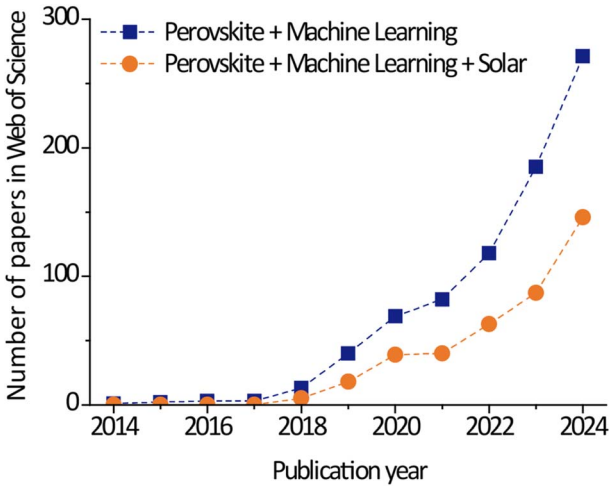

Fig. 2 Number of reports in the Web of Science platform with the topics: perovskite + machine learning (blue) and perovskite + machine learning + solar (orange).

and long-term stability. For example, supervised learning models are being trained on labelled datasets to predict power conversion efficiency (PCE), open-circuit voltage, or bandgaps, while unsupervised learning is applied to uncover hidden patterns in synthesis outcomes or degradation behaviour. Reinforcement learning and generative models are also emerging as tools for autonomous materials design and process control. To support these applications, a wide range of computational tools and libraries are employed for data preprocessing, feature extraction, model training, and interpretation. These platforms enable researchers to integrate diverse data sources, from molecular descriptors and crystal structures to device-level performance metrics, into predictive and interpretable ML workflows.

In this section, the discussion is organized around the typical machine learning workflow illustrated in Fig. 3, which outlines the key stages of ML-based application development in perovskite research, from data extraction and preprocessing to model selection, model validation, performance evaluation and results interpretation, providing a structured framework to connect the diverse case studies that follow.

3.1 Data extraction

Machine Learning (ML) models are only as reliable as the data used to train them, making data quality, diversity, and relevance fundamental to their success.[72] Balanced datasets are crucial for

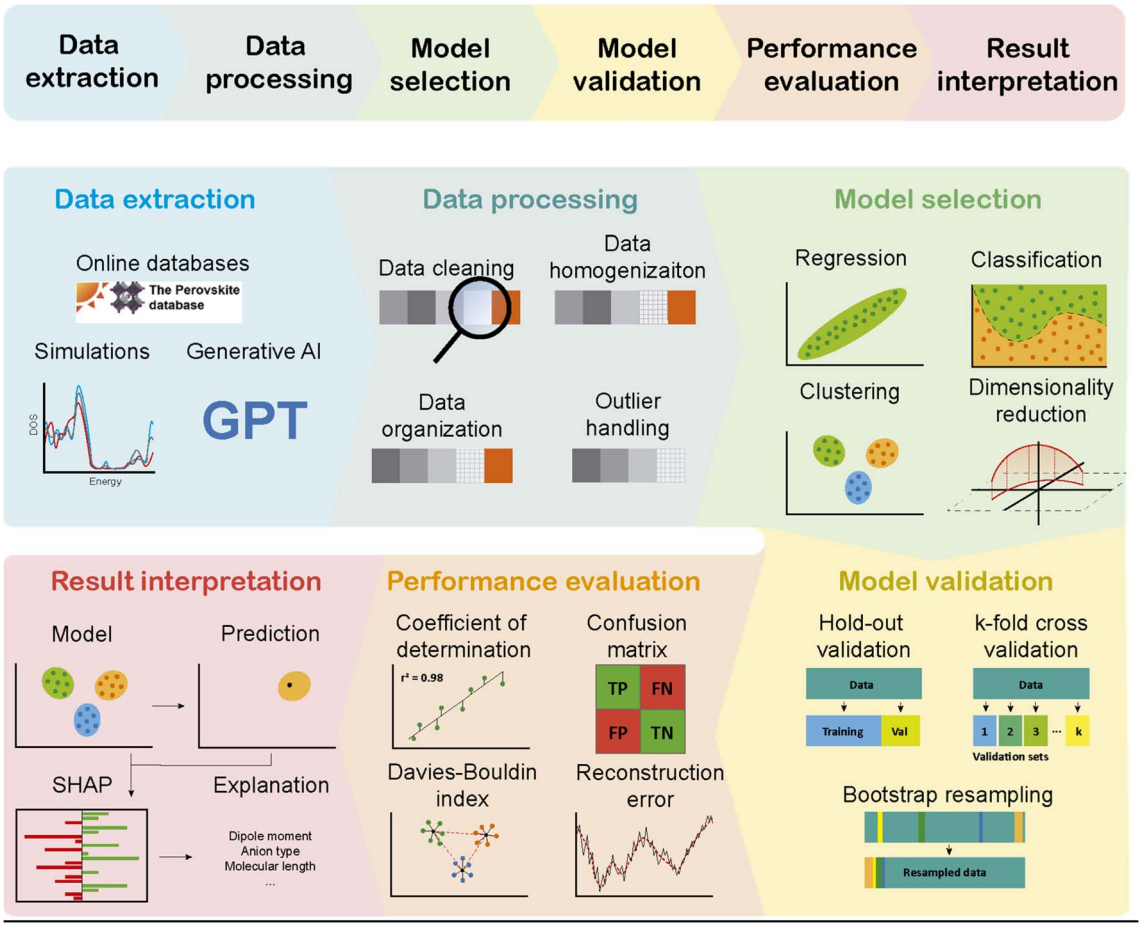

Fig. 3 Schematic representation of a typical workflow in ML development. Main steps involved in the development of a ML solution and examples of techniques associated with each step.

© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 04 1447. Downloaded on 25/11/47 10:58:40 .
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

enabling models to generalize and avoid bias, while irrelevant features can introduce spurious correlations and misleading conclusions. Likewise, inaccurate, inconsistent, or incomplete data compromise predictive performance and limit model applicability.[73,74] Therefore, access to high-quality, trustworthy experimental or computational datasets is essential for building robust ML solutions.[75,76]

In recent years, several open-source databases have been developed to address this need, providing data resources for ML applications in perovskite research (summarized in Table 1).

Detailed examples are the Perovskite Database Project,[77] which is described in Fig. 4 and offers data such as cell definition, device stack conformation, synthesis details and key metrics from more than 42 000 perovskite-based devices. This database stands out as one of the most comprehensive PSC-specific resources and is invaluable for benchmarking device performance and studying processing–performance relationships. However, while rich in device-level data, it remains limited in structural descriptors, which can constrain its direct integration into atomistic-level ML models.

The 2D perovskite database,[78] which contains data for over 840 2D perovskites, provides a focused dataset for exploring dimensionality effects. Although the dataset size is relatively small and its most recent structural addition dates back to 2023, its specialization makes it particularly relevant for studies targeting stability, excitonic effects, and layered perovskite architectures. Its limitation lies in its scale, which restricts the use of deep learning approaches requiring larger data volumes. On the other hand, the 2D HOIP platform[79] with 304 920 predicted structures and band gaps exemplifies high-throughput computational screening efforts. Such datasets are powerful for generating training sets for property prediction and identifying unexplored regions of chemical space. Nonetheless, the reliance on computational predictions introduces potential discrepancies with experimental reality, highlighting the need for validation workflows before direct application in device design. Following this line, the Materials Project,[80] including more that 169 000 materials, also offers computational data on material properties, including perovskite structures and electronic characteristics. Its strength lies in providing standardized, high-quality DFT-computed descriptors that enable cross-comparison across materials families. Its integration with tools like Pymatgen[81] and Matminer[82] makes it a cornerstone for feature extraction and model development. However, its focus on inorganic materials and DFT-derived properties may limit its direct applicability to hybrid perovskites, while useful, it is less detailed compared to specialized databases. From a different perspective the Cambridge Structural Database (CSD) is a comprehensive repository of crystallographic data for organic and organometallic compounds that contains experimentally

Table 1 Summary of databases for perovskite-related data

Database	Brief description	URL
AFLOW	Computational materials data including thermodynamic, electronic, and structural properties	https://aflowlib.org/
CSD	The Cambridge Structural Database is a repository of crystallographic data	https://www.ccdc.cam.ac.uk/
CMR	DFT-based computational data on material systems	https://cmr.fysik.dtu.dk/
COD	Free crystal structure database focused on inorganic materials	https://www.crystallography.net/cod/
HybriD^3 materials	Experimental and computational materials data for crystalline organic–inorganic compounds, predominantly based on the perovskite paradigm	https://materials.hybrid3.duke.edu/
ICSD	The world's largest database for completely identified inorganic crystal structures	https://icsd.products.fiz-karlsruhe.de/
ICDD	Powder diffraction files for material identification	https://www.icdd.com/
JARVIS	Repository designed to automate materials discovery and optimization using calculations and experiments	https://jarvis.nist.gov/
Materials Cloud	Open-science platform for computational materials data and workflows	https://materialscloud.org
Materials Project	Computed data on known and predicted materials	https://next-gen.materialsproject.org/
Perovskite Database	Open database with perovskite device data	https://www.perovskitedatabase.com/
Refractive index	Refractive index data for various materials, including perovskites and transport layers	https://refractiveindex.info/
Springer Materials	Physical, chemical, and structural properties of materials	https://materials.springer.com/
ZINC15	A free database of commercially-available compounds for virtual screening	https://zinc15.docking.org/

Open Access Article. Published on 04 1447. Downloaded on 25/11/47 10:58:40 .
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

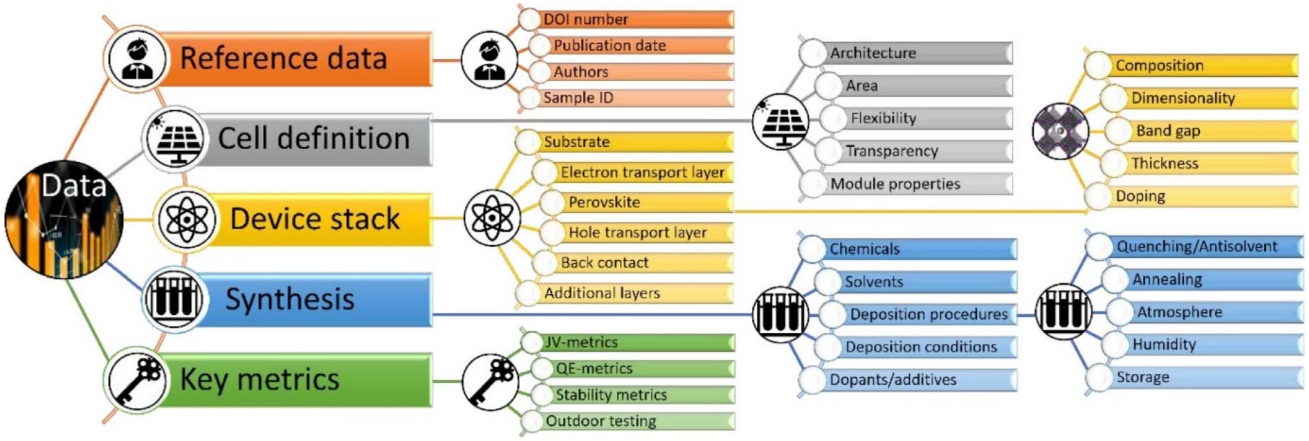

Fig. 4 Overview of data categories in the Perovskite Database. Reproduced from ref. 77 with permission from Springer Nature, under a Creative Commons Attribution 4.0 International License.

determined crystal structures of more than 1.24 M compounds.[83] It is a rich source for structural descriptors and molecular geometry data, particularly useful for modeling organic cations and interface materials in PSCs. Its general-purpose nature, however, means that perovskite-specific metadata (*e.g.*, device performance or synthesis conditions) is typically absent. This limitation becomes particularly evident when compared with for example the Perovskite Database Project, which emphasizes device architectures and performance metrics. Used they could offer a more complete picture, each compensating for the limitations of the other.

In addition to larger datasets, multiple smaller datasets have been collected, some examples in literature containing data of less than 800 devices are by She *et al.*,[84] Li *et al.*,[85] Ramazan *et al.*[86,87] and Liu *et al.*[88] Although their scale restricts the training of more complex ML models, these datasets are often highly consistent and well-documented, offering excellent case studies for proof-of-concept modeling and validation of methods.

Beyond databases and data repositories, the annual "Emerging PV Report"[89] highlights the latest advancements in the performance of emerging photovoltaic (e-PV) devices across various e-PV research areas. As good example highlighting the importance of data sources, Felix Mayr and Alessio Gagliardi explores seven open-source databases of perovskite-like materials and proposes a comprehensive comparison of structural fingerprint-based machine learning models.[90]

Overall, the landscape of databases in perovskite research reflects a balance between scope and specialization. Large-scale platforms such as the Materials Project and CSD offer wide coverage and versatility but require careful filtering to extract perovskite-specific insights. In contrast, perovskite-focused databases, such as the Perovskite Database Project and 2D perovskite datasets, provide highly relevant, context-specific data but are limited in scale. This duality highlights the need for hybrid strategies that combine the depth of specialized PSC datasets with the broad coverage of larger databases, along with enhanced approaches for data validation and interoperability.

3.2 Data exploration and preprocessing

Data preprocessing is a critical step in applying machine learning to perovskite solar cells, as it directly impacts the reliability of model predictions. PSC datasets often combine experimental results with simulated data, which can introduce inconsistencies in format, nomenclature, and quality.[76] To address this, preprocessing involves removing incomplete or duplicated entries, harmonizing descriptors (*e.g.*, device parameters or material compositions), and applying transformations such as normalization or encoding of categorical variables. Outliers, arising from measurement errors, synthesis variability, or reporting inconsistencies, must also be carefully handled to avoid misleading trends.[91] By ensuring that PSC datasets are accurate, consistent, and accessible, preprocessing provides the foundation for extracting meaningful relationships and improving model performance and generalization.

To fully utilize the data from the previously discussed and other databases, specialized computational tools are essential for retrieval, processing, and analysis, a summary of the most popular are presented in Table 2. Key ML libraries such as Scikit-learn,[92] Keras,[93] TensorFlow,[94] or PyTorch[95] and JAX[96] offer building blocks and algorithms for classification, regression, and deep learning, enabling advanced data-driven research. In particular, PyTorch is often preferred for its flexibility and ease of use, while JAX is often preferred for its high-performance speed. In addition, by using the molecule's SMILES,[97,98] as a as string input, Python packages, such as Mordred[99] and RDKit[100] can generate hundreds of molecular descriptors and fingerprints for use as feature vectors, enabling their integration into machine learning models and materials research.[101] Moreover, there are open-source Python-based platforms such as Matminer[82] that aims to facilitate data-driven methods for analyzing and predicting material properties. It enables users to retrieve large datasets from external databases like the Materials Project and Citrine, extract features using a library of descriptors, and create interactive visualizations.

© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 04 1447. Downloaded on 25/11/47 10:58:40 .
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Table 2 Summary of computational tools and libraries for ML applications

Tool/library	Type	Functionality	Use in perovskite solar cells (PSCs)
Scikit-learn	ML library	Classical ML algorithms (classification, regression, clustering)	Used to predict device performance metrics (*e.g.*, PCE, bandgap), and classify synthesis feasibility
Keras	Deep learning framework	High-level API for building neural networks	Applied in deep learning models for image-based crystal growth monitoring and device architecture design
TensorFlow	Deep learning framework	Scalable deep learning library with GPU support	Enables training of large models for predicting PSC efficiency and analysing fabrication data
PyTorch	Deep learning framework	Flexible deep learning library with dynamic computation graphs	Used in generative models and CNNs for structure–property prediction and synthesis control
SMILES	Molecular representation	Text-based representation of molecular structures	Input format for generating molecular descriptors for HTMs and interface materials in PSCs
Mordred	Descriptor generator	Computes molecular descriptors from SMILES	Generates features for ML models predicting HTM performance and stability in PSCs
RDKit	Cheminformatics toolkit	Molecular manipulation and descriptor calculation	Used for fingerprinting and feature extraction of organic components in PSCs
Matminer	Materials informatics	Feature extraction, dataset retrieval, visualization	Facilitates ML workflows for PSCs by integrating with databases and extracting features for PCE prediction
JAX	Deep learning framework	High-performance deep-learning/ computing library with functional programming style	Often used to speed up scientific simulations[102]

Finally, feature selection and feature extraction techniques are key preprocessing steps to optimize the efficiency and performance of the ML model to be trained.[103,104] These methods reduce the number of input features while retaining most of the relevant information, which helps to optimize computational time, avoid overfitting, and improve predictive accuracy and interpretability. While feature selection techniques[105,106] select a subset of relevant features from the original data set, feature extraction techniques, such as principal component analysis (PCA) and partial least squares (PLS), transform the original data into a new reduced space. PCA is an unsupervised method that offers a data projection of lower dimensionality in orthogonal axes (principal components) that capture the maximum variance in the data. This projection can then be used to classify the different datapoints into classes according to the similarities between the clustered datapoints. As practical example, Boubchir *et al.* used a multivariate technique based on principal component analysis (PCA) and the partial least squares regression (PLS-R) to predict mechanical properties of a set of 129 perovskites and propose 10 as the most promising.[107] This approach highlights PCA as a predictive tool for identifying promising material combinations, with future validation expected from high-throughput *ab initio* calculations. PLS on the other hand, is a supervised dimension reduction methodology, meaning that a quantifiable parameter is assigned to each datapoint. Dimensionality reduction is performed in a similar way to PCA by projecting the predicted and the observable variables to a new space by representing the relation between one or more latent variables and the response variable. Partial least squares discriminant analysis (PLS-DA) operates under the same principles, but a categorical label is assigned to the dataset instead of a quantitative variable, being the output of PLS-DA a classification of the supplied datapoints into the predefined categories. As an example, Imamura *et al.* used PLS to predict the Néel temperature of perovskites, whereas PLS-DA was used to organize the data into subgroups.[108] In addition, feature engineering can be used to create new features from existing ones if it can provide better insights from the existing data. In their manuscript, Q. Deng and B. Lin integrate feature engineering with automated machine learning to explore structure–composition relationships in cubic perovskite oxides. By identifying optimal descriptors through feature elimination, the study demonstrates that lattice constants are primarily determined by B-site ionic radii, offering a simple predictive expression.[109] This approach enhances interpretability, accelerates materials design, and provides a framework for data-driven materials discovery.

3.3 Model selection

Considering that different algorithms are better suited for specific tasks and data structures, the selection of the appropriate ML model is a critical factor to obtain impactful and reproducible results to advance and optimize Perovskite Solar

Open Access Article. Published on 04 1447. Downloaded on 25/11/47 10:58:40 .
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Cell (PSC) research. To guide this decision, a general flowchart for ML model selection is presented in Fig. 5 (adapted from ref. 69) broadly, ML algorithms can be classified according to their training approach: supervised, unsupervised, or reinforcement learning.[110,111] Supervised learning, the most widely applied in perovskite research, relies on labeled datasets and is particularly effective for predicting well-defined output variables such as bandgaps or device efficiency. It can be further subdivided into regression methods, which predict continuous values, and classification methods, which assign discrete categories. Unsupervised learning, in contrast, uncovers hidden structures in unlabeled datasets, while reinforcement learning enables models to optimize strategies through iterative trial-and-error interactions with their environment, making it promising for tasks such as process optimization and autonomous materials discovery.

In 2025, I. Mao and C. Xiang reviewed 119 research papers on perovskite-based studies and concluded that the most frequently used algorithms were random forest (36.1%), support vector machines (16.8%), and linear regression (15.1%).[19] The predominance of random forest reflects its robustness in handling relatively small and noisy datasets, which are typical in perovskite research, while the continued use of SVM and linear regression, particularly for materials discovery, highlights the community's reliance on well-established, interpretable models despite the increasing availability of more complex algorithms.

In this section, the principal ML algorithms employed in materials science and specifically perovskite research are summarized in Table 3, followed by a detailed discussion of the most widely applied approaches with representative examples.

Linear Regression (LR) is one of the most fundamental and widely used algorithms in regression analysis. It is primarily used to model the relationship between a dependent variable (also known as the response or target variable) and one or more independent variables (also known as predictors or features).

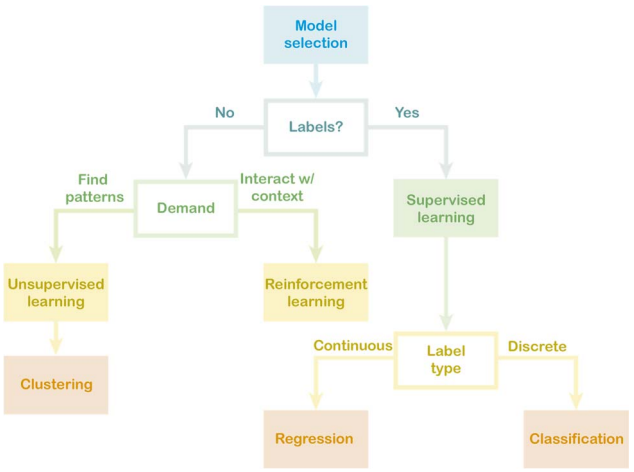

Fig. 5 General classification of ML models. Adapted from ref. 69 with permission from OAE Publishing Inc., under a Creative Commons Attribution 4.0 International License.

The goal of linear regression is to find the linear relationship between these variables and use this relationship to predict the unknown value of the dependent variable based on the values of the independent variables. To do this, the least squares method is used to find the best fitting line or hyperplane minimizing the sum of the squared differences between the observed and the predicted values. Although this model is relatively simple, fast and a good starting point to establish the baseline to be improved by other more sophisticated methods, it is only applicable to systems with linear relationships, unless modifications such as isotonic or ridge regression are implemented, and complex multiparametric problems may need to be simplified. As an example, Vakharia *et al.* used LR models as shown in Fig. 6, including isotonic and ElasticNet regression to predict the bandgaps of Cs-based perovskites compared to their theoretical values with a low RSME of 0.13 eV and high correlation of $R^2 = 0.98$.[112]

Support Vector Machines (SVM) are supervised machine learning models that find an optimal hyperplane in an *N*-dimensional space to classify data points. While similar to linear regression in concept, SVMs are not limited to linear relationships and can handle non-linear data through the use of kernel functions. These kernels transform the data into a higher-dimensional space, enabling the separation of data points that are not linearly separable in the original space by calculating pairwise similarities between data points. As an example, Pilania *et al.* use a SVM algorithm to find new perovskite halide compositions using a dataset composed of 185 known compounds. Their model allowed to predict several novel ABX_3 perovskites with high degree of confidence.[113] Advanced algorithms like sequential minimal optimization (SMO) can be used to overcome problems faced when training SVM models by using Lagrange multipliers and variational calculus the error function is minimized when finding the hyperplane.[114] For instance, Alsulami *et al.* used a SMO algorithm to assess the best materials to improve the stability of perovskite-based devices.[115] Their results showed that architecture-specific models, particularly for p–i–n devices ($R = 0.963$), outperformed general models, underscoring both the value of ML in revealing stability trends and the need for careful dataset segmentation to avoid obscuring architecture-dependent behaviours.

Gaussian Process Regression (GPR) is a probabilistic method that models data under a joint Gaussian distribution, using a covariance function to capture similarities between points. Unlike deterministic or black-box models, GPR provides not only predictions but also an estimate of their uncertainty, with the output variance reflecting the model's confidence. This makes GPR particularly valuable in PSC research, where understanding prediction reliability is as important as the prediction itself. F. Akhundova and co-workers applied GPR to link photoluminescence (PL) spectral features with non-radiative losses in wide-bandgap mixed-halide perovskites. By predicting PL quantum yield from spectral shapes, GPR identifies the key photophysical factors controlling PL quenching and charge recombination, enabling rapid, high-throughput optimization of film morphology. This demonstrates how GPR

© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 04 1447. Downloaded on 25/11/47 10:58:40 .
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Table 3 Main types of ML/AI algorithms by type and training, description and related algorithms

Algorithm type	Type	Training	Description	Related algorithms
Linear regression	Regression	Supervised	A dependent variable and one or more independent variables are correlated using a linear equation. It is used for predicting continuous values	Ridge regression, lasso regression, polynomial regression
Logistic regression	Classification	Supervised	It estimates probabilities using a logistic function. Probabilities are used for classification of binary and multi-class systems	Softmax regression, probit regression
Principal Components Analysis (PCA)	Dimensionality reduction	Unsupervised	An unsupervised technique used for reducing the dimensionality of large datasets while preserving as much variance as possible	Independent Component Analysis (ICA), Singular Value Decomposition (SVD), autoencoders
Partial Least Squares (PLS)	Regression	Supervised	Used when predictor variables are highly collinear. It combines features of PCA and linear regression for dimensionality reduction and prediction	Canonical Correlation Analysis (CCA)
Support Vector Machines (SVM)	Classification Regression	Supervised	Finds the optimal hyperplane that maximally separates classes in high-dimensional space. It is effective in both linear and non-linear classification	Support Vector Regression (SVR), kernel SVM, linear SVM
Decision Trees (DT)	Classification Regression	Supervised	A rule-based model that splits data into branches based on feature values to make decisions. It can be used for classification (categorical targets) or regression (continuous targets)	Boosted trees, conditional inference trees, Classification and Regression Trees (CART)
Random Forest (RF)	Classification Regression	Supervised	Ensemble learning method that builds multiple decision trees and averages their outputs to improve accuracy and reduce overfitting	Extra trees, gradient boosting trees
Gradient Boosting (GB)	Classification Regression	Supervised	Gradient boosting builds models sequentially using weak learners collectively, correcting the errors of previous models	XGBoost, LightGBM, AdaBoost
K-Nearest Neighbors (KNN)	Classification Regression	Supervised Unsupervised	Labels are based on the majority class among *k*-nearest data points. It is used for classification, regression (supervised), and clustering (unsupervised)	Weighted KNN, Density-Based Spatial Clustering of Applications with Noise (DBSCAN)
Neural Networks (NN)	Classification Regression	Supervised Unsupervised	Composed of interconnected neurons that learn patterns in data. They are the foundation of deep learning models	Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), transformers

can serve as a fast, structure-sensitive tool for guiding material processing toward higher-efficiency PSCs.[116]

The *K*-Nearest Neighbors (KNN) method is a non-parametric supervised machine learning algorithm. KNN relies on a "lazy learning" approach, meaning that it only stores a training dataset, and the computation occurs when a classification or prediction is performed. It is based on the principle that less distant points in the dataset share the same category (classification) or value (regression) as the queried datapoint. KNN is easy to implement, is easy to update with new data, and requires few hyperparameters but relies heavily on memory storage which makes it harder to scale and is prone to overfitting when testing lower values of k (or neighbors) and underfitting when testing higher K values. Choe *et al.* used KNN along with time-correlated single photon measurements to study complex energy transfer processes in perovskite nanocrystals. This approach not only provided prediction capabilities over a dataset but also yielded new insights to elucidate the physical phenomena behind this complex process.[117]

Decision Trees (DT) are non-parametric, supervised machine learning algorithms in which variables (for regression) or classes (for classification) are assigned to discrete values represented by leaves, connected through branches representing the features leading to each outcome. While DTs are straightforward to implement and interpret, they are prone to overfitting and can be unstable, as small changes in the dataset may lead to substantially different trees. To address these limitations, ensemble methods such as Random Forest (RF) and

Open Access Article. Published on 04 1447. Downloaded on 25/11/47 10:58:40 .
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

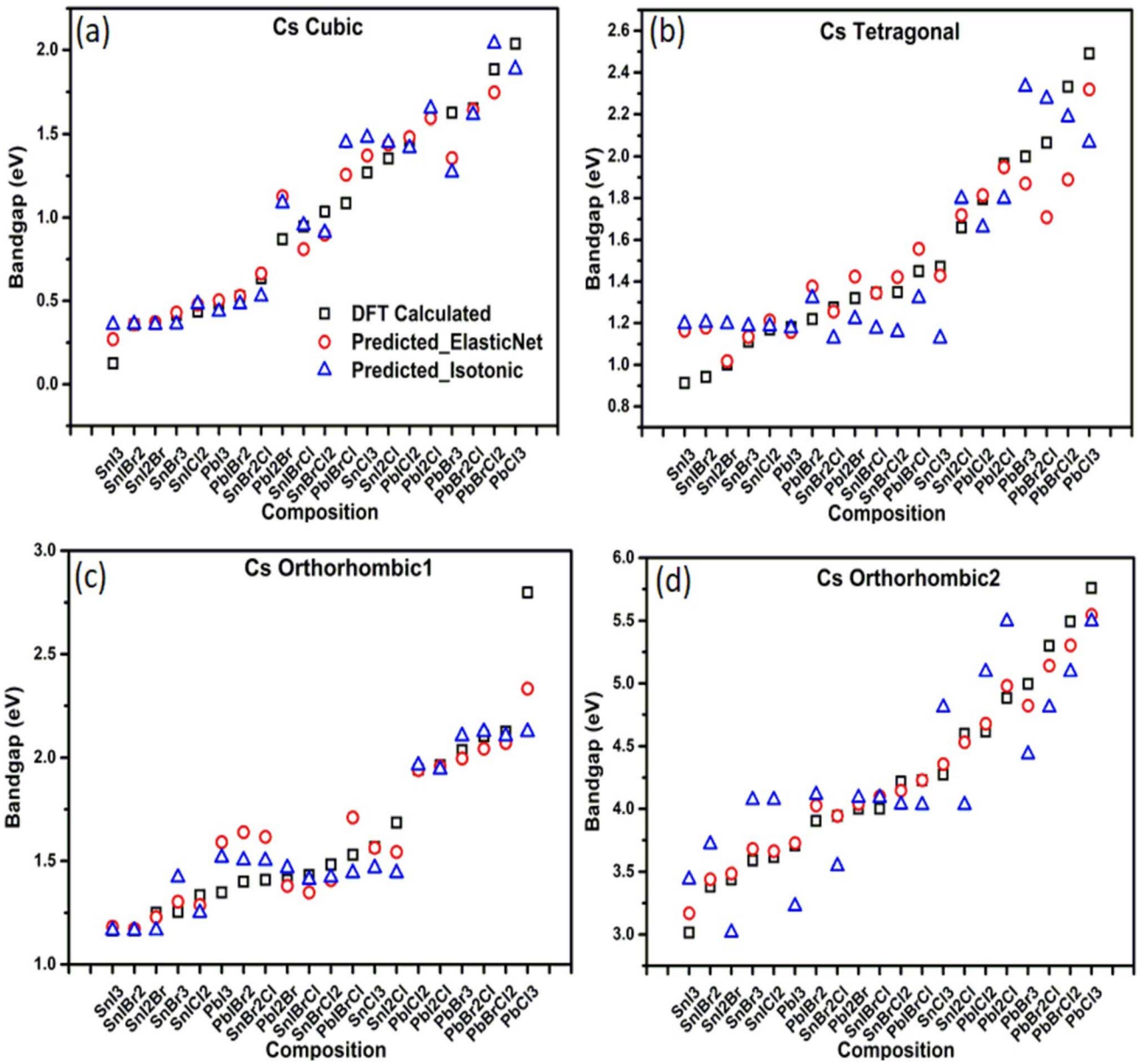

Fig. 6 Predicted and calculated bandgap values the Cs-based perovskites (a) cubic, (b) tetragonal, (c) orthorhombic1 and (d) orthorhombic2 symmetries using ElasticNet and isotonic regression algorithms. Reproduced from ref. 112 with permission from Elsevier, copyright 2022.

Extremely Randomized Trees (Extra Trees, ET) have been used. These methods combine multiple trees to improve robustness, reduce variance, and handle complex datasets. On this line, Mammeri *et al.*, trained an extra trees model on 1050 perovskite device samples with varied materials, deposition methods, and storage conditions.[118] This model was used to identify the optimal combinations of layers and processing parameters for both regular and inverted cells. By examining the constituent decision trees and applying feature importance analysis, the study demonstrated the ability of ensemble ML methods to capture non-linear relationships between materials, fabrication processes, and long-term stability, highlighting the critical role of manufacturing techniques in achieving high-performance, durable PSCs. In addition, as the interest in Indoor Perovskite Solar Cells (IPSCs) is increasing due to their potential to efficiently power IoT devices, more research in this area is done. For example, Mishra *et al.* developed a machine learning bandgap prediction model (BPM) to identify optimal wide-bandgap perovskite materials.[119] All six algorithms used (LR, XGB, AdB, KNN, SVR, and RF) achieved low RMSE values, with RF showing the highest r value with lowest RMSE. Simulations using the selected materials propose excellent indoor efficiencies exceeding 35%, showing that ML methodologies can effectively guide the discovery of new perovskite compositions for high-performance IPSCs.

Gradient Boosting (GB) models work as an ensemble of weak prediction learners, usually sequentially ordered DTs, where each new tree corrects the errors of the previous ones, thereby

© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 04 1447. Downloaded on 25/11/47 10:58:40 .
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

improving overall predictive accuracy. Different GB algorithms have been adapted in the literature to improve the performance of ML models related with perovskite research. For instance, CatBoost models, which adds optimized handling of categorical variables and regularization to reduce overfitting, have demonstrated superior predictive performance for bandgap estimation of ABX_3 perovskites for large datasets. In the testing phase, CatBoost achieved the lowest prediction errors and the highest coefficients of determination ($R^2 \geq 0.88$) compared to other methods such as XGBoost, RF, CompoundNet, LightGBM, and decision trees, as reported by D. O. Obada *et. al.*[120] This improved accuracy, however, comes at the cost of significantly longer training times. In contrast, Adaptive Boosting (AdaBoost), focus on misclassified instances and refine the subsequent DTs according to the error of the current prediction. It is commonly known as a one of the fastest GB models to implement although its predictive performance is usually lower than other methods. Finally, XGBoost has emerged as the most widely used GB algorithm in PSC studies due to its strong predictive performance combined with short training and execution times. In their manuscript, N. Shrivastav and coworkers studied six cesium-based PSCs using SCAPS simulations and machine learning models (LR, SVR, NN, RF, XGBoost) trained on 2160 datasets varying absorber material, thickness, and defect density. XGBoost achieved the highest predictive accuracy ($R^2 = 0.9999$), with SHAP analysis revealing that the absorber type and thickness most strongly influence efficiency, highlighting $CsPbI_3$ as the most promising candidate for stable, high-performance PSCs.[121]

Collectively, these studies highlight the importance of selecting the appropriate GB variant depending on the research goal, with CatBoost excelling in accuracy-driven applications and XGBoost providing a practical balance for high-throughput screening and device design.

Neural Networks (NN): neural networks are a group of machine learning algorithms inspired by the human brain that mimic how the neurons work collectively to obtain information from different inputs. This model is based on nodes or neurons, organized in a layered structure, that are interconnected by edges that represent the human synapses. Each neuron receives signals from other neurons, processes these signals using an activation function (typically non-linear), and passes the result to the next layer of neurons thorough the corresponding synapses. During the training phase, the weights of these connections (synapses) are adjusted through optimization techniques to minimize the error in predictions, allowing the network to learn from the data. In the field of perovskite research, Bak *et al.* developed a deep neural network to accurately predict the structure of high-performance leadfree perovskite-based solar cells.[122]

Fig. 7 compares the performance of a neural network, a random forest model, and a GB algorithm to predict the efficiency of PSCs. According to the researchers, the GB algorithm outperformed other approaches.[123,124] In a related study, the authors also performed a SHAP analysis to determine the impact of each variable, being the absorber layer and thickness the most relevant parameters in predicting the PCE of PSCs.[121] Although well-established approaches like GB can often outperform NNs, especially when working with relatively small datasets, the development of NN solutions in materials science remains highly relevant. These algorithms are particularly effective at handling high-dimensional descriptors and capturing complex nonlinear relationships.

As section summary, model choice in PSC studies should reflect the data type and the research question. For descriptor-based datasets (composition, device-stack and processing descriptors), tree ensembles (random forest; gradient boosting such as XGBoost/LightGBM) are dependable first picks because they capture non-linear interactions, accommodate mixed descriptor types with minimal scaling, and provide interpretable behavior; regularized linear models (ridge/lasso/elastic-net) remain fast and transparent baselines but typically underperform when relationships are strongly non-linear. Kernel methods (SVM; Gaussian processes) are competitive on medium-sized datasets; Gaussian processes additionally offer calibrated uncertainty (*i.e.*, they quantify confidence in predictions) at the cost of higher computational demands. Deep networks become advantageous when learning directly from images or spectra/time-series (*e.g.*, SEM/AFM, PL/UV-vis, *JV* or degradation traces), provided careful regularization (constraints to prevent overfitting) and data augmentation (label-preserving transformations) are used. For stability classification, ensemble methods are widely used and tend to perform strongly relative to simpler baselines.

As practical guidance, gradient boosting and random forest frequently emerge as effective options for composition-to-property prediction and for power-conversion-efficiency modeling from device or processing descriptors; Gaussian processes are often employed when calibrated uncertainty is important; convolutional neural networks are typically well suited to imaging data; one-dimensional convolutional networks or Gaussian process regression (or partial least squares when datasets are very small) are commonly used for spectral signals; and linear or PLS models remain useful as transparent baselines in data-scarce regimes. In terms of accuracy-compute trade-offs, computational cost generally increases from linear/PLS (lowest) through trees/random forest (low-moderate) to gradient boosting and SVM (moderate-high), with Gaussian processes and deep networks typically requiring the most training effort, though deep models can be efficient at inference once trained.

Model training and validation: the training and validation stage is where machine learning models establish relationships between input data and target tasks, ultimately determining their reliability for real-world applications. While the technical workflow of parameter initialization, iterative optimization and validation testing, is well established, its implementation in perovskite research poses unique challenges. PSC datasets are often small, heterogeneous, and can also be with noisy data, that means that choices in hyperparameter tuning or optimization strategies can disproportionately affect predictive accuracy and reproducibility. This underscores the need for careful benchmarking and transparent reporting of training protocols,

Open Access Article. Published on 04 1447. Downloaded on 25/11/47 10:58:40 .
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

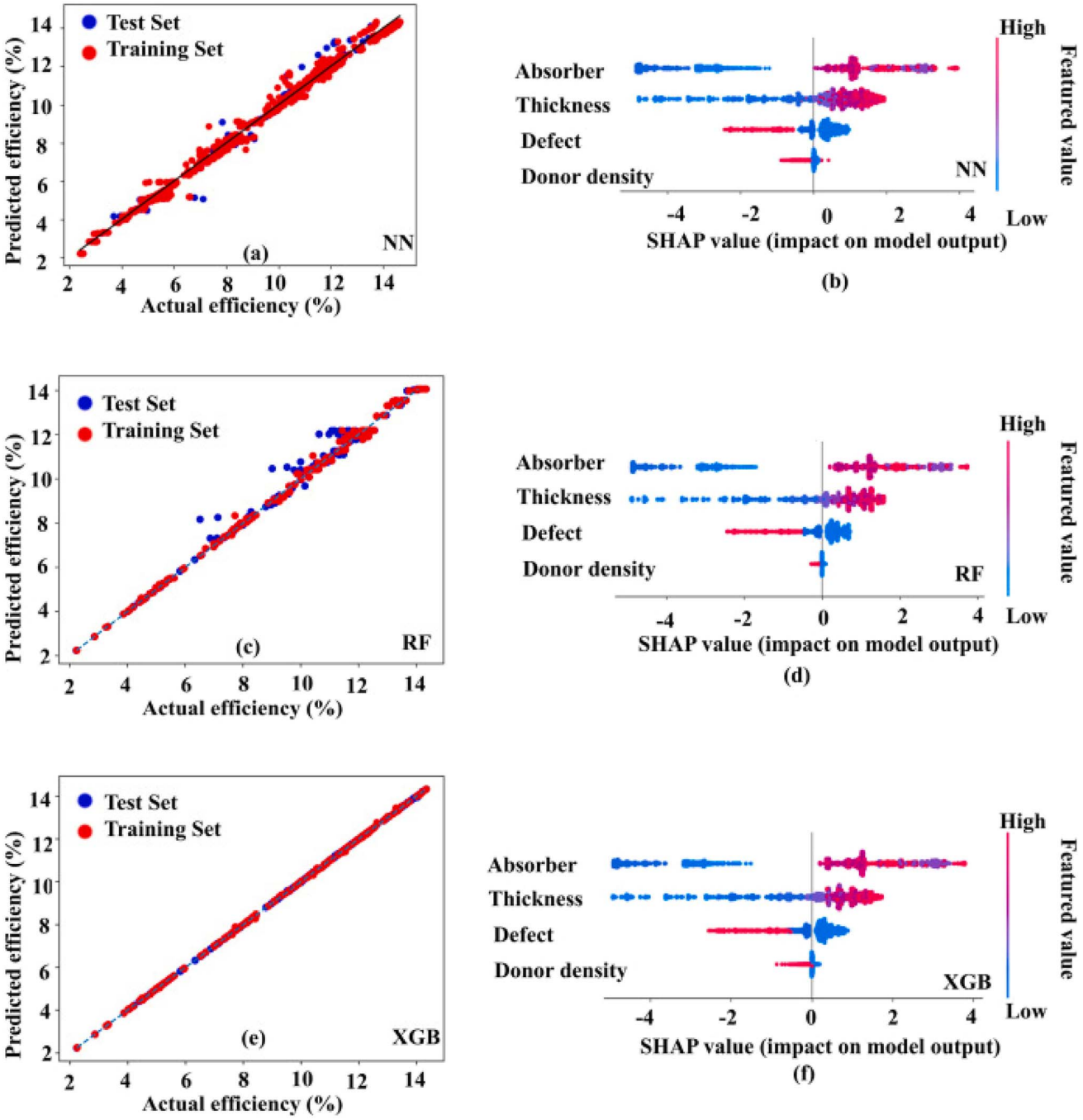

Fig. 7 Predicted efficiency *versus* actual efficiency of Cs-based perovskite-based solar cells and associated SHAP value for three different algorithms. From top to bottom: neural network (a,b), random forest (c,d), and XGBoost (e,f). Reproduced from ref. 121 with permission from Elsevier, copyright 2024.

as seemingly minor changes can lead to large discrepancies in model outputs.

On this regard, validation is particularly critical in PSC research, where overfitting is a frequent risk due to the limited size of available datasets. Models that perform exceptionally on training data may fail to generalize when tested on unseen devices or material compositions, limiting their utility for discovery or optimization tasks. Conversely, underfitting reflects insufficient model complexity to capture the intricate nonlinearities of PSC behavior, while data imbalance, for instance in stability datasets where unstable devices dominate, can bias predictions toward the majority class. Corrective strategies such as cross-validation, resampling, data augmentation, and adjusting model complexity can mitigate these issues, but their application requires domain knowledge to avoid artificially inflating performance.

Performance evaluation: assessing performance in machine learning models for PSC research goes far beyond reporting

© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 04 1447. Downloaded on 25/11/47 10:58:40 .
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

accuracy, as different metrics capture complementary aspects of model quality and can even lead to conflicting conclusions if used in isolation. For instance, in regression models, the coefficient of determination (R^2) may be used to determine proportion of the variance in the dependent variable that is predictable from the independent variables, whereas the root mean square error (RMSE) provides an estimation of how accurately the model can predict the target value by measuring the average difference between the actual values and those predicted by the model. While R^2 is often used to demonstrate the explanatory power of a model, a high R^2 with a large RMSE may still be scientifically meaningless if the absolute errors are too large to guide material optimization. In PSC research, where experimental uncertainties are already significant, RMSE often provides a more practical assessment of predictive utility. In classification tasks, accuracy alone is particularly misleading in the presence of imbalanced datasets. For example, using a dataset of stability studies where most devices fail rapidly. Here, metrics like precision, recall, and F_1-score are more informative. For example, high recall but low precision might suggest that the model correctly identifies most unstable devices but at the expense of many false alarms, which could slow down screening pipelines. Conversely, high precision but low recall could overlook promising candidates. Balancing these metrics through the F_1-score or domain-driven weighting is essential. For unsupervised tasks such as clustering, indices like the silhouette score and Davies–Bouldin index quantify how well materials are grouped based, for example, on structural or electronic descriptors. However, in PSC research these indices must be interpreted cautiously: a numerically "optimal" clustering may not correspond to chemically meaningful aggrupation (*e.g.*, grouping compounds by synthesis route rather than intrinsic properties). Thus, coupling clustering metrics with expert domain knowledge is vital for optimal interpretation. Finally, in dimensionality reduction tasks,[18] explained variance and reconstruction error assess how much structural or compositional information is retained when reducing feature space. Yet, in PSC research, retaining interpretability may be as important as preserving variance: a low-dimensional representation that highlights chemically meaningful trends (*e.g.*, tolerance factor *vs.* bandgap) may be more valuable than one that optimizes statistical variance but obscures physical relationships.

Overall, the choice of evaluation metrics is not merely a technical detail but a scientific decision. Each metric emphasizes different trade-offs, such as generalization *versus* interpretability, sensitivity *versus* specificity, or variance explained *versus* error magnitude. Selecting the right metrics is therefore essential to ensure that ML models not only perform well statistically but also generate insights that are physically meaningful and truly useful for advancing perovskite solar cell research.

Result interpretation: a persistent challenge in applying ML models is that most of them behave as "black boxes," producing predictions without clear insight into the underlying physical mechanisms. In the case of perovskites, this lack of interpretability limits their usefulness in guiding rational material or device design, since knowing why a model makes a prediction is often as important as the prediction itself.[125] To shine some light on this, explanation methods such as SHAP (Shapley Additive exPlanations) have been increasingly adopted. SHAP assigns quantitative weights to individual features. Importantly, SHAP is model-agnostic and additive, ensuring consistent and transparent feature attribution across different algorithms. Within the context of PSCs, such interpretability tools have proven especially valuable. For example, as shown in Fig. 8, Mishra *et al.* used correlation matrixes and SHAP analyses to establish the relative importance of the evaluated features on the prediction of perovskite performance comparing different algorithms including Decision Tree Regression (DTR), *K*-Nearest Neighbors (KNN), Light Gradient Boosting Machine (LightGBM), Random Forest (RF), and XGBoost.[126] Here, they not only discussed which models performed best but also which descriptors influenced efficiency predictions most strongly. This type of analysis goes beyond reporting accuracy metrics: it provides mechanistic insights that can inform material synthesis or device engineering.

While SHAP has become the most widely adopted interpretability tool in perovskite solar cell research due to its robustness and ability to capture non-linear effects, it is not the only option available. Other approaches, such as Local Interpretable Model-agnostic Explanations (LIME), Partial Dependence Plots (PDPs), and feature importance rankings also provide complementary perspectives but differ in scope, reliability, and computational demands. No single interpretability method is universally sufficient: SHAP offers detailed and consistent explanations but can be computationally intensive; LIME is lightweight and effective for local analyses, though often less stable; and PDPs or feature importance give intuitive global insights but may miss complex feature interactions, that are critical to PSC performance.

4. Application cases of machine learning in perovskite research

Many factors influence the performance and stability of perovskites and perovskite-based devices, being a complex multifactorial problem with many parameters to be optimized to maximize their efficiency while minimizing the associated pollution, waste, and expenses. All the processes in the device's life cycle are subject to optimization including the synthesis of the components, assembly, storage, deployment, maintenance, and recycling. Fig. 9 is an example of a flowchart for a two-step ML solution for perovskite discovery. Nonetheless, a similar workflow can be adapted to other aspects of perovskite research.[84]

Given the complexity of the challenges associated with perovskite research, ML has emerged as a powerful tool.[17,127–132] Actual reports focus on the development of ML protocols for predicting properties of halide perovskites[133] and the discovery of new perovskites[134] for solar power conversion,[135,136] light emitting diodes,[137] photodetectors,[138] and of perovskite sensors for monitoring the gas in lithium-ion battery.[139] The following

Open Access Article. Published on 04 1447. Downloaded on 25/11/47 10:58:40 .
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

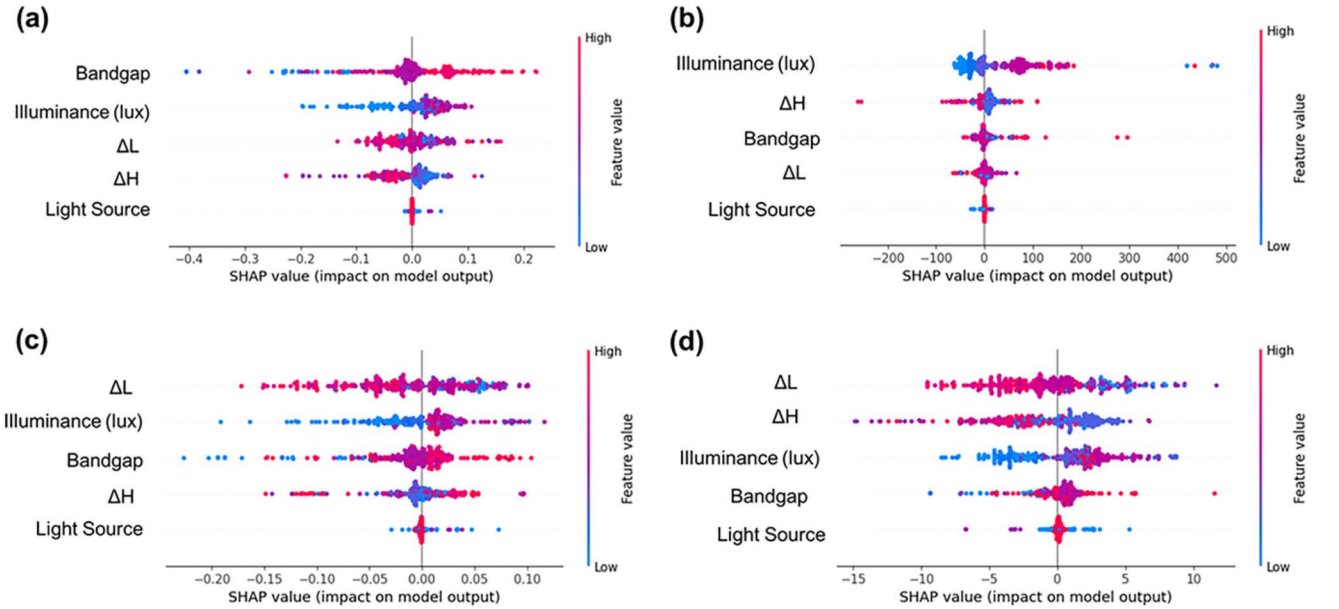

Fig. 8 SHAP analysis for the main performance parameters; (a) V_{OC}, (b) J_{SC}, (c) FF, and (d) PCE, predicted using a decision tree regression algorithm. Reproduced from ref. 126 with permission from John Wiley and Sons, copyright 2024.

section aims to highlight the main ML-based solutions for different aspects of the production and life cycle of perovskites for solar power generation.

One of the most exciting applications of ML to perovskite research is the use of advanced algorithms to discover potential candidates for high-efficiency perovskite compositions. Due to experimental limitations and time constraints, the number of compositions that can be tested and compared in the same experimental conditions is limited. Furthermore, most of the compositions will underperform the current literature, making the discovery of new perovskites costly and time-consuming. ML can help researchers to screen hundreds or even thousands of potential compositions and proceed only with the most promising formulations.[140–142] Yang *et al.* applied a multi-fidelity RF regression model and a genetic algorithm (GA) trained with a high-throughput DFT dataset of halide perovskite alloys to obtain thousands of promising materials with low decomposition energy. Here the researchers combined theoretical decomposition, energies bandgaps, and photovoltaic efficiencies with experimental data collected from the literature. In this case, both models are employed sequentially; the RF model is trained for property prediction while a GA is employed for inverse design of new perovskite compositions.[143] This work demonstrates the potential of mixed datasets of both computational and experimental data to expand the pool of available stable and efficient materials. Alternatively, other ML screening approaches are being adopted to elucidate new compositions targeting a single desired property without inverse design. For example, Kumar *et al.* trained a RF algorithm with open-source data focusing on bandgap prediction with formability and stability filters to predict 6855 new candidates and 7 prototype structures (Fig. 10).[144] In addition, novel narrow band gap inorganic halide perovskites were discovered using XGBoost by Li *et al.* with an accuracy higher than 0.9.[145] To do so, the authors employed the open-source Matminer Python package. Importantly, by performing a SHAP analysis, the authors linked the electronegativity range to the possibility of obtaining new perovskites with narrow bandgaps. Other gradient boosting algorithms were also employed to discover new perovskites. Specifically, LGBM showed a F_1-score of 90% on an unseen set and 176 promising perovskite compositions in terms of stability and band gap were identified, of which 153 were not previously reported.[146] Convolutional Neural Networks (CNN) can also be used to predict novel perovskite compositions. In this regard, a CNN was trained with available data in a public repository to predict new lead-free halide perovskite compositions. The candidates were examined with DFT calculations to assess their stability and theoretical band gap. Here the versatility of research workflows integrating ML approaches is demonstrated, as time-intensive DFT calculations can be minimized to the most viable candidates.[147] Also, different models can be used in different stages of the material discovery research, for example Lu *et al.* used GB classification and regression models to discover 151 ferroelectric perovskites with optimal band gaps for photovoltaics, achieving >90% accuracy. Here, the authors started with a large pool of unexplored candidate compounds, of which more than 1000 perovskite and non-perovskite materials were employed as the training set. A classification algorithm was used to determine whether the unexplored compounds are perovskites or non-perovskite materials and regression models were employed on the evaluated set to determine the structure and band gap.[148]

It is worth to note that some properties like the dimensionality of new perovskites may be directly linked to their applicability in different scenarios. In this line, Lyu *et al.* developed a ML-assisted approach to predict the dimensionality and

© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 04 1447. Downloaded on 25/11/47 10:58:40 .
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

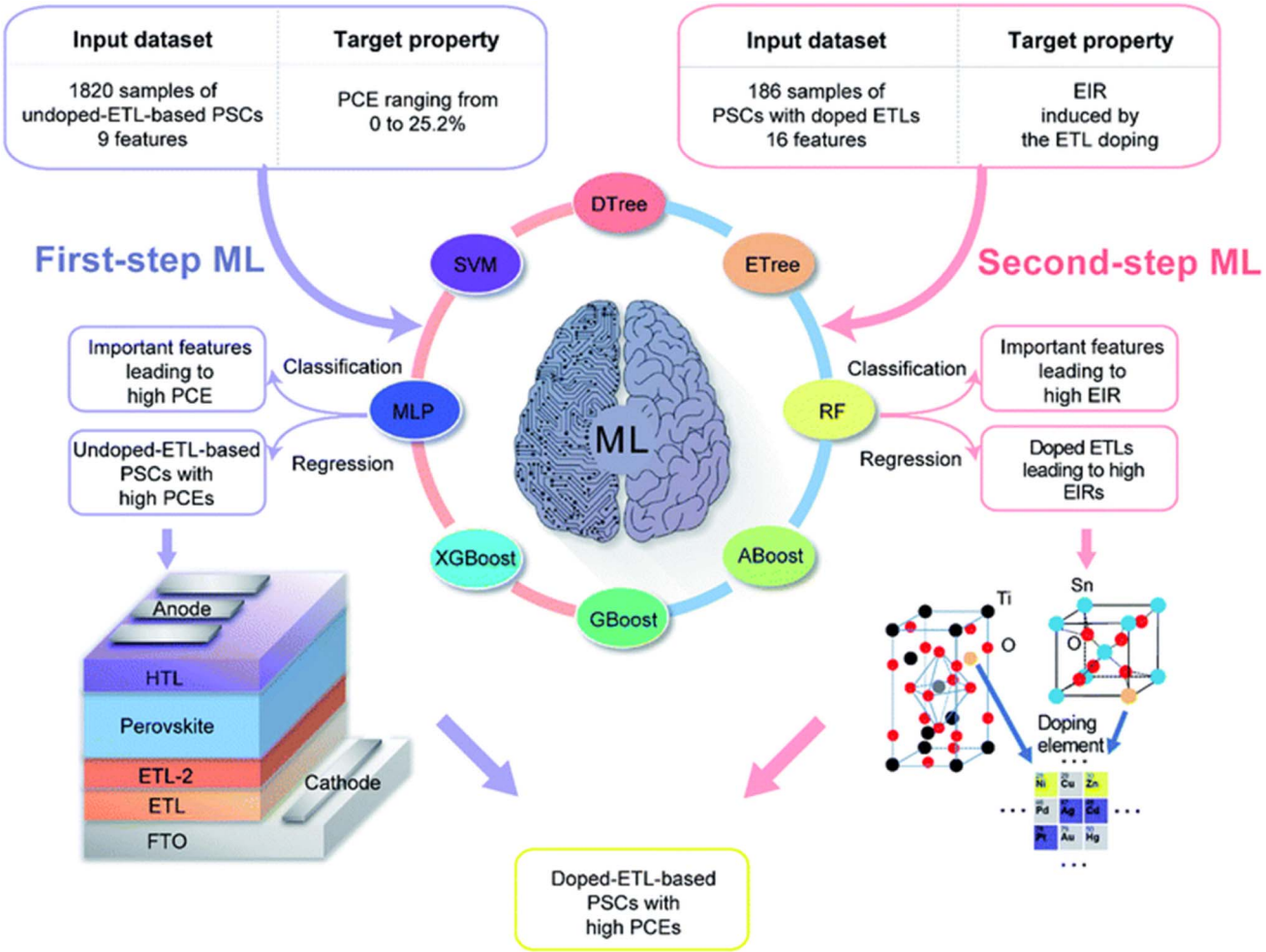

Fig. 9 Flowchart of a two-step ML solution for guided high-efficiency perovskite discovery. Reproduced from ref. 84 with permission from the Royal Society of Chemistry, copyright 2021.

crystal structure of mixed halide perovskites using classification algorithms. Notably, the model was able to predict the dimensionality of both newly synthesized perovskites, and examples extracted from the literature with a 79% prediction accuracy.[149] Finally, identifying candidates with feasible preparation conditions in an industrial setting is decisive in the development of new perovskites for solar applications. In this regard, Pendleton *et al.* explored whether ML could predict halide perovskite crystallization without explicit physicochemical data, using GB, KNN, and linear SVM algorithms.[150] This work represents a leap forward towards ML-assisted preparation of novel materials in relevant scenarios.

The previous examples of ML approaches on perovskite discovery are mainly based on datasets collected from experimental evidence, theoretical computational data, or available information from the literature or repositories and specifically trained for a predefined purpose. Alternatively, general-purpose large language models (LLM) are primarily generalist models trained with undisclosed datasets to provide the users with responses to virtually any prompt. Recently, generative AI (in particular, Chat GPT) has been used by Chen and co-workers to generate candidates that were further filtered and tested using computational approaches to discover new high-performance perovskites and perovskite-based devices.[151] While this work illustrates a promising trend in materials research, this approach remains largely limited by the lack of explicit physicochemical grounding, the opacity of the dataset and the lack of transparency in feature attributions. However, these hurdles may be tackled in the future by developing domain-specific generative models in materials science.

Considering all above mentioned, machine learning has emerged as a transformative tool in perovskite research, enabling rapid screening of vast compositional spaces and guiding the discovery of novel materials for solar cells and other optoelectronic devices. While models such as RF, GF, and CNN have demonstrated high predictive accuracy for band gap and stability, their applicability remains constrained by data biases, limited experimental validation, and narrow chemical coverage. Recent advances, including generative AI for candidate ideation, signal a shift toward human–AI co-design, though care must be taken to ensure chemical plausibility and sustainability. To fully realize the potential of ML in perovskite

Open Access Article. Published on 04 1447. Downloaded on 25/11/47 10:58:40 .
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

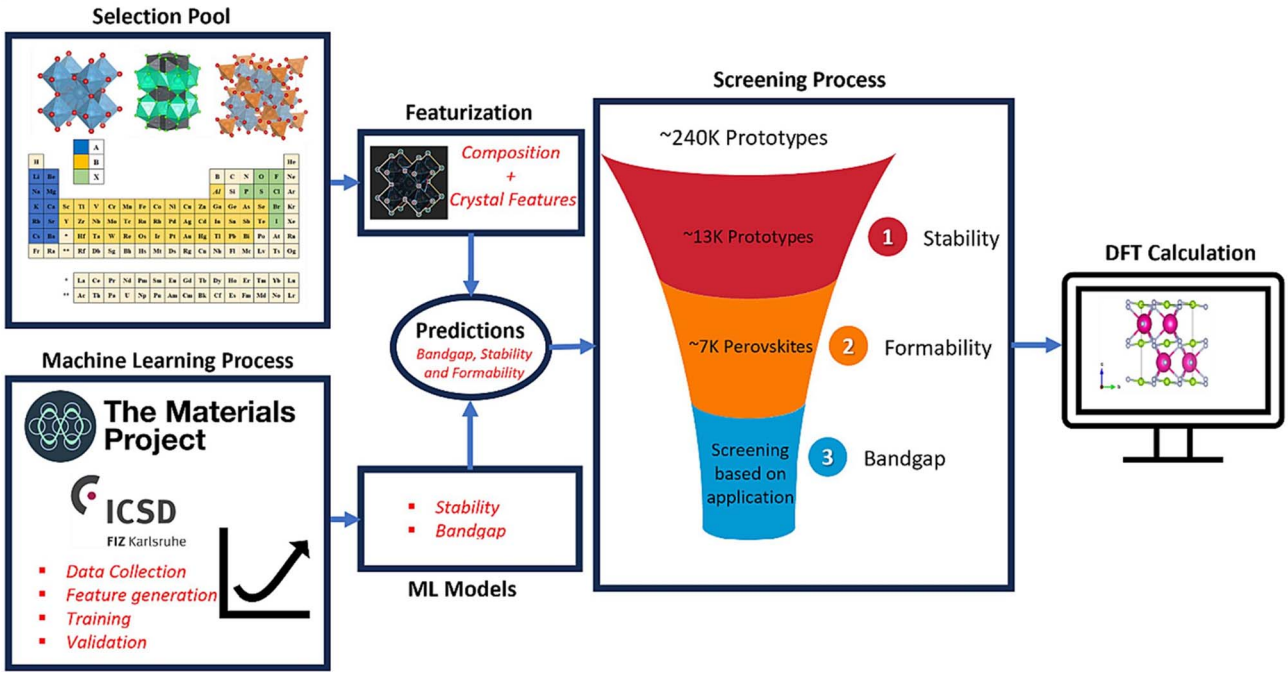

Fig. 10 Workflow of a ML process for the discovery of new perovskite materials for energy applications. More than 240 000 protype perovskite were screened in term of thermodynamic stability and formability. Then, their band gap is predicted using a RF algorithm trained with an open-source materials library. The predictions are then validated using DFT calculations. Reproduced from ref. 144 with permission from Elsevier, copyright 2023.

research, future efforts should focus on integrating predictive models with experimental workflows, curating diverse and standardized datasets, and developing evaluation metrics that reflect practical device performance.

Once the candidate structures are determined, the synthesis of the desired components remains a challenge. Similarly, the adaptation of the processes carried out in a laboratory setting are usually difficult to test and implement in a large-scale industrial environment. Fig. 11 shows an example of a support vector classification (SVC) algorithm to predict the synthesis feasibility of different perovskites based on their components. A SHAP analysis is also performed to determine the most relevant descriptors in the prediction of synthesis feasibility.[152] The use of ML for the discovery of new perovskite and device preparation procedures with higher yield, safer solvents and antisolvents, or synthesis protocols that yield byproducts that are easier to handle is a current field of study. SVM algorithms have been developed to assess the feasibility of proposed synthesis methods for the preparation of two-dimensional perovskites, and the optimization of the experimental conditions.[152–154] Kirman *et al.* employed high-throughput screening using a CNN and a RF regressor to accelerate single-crystal perovskite synthesis, successfully identifying optimal growth conditions. In this approach, the authors use dual-model approach. A CNN is trained to discriminate between crystals and non-crystals, and a ML regressor based on RF relates the experimental parameters to the likelihood of successful crystal growth.[155] Gaussian processing with a NN and a RF classifier was successfully implemented to predict the best experimental conditions to optimize the synthesis of perovskite nanocrystals in a computationally affordable manner. Using this approach, the authors were able to optimize the preparation of 2–6 monolayer thick nanoplatelets with superior photoluminescence characteristics.[156] Similarly, a ML framework was used to evaluate the formation of perovskite 2D nanosheets. In this case, the authors explore a joint spectral-kinetic model trained with both DFT and experimental data and identify two mechanisms involved in the complex formation processes.[157] Also, different ML algorithms including KNN, SVM, NN, and GB methods were applied for the prediction of the crystal structure of perovskite materials. In this context, the authors demonstrated superior performance of XGBoost for this application.[158] In the same line, XGBoost provided the best accuracy on the prediction of the experimental conditions to obtain Ruddlesden–Popper and Dion–Jacobson 2D lead halide perovskites among 26 ML models including variations of different RF, KNN, SVM, and GB algorithms.[159] Regarding the preparation of optoelectronic perovskite-based devices, Wang *et al.* combined a RF model with a genetic algorithm to determine the optimal device architecture and fabrication conditions for the preparation of vapor-deposited solar cells. A SHAP analysis was also performed, showing that the ratio of cations to anions in the perovskite layer and the annealing temperature are the leading contributions to power conversion efficiency.[160] Also, ML approaches were adapted to screen new interface materials and passivation materials in p–i–n type PSCs.[161–163]

© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 04 1447. Downloaded on 25/11/47 10:58:40 .
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

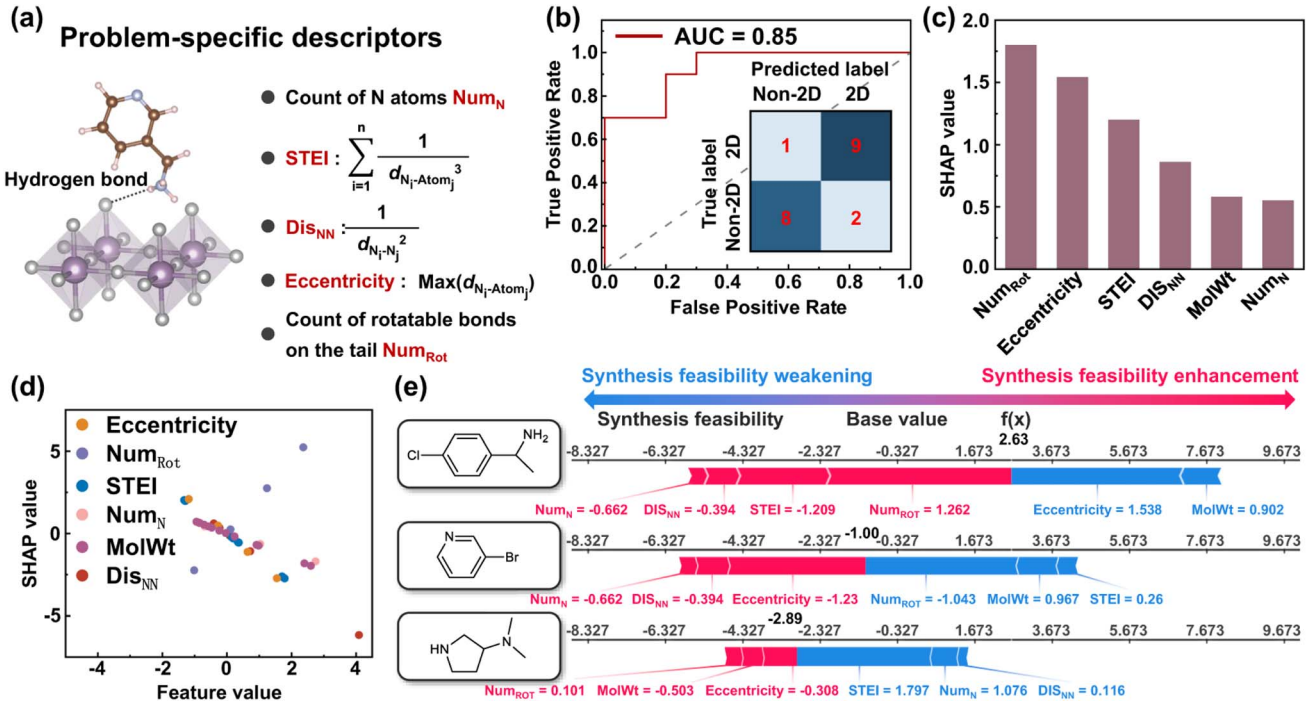

Fig. 11 Summary of an application for the evaluation of the synthesis feasibility of different 2D perovskites. Descriptors used in the SVC algorithm (a), confusion matrix for the feasibility of 2D perovskites synthesis (b), and SHAP values for the different descriptors (c and d). SHAP analysis representing the positive and negative contributions on the synthesis of selected 2D perovskites and overall feasibility bolded black (e). Reproduced from ref. 152 with permission from Springer Nature, under a Creative Commons Attribution 4.0 International License.

Machine learning is increasingly being applied to address the challenges of perovskite synthesis and device manufacturing, particularly in scaling laboratory protocols to industrial settings. Algorithms such as SVMs, CNNs, and Gaussian processes have demonstrated utility in predicting synthesis feasibility, controlling crystal growth, and optimizing fabrication parameters. However, current approaches often rely on limited datasets and oversimplified metrics, which may not capture the complexity of real-world synthesis or device performance. To enhance reliability and scalability, future efforts should focus on curating comprehensive datasets, integrating ML with automated synthesis platforms, and developing evaluation strategies that account for environmental and long-term stability considerations.

In this section we discuss how ML models are used to predict and optimize the performance of perovskite-based devices. Various algorithms, including XGBoost, RF, CatBoost, and deep learning, have been used to decrease the dependence on traditional trial and error methods. In this line, performance is a key factor to take into consideration to determine the applicability of perovskite-based devices in the current and highly competitive market. However, testing experiments are time-consuming and expensive and relies on highly specialized equipment and trained personnel to test the synthesized materials and devices. In this regard, ML approaches allow scientists and research centers to focus on the most promising materials.[164]

Considering that efficiency is one of the most important features of new materials and devices, multiple studies have been focused on trying to understand and predict it. Fig. 12 shows an example of the implementation of a XGBoost algorithm to accurately predict the photoelectric conversion efficiency, open-circuit voltage, short-circuit current, and fill factor of perovskite-based devices.[165] Similarly, XGBoost was employed by Yılmaz *et al.* to predict the efficiency of solar cells with small discrepancies between the predictions and the observed values.[166] Pindolia *et al.* determined the theoretical efficiency of $KSnI_3$-based PSCs using a RF algorithm,[167] CatBoost regression was employed by Khan *et al.*[168] In these manuscripts, the authors used different algorithms to predict PCE, demonstrating the feasibility of machine learning in predicting the device performance before experiments and opening the door to reverse experimental design of highly efficient PSCs. Building on this idea, an important next step was to systematically compare how different algorithms perform in this predictive task. In this context, Yang A. *et al.* evaluated and compared the PCE prediction capabilities of several widely used algorithms, including LR, SVR, ANN, DT, RF, LGBM, XGBoost, and CatBoost. Among these, RF and gradient boosting techniques (LGBM, XGBoost, and CatBoost) demonstrated superior performance based on RMSE and R^2 metrics.[169]

Other relevant parameters in perovskite research have been predicted using ML approaches. For example, RF has been used to predict the band gaps (E_g) of two-dimensional halide perovskites,[170] while XGBoost displayed superior capabilities for the prediction of the band gap of organic–inorganic hybrid perovskites.[171] On this line, Taeseo, *et al.* presented a new approach for accurately predicting experimental band gap values using

Open Access Article. Published on 04 1447. Downloaded on 25/11/47 10:58:40 .
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

machine learning models enhanced with transfer learning. They developed surrogate models based on E_g values calculated using the GGA functional, incorporating easily accessible features from chemical formulas. These models demonstrated superior predictive performance, achieving a coefficient of determination (R^2) of 0.817 and a mean absolute error of 0.289 eV.[172] Yang *et al.* used a dataset containing 2079 experimental PSCs to predict PCE values using SHapley Additive exPlanations (SHAP), achieving a coefficient of determination (R^2) value of 0.76, they reported that by following the optimization strategy suggested by the model, they successfully enhanced the device's PCE to 25.01%.[173] Mamunur Rashid *et al.* used DFT and Monte Carlo simulations to train a ML model and obtain crucial theoretical insights into the molecular dynamics governing hole mobility in PSCs, aiming to understand its role in developing efficient hole-transporting materials (HTMs) for enhanced perovskite solar cell performance.[174] Moreover, researchers from Karlsruhe Institute and Helmholtz AI published a study in where they present three use cases of how deep learning augments complex experimental data analysis to process monitoring of scalable perovskite thin-film fabrication. Specifically, the reported approach allows to: (1) monitor material composition based on the precursors ensuring consistency during fabrication, (2) predict thin-film quality and preliminary device performance, and (3) recommend process control based on the final performance forecasts.[175]

Stability of perovskites in the operational conditions is another key factor to take into consideration to gauge the applicability of perovskite-based solar power devices.[176,177] More stable perovskites will lead to longer lifetimes, reducing the waste and pollution linked to the synthesis, transportation and installation of new cells. Also, by understanding the processes behind perovskite degradation, new recycling protocols can be discovered.[178] In this regard, researchers from University of Washington have developed physiochemical machine learning models to predict operational lifetimes of $MAPbI_3$ solar cells. Also, different regression algorithms are compared to predict the stability of perovskite based solar cells with lasso regression showing the best predictive performance as shown in Fig. 13. The model can be used to determine the parameters responsible for better operational lifetimes.[179] XGBoost has been reported as a powerful tool for the thermodynamic stability of perovskites. Notably, by analyzing feature attributions, authors were able to identify a strong dependance between the predicted stability and the A-site elements, providing valuable insights towards the preparation of stable perovskite materials.[180] Other gradient boosting algorithms like LightGBM also showed up to 92% classification accuracy in the prediction of

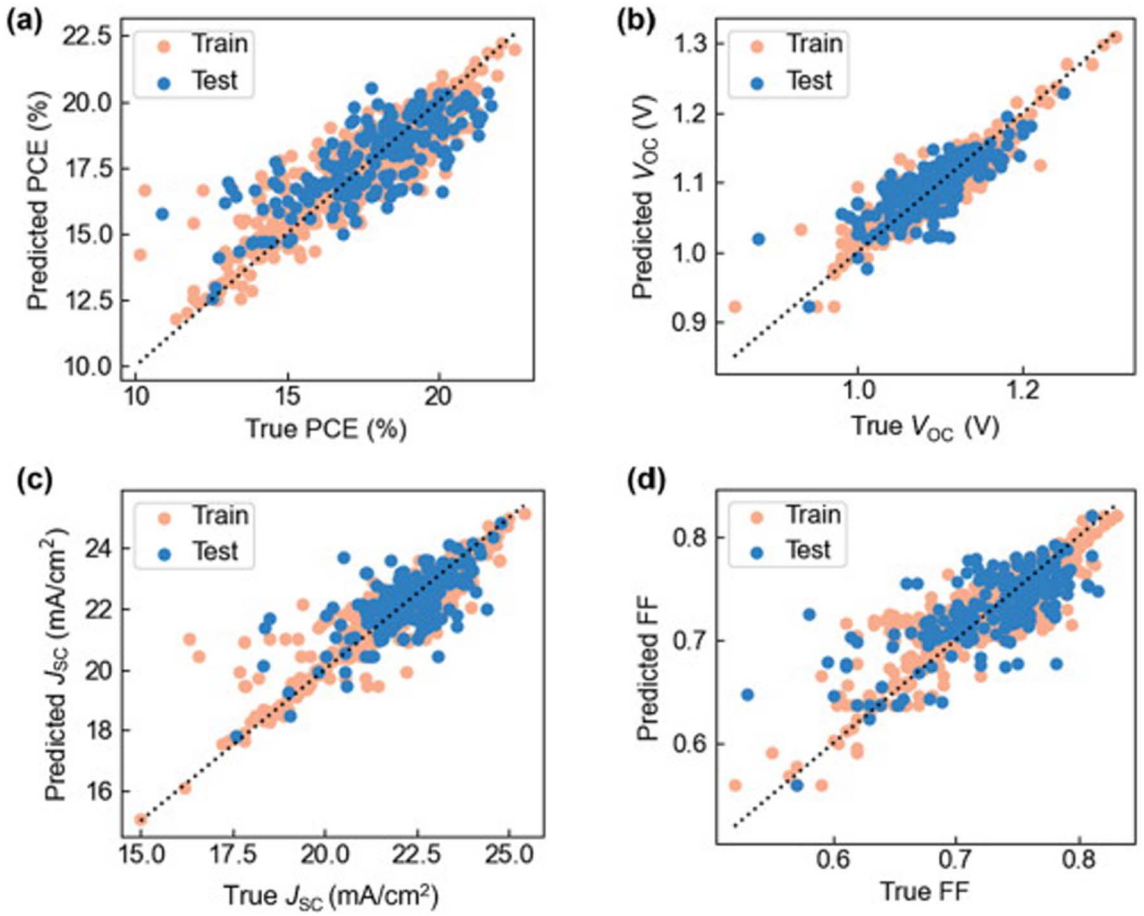

Fig. 12 Comparison of predicted performance features including photoelectric conversion efficiency (a), open-circuit voltage (b), short-circuit current (c), and fill factor (d) with their actual values using XGBoost. Reproduced from ref. 165 with permission from Elsevier, copyright 2023.

© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 04 1447. Downloaded on 25/11/47 10:58:40 .
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

thermodynamic stability metrics of double perovskites. Interestingly, the authors of this work arrive to a similar conclusion on the importance of cation composition in the design of highly-stable perovskites for energy applications.[181] Importantly, the final stability of perovskite-based devices may not be limited by the stability of the perovskite materials instead. Regarding the stability of PSC, recent studies by Mammeri M. *et al.* reported the use of DT-based algorithms and NN to determine the impact of environmental conditions on the degradation of PSCs. For instance, ML allowed the researchers to identify the use of hydrophobic layers to greatly influence the stability of PSC.[182,183]

Considering that surfaces and interfaces play a crucial role in the stability of PSCs,[57,59,184–186] integrating machine learning techniques such as photoelectron spectroscopy (PES) analysis can further enhance surface characterization by identifying complex trends in large datasets, improving signal interpretation, and enabling predictive modeling of perovskite behavior under different environmental conditions.[187] This approach can accelerate the optimization of PSCs by uncovering new correlations between material properties and stability under illumination.[188,189] The screening of the thermal and mechanical stability of perovskites-base materials is also relevant.[190–194] Jaafreh R. *et al.* developed a machine learning-assisted approach to a rapid and reliable screening for mechanically stable perovskite-based materials.[195] By employing ML algorithms it is possible to extrapolate relevant information that can be used for material selection. ML force fields (FF) have been successfully implemented to reduce the computational costs of simulations related to the strain-induced grain boundary stabilization.[196] Moreover, PCA and PLS were used by Boubchir *et al.* to predict which perovskites and inverse perovskites are promising in terms of hardness and fracture toughness for their use as thermal barrier coatings.[197]

ML algorithms can be trained with big sets of experimental data and linked to conclusions drawn from experts across different fields and backgrounds in different ML-guided research approaches. With careful design and training, it is possible to obtain ML algorithms capable of drawing meaningful conclusions and suggest pertinent actions according to real-time inputs. In this regard, ML is an important tool in the fields of applied and fundamental research. For example, a RF system has been trained to help identifying relationships between halide perovskite properties extracted from X-ray detector data. The system helped the authors to identify the main factors influencing the bandgap and performance of the synthesized materials.[198] NN have been also used to perform grain-analysis from the raw experimental data to quantitative numerical data, as shown in Fig. 14, this information was further used to investigate the relationship between the microscopic grain structure and the device performance.[199] Also, a RF model coupled with appropriate features including geometry-driven and key structural modes was employed to accurately predict cation ordering in double perovskites.[200] Finally, ML was employed for the interpretation of complex electrochemical impedance spectroscopy (EIS) data and accurately predict the low-frequency (50 Hz to 300 mHz) EIS response of new materials.[201] The developments recently reported on the field of ML applied to perovskite research, new expert systems are expected to help scientists and technologists in relevant decision-making tasks while dealing with procedures on complex systems such as perovskite-based optoelectronic devices.

5. Challenges and limitations of machine learning in perovskite solar cells research

When integrated into the research workflow, ML has shown great potential in accelerating the discovery and optimization of perovskites and PSCs. This review highlights some of the main tools and applications for integrating ML techniques in

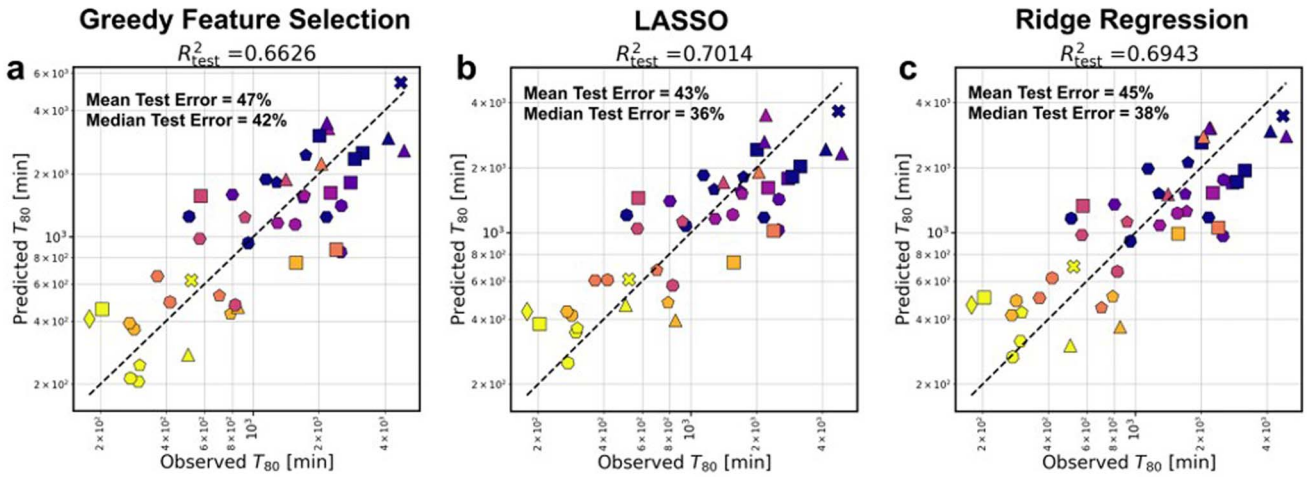

Fig. 13 Predicted *versus* observed stability of perovskite-based solar cells at different operating conditions with varying temperature and relative humidity using different algorithms: Greedy Feature Selection (a), LASSO (b), and Ridge Regression (c). Reproduced from ref. 179 with permission from the Royal Society of Chemistry, copyright 2024.

Open Access Article. Published on 04 1447. Downloaded on 25/11/47 10:58:40 .
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

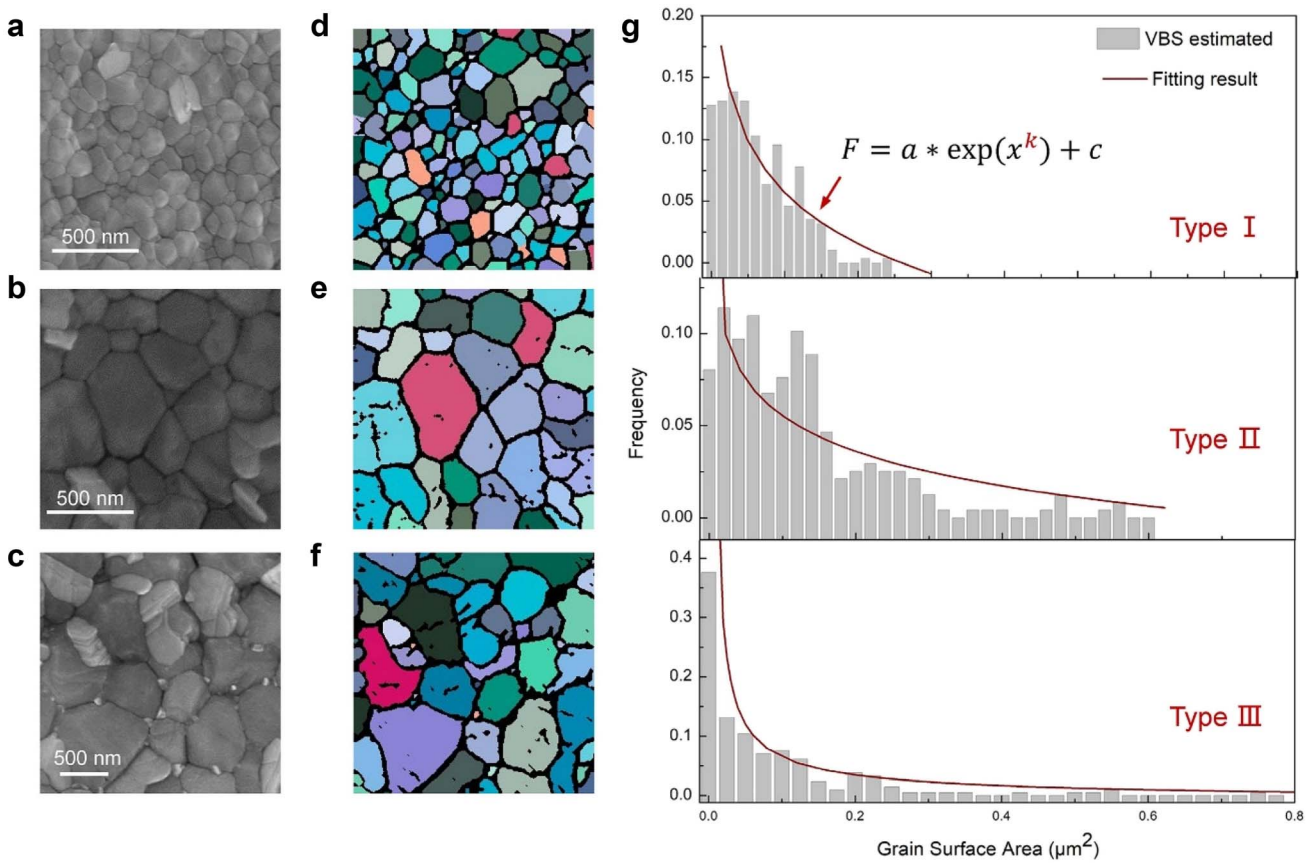

Fig. 14 SEM images of perovskite surfaces with different grain structure (a–c), results of the CNN-based grain extractor tool corresponding to the SEM images (d–f), statistical grain surface areas (GSA) estimated by the CNN algorithm and (g) fitting corresponding to a Weibull distribution. Reproduced from ref. 199 with permission from Elsevier, copyright 2024.

materials science research. However, several challenges and limitations challenge the implementation of these techniques on a larger scale. Hence, these hurdles must be properly identified and addressed to continue with its effective implementation in this field.

One major challenge is the availability and quality of data. ML models require large, diverse, and high-quality datasets to generate reliable predictions. However, existing PSC datasets are often limited in size, inconsistent in measurement conditions, or biased toward specific material compositions and device architectures. This lack of standardized, high-quality data hinders the generalizability of ML models. Large amounts of data can be extracted from traditional computational methods like DFT. However, these methods are generally time-consuming and computationally expensive, which in turn limits their scalability for high-throughput data generation. ML approaches present the potential to greatly reduce the time and computational burden of DFT and other computational approaches by learning surrogate models that can generate large amounts of data without the need for first-principles calculations. Another limitation is the computational cost associated with training complex ML models, particularly for deep learning architectures that require extensive training on large datasets. Balancing accuracy and computational efficiency remain an ongoing challenge. Some alternatives to tackle this problem include transfer learning from models trained in related datasets, active learning by using ML to identify the most informative data for the intended application, or hybrid models combining ML with physical constraints. Alternatively emerging approaches such as cloud computing and parallelization can speed up the handling of large datasets.

Interpretability of ML models is another critical issue. Some ML techniques, especially deep learning, function as "black boxes," making it difficult to understand how predictions are made. In materials science, interpretability is crucial for gaining insights into the underlying physical and chemical principles governing perovskite stability and efficiency. Tools like SHAP have been proven effective towards identifying the key physicochemical parameters that can be further guide the experimental design and development of new materials and architectures. In the case of deep learning, approaches like autoencoder latent space visualization can help to explore patterns used by the model by projecting latent variables. However, feature attribution remains a challenge in fully unsupervised models.

Also, extrapolation to new materials and conditions remains challenging, most ML models perform well within the range of data they have been trained on but struggle when predicting the properties of novel perovskite compositions or device structures. Ensuring that models can generalize beyond known data

© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 04 1447. Downloaded on 25/11/47 10:58:40 .
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

is essential for accelerating materials discovery. In this context, leveraging knowledge from pre-trained models or applying cross-domain learning strategies could provide valuable insights and enhance predictive capabilities in this field.

Finally, integration with experimental workflows also present several challenges, while ML can suggest promising materials or device architectures, translating these predictions into practical experiments requires collaboration between computational and experimental researchers. In this line, the iterative process of ML model validation and refinement through experiments is still largely underdeveloped in PSC research and can help with the integration of ML tools in common research practices.

Addressing these challenges requires the development of standardized datasets, improved model interpretability, efficient computational strategies, and stronger collaboration between data scientists and experimentalists. However, overcoming these limitations is key to unlock the full potential of ML in advancing PSC technology.

6. Future perspectives and emerging trends in machine learning for materials scientists

The integration of ML in perovskite solar cell (PSC) research is rapidly evolving, with several emerging trends shaping the future of this field.[202] As computational techniques and experimental methodologies advance, ML-driven approaches are expected to play an even greater role in materials' discovery, device optimization, and stability improvement.

In this line, the combination of ML with robotic systems may pave the way for autonomous experimentation schemes, where experiments are designed and conducted automatically. The concept of "self-driving lab" combines closed-loop systems where ML designs experiments with robots to execute them. These systems are envisioned as active learning strategies, where the data is fed back to refine the ML models, allowing for iterative loops that can be tailored to optimize an experimental workflow according to predefined goals. The use of cloud computing in this field must be encouraged to provide access to vast resources and shared data, fostering collaboration to accelerate research by facilitating the use of transfer learning techniques.

Future autonomous labs will go beyond simply automating tasks. They could integrate advanced sensing techniques like hyperspectral imaging, and *in situ* characterization allowing for the collection of richer data that will lead to more accurate ML predictions. Also, the use of augmented and mixed reality can also be adapted to experimental routines by providing scientists, engineers and technicians with relevant contextual information in real time. Furthermore, if we keep following this trend, robots will become more intelligent, utilizing AI algorithms, such as reinforcement learning, making autonomous decisions and navigating complex material design spaces.

Labs will also become more flexible with modular robotic systems that can be easily adapted for different materials and experiments. Another key advancement will be the integration of multiscale modeling, predicting material properties from the atomic level up to the macroscopic scale, which will greatly enhance our understanding and predictive capabilities. Finally, we could also expect the rise of digital twins (virtual replicas of materials and processes) enabling researchers to conduct experiments in a digital environment, reducing the need for physical trials and speeding up the development process.

In summary, AI and ML tools are expected to play an increasingly important role in shaping the workflows of future materials scientists and technologists. Their applications range from assisting in data collection and interpretation to supporting ML-guided experimental design and providing specialized tools that can streamline various processes in materials research.

7. Conclusions

ML has emerged as a powerful tool for accelerating advancements in PSCs, offering innovative approaches to materials discovery, device optimization, and stability enhancement. By leveraging large datasets and predictive algorithms, ML enables more efficient material screening, improved synthesis protocols, and enhanced performance predictions.

Here, the main algorithms used in materials research were introduced, along with the general workflows and main applications where ML have been proven advantageous. In this regard, the intersection of ML and perovskite research presents both opportunities and challenges to materials scientists. While ML can significantly enhance efficiency and help discover new materials, a deeper understanding of computational methods and interdisciplinary collaboration is essential to harness its full potential. Future efforts should focus on expanding high-quality standardized databases, refining ML algorithms tailored for perovskite research, and developing explainable AI models to bridge the gap between computational predictions and experimental validation.

As the field evolves, ML-driven approaches are expected to play a crucial role in overcoming the limitations of traditional trial-and-error methods, paving the way for more stable, efficient, and scalable perovskite solar technologies. By addressing current challenges and embracing the potential of ML, the materials science community can drive forward the next generation of PSCs and contribute to the broader transition toward sustainable energy solutions.

Conflicts of interest

There are no conflicts to declare.

Data availability

No primary research results, software or code have been included and no new data were generated or analysed as part of this review.

Open Access Article. Published on 04 1447. Downloaded on 25/11/47 10:58:40 .
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Acknowledgements

We thank the Swedish Research Council (Grant No. VR 2022-03168) and the Göran Gustafsson foundation for funding. This work was partially supported by the Wallenberg Initiative Materials Science for Sustainability (WISE) funded by the Knut and Alice Wallenberg Foundation. We also acknowledge Project PID2023-147404OB-I00 funded by MCIN/AEI/10.13039/501100011033/ERDF, UE, and Xunta de Galicia (Grant ED431C 2022/44). A. G.-F. acknowledges support from a Beatriz Galindo junior fellowship (BG23/00033) from the Spanish Ministry of Science and Innovation.

References

1 H. S. Jung and N. G. Park, Perovskite Solar Cells: From Materials to Devices, *Small*, 2015, **11**(1), 10–25, DOI: 10.1002/SMLL.201402767.

2 M. Noman, Z. Khan and S. T. Jan, A Comprehensive Review on the Advancements and Challenges in Perovskite Solar Cell Technology, *RSC Adv.*, 2024, **14**(8), 5085–5131, DOI: 10.1039/D3RA07518D.

3 J. Huang, Y. Yuan, Y. Shao and Y. Yan, Understanding the Physical Properties of Hybrid Perovskites for Photovoltaic Applications, *Nat. Rev. Mater.*, 2017, **2**(7), 1–19, DOI: 10.1038/natrevmats.2017.42.

4 T. Wu, Z. Qin, Y. Wang, Y. Wu, W. Chen, S. Zhang, M. Cai, S. Dai, J. Zhang, J. Liu, Z. Zhou, X. Liu, H. Segawa, H. Tan, Q. Tang, J. Fang, Y. Li, L. Ding, Z. Ning, Y. Qi, Y. Zhang and L. Han, The Main Progress of Perovskite Solar Cells in 2020–2021, *Nano-Micro Lett.*, 2021, **13**(1), 1–18, DOI: 10.1007/S40820-021-00672-W/FIGURES/4.

5 C. Yang, W. Hu, J. Liu, C. Han, Q. Gao, A. Mei, Y. Zhou, F. Guo and H. Han, Achievements, Challenges, and Future Prospects for Industrialization of Perovskite Solar Cells, *Light: Sci. Appl.*, 2024, **13**(1), 1–48, DOI: 10.1038/s41377-024-01461-x.

6 The National Renewable Energy Laboratory, *Best Research-Cell Efficiencies*, https://www.nrel.gov/pv/interactive-cell-efficiency.html, accessed 2025-03-17.

7 S. L. Hamukwaya, H. Hao, Z. Zhao, J. Dong, T. Zhong, J. Xing, L. Hao and M. M. Mashingaidze, A Review of Recent Developments in Preparation Methods for Large-Area Perovskite Solar Cells, *Coatings*, 2022, **12**(2), 252, DOI: 10.3390/COATINGS12020252.

8 D. Li, D. Zhang, K.-S. Lim, Y. Hu, Y. Rong, A. Mei, N.-G. Park, H. Han, D. Li, D. Zhang, Y. Hu, Y. Rong, A. Mei, H. Han, K. Lim and N. Park, A Review on Scaling Up Perovskite Solar Cells, *Adv. Funct. Mater.*, 2021, **31**(12), 2008621, DOI: 10.1002/ADFM.202008621.

9 Y. Ma and Q. Zhao, A Strategic Review on Processing Routes towards Scalable Fabrication of Perovskite Solar Cells, *J. Energy Chem.*, 2022, **64**, 538–560, DOI: 10.1016/J.JECHEM.2021.05.019.

10 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 and T. Riedl, Perovskite–Organic Tandem Solar Cells, *Nat. Rev. Mater.*, 2024, **9**(3), 202–217, DOI: 10.1038/s41578-023-00642-1.

11 E. Aydin, T. G. Allen, M. D. Bastiani, A. Razzaq, L. Xu, E. Ugur, J. Liu and S. De Wolf, Pathways toward Commercial Perovskite/Silicon Tandem Photovoltaics, *Science*, 2024, **383**(6679), 1–13, DOI: 10.1126/SCIENCE.ADH3849.

12 M. J. Talukder, M. Akteruzzaman, Z. Hasan and E. Haque, Advances in high-efficiency solar photovoltaic materials: a comprehensive review of perovskite and tandem cell technologies, *American Journal of Advanced Technology and Engineering Solutions*, 2025, **1**(01), 201–225, DOI: 10.63125/5AMNVB37.

13 T. A. Chowdhury, M. A. Bin Zafar, M. Sajjad-Ul Islam, M. Shahinuzzaman, M. A. Islam and M. U. Khandaker, Stability of Perovskite Solar Cells: Issues and Prospects, *RSC Adv.*, 2023, **13**(3), 1787–1810, DOI: 10.1039/D2RA05903G.

14 R. Sharma, A. Sharma, S. Agarwal and M. S. Dhaka, Stability and Efficiency Issues, Solutions and Advancements in Perovskite Solar Cells: A Review, *Sol. Energy*, 2022, **244**, 516–535, DOI: 10.1016/J.SOLENER.2022.08.001.

15 Q. Tao, P. Xu, M. Li and W. Lu, Machine Learning for Perovskite Materials Design and Discovery, *npj Comput. Mater.*, 2021, **7**(1), 1–18, DOI: 10.1038/s41524-021-00495-8.

16 J. P. Correa-Baena, K. Hippalgaonkar, J. van Duren, S. Jaffer, V. R. Chandrasekhar, V. Stevanovic, C. Wadia, S. Guha and T. Buonassisi, Accelerating Materials Development via Automation, Machine Learning, and High-Performance Computing, *Joule*, 2018, **2**(8), 1410–1420, DOI: 10.1016/j.joule.2018.05.009.

17 M. Chen, Z. Yin, Z. Shan, X. Zheng, L. Liu, Z. Dai, J. Zhang, S. Liu and Z. Xu, Application of Machine Learning in Perovskite Materials and Devices: A Review, *J. Energy Chem.*, 2024, **94**, 254–272, DOI: 10.1016/J.JECHEM.2024.02.035.

18 T. Mueller, A. G. Kusne and R. Ramprasad, Machine Learning in Materials Science, *Rev. Comput. Chem.*, 2016, **29**, 186–273, DOI: 10.1002/9781119148739.CH4.

19 L. Mao and C. Xiang, A Comprehensive Review of Machine Learning Applications in Perovskite Solar Cells: Materials Discovery, Device Performance, Process Optimization and Systems Integration, *Mater. Today Energy*, 2025, **47**, 101742, DOI: 10.1016/J.MTENER.2024.101742.

20 S. Subba, P. Rai and S. Chatterjee, Machine Learning Approaches in Advancing Perovskite Solar Cells Research, *Adv. Theory Simul.*, 2025, **8**(3), 2400652, DOI: 10.1002/ADTS.202400652.

21 K. Leng, W. Fu, Y. Liu, M. Chhowalla and K. P. Loh, From Bulk to Molecularly Thin Hybrid Perovskites, *Nat. Rev. Mater.*, 2020, **5**(7), 482–500, DOI: 10.1038/s41578-020-0185-1.

22 G. Grancini and M. K. Nazeeruddin, Dimensional Tailoring of Hybrid Perovskites for Photovoltaics, *Nat. Rev. Mater.*, 2018, **4**(1), 4–22, DOI: 10.1038/s41578-018-0065-0.

23 A. Garcia-Fernandez, E. J. Juarez-Perez, S. Castro-Garcia, M. Sanchez-Andujar and M. A. Señaris-Rodriguez, Hybrid

© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 04 1447. Downloaded on 25/11/47 10:58:40 .
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Organic-Inorganic Perovskites: A Spin-off of Oxidic Perovskites, in *Perovskites and Other Framework Structures Crystalline Materials*, Collaborating Academics, 2021, vol. 1, pp. 143–174, DOI: 10.23647/ca.md20202205.

24 J. Han, K. Park, S. Tan, Y. Vaynzof, J. Xue, E. W.-G. Diau, M. G. Bawendi, J.-W. Lee and I. Jeon, Perovskite Solar Cells, *Nat. Rev. Methods Primers*, 2025, **5**(1), 1–27, DOI: 10.1038/s43586-024-00373-9.

25 A. S. R. Bati, Y. L. Zhong, P. L. Burn, M. K. Nazeeruddin, P. E. Shaw and M. Batmunkh, Next-Generation Applications for Integrated Perovskite Solar Cells, *Commun. Mater.*, 2023, **4**(1), 1–24, DOI: 10.1038/s43246-022-00325-4.

26 Z. Zhang, Z. Li, L. Meng, S. Y. Lien and P. Gao, Perovskite-Based Tandem Solar Cells: Get the Most Out of the Sun, *Adv. Funct. Mater.*, 2020, **30**(38), 2001904, DOI: 10.1002/ADFM.202001904.

27 P. Subudhi and D. Punetha, Progress, Challenges, and Perspectives on Polymer Substrates for Emerging Flexible Solar Cells: A Holistic Panoramic Review, *Prog. Photovoltaics Res. Appl.*, 2023, **31**(8), 753–789, DOI: 10.1002/PIP.3703.

28 N. G. Park, Perovskite Solar Cells: An Emerging Photovoltaic Technology, *Mater. Today*, 2015, **18**(2), 65–72, DOI: 10.1016/J.MATTOD.2014.07.007.

29 S. Foo, M. Thambidurai, P. Senthil Kumar, R. Yuvakkumar, Y. Huang and C. Dang, Recent Review on Electron Transport Layers in Perovskite Solar Cells, *Int. J. Energy Res.*, 2022, **46**(15), 21441–21451, DOI: 10.1002/ER.7958.

30 A. García-Fernández, B. Kammlander, S. Riva, D. Kühn, S. Svanström, H. Rensmo and U. B. Cappel, Interface Energy Alignment between Lead Halide Perovskite Single Crystals and TIPS-Pentacene, *Inorg. Chem.*, 2023, **62**(38), 15412–15420, DOI: 10.1021/ACS.INORGCHEM.3C01482.

31 Y. Yao, C. Cheng, C. Zhang, H. Hu, K. Wang, S. De Wolf, Y. Yao, C. Cheng, C. Zhang, K. Wang, H. Hu and S. De Wolf, Organic Hole-Transport Layers for Efficient, Stable, and Scalable Inverted Perovskite Solar Cells, *Adv. Mater.*, 2022, **34**(44), 2203794, DOI: 10.1002/ADMA.202203794.

32 H. Pan, X. Zhao, X. Gong, H. Li, N. H. Ladi, X. L. Zhang, W. Huang, S. Ahmad, L. Ding, Y. Shen, M. Wang and Y. Fu, Advances in Design Engineering and Merits of Electron Transporting Layers in Perovskite Solar Cells, *Mater. Horiz.*, 2020, **7**(9), 2276–2291, DOI: 10.1039/D0MH00586J.

33 S. Svanström, T. J. Jacobsson, G. Boschloo, E. M. J. Johansson, H. Rensmo and U. B. Cappel, Degradation Mechanism of Silver Metal Deposited on Lead Halide Perovskites, *ACS Appl. Mater. Interfaces*, 2020, **12**(6), 7212–7221, DOI: 10.1021/ACSAMI.9B20315.

34 S. Svanström, A. García-Fernández, T. J. Jacobsson, I. Bidermane, T. Leitner, T. Sloboda, G. J. Man, G. Boschloo, E. M. J. Johansson, H. Rensmo and U. B. Cappel, The Complex Degradation Mechanism of Copper Electrodes on Lead Halide Perovskites, *ACS Mater. Au*, 2022, **2**(3), 301–312, DOI: 10.1021/ACSMATERIALSAU.1C00038.

35 F. Meng, D. Wang, J. Chang, J. Li and G. Wang, Application of Carbon Materials in Conductive Electrodes for Perovskite Solar Cells, *Sol. RRL*, 2024, **8**(6), 2301030, DOI: 10.1002/SOLR.202301030.

36 J. He, Y. Bai, Z. Luo, R. Ran, W. Zhou, W. Wang and Z. Shao, Advanced Carbon-Based Rear Electrodes for Low-Cost and Efficient Perovskite Solar Cells, *Energy Environ. Sci.*, 2025, **18**(5), 2136–2164, DOI: 10.1039/D4EE05462H.

37 J. Yang, Q. Cao, T. Wang, B. Yang, X. Pu, Y. Zhang, H. Chen, I. Tojiboyev, Y. Li, L. Etgar, X. Li and A. Hagfeldt, Inhibiting Metal-Inward Diffusion-Induced Degradation through Strong Chemical Coordination toward Stable and Efficient Inverted Perovskite Solar Cells, *Energy Environ. Sci.*, 2022, **15**(5), 2154–2163, DOI: 10.1039/D1EE04022G.

38 M. Ghasemi, B. Guo, K. Darabi, T. Wang, K. Wang, C. W. Huang, B. M. Lefler, L. Taussig, M. Chauhan, G. Baucom, T. Kim, E. D. Gomez, J. M. Atkin, S. Priya and A. Amassian, A Multiscale Ion Diffusion Framework Sheds Light on the Diffusion–Stability–Hysteresis Nexus in Metal Halide Perovskites, *Nat. Mater.*, 2023, **22**(3), 329–337, DOI: 10.1038/s41563-023-01488-2.

39 M. Luo, A. Tarasov, H. Zhang and J. Chu, Hybrid Perovskites Unlocking the Development of Light-Emitting Solar Cells, *Nat. Rev. Mater.*, 2024, **9**(5), 295–297, DOI: 10.1038/s41578-024-00675-0.

40 C. Wu, R. Wang, Z. Lin, N. Yang, Y. Wu and X. Ouyang, Simultaneous Dual-Interface Modification Based on Mixed Cations for Efficient Inverted Perovskite Solar Cells with Excellent Stability, *Chem. Eng. J.*, 2024, **493**, 152899, DOI: 10.1016/J.CEJ.2024.152899.

41 M. Saliba, T. Matsui, J. Y. Seo, K. Domanski, J. P. Correa-Baena, M. K. Nazeeruddin, S. M. Zakeeruddin, W. Tress, A. Abate, A. Hagfeldt and M. Grätzel, Cesium-Containing Triple Cation Perovskite Solar Cells: Improved Stability, Reproducibility and High Efficiency, *Energy Environ. Sci.*, 2016, **9**(6), 1989–1997, DOI: 10.1039/C5EE03874J.

42 S. Wang, S. Pang, D. Chen, W. Zhu, H. Xi and C. Zhang, Improving Perovskite Solar Cell Performance by Compositional Engineering via Triple-Mixed Cations, *Sol. Energy*, 2021, **220**, 412–417, DOI: 10.1016/J.SOLENER.2021.03.036.

43 S. Ahmed, M. A. Gondal, A. S. Alzahrani, M. Parvaz, A. Ahmed and S. Hussain, Recent Trends and Challenges in Lead-Free Perovskite Solar Cells: A Critical Review, *ACS Appl. Energy Mater.*, 2024, **7**(4), 1382–1397, DOI: 10.1021/ACSAEM.3C02327.

44 W. Yu, Y. Zou, H. Wang, S. Qi, C. Wu, X. Guo, Y. Liu, Z. Chen, B. Qu and L. Xiao, Breaking the Bottleneck of Lead-Free Perovskite Solar Cells through Dimensionality Modulation, *Chem. Soc. Rev.*, 2024, **53**(4), 1769–1788, DOI: 10.1039/D3CS00728F.

45 A. García-Fernández, I. Marcos-Cives, C. Platas-Iglesias, S. Castro-García, D. Vázquez-García, A. Fernández and M. Sánchez-Andújar, Diimidazolium Halobismuthates $[Dim]_2[Bi_2X_{10}]$ (X = Cl^-, Br^-, or I^-): A New Class of Thermochromic and Photoluminescent Materials, *Inorg.*

Open Access Article. Published on 04 1447. Downloaded on 25/11/47 10:58:40 .
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Chem., 2018, **57**(13), 7655–7664, DOI: **10.1021/acs.inorgchem.8b00629**.

46 A. García-Fernández, E. J. Juarez-Perez, J. M. Bermúdez-García, A. L. Llamas-Saiz, R. Artiaga, J. J. López-Beceiro, M. A. Señarís-Rodríguez, M. Sánchez-Andújar and S. Castro-García, Hybrid Lead Halide $[(CH_3)_2NH_2]PbX_3$ (X = Cl^- and Br^-) Hexagonal Perovskites with Multiple Functional Properties, *J. Mater. Chem. C*, 2019, **7**(32), 10008–10018, DOI: **10.1039/c9tc03543e**.

47 A. García-Fernández, J. M. Bermúdez-García, S. Castro-García, A. L. Llamas-Saiz, R. Artiaga, J. J. López-Beceiro, M. Sánchez-Andújar and M. A. Señarís-Rodríguez, $_7Pb_4X_{15}$ (X = Cl^- and Br^-), 2D-Perovskite Related Hybrids with Dielectric Transitions and Broadband Photoluminiscent Emission, *Inorg. Chem.*, 2018, **57**(6), 3215–3222, DOI: **10.1021/acs.inorgchem.7b03217**.

48 A. García-Fernández, J. M. Bermúdez-García, S. Castro-García, A. L. Llamas-Saiz, R. Artiaga, J. López-Beceiro, S. Hu, W. Ren, A. Stroppa, M. Sánchez-Andújar and M. A. Señarís-Rodríguez, Phase Transition, Dielectric Properties, and Ionic Transport in the $[(CH_3)_2NH_2]PbI_3$ Organic-Inorganic Hybrid with 2H-Hexagonal Perovskite Structure, *Inorg. Chem.*, 2017, **56**(9), 4918–4927, DOI: **10.1021/acs.inorgchem.6b03095**.

49 O. Nazarenko, M. R. Kotyrba, M. Wörle, E. Cuervo-Reyes, S. Yakunin and M. V. Kovalenko, Luminescent and Photoconductive Layered Lead Halide Perovskite Compounds Comprising Mixtures of Cesium and Guanidinium Cations, *Inorg. Chem.*, 2017, **56**(19), 11552–11564, DOI: **10.1021/ACS.INORGCHEM.7B01204**.

50 F. B. Minussi, R. M. Silva, J. F. Carvalho and E. B. Araújo, Thermal Degradation in Methylammonium–Formamidinium–Guanidinium Lead Iodide Perovskites, *J. Mater. Chem. C*, 2024, **12**(14), 5138–5149, DOI: **10.1039/D4TC00395K**.

51 C. W. Lin, F. Liu, T. Y. Chen, K. H. Lee, C. K. Chang, Y. He, T. L. Leung, A. M. C. Ng, C. H. Hsu, J. Popović, A. Djurišić and H. Ahn, Structure-Dependent Photoluminescence in Low-Dimensional Ethylammonium, Propylammonium, and Butylammonium Lead Iodide Perovskites, *ACS Appl. Mater. Interfaces*, 2020, **12**(4), 5008–5016, DOI: **10.1021/ACSAMI.9B17881**.

52 Y. Huang, Y. Jiang, S. Zou, Z. Zhang, J. Jin, R. He, W. Hu, S. Ren and D. Zhao, Substitution of Ethylammonium Halides Enabling Lead-Free Tin-Based Perovskite Solar Cells with Enhanced Efficiency and Stability, *ACS Appl. Mater. Interfaces*, 2023, **15**(12), 15775–15784, DOI: **10.1021/ACSAMI.3C00299**.

53 H. L. Kagdada, B. Roondhe, V. Roondhe, S. D. Dabhi, W. Luo, D. K. Singh and R. Ahuja, Exploring A-Site Cation Variations in Dion–Jacobson Two-Dimensional Halide Perovskites for Enhanced Solar Cell Applications: A Density Functional Theory Study, *Adv. Energy Sustainability Res.*, 2024, **5**(1), 2300147, DOI: **10.1002/AESR.202300147**.

54 S. Jin, Can We Find the Perfect A-Cations for Halide Perovskites?, *ACS Energy Lett.*, 2021, **6**(9), 3386–3389, DOI: **10.1021/ACSENERGYLETT.1C01806**.

55 M. Lira-Cantu and K. Tabah Tanko, Perovskite Solar Cells: What Do You Mean When You Say "Stable.", *APL Energy*, 2024, **2**(3), DOI: **10.1063/5.0239002**.

56 P. Holzhey and M. A. Saliba, Full Overview of International Standards Assessing the Long-Term Stability of Perovskite Solar Cells, *J. Mater. Chem. A*, 2018, **6**(44), 21794–21808, DOI: **10.1039/C8TA06950F**.

57 P. Schulz, D. Cahen and A. Kahn, Halide Perovskites: Is It All about the Interfaces?, *Chem. Rev.*, 2019, **119**(5), 3349–3417, DOI: **10.1021/acs.chemrev.8b00558**.

58 S. Shao and M. A. Loi, The Role of the Interfaces in Perovskite Solar Cells, *Adv. Mater. Interfaces*, 2020, **7**(1), 1901469, DOI: **10.1002/admi.201901469**.

59 D. Luo, X. Li, A. Dumont, H. Yu and Z. H. Lu, Recent Progress on Perovskite Surfaces and Interfaces in Optoelectronic Devices, *Adv. Mater.*, 2021, **33**(30), 2006004, DOI: **10.1002/adma.202006004**.

60 S. Svanström, A. García Fernández, T. Sloboda, T. J. Jacobsson, H. Rensmo and U. B. Cappel, X-Ray Stability and Degradation Mechanism of Lead Halide Perovskites and Lead Halides, *Phys. Chem. Chem. Phys.*, 2021, **23**(21), 12479–12489, DOI: **10.1039/d1cp01443a**.

61 S. S. Dipta, M. A. Rahim and A. Uddin, Encapsulating Perovskite Solar Cells for Long-Term Stability and Prevention of Lead Toxicity, *Appl. Phys. Rev.*, 2024, **11**(2), 21301, DOI: **10.1063/5.0197154/3280360**.

62 J. Suo, H. Pettersson and B. Yang, Sustainable Approaches to Address Lead Toxicity in Halide Perovskite Solar Cells: A Review of Lead Encapsulation and Recycling Solutions, *EcoMat*, 2025, **7**(1), e12511, DOI: **10.1002/EOM2.12511**.

63 T. Wu, S. Mariotti, P. Ji, L. K. Ono, T. Guo, I. N. Rabehi, S. Yuan, J. Zhang, C. Ding, Z. Guo and Y. Qi, Self-Assembled Monolayer Hole-Selective Contact for Up-Scalable and Cost-Effective Inverted Perovskite Solar Cells, *Adv. Funct. Mater.*, 2024, **34**(32), 2316500, DOI: **10.1002/ADFM.202316500**.

64 A. Khorasani, F. Mohamadkhani, M. Marandi, H. Luo and M. Abdi-Jalebi, Opportunities, Challenges, and Strategies for Scalable Deposition of Metal Halide Perovskite Solar Cells and Modules, *Adv. Energy Sustainability Res.*, 2024, **5**(7), 2300275, DOI: **10.1002/AESR.202300275**.

65 P. Liu, H. Wang, T. Niu, L. Yin, Y. Du, L. Lang, Z. Zhang, Y. Tu, X. Liu, X. Chen, S. Wang, N. Wu, R. Qin, L. Wang, S. Yang, C. Zhang, X. Pan, S. Frank Liu and K. Zhao, Ambient Scalable Fabrication of High-Performance Flexible Perovskite Solar Cells, *Energy Environ. Sci.*, 2024, **17**(19), 7069–7080, DOI: **10.1039/D4EE02925A**.

66 L. Klein, S. Ziegler, F. Laufer, C. Debus, M. Götz, K. Maier-Hein, U. W. Paetzold, F. Isensee and P. F. Jäger, Discovering Process Dynamics for Scalable Perovskite Solar Cell Manufacturing with Explainable AI, *Adv. Mater.*, 2024, **36**(7), 2307160, DOI: **10.1002/ADMA.202307160**.

© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 04 1447. Downloaded on 25/11/47 10:58:40 .
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

67 M. I. Jordan and T. M. Mitchell, Machine Learning: Trends, Perspectives, and Prospects, *Science*, 2015, **349**(6245), 255–260, DOI: 10.1126/SCIENCE.AAA8415.

68 S. Subba, P. Rai and S. Chatterjee, Machine Learning Approaches in Advancing Perovskite Solar Cells Research, *Adv. Theory Simul.*, 2025, **8**(3), 2400652, DOI: 10.1002/ADTS.202400652.

69 J. Liang, T. Wu, Z. Wang, Y. Yu, L. Hu, H. Li, X. Zhang, X. Zhu and Y. Zhao, Accelerating Perovskite Materials Discovery and Correlated Energy Applications through Artificial Intelligence, *Energy Mater.*, 2022, **2**(3), 200016, DOI: 10.20517/ENERGYMATER.2022.14.

70 I. H. Sarker, Machine Learning: Algorithms, Real-World Applications and Research Directions, *SN Computer Science*, 2021, **2**(3), 1–21, DOI: 10.1007/S42979-021-00592-X.

71 D. Pandey, K. Niwaria and B. Chourasia, Machine Learning Algorithms-A Review, *Mach. Learn.*, 2019, **6**(2), 916–922.

72 A. Jain, H. Patel, L. Nagalapatti, N. Gupta, S. Mehta, S. Guttula, S. Mujumdar, S. Afzal, R. Sharma Mittal and V. Munigala, Overview and Importance of Data Quality for Machine Learning Tasks, *Afzal Proceedings of the 26th ACM SIGKDD International Conference on Knowledge* 2020, pp. 3561–3562, DOI: 10.1145/3394486.3406477.

73 T. Hagendorff, Linking Human And Machine Behavior: A New Approach to Evaluate Training Data Quality for Beneficial Machine Learning, *Minds Mach.*, 2021, **31**(4), 563–593, DOI: 10.1007/S11023-021-09573-8.

74 M. Y. Ng, A. Youssef, A. S. Miner, D. Sarellano, J. Long, D. B. Larson, T. Hernandez-Boussard and C. P. Langlotz, Perceptions of Data Set Experts on Important Characteristics of Health Data Sets Ready for Machine Learning: A Qualitative Study, *JAMA Netw. Open*, 2023, **6**(12), e2345892, DOI: 10.1001/jamanetworkopen.2023.45892.

75 D. Kirk, Good Practices and Common Pitfalls of Machine Learning in Nutrition Research, *Proc. Nutr. Soc.*, 2024, 1–14, DOI: 10.1017/S0029665124007638.

76 J. Stausberg and S. Harkener, Data Quality and Data Quantity: Complements or Contradictions?, *Stud. Health Technol. Inf.*, 2023, **305**, 24–27, DOI: 10.3233/SHTI230414.

77 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, C. Rehermann, D. Ramirez, D. Fairen-Jimenez, D. Di Girolamo, D. Jia, E. Avila, E. J. Juarez-Perez, F. Baumann, F. Mathies, G. S. A. González, G. Boschloo, G. Nasti, G. Paramasivam, G. Martínez-Denegri, H. Näsström, H. Michaels, H. Köbler, H. Wu, I. Benesperi, M. I. Dar, I. Bayrak Pehlivan, I. E. Gould, J. N. Vagott, J. Dagar, J. Kettle, J. Yang, J. Li, J. A. Smith, J. Pascual, J. J. Jerónimo-Rendón, J. F. Montoya, J.-P. Correa-Baena, J. Qiu, J. Wang, K. Sveinbjörnsson, K. Hirselandt, K. Dey, K. Frohna, L. Mathies, L. A. Castriotta, M. H. Aldamasy, M. Vasquez-Montoya, M. A. Ruiz-Preciado, M. A. Flatken, M. V. Khenkin, M. Grischek, M. Kedia, M. Saliba, M. Anaya, M. Veldhoen, N. Arora, O. Shargaieva, O. Maus, O. S. Game, O. Yudilevich, P. Fassl, Q. Zhou, R. Betancur, R. Munir, R. Patidar, S. D. Stranks, S. Alam, S. Kar, T. Unold, T. Abzieher, T. Edvinsson, T. W. David, U. W. Paetzold, W. Zia, W. Fu, W. Zuo, V. R. F. Schröder, W. Tress, X. Zhang, Y.-H. Chiang, Z. Iqbal, Z. Xie and E. Unger, An Open-Access Database and Analysis Tool for Perovskite Solar Cells Based on the FAIR Data Principles, *Nat. Energy*, 2022, **7**(1), DOI: 10.1038/s41560-021-00941-3.

78 E. I. Marchenko, S. A. Fateev, A. A. Petrov, V. V. Korolev, A. Mitrofanov, A. V. Petrov, E. A. Goodilin and A. B. Tarasov, Database of Two-Dimensional Hybrid Perovskite Materials: Open-Access Collection of Crystal Structures, Band Gaps, and Atomic Partial Charges Predicted by Machine Learning, *Chem. Mater.*, 2020, **32**(17), 7383–7388, DOI: 10.1021/ACS.CHEMMATER.0C02290.

79 A. Chen, Z. Wang, J. Gao, Y. Han, J. Cai, S. Ye and J. Li, A Data-Driven Platform for Two-Dimensional Hybrid Lead-Halide Perovskites, *ACS Nano*, 2023, **17**(14), 13348–13357, DOI: 10.1021/ACSNANO.3C01442.

80 A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder and K. A. Persson, Commentary: The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation, *APL Mater.*, 2013, **1**(1), 11002, DOI: 10.1063/1.4812323/119685.

81 S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson and G. Ceder, Python Materials Genomics (Pymatgen): A Robust, Open-Source Python Library for Materials Analysis, *Comput. Mater. Sci.*, 2013, **68**, 314–319, DOI: 10.1016/J.COMMATSCI.2012.10.028.

82 L. Ward, A. Dunn, A. Faghaninia, N. E. R. Zimmermann, S. Bajaj, Q. Wang, J. Montoya, J. Chen, K. Bystrom, M. Dylla, K. Chard, M. Asta, K. A. Persson, G. J. Snyder, I. Foster and A. Jain, Matminer: An Open Source Toolkit for Materials Data Mining, *Comput. Mater. Sci.*, 2018, **152**, 60–69, DOI: 10.1016/J.COMMATSCI.2018.05.018.

83 C. R. Groom, I. J. Bruno, M. P. Lightfoot and S. C. Ward, The Cambridge Structural Database, *Struct. Sci.*, 2016, **72**(2), 171–179, DOI: 10.1107/S2052520616003954.

84 C. She, Q. Huang, C. Chen, Y. Jiang, Z. Fan and J. Gao, Machine Learning-Guided Search for High-Efficiency Perovskite Solar Cells with Doped Electron Transport Layers, *J. Mater. Chem. A*, 2021, **9**(44), 25168–25177, DOI: 10.1039/D1TA08194B.

85 J. Li, B. Pradhan, S. Gaur and J. Thomas, Predictions and Strategies Learned from Machine Learning to Develop High-Performing Perovskite Solar Cells, *Adv. Energy Mater.*, 2019, **9**(46), 1901891, DOI: 10.1002/aenm.201901891.

86 B. Yılmaz, Ç. Odabaşı and R. Yıldırım, Efficiency and Stability Analysis of 2D/3D Perovskite Solar Cells Using Machine Learning, *Energy Technol.*, 2022, **10**(3), 2100948, DOI: 10.1002/ente.202100948.

87 Ç. Odabaşı Özer and R. Yıldırım, Performance Analysis of Perovskite Solar Cells in 2013–2018 Using Machine-

Open Access Article. Published on 04 1447. Downloaded on 25/11/47 10:58:40 .
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Learning Tools, *Nano Energy*, 2019, **56**, 770–791, DOI: 10.1016/J.NANOEN.2018.11.069.

88 Y. Liu, W. Yan, S. Han, H. Zhu, Y. Tu, L. Guan and X. Tan, How Machine Learning Predicts and Explains the Performance of Perovskite Solar Cells, *Sol. RRL*, 2022, **6**(6), 2101100, DOI: 10.1002/solr.202101100.

89 O. Almora, G. C. Bazan, C. I. Cabrera, L. A. Castriotta, S. Erten-Ela, K. Forberich, K. Fukuda, F. Guo, J. Hauch, A. W. Y. Ho-Baillie, T. J. Jacobsson, R. A. J. Janssen, T. Kirchartz, R. R. Lunt, X. Mathew, D. B. Mitzi, M. K. Nazeeruddin, J. Nelson, A. F. Nogueira, U. W. Paetzold, B. P. Rand, U. Rau, T. Someya, C. Sprau, L. Vaillant-Roca, C. J. Brabec, O. Almora, C. J. Brabec, G. C. Bazan and L. A. Castriotta, Device Performance of Emerging Photovoltaic Materials (Version 5), *Adv. Energy Mater.*, 2024, **15**(12), 2404386, DOI: 10.1002/aenm.202404386.

90 F. Mayr and A. Gagliardi, Global Property Prediction: A Benchmark Study on Open-Source, Perovskite-like Datasets, *ACS omega*, 2025, **6**(19), 12722–12732, DOI: 10.1021/acsomega.1c00991.

91 C. Fan, M. Chen, X. Wang, J. Wang and B. Huang, A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data, *Front. Energy Res.*, 2021, **9**, 652801, DOI: 10.3389/FENRG.2021.652801.

92 F. Pedregosa, V. Michel, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, J. Vanderplas, D. Cournapeau, G. Varoquaux, A. Gramfort, B. Thirion, V. Dubourg, A. Passos, M. Brucher, M. Perrot and D. Édouard, Scikit-Learn: Machine Learning in Python, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.

93 *Deep Learning with Keras - Antonio Gulli, Sujit Pal - Google Libros*, accessed 2025-09-02.

94 M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu and X. Zheng, TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems, *arXiv*, 2016, preprint, arXiv:1603.04467, DOI: 10.48550/arXiv.1603.04467.

95 A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga and A. Desmaison, Pytorch: An imperative style, high-performance deep learning library, *Advances in neural information processing systems*, 2019, p. 32.

96 *GitHub - jax-ml/jax: Composable transformations of Python+NumPy programs: differentiate, vectorize, JIT to GPU/TPU, and more*, https://github.com/jax-ml/jax, accessed 2025-09-02.

97 W. F. Lunnon, J. Brunvoll, S. J. Cyvin, B. N. Cyvin and A. T. Balaban, SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules, *J. Chem. Inf. Comput. Sci.*, 1988, **28**(1), 31–36, DOI: 10.1021/CI00057A005.

98 M. E. Mswahili and Y. S. Jeong, Transformer-Based Models for Chemical SMILES Representation: A Comprehensive Literature Review, *Heliyon*, 2024, **10**(20), 2405–8440, DOI: 10.1016/J.HELIYON.2024.E39038.

99 H. Moriwaki, Y. S. Tian, N. Kawashita and T. Takagi, Mordred: A Molecular Descriptor Calculator, *J. Cheminf.*, 2018, **10**(1), 4, DOI: 10.1186/S13321-018-0258-Y.

100 G. Landrum, *RDKit: Open-Source Cheminformatics. Release 2014.03.1*, 2014, DOI: 10.5281/ZENODO.10398.

101 J. Sieg, C. W. Feldmann, J. Hemmerich, C. Stork, F. Sandfort, P. Eiden and M. Mathea, MolPipeline: A Python Package for Processing Molecules with RDKit in Scikit-Learn, *J. Chem. Inf. Model.*, 2024, **64**(24), 9027–9033, DOI: 10.1021/ACS.JCIM.4C00863.

102 S. Mann, E. Fadel, S. S. Schoenholz, E. D. Cubuk, S. G. Johnson and G. Romano, ∂PV: An End-to-End Differentiable Solar-Cell Simulator, *Comput. Phys. Commun.*, 2022, **272**, 108232, DOI: 10.1016/J.CPC.2021.108232.

103 C. O. S. Sorzano, J. Vargas and A. P. Montano, A Survey of Dimensionality Reduction Techniques, *arXiv*, 2014, preprint, arXiv:1403.2877, DOI: 10.48550/1403.2877.

104 D. Coelho, A. Madureira, I. Pereira and R. Gonçalves, A Review on Dimensionality Reduction for Machine Learning, *Lecture Notes in Networks and Systems*, 2023, vol. 649, pp. 287–296, DOI: 10.1007/978-3-031-27499-2_27.

105 V. Bolón-Canedo, N. Sánchez-Maroño and A. Alonso-Betanzos, A Review of Feature Selection Methods on Synthetic Data, *Knowl. Inf. Syst.*, 2013, **34**(3), 483–519, DOI: 10.1007/S10115-012-0487-8/METRICS.

106 P. Dhal and C. Azad, A Comprehensive Survey on Feature Selection in the Various Fields of Machine Learning, *Appl. Intell.*, 2021, **52**(4), 4543–4581, DOI: 10.1007/S10489-021-02550-9.

107 M. Boubchir, R. Boubchir and H. Aourag, The Principal Component Analysis as a Tool for Predicting the Mechanical Properties of Perovskites and Inverse Perovskites, *Chem. Phys. Lett.*, 2022, **798**, 139615, DOI: 10.1016/j.cplett.2022.139615.

108 N. Imamura, T. Mizoguchi, H. Yamauchi and M. Karppinen, Multivariate Data Analysis Approach to Understand Magnetic Properties of Perovskite Manganese Oxides, *J. Solid State Chem.*, 2008, **181**(5), 1195–1203, DOI: 10.1016/j.jssc.2008.02.025.

109 Q. Deng and B. Lin, Exploring Structure-Composition Relationships of Cubic Perovskite Oxides via Extreme Feature Engineering and Automated Machine Learning, *Mater. Today Commun.*, 2021, **28**, 102590, DOI: 10.1016/J.MTCOMM.2021.102590.

110 E. F. Morales and H. J. Escalante, A Brief Introduction to Supervised, Unsupervised, and Reinforcement Learning, *Biosignal Processing and Classification Using Computational Learning and Intelligence: Principles, Algorithms, and*

© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 04 1447. Downloaded on 25/11/47 10:58:40 .
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Applications, 2022, pp. 111–129, DOI: 10.1016/B978-0-12-820125-1.00017-8.

111 P. Kulkarni, *Reinforcement and Systemic Machine Learning for Decision Making*, 2012, p. 422.

112 V. Vakharia, I. E. Castelli, K. Bhavsar and A. Solanki, Bandgap Prediction of Metal Halide Perovskites Using Regression Machine Learning Models, 2021.

113 G. Pilania, P. V. Balachandran, C. Kim and T. Lookman, Finding New Perovskite Halides via Machine Learning, *Frontiers in Materials*, 2016, **3**, 19, DOI: 10.3389/fmats.2016.00019.

114 T. W. David, H. Anizelli, T. J. Jacobsson, C. Gray, W. Teahan and J. Kettle, Enhancing the Stability of Organic Photovoltaics through Machine Learning, *Nano Energy*, 2020, **78**, 105342, DOI: 10.1016/j.nanoen.2020.105342.

115 B. N. N. Alsulami, T. W. David, A. Essien, S. Kazim, S. Ahmad, T. J. Jacobsson, A. Feeney and J. Kettle, Application of Large Datasets to Assess Trends in the Stability of Perovskite Photovoltaics through Machine Learning, *J. Mater. Chem. A*, 2024, **12**(5), 3122–3132, DOI: 10.1039/d3ta05966a.

116 F. Akhundova, L. Lüer, A. Osvet, J. Hauch, I. M. Peters, K. Forberich, N. Li and C. Brabec, Building Process Design Rules for Microstructure Control in Wide-Bandgap Mixed Halide Perovskite Solar Cells by a High-Throughput Approach, *Appl. Phys. Lett.*, 2021, **118**(24), 1ENG, DOI: 10.1063/5.0049010/238983.

117 H. Choe, H. Jin, S. J. Lee and J. Cho, Machine Learning-Directed Predictive Models: Deciphering Complex Energy Transfer in Mn-Doped CsPb(Cl1-YBry)3 Perovskite Nanocrystals, *Chem. Mater.*, 2023, **35**(14), 5401–5411, DOI: 10.1021/ACS.CHEMMATER.3C00731.

118 M. Mammeri, L. Dehimi, H. Bencherif and F. Pezzimenti, Paths towards High Perovskite Solar Cells Stability Using Machine Learning Techniques, *Sol. Energy*, 2023, **249**, 651–660, DOI: 10.1016/J.SOLENER.2022.12.002.

119 S. Mishra, B. Boro, N. K. Bansal and T. Singh, Machine Learning-Assisted Design of Wide Bandgap Perovskite Materials for High-Efficiency Indoor Photovoltaic Applications, *Mater. Today Commun.*, 2023, **35**, 106376, DOI: 10.1016/J.MTCOMM.2023.106376.

120 D. O. Obada, E. Okafor, S. A. Abolade, A. M. Ukpong, D. Dodoo-Arhin and A. Akande, Explainable Machine Learning for Predicting the Band Gaps of ABX3 Perovskites, *Mater. Sci. Semicond. Process.*, 2023, **161**, 107427, DOI: 10.1016/J.MSSP.2023.107427.

121 N. Shrivastav, J. Madan and R. Pandey, Predicting Photovoltaic Efficiency in Cs-Based Perovskite Solar Cells: A Comprehensive Study Integrating SCAPS Simulation and Machine Learning Models, *Solid State Commun.*, 2024, **380**, 115437, DOI: 10.1016/J.SSC.2024.115437.

122 T. Bak, K. Kim, E. Seo, J. Han, H. Sung, I. Jeon and I. D. Jung, Accelerated Design of High-Efficiency Lead-Free Tin Perovskite Solar Cells via Machine Learning, *Int. J. Precis. Eng. Manuf.*, 2023, **10**(1), 109–121, DOI: 10.1007/S40684-022-00417-Z.

123 B. D. Nguyen, P. Potapenko, A. Demirci, K. Govind, S. Bompas and S. Sandfeld, Efficient Surrogate Models for Materials Science Simulations: Machine Learning-Based Prediction of Microstructure Properties, *Machine Learning with Applications*, 2024, **16**, 100544, DOI: 10.1016/j.mlwa.2024.100544.

124 K. Choudhary, B. DeCost, C. Chen, A. Jain, F. Tavazza, R. Cohn, C. W. Park, A. Choudhary, A. Agrawal, S. J. L. Billinge, E. Holm, S. P. Ong and C. Wolverton, Recent Advances and Applications of Deep Learning Methods in Materials Science, *npj Comput. Mater.*, 2022, **8**(1), 1–26, DOI: 10.1038/s41524-022-00734-6.

125 I. U. Ekanayake, D. P. P. Meddage and U. Rathnayake, A Novel Approach to Explain the Black-Box Nature of Machine Learning in Compressive Strength Predictions of Concrete Using Shapley Additive Explanations (SHAP), *Case Stud. Constr. Mater.*, 2022, **16**, e01059, DOI: 10.1016/J.CSCM.2022.E01059.

126 S. Mishra, S. B. Gaikwad and T. Singh, Machine Learning Guided Strategies to Develop High Efficiency Indoor Perovskite Solar Cells, *Adv. Theory Simul.*, 2024, **7**(5), 2301193, DOI: 10.1002/adts.202301193.

127 Z. Chen, S. Pan, J. Wang, Y. Min, Y. Chen and Q. Xue, Machine Learning Will Revolutionize Perovskite Solar Cells, *Innovation*, 2024, **5**, 100602, DOI: 10.1016/j.xinn.2024.100602.

128 Y. Liu, X. Tan, J. Liang, H. Han, P. Xiang and W. Yan, Machine Learning for Perovskite Solar Cells and Component Materials: Key Technologies and Prospects, *Adv. Funct. Mater.*, 2023, **33**(17), 2214271, DOI: 10.1002/ADFM.202214271.

129 C. Chen, A. Maqsood and T. J. Jacobsson, The Role of Machine Learning in Perovskite Solar Cell Research, *J. Alloys Compd.*, 2023, **960**, 170824, DOI: 10.1016/J.JALLCOM.2023.170824.

130 Z. Hui, M. Wang, X. Yin, Y. Wang and Y. Yue, Machine Learning for Perovskite Solar Cell Design, *Comput. Mater. Sci.*, 2023, **226**, 112215, DOI: 10.1016/J.COMMATSCI.2023.112215.

131 C. W. Myung, A. Hajibabaei, J. H. Cha, M. Ha, J. Kim and K. S. Kim, Challenges, Opportunities, and Prospects in Metal Halide Perovskites from Theoretical and Machine Learning Perspectives, *Adv. Energy Mater.*, 2022, **12**(45), 2202279, DOI: 10.1002/AENM.202202279.

132 R. Chakraborty and V. Blum, Curated Materials Data of Hybrid Perovskites: Approaches and Potential Usage, *Trends Chem.*, 2023, **5**(10), 720–733, DOI: 10.1016/J.TRECHM.2023.08.005.

133 M. Srivastava, A. R. Hering, Y. An, J. P. Correa-Baena and M. S. Leite, Machine Learning Enables Prediction of Halide Perovskites' Optical Behavior with >90% Accuracy, *ACS Energy Lett.*, 2023, **8**(4), 1716–1722, DOI: 10.1021/ACSENERGYLETT.2C02555.

134 J. Yang and A. Mannodi-Kanakkithodi, High-Throughput Computations and Machine Learning for Halide Perovskite Discovery, *MRS Bull.*, 2022, **47**(9), 940–948, DOI: 10.1557/S43577-022-00414-2.

Open Access Article. Published on 04 1447. Downloaded on 25/11/47 10:58:40 .
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

135 N. K. Bansal, S. Mishra, H. Dixit, S. Porwal, P. Singh and T. Singh, Machine Learning in Perovskite Solar Cells: Recent Developments and Future Perspectives, *Energy Technol.*, 2023, **11**(12), 2300735, DOI: 10.1002/ENTE.202300735.

136 S. Touati, A. Benghia, Z. Hebboul, I. K. Lefkaier, M. B. Kanoun and S. Goumri-Said, Predictive Machine Learning Approaches for Perovskites Properties Using Their Chemical Formula: Towards the Discovery of Stable Solar Cells Materials, *Neural Comput. Appl.*, 2024, **36**(26), 16319–16329, DOI: 10.1007/S00521-024-09992-5.

137 L. Zhang, F. Lu, G. Tao, M. Li, Z. Yang, A. Wang, W. Zhu, Y. Cao, Y. Jin, L. Zhu, W. Huang and J. Wang, Prediction of Operational Lifetime of Perovskite Light Emitting Diodes by Machine Learning, *Advanced Intelligent Systems*, 2024, **6**(6), 2300772, DOI: 10.1002/AISY.202300772.

138 S. V. Pandey, N. Parikh, A. Kalam, D. Prochowicz, S. Satapathi, S. Akin, M. M. Tavakoli and P. Yadav, A Machine Learning Framework for Predicting Device Performance in 2D Metal Halide Perovskite Photodetector, *Sol. Energy*, 2024, **270**, 112399, DOI: 10.1016/J.SOLENER.2024.112399.

139 D. Hu, Z. Yang and S. Huang, Machine Learning Prediction of Perovskite Sensors for Monitoring the Gas in Lithium-Ion Battery, *Sens. Actuators, A*, 2024, **369**, 115162, DOI: 10.1016/J.SNA.2024.115162.

140 W. Yan, Y. Liu, Y. Zang, J. Cheng, Y. Wang, L. Chu, X. Tan, L. Liu, P. Zhou, W. Li and Z. Zhong, Machine Learning Enabled Development of Unexplored Perovskite Solar Cells with High Efficiency, *Nano Energy*, 2022, **99**, 107394, DOI: 10.1016/J.NANOEN.2022.107394.

141 X. Cai, F. Liu, A. Yu, J. Qin, M. Hatamvand, I. Ahmed, J. Luo, Y. Zhang, H. Zhang and Y. Zhan, Data-Driven Design of High-Performance MASnxPb1-XI3 Perovskite Materials by Machine Learning and Experimental Realization, *Light: Sci. Appl.*, 2022, **11**(1), 1–12, DOI: 10.1038/s41377-022-00924-3.

142 Y. Liu, W. Yan, S. Han, H. Zhu, Y. Tu, L. Guan and X. Tan, How Machine Learning Predicts and Explains the Performance of Perovskite Solar Cells, *Sol. RRL*, 2022, **6**(6), 2101100, DOI: 10.1002/SOLR.202101100.

143 J. Yang, P. Manganaris and A. Mannodi-Kanakkithodi, Discovering Novel Halide Perovskite Alloys Using Multi-Fidelity Machine Learning and Genetic Algorithm, *J. Chem. Phys.*, 2024, **160**(6), 58, DOI: 10.1063/5.0182543/3263011.

144 S. Kumar, S. Dutta, R. Jaafreh, N. Singh, A. Sharan, K. Hamad and D. H. Yoon, Accelerated Discovery of Perovskite Materials Guided by Machine Learning Techniques, *Mater. Lett.*, 2023, **353**, 135311, DOI: 10.1016/J.MATLET.2023.135311.

145 G. Li, C. Wang, J. Huang, L. Huang and Y. Zhu, Machine Learning Guided Rapid Discovery of Narrow-Bandgap Inorganic Halide Perovskite Materials, *Appl. Phys. A:Mater. Sci. Process.*, 2024, **130**(2), 1–11, DOI: 10.1007/S00339-023-07187-8.

146 Y. Chen, H. Liu, X. Fang, Y. Li, J. Chen, L. Peng, X. Liu and J. Lin, A Machine Learning Workflow for Large-Scale Discovery of Direct Bandgap Double Perovskites, *Sol. Energy Mater. Sol. Cells*, 2025, **282**, 113402, DOI: 10.1016/J.SOLMAT.2025.113402.

147 U. Kumar, H. W. Kim, G. K. Maurya, B. B. Raj, S. Singh, A. K. Kushwaha, S. B. Cho and H. Ko, Machine Learning-Enhanced Design of Lead-Free Halide Perovskite Materials Using Density Functional Theory, *Curr. Appl. Phys.*, 2025, **69**, 1–7, DOI: 10.1016/J.CAP.2024.10.012.

148 S. Lu, Q. Zhou, L. Ma, Y. Guo and J. Wang, Rapid Discovery of Ferroelectric Photovoltaic Perovskites and Material Descriptors via Machine Learning, *Small Methods*, 2019, **3**(11), 1900360, DOI: 10.1002/SMTD.201900360.

149 R. Lyu, C. E. Moore, T. Liu, Y. Yu and Y. Wu, Predictive Design Model for Low-Dimensional Organic-Inorganic Halide Perovskites Assisted by Machine Learning, *J. Am. Chem. Soc.*, 2021, **143**(32), 12766–12776, DOI: 10.1021/JACS.1C05441.

150 I. M. Pendleton, M. K. Caucci, M. Tynes, A. Dharna, M. A. N. Nellikkal, Z. Li, E. M. Chan, A. J. Norquist and J. Schrier, Can Machines "Learn" Halide Perovskite Crystal Formation without Accurate Physicochemical Features?, *J. Phys. Chem. C*, 2020, **124**(25), 13982–13992, DOI: 10.1021/ACS.JPCC.0C01726.

151 C. Chen, A. Maqsood, Z. Zhang, X. Wang, L. Duan, H. Wang, T. Chen, S. Liu, Q. Li, J. Luo and T. J. Jacobsson, The Use of ChatGPT to Generate Experimentally Testable Hypotheses for Improving the Surface Passivation of Perovskite Solar Cells, *Cell Rep. Phys. Sci.*, 2024, **5**(7), DOI: 10.1016/j.xcrp.2024.102058.

152 Y. Wu, C. F. Wang, M. G. Ju, Q. Jia, Q. Zhou, S. Lu, X. Gao, Y. Zhang and J. Wang, Universal Machine Learning Aided Synthesis Approach of Two-Dimensional Perovskites in a Typical Laboratory, *Nat. Commun.*, 2024, **15**(1), 1–10, DOI: 10.1038/s41467-023-44236-5.

153 E. J. Braham, J. Cho, K. M. Forlano, D. F. Watson, R. Arròyave and S. Banerjee, Machine Learning-Directed Navigation of Synthetic Design Space: A Statistical Learning Approach to Controlling the Synthesis of Perovskite Halide Nanoplatelets in the Quantum-Confined Regime, *Chem. Mater.*, 2019, **31**(9), 3281–3292, DOI: 10.1021/ACS.CHEMMATER.9B00212.

154 T. Pretto, F. Baum, R. A. Gouvêa, A. G. Brolo and M. J. L. Santos, Optimizing the Synthesis Parameters of Double Perovskites with Machine Learning Using a Multioutput Regression Model, *J. Phys. Chem. C*, 2024, **128**(17), 7041–7052, DOI: 10.1021/ACS.JPCC.3C06801.

155 J. Kirman, A. Johnston, D. A. Kuntz, M. Askerka, Y. Gao, P. Todorović, D. Ma, G. G. Privé and E. H. Sargent, Machine-Learning-Accelerated Perovskite Crystallization, *Matter*, 2020, **2**(4), 938–947, DOI: 10.1016/J.MATT.2020.02.012.

156 C. Lampe, I. Kouroudis, M. Harth, S. Martin, A. Gagliardi, A. S. Urban, C. Lampe, S. Martin, A. S. Urban, I. Kouroudis, M. Harth and A. Gagliardi, Rapid Data-Efficient Optimization of Perovskite Nanocrystal

© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 04 1447. Downloaded on 25/11/47 10:58:40 .
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Syntheses through Machine Learning Algorithm Fusion, *Adv. Mater.*, 2023, **35**(16), 2208772, DOI: 10.1002/ADMA.202208772.

157 J. C. Dahl, S. Niblett, Y. Cho, X. Wang, Y. Zhang, E. M. Chan and A. P. Alivisatos, Scientific Machine Learning of 2D Perovskite Nanosheet Formation, *J. Am. Chem. Soc.*, 2023, **145**(42), 23076–23087, DOI: 10.1021/JACS.3C05984.

158 R. Priyadarshini, H. Joardar, S. K. Bisoy and T. Badapanda, Crystal Structural Prediction of Perovskite Materials Using Machine Learning: A Comparative Study, *Solid State Commun.*, 2023, **361**, 115062, DOI: 10.1016/J.SSC.2022.115062.

159 Z. Z. Zhang, T. M. Guo, Z. G. Li, F. F. Gao, W. Li, F. Wei and X. H. Bu, Machine Learning Assisted Synthetic Acceleration of Ruddlesden-Popper and Dion-Jacobson 2D Lead Halide Perovskites, *Acta Mater.*, 2023, **245**, 118638, DOI: 10.1016/J.ACTAMAT.2022.118638.

160 J. Wang, Y. Qi, H. Zheng, R. Wang, S. Bai, Y. Liu, Q. Liu, J. Xiao, D. Zou and S. Hou, Advancing Vapor-Deposited Perovskite Solar Cells via Machine Learning, *J. Mater. Chem. A*, 2023, **11**(25), 13201–13208, DOI: 10.1039/D3TA00027C.

161 W. Liu, N. Meng, X. Huo, Y. Lu, Y. Zhang, X. Huang, Z. Liang, S. Zhao, B. Qiao, Z. Liang, Z. Xu and D. Song, Machine Learning Enables Intelligent Screening of Interface Materials towards Minimizing Voltage Losses for P-i-n Type Perovskite Solar Cells, *J. Energy Chem.*, 2023, **83**, 128–137, DOI: 10.1016/J.JECHEM.2023.04.015.

162 D. Huang, C. Guo, Z. Li, H. Zhou, X. Zhao, Z. Feng, R. Zhang, M. Liu, J. Liang, L. Zhao and J. Meng, Machine Learning-Assisted Screening of Effective Passivation Materials for P–I–N Type Perovskite Solar Cells, *J. Mater. Chem. C*, 2023, **11**(28), 9602–9610, DOI: 10.1039/D3TC01140B.

163 W. Liu, Y. Lu, D. Wei, X. Huo, X. Huang, Y. Li, J. Meng, S. Zhao, B. Qiao, Z. Liang, Z. Xu and D. Song, Screening Interface Passivation Materials Intelligently through Machine Learning for Highly Efficient Perovskite Solar Cells, *J. Mater. Chem. A*, 2022, **10**(34), 17782–17789, DOI: 10.1039/D2TA04788H.

164 Z. S. Ismail, E. F. Sawires, F. Z. Amer and S. O. Abdellatif, Perovskites Informatics: Studying the Impact of Thicknesses, Doping, and Defects on the Perovskite Solar Cell Efficiency Using a Machine Learning Algorithm, *Int. J. Numer. Model. Electron. Network. Dev. Field.*, 2024, **37**(2), e3164, DOI: 10.1002/JNM.3164.

165 Y. Lu, D. Wei, W. Liu, J. Meng, X. Huo, Y. Zhang, Z. Liang, B. Qiao, S. Zhao, D. Song and Z. Xu, Predicting the Device Performance of the Perovskite Solar Cells from the Experimental Parameters through Machine Learning of Existing Experimental Results, *J. Energy Chem.*, 2023, **77**, 200–208, DOI: 10.1016/J.JECHEM.2022.10.024.

166 B. Yılmaz, Ç. Odabaşı and R. Yıldırım, Efficiency and Stability Analysis of 2D/3D Perovskite Solar Cells Using Machine Learning, *Energy Technol.*, 2022, **10**(3), 2100948, DOI: 10.1002/ENTE.202100948.

167 G. Pindolia and S. M. Shinde, Prediction of Efficiency for KSnI3 Perovskite Solar Cells Using Supervised Machine Learning Algorithms, *J. Electron. Mater.*, 2024, **53**(6), 3268–3275, DOI: 10.1007/S11664-024-10988-Z.

168 A. Khan, J. Kandel, H. Tayara and K. T. Chong, Predicting the Bandgap and Efficiency of Perovskite Solar Cells Using Machine Learning Methods, *Mol. Inf.*, 2024, **43**(2), e202300217, DOI: 10.1002/MINF.202300217.

169 A. Yang, Y. Sun, J. Zhang, F. Wang, C. Zhong, C. Yang, H. Hu, J. Liu and X. Lin, Enhancing Power Conversion Efficiency of Perovskite Solar Cells Through Machine Learning Guided Experimental Strategies, *Adv. Funct. Mater.*, 2025, **35**(4), 2410419, DOI: 10.1002/ADFM.202410419.

170 W. Hu and L. Zhang, High-Throughput Calculation and Machine Learning of Two-Dimensional Halide Perovskite Materials: Formation Energy and Band Gap, *Mater. Today Commun.*, 2023, **35**, 105841, DOI: 10.1016/J.MTCOMM.2023.105841.

171 S. Feng and J. Wang, Prediction of Organic–Inorganic Hybrid Perovskite Band Gap by Multiple Machine Learning Algorithms, *Molecules*, 2024, **29**(2), 499, DOI: 10.3390/MOLECULES29020499.

172 T. Ko, T. Park, M. Kim and K. Min, Enhancing Predictions of Experimental Band Gap Using Machine Learning and Knowledge Transfer, *Mater. Today Commun.*, 2024, **41**, 110717, DOI: 10.1016/J.MTCOMM.2024.110717.

173 A. Yang, Y. Sun, J. Zhang, F. Wang, C. Zhong, C. Yang, H. Hu, J. Liu and X. Lin, Enhancing Power Conversion Efficiency of Perovskite Solar Cells Through Machine Learning Guided Experimental Strategies, *Adv. Funct. Mater.*, 2024, **35**(4), 2410419, DOI: 10.1002/ADFM.202410419.

174 M. A. M. Rashid, S. Lee, K. H. Kim, J. Kim and K. Jeong, Machine Learning Approach for Predicting the Hole Mobility of the Perovskite Solar Cells, *Adv. Theory Simul.*, 2024, **7**(6), 2300978, DOI: 10.1002/adts.202300978.

175 F. Laufer, M. Götz and U. W. Paetzold, Deep Learning for Augmented Process Monitoring of Scalable Perovskite Thin-Film Fabrication, *Energy Environ. Sci.*, 2025, **18**(4), 1767–1782, DOI: 10.1039/D4EE03445G.

176 S. Svanström, A. García Fernández, T. Sloboda, T. J. Jacobsson, F. Zhang, F. O. L. Johansson, D. Kühn, D. Céolin, J. P. Rueff, L. Sun, K. Aitola, H. Rensmo and U. B. Cappel, Direct Measurements of Interfacial Photovoltage and Band Alignment in Perovskite Solar Cells Using Hard X-Ray Photoelectron Spectroscopy, *ACS Appl. Mater. Interfaces*, 2023, **15**(9), 12485–12494, DOI: 10.1021/ACSAMI.2C17527.

177 J. Qin, Z. Che, Y. Kang, C. Liu, D. Wu, H. Yang, X. Hu and Y. Zhan, Towards Operation-Stabilizing Perovskite Solar Cells: Fundamental Materials, Device Designs, and Commercial Applications, *InfoMat*, 2024, **6**(4), e12522, DOI: 10.1002/INF2.12522.

178 C. I. Sprague, V. de la Asunción-Nadal and A. García-Fernández Implementing AI in Advanced Recycling of Perovskite Solar Cells, *9th International Conference on*

Open Access Article. Published on 04 1447. Downloaded on 25/11/47 10:58:40 .
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Smart and Sustainable Technologies (SpliTech), Bol and Split, Croatia, 2024, pp. 1–4, DOI: 10.23919/SpliTech61897.2024.10612339.

179 W. A. Dunlap-Shohl, Y. Meng, P. P. Sunkari, D. A. C. Beck, M. Meilă and H. W. Hillhouse, Physiochemical Machine Learning Models Predict Operational Lifetimes of CH3NH3PbI3 Perovskite Solar Cells, *J. Mater. Chem. A*, 2024, **12**(16), 9730–9746, DOI: 10.1039/D3TA06668A.

180 Y. Zhu, J. Zhang, Z. Qu, S. Jiang, Y. Liu, Z. Wu, F. Yang, W. Hu, Z. Xu and Y. Dai, Accelerating Stability of ABX3 Perovskites Analysis with Machine Learning, *Ceram. Int.*, 2024, **50**(4), 6250–6258, DOI: 10.1016/J.CERAMINT.2023.11.349.

181 Y. Zhan, X. Ren, S. Zhao and Z. Guo, Improving Thermodynamic Stability of Double Perovskites with Machine Learning: The Role of Cation Composition, *Sol. Energy*, 2024, **279**, 112839, DOI: 10.1016/J.SOLENER.2024.112839.

182 M. Mammeri, H. Bencherif, L. Dehimi, A. Hajri, P. Sasikumar, A. Syed and H. A. AL-Shwaiman, Stability Forecasting of Perovskite Solar Cells Utilizing Various Machine Learning and Deep Learning Techniques, *J. Opt.*, 2024, 1–9, DOI: 10.1007/S12596-024-01819-9/FIGURES/7.

183 M. Mammeri, L. Dehimi, H. Bencherif and F. Pezzimenti, Paths towards High Perovskite Solar Cells Stability Using Machine Learning Techniques, *Sol. Energy*, 2023, **249**, 651–660, DOI: 10.1016/J.SOLENER.2022.12.002.

184 A. García-Fernández, S. Svanström, C. M. Sterling, A. Gangan, A. Erbing, C. Kamal, T. Sloboda, B. Kammlander, G. J. Man, H. Rensmo, M. Odelius and U. B. Cappel, Experimental and Theoretical Core Level and Valence Band Analysis of Clean Perovskite Single Crystal Surfaces, *Small*, 2022, **18**(13), 2106450, DOI: 10.1002/smll.202106450.

185 A. García-Fernández, B. Kammlander, S. Riva, H. Rensmo and U. B. Cappel, Composition Dependence of X-Ray Stability and Degradation Mechanisms at Lead Halide Perovskite Single Crystal Surfaces, *Phys. Chem. Chem. Phys.*, 2024, **26**(2), 1000–1010, DOI: 10.1039/D3CP05061K.

186 S. Svanström, A. García Fernández, T. Sloboda, T. J. Jacobsson, H. Rensmo and U. B. Cappel, X-Ray Stability and Degradation Mechanism of Lead Halide Perovskites and Lead Halides, *Phys. Chem. Chem. Phys.*, 2021, **23**(21), 12479–12489, DOI: 10.1039/D1CP01443A.

187 G. Drera, C. M. Kropf and L. Sangaletti, Deep Neural Network for X-Ray Photoelectron Spectroscopy Data Analysis, *Machine Learning: Science and Technology*, 2020, **1**(1), 015008, DOI: 10.1088/2632-2153/AB5DA6.

188 Q. Zhang, H. Wang, Q. Zhao, A. Ullah, X. Zhong, Y. Wei, C. Zhang, R. Xu, S. De Wolf and K. Wang, Machine-Learning-Assisted Design of Buried-Interface Engineering Materials for High-Efficiency and Stable Perovskite Solar Cells, *ACS Energy Lett.*, 2024, 5924–5934, DOI: 10.1021/ACSENERGYLETT.4C02610.

189 N. T. P. Hartono, J. Thapa, A. Tiihonen, F. Oviedo, C. Batali, J. J. Yoo, Z. Liu, R. Li, D. F. Marrón, M. G. Bawendi, T. Buonassisi and S. Sun, How Machine Learning Can Help Select Capping Layers to Suppress Perovskite Degradation, *Nat. Commun.*, 2020, **11**(1), 1–9, DOI: 10.1038/s41467-020-17945-4.

190 A. García-Fernández, E. J. Juarez-Perez, S. Castro-García, M. Sánchez-Andújar, L. K. Ono, Y. Jiang and Y. Qi, Benchmarking Chemical Stability of Arbitrarily Mixed 3D Hybrid Halide Perovskites for Solar Cell Applications, *Small Methods*, 2018, **2**(10), 1800242, DOI: 10.1002/smtd.201800242.

191 A. García-Fernández, Z. Moradi, J. M. Bermúdez-García, M. Sánchez-Andújar, V. A. Gimeno, S. Castro-García, M. A. Senarís-Rodríguez, E. Mas-Marzá, G. Garcia-Belmonte and F. Fabregat-Santiago, Effect of Environmental Humidity on the Electrical Properties of Lead Halide Perovskites, *J. Phys. Chem. C*, 2019, **123**(4), 2011–2018, DOI: 10.1021/acs.jpcc.8b03915.

192 B. Kammlander, S. Svanström, D. Kühn, F. O. L. Johansson, S. Sinha, H. Rensmo, A. G. Fernández and U. B. Cappel, Thermal Degradation of Lead Halide Perovskite Surfaces, *Chem. Commun.*, 2022, **58**(97), 13523–13526, DOI: 10.1039/D2CC04867A.

193 Y. Shi, Y. Zheng, X. Xiao, Y. Li, D. Feng, G. Zhang, Y. Zhang, T. Li and Y. Shao, Improving Thermal Stability of Perovskite Solar Cells by Suppressing Ion Migration, *Small Struct.*, 2024, **5**(10), 2400132, DOI: 10.1002/SSTR.202400132.

194 S. Kim, S. Sabury, C. A. R. Perini, T. Hossain, A. O. Yusuf, X. Xiao, R. Li, K. R. Graham, J. R. Reynolds and J. P. Correa-Baena, Enhancing Thermal Stability of Perovskite Solar Cells through Thermal Transition and Thin Film Crystallization Engineering of Polymeric Hole Transport Layers, *ACS Energy Lett.*, 2024, **16**, 4501–4508, DOI: 10.1021/ACSENERGYLETT.4C01546.

195 R. Jaafreh, A. Sharan, M. Sajjad, N. Singh, K. Hamad, R. Jaafreh, K. Hamad, A. Sharan, M. Sajjad and N. Singh, A Machine Learning-Assisted Approach to a Rapid and Reliable Screening for Mechanically Stable Perovskite-Based Materials, *Adv. Funct. Mater.*, 2023, **33**(1), 2210374, DOI: 10.1002/ADFM.202210374.

196 D. Liu, Y. Wu, M. R. Samatov, A. S. Vasenko, E. V. Chulkov and O. V. Prezhdo, Compression Eliminates Charge Traps by Stabilizing Perovskite Grain Boundary Structures: An Ab Initio Analysis with Machine Learning Force Field, *Chem. Mater.*, 2024, **36**(6), 2898–2906, DOI: 10.1021/ACS.CHEMMATER.3C03261.

197 M. Boubchir, R. Boubchir and H. Aourag, The Principal Component Analysis as a Tool for Predicting the Mechanical Properties of Perovskites and Inverse Perovskites, *Chem. Phys. Lett.*, 2022, **798**, 139615, DOI: 10.1016/J.CPLETT.2022.139615.

198 B. Zhang, E. Huang, X. Du, X. Ma, L. Zhang, J. You, A. K. Y. Jen and S. Liu, Machine Learning Aided Parameter Analysis in Perovskite X-Ray Detector, *arXiv*, 2024, preprint, arXiv:2405.04729, DOI: 10.48550/2405.04729.

199 Y. Zhang and Y. Zhou, Machine Learning Quantification of Grain Characteristics for Perovskite Solar Cells, *Matter*, 2024, **7**(1), 255–265, DOI: 10.1016/j.matt.2023.10.032.

© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 04 1447. Downloaded on 25/11/47 10:58:40 .
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

200 A. Ghosh, G. Palanichamy, D. P. Trujillo, M. Shaikh and S. Ghosh, Insights into Cation Ordering of Double Perovskite Oxides from Machine Learning and Causal Relations, *Chem. Mater.*, 2022, **34**(16), 7563–7578, DOI: **10.1021/ACS.CHEMMATER.2C00217**.

201 N. Parikh, S. Akin, A. Kalam, D. Prochowicz and P. Yadav, Probing the Low-Frequency Response of Impedance Spectroscopy of Halide Perovskite Single Crystals Using Machine Learning, *ACS Appl. Mater. Interfaces*, 2023, **15**(23), 27801–27808, DOI: **10.1021/ACSAMI.3C00269**.

202 A. Maqsood, C. Chen and T. J. Jacobsson, The Future of Material Scientists in an Age of Artificial Intelligence, *Adv. Sci.*, 2024, **11**(19), 2401401, DOI: **10.1002/ADVS.202401401**.