Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Integrating Density Functional Theory, Machine Learning, and Device Simulation for Perovskite Photovoltaics: A Review

Sameer Pandey

Sameerpandey326@gmail.com

Vikrant University

Systematic Review

Keywords: Perovskite solar cells, Density Functional Theory, Machine learning, Materials informatics

Posted Date: October 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10854613/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Integrating Density Functional Theory, Machine Learning, and Device Simulation for Perovskite Photovoltaics: A Review

Sameer Pandey

Department of Natural and Applied Science, Vikrant University, 474006, (M.P), INDIA

Corresponding author: Sameer Pandey, Sameerpandey326@gmail.com

Abstract

Metal halide and lead-free double perovskites are promising materials for next-generation photovoltaics. But it is a challenge to navigate their huge compositional space and to simultaneously control long-term stability and non-radiative recombination. This review provides a critical synthesis of the cross-scale computational paradigm that brings together Density Functional Theory (DFT), machine learning (ML) and macroscopic device simulations like SCAPS-1D. We discuss how to transform first-principles atomistic parameters such as band edge alignments, dielectric constants and effective masses into meaningful device parameters without physical oversimplification. This work does not see parameter transfer as a "plug-and-play" process, but rather explains the bridging mechanisms that must be developed between microscopic energetics and macroscopic transport equations. We critically assess functional dependencies, active learning efficiencies, device-level limitations, and uncertainties related to ion migration and transient dynamics. This review provides actionable roadmap towards high-throughput physics-informed computational discovery of robust perovskite solar cells.

Keywords: Perovskite solar cells; Density Functional Theory; Machine learning; Materials informatics

1. Introduction

It has been discovered that metal halide perovskites have quickly become the top choices in next-generation photovoltaic technologies with certified power conversion efficiencies now over 26%. Besides extremely excellent optoelectronic behavior, they also possess the capability for extremely high fidelity of their writable compositional space, which has allowed for systematic optimization of stability, defect tolerance and bandgap engineering. Nevertheless, the costly nature of exploring this multidimensional compositional space experimentally is why its exploration is still prohibitively slow, which spurs the increased use of computational design techniques in accelerating exploration and optimization of its performance.

Density Functional Theory (DFT) has in the past few years emerged as the primary tool in the production of high-fidelity descriptors of perovskite materials, such as band-edge energies, formation on thermodynamics, dielectric response, and defect formation energies. These descriptors can be used to give a quantitative understanding of the stability trends of mixed-cation and mixed-halide compositions and describe the experimentally observed defect tolerance. At the same time, the availability of large-scale DFT datasets and open materials libraries has been able to do perovskite family modeling in a reproducible, data-driven manner to overcome previous data availability and consistency constraints (for oxide databases, see Bare et al., 2023; Mayr, 2021).

The first-principles dataset-trained machine learning (ML) models have further revolutionized perovskite studies by allowing one to swiftly make predictions of properties and screen distributions through large unexplored compositional planes. This benchmark work, recently carried out, has revealed that the present ML models suffer from a large spread in their declarations of bandgaps and thermodynamic stability as well as that an active learning approach reduces the need for DFT calculations significantly while having a negligible impact on the quality of the models (Tao et al., 2021; Wang et al., 2023). Interestingly, the most recent reviews put emphasis on the trend towards more and more solving devices based on ML being used in combination with solvers that directly

operate on the materials descriptors and map them to the performance of a photovoltaic device (Lu et al., 2024).

Advances in numerical device Simulations, which make it possible to use DFT-based material parameters and ML-based material parameters in drift-diffusion models of perovskite solar cells, have allowed for the quantitative assessment of recombination, interface energetics and defect-assisted transport limit cycle losses. The reliability of computational design schemes across multiple scales (Roberts, 2024) is now being bolstered by open-source toolchains and community databases for reproducible ML-to-device workflows.

All of these have revolutionized perovskite photovoltaics, which used to be employed to solve isolated properties to multiscale-integrated computational optimization. Nevertheless, issues of model transferability between codes of DFT, consistency of training data and the ability to model realistic device degradation physics have not been completely addressed. These gaps must be overcome by establishing a linkage between DFT-computed descriptors, ML-based screening and the assessment of the performance level on a device level. This integrated computational approach is demonstrated in this work to highlight the new opportunities and challenges in rational design of perovskite photovoltaic materials.

Scope and Novel Contribution

In the present paper, the three distinct branches of computational research in the field of perovskite photovoltaic materials – (i) Derivation of basic material properties based on Density Functional Theory (DFT) calculations, (ii) Computation of material properties and screening through Machine Learning (ML), and (iii) Numerical modelling of photovoltaic devices by optimization – are specifically interconnected. Rather than providing an overview of the research performed in the field of perovskites, the current effort will be directed towards understanding how such computational layers can be integrated with the rest of the atomic-scale physics and the scale of metrics associated with the macroscopic device.

A unique feature of this paper is a systematic mapping of DFT-derived quantities (bandgaps, effective masses, dielectric constants and defect formation energies) to ML descriptors and a mapping of currently predictable outputs of ML descriptors into the input of performance evaluation of device simulations. This end-to-end approach brings out data transferability opportunities and restrictions, model consistency, and physics scaling simplifications. In this article, a comprehensive summary of recent technologies in integrated computational design strategies offers a systematic model of rational discovery and optimization of stable and high-efficiency perovskite photovoltaic materials. Although our previous research investigated computational methods based on individual compositions and initial materials workflows (Pandey et al., 2026a and Pandey et al., 2026b), the current review paves the way for a cross-scale parameter-bridging framework that can be applied to a range of perovskite families. It critically analyzes code-to-code functional dependency, defect parameter mapping and macroscopic limitation on the devices.

Literature Selection: A group of keywords were chosen: perovskite solar cells, density functional theory, machine learning materials screening, and SCAPS-1D drift-diffusion simulation, and the above-mentioned databases were searched with these keywords. Focus is on primary studies and benchmark datasets published within the past eight years (2018-2026) with a particular interest in the translation of parameters across scales and repeatable workflows.

2. Perovskite Materials: Structure, Chemistry and Functional Properties

Most of the PV applications of perovskite materials are related to the more general chemical formula of the material ABX_3, where A is a monovalent cation, B is a divalent metal cation, and X is a halide anion. This crystal structure comprises of corner-sharing BX_6 octahedra, with the cuboctahedral vacancies between them filled by the A-site cation. This motif is structurally the simplest, but also

highly chemically versatile, meaning it can be completely versatile, both as to compositions and properties can be engineered (Kojima et al., 2009; Snaith, 2013).

2.1 Crystal Structure and Chemical Composition

The A-site in the metal-halide perovskites can be occupied by an organic cation like methylammonium (MA+) or formamidinium (FA+) or an inorganic ion like cesium (Cs+). The B-site is generally filled by divalent metal ions like Pb^{2+} or Sn^{2+}, and the X-site is usually filled by halide ions like I^-, Br^-, or Cl^-. Such substitution/mixing at these crystallographic positions can change the lattice parameters, the lattice structural symmetry, the octahedral tilting, the phase stability, and the electronic properties of the perovskite materials formed. Specifically, tuning the structural and optoelectronic properties of metal-halide perovskites by using A-site cation engineering has been extensively explored, such as in Sn-based perovskite compositions (Gao et al., 2020; Huang et al., 2019).

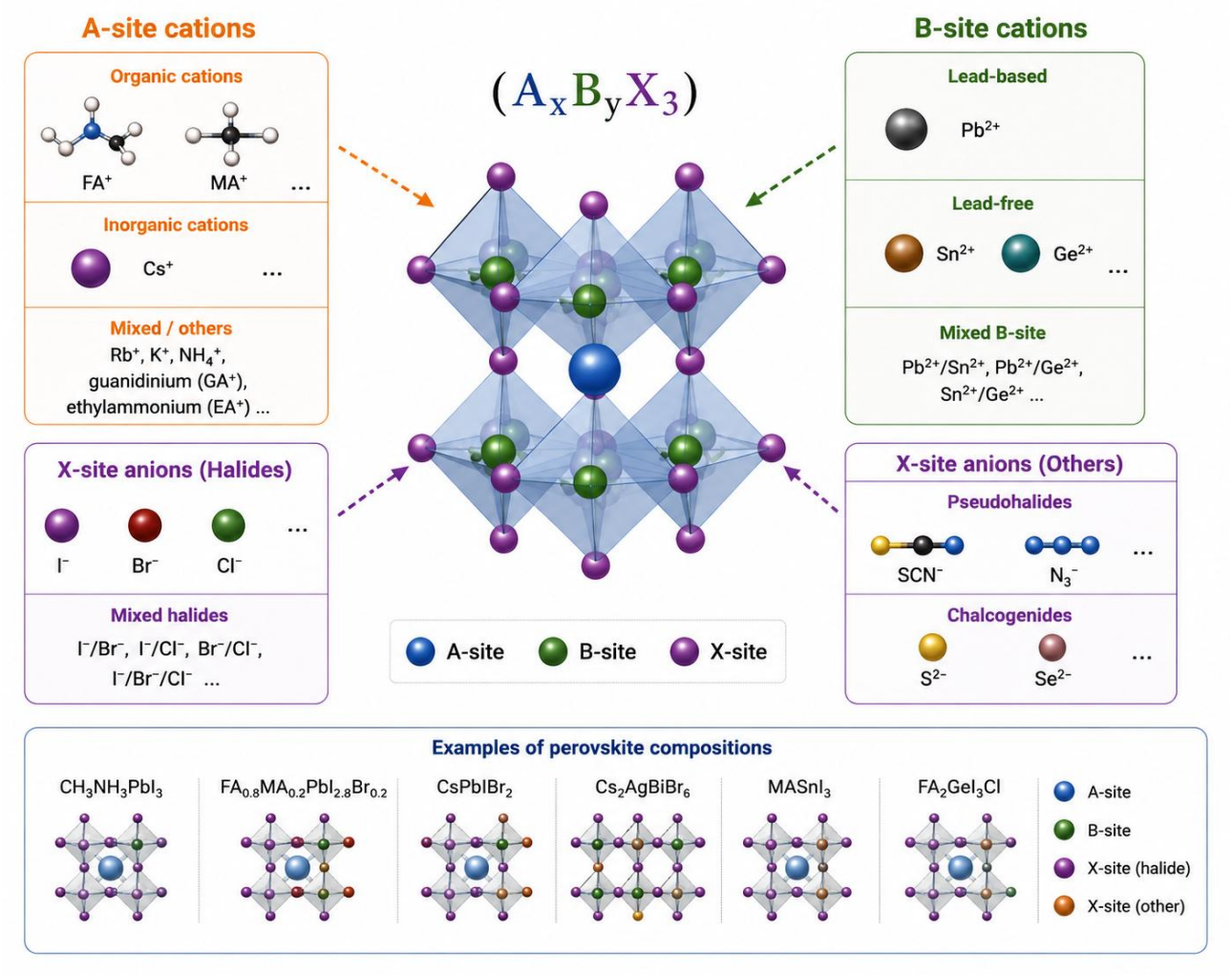


Figure 1. Crystal structure and compositional versatility of metal-halide and double perovskites ($A_xB_yX_3$). The schematic illustrates corner-sharing BX_6 octahedra with cuboctahedral interstitial A-sites, highlighting substitution pathways across organic/inorganic cations, divalent B-site metals, and halide/pseudohalide anions. Note: Chalcogenide substitutions (S^{2-}, Se^{2-}) represent mixed-anion systems requiring appropriate B-site valence adjustments to preserve charge neutrality.

The Goldschmidt tolerance factor and the octahedral factor are frequently used in the early stages of evaluation of the structural formability of perovskite compositions as geometrical descriptors. The above considerations should however be treated not as determining phase stability but as screening criteria as structural stability can also be determined by chemical bonding, thermodynamic effects, lattice dynamics, and compositional disorder. The applicability of these geometrical descriptors has been expanded to double perovskite systems and they have been shown to be useful for computational screening of new compositions (Fedorovskiy et al., 2020).

Conventional ABX_3 perovskites have been a hot topic of interest, as have similarly formulated halide double perovskites, $A_2BB'X_6$, which might serve as lead-free alternatives. In these materials, the two different cations are ordered into the B-sites, that increase the tunability in terms of chemical and electronic properties compared to the conventional ABX_3 perovskites. They have been developed in part to remove the toxicity of lead-containing materials, while preserving their useful optoelectronic properties (Meyer et al., 2018; Bibi et al., 2021; Wolf et al., 2021).

Such a chemical space, created by the resulting compositional diversity, is much greater with the perovskite materials (A-site substitution, alloying or ordering of the B-site, mixed-halide compositions). Concurrently, there are opportunities to perform computational screening via first-principles calculations, machine learning, and data-driven materials-design approaches (Ghasemi et al., 2021; Ji et al., 2021), while systematic exploration of this enlarged design space becomes increasingly challenging.

2.2 Photovoltaics-Relevant Functional Properties

These materials exhibit exceptional optoelectronic characteristics favorable for solar energy conversion, including large optical absorption coefficients ($> 10^5$ cm^{-1}), long carrier diffusion lengths, low exciton binding energies, and defect tolerance originating from the antibonding characteristics of the valence band maximum (Yin et al., 2014). Compositional engineering can be used to continually tune bandgaps between the ranges of about 1.2 to 2.3 eV, which allows the optimization of single-junction and tandem solar cell designs (Jeong et al., 2020).
The thermal or electrical stress phase instability, ion migration due to the same soft ionic lattice that gives the substance its defect tolerance, as well as moisture sensitivity, however, all come at a cost. The discrepancies and opposing effects dictate that the study of the structure-property relations be carried out at the atomic level, as well as the mesoscale level. Thus, the perovskite materials have been suggested to be the perfect surface for mapping out the connection of crystal chemistry to electronic structure and macroscopic device performance also in other parts.

Table 1. Common Perovskite Compositions and Their Functional Implications

Category	**Representative Composition**	**Key Functional Impact**
Hybrid perovskite	$MAPbI_3$	High efficiency, poor moisture stability
Mixed-cation	$FA_{0.85}Cs_{0.15}Pb(I_{0.8}Br_{0.2})_3$	Enhanced thermal and phase stability
All-inorganic	$CsPbBr_3$	Improved environmental stability, wider bandgap
Lead-free	$CsSnI_3$	Reduced toxicity, lower stability
Double perovskite	$Cs_2AgBiBr_6$	Enhanced stability, indirect bandgap

3. DFT-Derived Descriptors of Perovskite Materials

Density functional theory (DFT) calculations are important tools to obtain physically meaningful descriptors for the computational design of perovskite materials. Different compositions of perovskites can be studied by DFT calculations to investigate structural parameters, electronic band structures, band-edge positions, thermodynamic stability, dielectric response, etc., and even defect energetics. The exchange–correlation functional and spin–orbit coupling, in particular for compounds with heavy elements like Pb and Bi, are important factors to be taken into account in the accurate calculation of the electronic structure of halide perovskites. Thus, the computed bandgap and electronic characteristics may significantly vary from one computational method to another. To choose descriptors for subsequent materials screening, comparisons between computational approaches and experiment are necessary, instead of assigning a universal numerical accuracy to DFT bandgaps (Meggiolaro et al., 2018; Xue et al., 2022).To mention a few examples, the self-interaction error of the semi-local GGA-PBE family systematically underestimates the band gap, and failure to include spin–orbit coupling (SOC) artificially lowers the conduction band edges based on the dominant heavy Pb-6p or Bi-6p in the calculations. For accurate positioning of defect charge-transition levels, therefore, calibrated hybrid functionals, in the presence of relativistic SOC corrections, are required.

Other thermodynamic parameters, such as formation energies and decomposition energies, can also be obtained from DFT total-energy calculations and be useful for information of thermodynamic stability. These descriptors may be used to search for chemically desirable areas in a large composition space or to compare different compositions for their merits. Additionally, supercell calculations of intrinsic defects calculate the defect formation energies and charge-transition energies that aid in the identification of the dominant defect species and also help assess if the defects will result in shallow or deep electronic states. For instance, by making first principles calculations systematically for a variety of hybrid and inorganic halide perovskite materials, the formation energies of defects and charge-transition levels are seen to be significantly different for different A-site cations, halide species and the chemical environments (Huang et al., 2018; Xue et al., 2022).
The DFT calculations can also be extrapolated to the perovskite absorber/charge-transport material interfaces. These calculations can be used to obtain information regarding the structure of the interfaces, the band alignment, charge redistribution, and electronic coupling, which are important for understanding charge extraction and interfacial recombination in photovoltaic devices. But the properties at the interface are dependent on structural configuration, composition, dimensionality and temperature and should not be considered as the material constants (Gao et al., 2020). These quantities derived from DFT can then be used as physically meaningful descriptors to train machine-learning models. These descriptors, combined with compositional and structural properties, can then be employed by ML approaches to speed up the process of predicting a property and to screen much larger chemical spaces that would be hard to cover with direct first-principles calculations (Tao et al., 2021; Yang et al., 2021).

Table 2. DFT-Derived Descriptors and Their Relevance to Machine Learning and Device Simulation

DFT-Derived Descriptor	Physical Meaning	Typical ML Usage	Relevance to Device Simulation	Potential Photovoltaic Impact

Bandgap (Eg)	Energy difference between the valence-band maximum (VBM) and conduction-band minimum (CBM)	Target variable for bandgap prediction and materials screening	Used as an absorber-layer parameter in drift–diffusion models such as SCAPS-1D	Influences the absorption range, theoretical voltage limit, and overall photovoltaic performance
Formation energy / enthalpy	Thermodynamic tendency and relative stability of a material or composition	Stability prediction and classification/screening of candidate materials	May support the selection of physically plausible absorber materials; not directly entered as a standard SCAPS parameter	Provides information about material stability and possible degradation susceptibility
Effective mass (m)*	Describes the response of charge carriers to external forces and is related to their transport characteristics	Input descriptor for predicting transport-related properties	Can inform the selection or estimation of transport-related parameters, but does not directly determine carrier mobility in SCAPS-1D	Can influence carrier transport, recombination, fill factor (FF), and current density (Jsc)
Dielectric constant (ε)	Describes the electronic and ionic screening response of the material	Input descriptor for predicting electronic and optoelectronic properties	Can provide physical guidance for modelling electrostatic and recombination-related parameters; not a direct predictor of device performance	Can influence carrier interactions, screening, recombination, and charge-transport behaviour
Defect formation energy	Energetic tendency for the formation of intrinsic defects	Descriptor or target for defect-related stability and defect-tolerance screening	Provides information for selecting realistic defect species and concentrations; defect density must be parameterized separately in SCAPS-1D	Defect populations can influence trap-assisted recombination, carrier lifetime, and device efficiency
Band-edge positions (VBM/CBM)	Absolute energetic positions of the valence- and conduction-band edges	Descriptors for band alignment and interface-property prediction	Can guide the selection of energy levels and band offsets at absorber/ETL/HTL interfaces	Influences carrier extraction, interfacial recombination, and voltage losses
Ion-migration barrier	Activation energy associated with ionic motion through the crystal lattice	Descriptor or target for predicting ion-migration tendency and operational stability	Not directly represented in standard steady-state SCAPS-1D simulations; can be used to assess limitations of static device models	Provides insight into hysteresis, compositional instability, degradation, and long-term operational stability

4. Property Prediction and Screening models Using Machine Learning

The machine-learning models based on DFT-derived data sets have been increasingly leveraged to predict the key properties of perovskite materials such as bandgap, thermodynamic stability and other structure–property correlations. They require a good quality, chemically diverse training set, the selection of descriptors, and validation procedure to be predictive. Machine-learning methods can be particularly useful at identifying nonlinear relationships between compositional or structural descriptors and target properties that are not easily represented with simple analytical relationships (Tao et al., 2021; Wang et al., 2023). For instance, recent work using a hybrid of DFT calculations and machine learning has shown how to efficiently explore mixed-cation and mixed-halide perovskite compositions, with machine learning models helping to identify those that combine desirable optical properties and stability (Filipoiu et al., 2021).

Another way to lower the number of expensive first-principles calculations is to use active-learning strategies, which choose calculations based on model uncertainty or the likelihood of gaining information. The computational savings of active learning does not need to be quantified as a fixed

percentage of the total DFT calculation, but should be stated in terms of the data set, sampling strategy and property sought in the study. Active learning has been shown recently to be an effective iterative method for finding promising candidates and strategically requesting more high-fidelity calculations, enhancing the efficiency of the computational screening process (Merchant et al., 2023).

The application of machine learning techniques is not limited to predicting properties of materials; it has also been applied to analyze the performance of photovoltaic devices. Experimental device datasets can be used to build ML models that can establish relationships between material composition, processing conditions, device architecture, and PV metrics like PCE and stability. Yılmaz et al. (2022) created ML models for the prediction of PCE and stability analysis of 2D/3D perovskites based on the dataset recorded from 146 publications and containing 599 observations. Likewise, composition–stability relationships have been studied using high-throughput experiments and ML, where the composition–stability relationship is highly sensitive to the measurement temperature and operating conditions (Yılmaz et al., 2022; Zhao et al., 2021a).

The properties predicted by the ML can then be transferred to the device-level simulations in a cross-scale computational workflow, but this should not be seen as automatic. For instance, the predicted bandgap can be used as an input or screening parameter to a model of an absorber, while the definition of the effective mass, defect energetics, carrier mobility, and trap density typically require additional physical assumptions or parameterization on which to base them prior to being able to include these in drift-diffusion simulations. Hence, the DFT- ML-device mapping is more of a parameter mapping than a one-to-one conversion of all the outputs of the ML into device simulators. This distinction is necessary to ensure physical consistency and not over-interpretation of the predicted photovoltaic performance (Tao et al., 2021; Srivastava et al., 2021).

Table 3. Machine-learning models and their roles in cross-scale perovskite materials design.

ML Model / Strategy	**Typical Input Descriptors**	**Predicted Output Property**	**Role in Device-Level Workflow**	**Key Advantage**
Gaussian Process Regression (GPR)	DFT-derived and compositional descriptors	Material properties such as bandgap, stability-related properties, or other continuous targets	Provides predictions with uncertainty estimates that can support candidate selection and subsequent device modelling	Probabilistic prediction and uncertainty quantification
Random Forest (RF)	Compositional, structural, and DFT-derived descriptors	Bandgap, stability, defect-related properties, or photovoltaic-performance indicators	Supports candidate screening and identification of materials for further device-level evaluation; does not directly determine SCAPS trap parameters	Handles nonlinear relationships and heterogeneous descriptors
Gradient-Boosting Trees	DFT-derived, compositional, and structural features	Property or photovoltaic-performance prediction	Preselects promising candidates for subsequent device-level modelling	Captures complex nonlinear relationships
Neural Networks (ANN/CNN)	High-dimensional compositional, structural, or multimodal descriptors	One or multiple material/device properties	Enables rapid prediction and can support multi-parameter optimization when sufficiently large and representative datasets are available	Suitable for complex and high-dimensional relationships

Active Learning	Model predictions, uncertainty estimates, and candidate descriptors	Identifies candidates requiring additional high-fidelity calculations or experiments	Iteratively improves the DFT–ML–device workflow by selecting informative new data points	Reduces unnecessary high-fidelity calculations and improves dataset efficiency

Table 4. Device-simulation parameters informed by DFT and machine-learning workflows.

Device-Simulation Parameter / Property	**Possible Source**	**Device-Simulation Application**	**Potentially Affected PV Metric**
Absorber bandgap (Eg)	DFT calculation or ML prediction trained on DFT/experimental data	Defines the absorber electronic structure and absorption characteristics	Voc, J_{sc}, PCE
Carrier mobility	Experimental data, transport calculations, or ML estimation supported by appropriate physical models	Used in carrier-transport equations	FF, J_{sc}, PCE
Defect density / trap density	Experimental characterization or physically parameterized estimates informed by DFT defect energetics and ML	Bulk and interface defect modelling	Recombination, Voc, FF, PCE
Dielectric permittivity	DFT calculation, experimental data, or validated ML prediction	Used in electrostatic and charge-transport/recombination modelling were supported by the simulator	Electric-field distribution and recombination behavior
Band offsets / energy-level alignment	DFT interface calculations, experimental measurements, or validated estimates	Defines energetic alignment between absorber and transport layers	Charge extraction, interfacial recombination, Voc
Ion-migration parameters	DFT migration-barrier calculations, experimental measurements, or validated ML models	Provides information for time-dependent stability or ion-migration models; not directly represented in standard steady-state SCAPS-1D calculations	Hysteresis, degradation and operational stability

5. Research Perspective on DFT–ML Integration

The developments highlight a trend in materials research for perovskite materials from prediction of individual properties to data-driven integrated materials discovery and optimization. In this nascent paradigm, DFT calculations offer physically intuitive descriptors, machine-learning models allow the exploration of large compositional and structural spaces in a timely manner, and device-level simulations offer a bridge between material properties and photovoltaic performance. For the development of reliable end-to-end computational workflows, it is important to keep the same generation and representation of DFT data, choose physically meaningful ML descriptors, thoroughly validate the model predictions, and enhance the treatment of defect and recombination mechanisms in device simulation (Tao et al., 2021; Srivastava et al., 2021; Wang et al., 2023). These issues will have to be resolved to design computational pipelines that are transferable and physically consistent, and that can guide the design of perovskite photovoltaic materials and devices reliably.

6. Machine Learning of Perovskite Materials Science

The vast number of compositions and structures in the perovskite family has made it difficult to explore them systematically via the traditional trial and error method. While density functional theory (DFT) can yield physically relevant information regarding the thermodynamics, electronic, and structural properties of perovskites, the computational cost of performing high throughput

calculations based on first principles can be prohibitively large for large chemical spaces, complex defects, interfaces and/or computationally challenging electronic-structure methods. Machine learning (ML) can thus be used to learn the relationship between material properties and properties of interest from available experimental or computational data, and then use the model for rapid screening of materials with properties that have not yet been studied (Tao et al., 2021; Srivastava et al., 2021; Chen et al., 2024). These studies introduce the concept of ML as a means to discover materials that can accelerate the perovskite materials discovery process and lower the number of costly calculations or experiments needed for screening.Typical machine learning workflows for perovskite materials include the creation of a dataset, the selection or generation of descriptors, training of the machine learning model, validation, and prediction of target properties. Input descriptors can be used to represent chemical composition, elemental properties, crystal structure, etc., and target variables could be used to represent the bandgap, formation energy, structural stability, defect-related properties, and photovoltaic performance, etc. The selection of descriptors is crucial since redundant or badly chosen descriptors can decrease the predictive performance and make the model less interpretable. More recently, some researchers have focused on feature-selection approaches for perovskite materials and stressed the need of finding informative descriptors and discarding irrelevant dimensions (Wang et al., 2023).

The perovskite materials and devices have been the subject of numerous studies of different ML algorithms. Linear and regularized regression models can give relatively interpretable relationships between material descriptors and material properties, while non-linear techniques like random forests, support vector regression, kernel methods and gradient-boosting algorithms can represent more complex associations between the input features and material properties. Advanced representation-learning methods, such as neural networks, can also be used if data sets are larger and structurally more diverse. ML methods for predicting material properties, finding new compositions, optimizing device structures, and analyzing photovoltaic-performance data have been reviewed in perovskites by Tao et al., Bansal et al., Liu et al., and Chen et al., respectively (Tao et al., 2021; Bansal et al., 2023; Liu et al., 2023; Chen et al., 2024).

The major benefit of the ML is that, once a large and representative set of training data has been created, it can speed up the process of high throughput screening. A trained model is capable of making a quick estimate for many hypothetical compositions, compared to the time it would take to do a first-principles calculation independently for each one. It has already been shown for perovskite materials such as mixed-cation perovskite and mixed-halide perovskite, where the ML models have been employed to study the composition-dependent optoelectronic properties and stability (Filipoiu et al., 2021; Yang et al., 2021). It should not be assumed, however, that the greater the number of materials screened in a test, the greater the reliability of the prediction. The accuracy of the models can be compromised if the candidate's chemical or structural structure is not built on the training set, thus emphasizing the importance of chemical diversity, proper validation and uncertainty estimation for reliable ML-based screening. The use of "active learning" represents another approach for increasing the efficiency of computational materials discovery. The active-learning workflow can be used to dynamically select compositions for which further high-fidelity calculations or experiments are likely to yield the most information, as opposed to a fixed dataset constructed prior to model development. The newly generated data then get added to the training set, and the model's representation of the chemical space will continuously get better with the inclusion of more data. This iterative approach of finding the optimal material properties is especially appealing for perovskites, as the compositional range that can be synthesized is significantly larger than the number of materials that has been experimentally or theoretically characterized (Srivastava et al., 2021; Tao et al., 2021). The effectiveness of active learning, however, is still domain-specific, as shown by Yu (2026) who found for lead-free halide perovskites that certain active-learning strategies were not always superior to random sampling, underscoring the importance of the representation of the descriptors.

However, there are some obstacles that hinder the accuracy and applicability of ML models in predicting perovskite materials. These include limited and heterogeneous data sets, data produced by the computational methods being inconsistent, feature redundancy, data leakage, extrapolation beyond the training domain, and lack of proper representation of defects, interfaces, phase transitions and degradation processes. Recent reviews and feature-selection studies have highlighted the need for the careful construction of descriptors and validation procedures to minimize spurious relationships and enhance model generalization (Wang et al., 2023; Bansal et al., 2023; Chen et al., 2024). A growing focus is thus taking place in the opposite direction, namely embedding physical knowledge into data-based models. The physically meaningful descriptors that are derived from DFT can also be introduced directly into the ML workflow to aid in the interpretability of the model and help to minimize the learning of physically meaningless correlations. This is a natural bridge between data-driven materials screening and first-principles calculations in the study of perovskites. Incorporating the necessary validation and uncertainty quantification, these DFT-supported ML methods can offer a scalable solution to explore composition–structure–property relationships and identify new candidates that can then be investigated in first-principles calculations or experiments (Tao et al., 2021; Wang et al., 2023; Liu et al., 2023).

Table 5. Common Machine Learning Algorithms Used in Perovskite Materials Research

Machine-Learning Algorithm	**Typical Application**	**Key Advantage**
Linear / Ridge Regression	Bandgap trends, stability-related property prediction, and screening	Interpretability and simple relationships
Random Forest (RF)	Bandgap, stability, defect-related, and photovoltaic-property prediction	Handles nonlinear relationships and heterogeneous features
Support Vector Regression (SVR)	Property prediction for small- to medium-sized datasets	Effective with limited training data and nonlinear relationships
Gaussian Process Regression (GPR)	Property prediction with uncertainty estimation	Probabilistic prediction and uncertainty quantification
Neural Networks	High-dimensional and complex structure–property relationships	Captures complex nonlinear patterns and supports large datasets

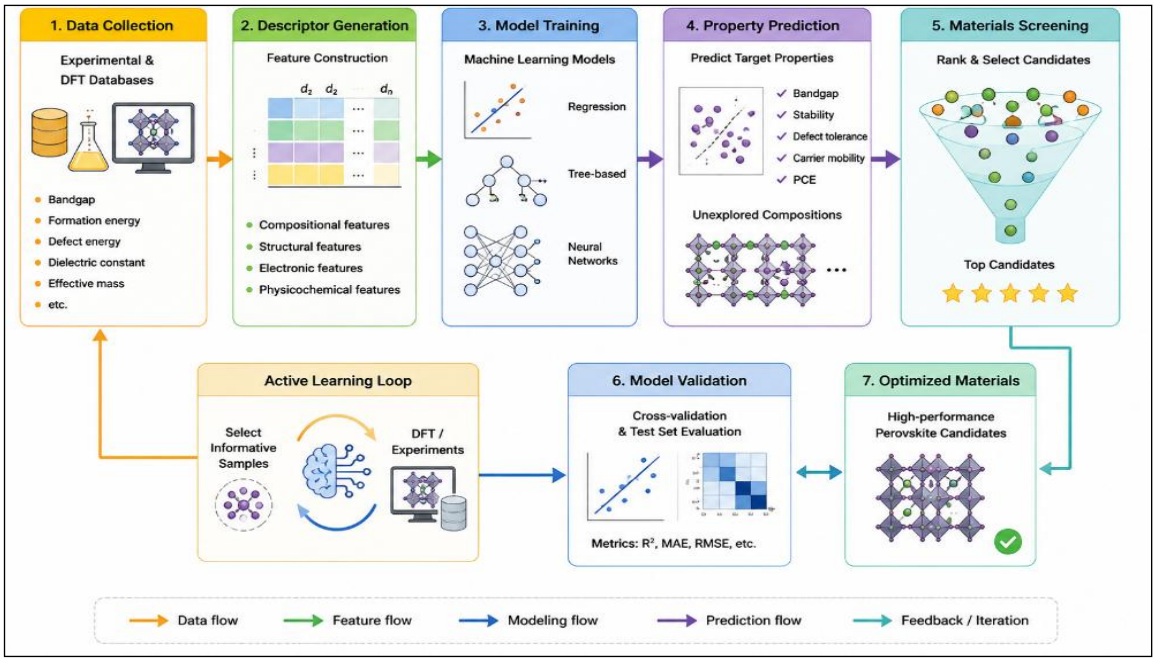


Figure 2. End-to-end machine-learning workflow for accelerated perovskite materials discovery. The framework depicts the sequential progression from DFT/experimental data curation and feature representation (compositional, structural, and electronic descriptors) to multi-algorithm model training (linear, tree-based ensembles, neural networks) and high-throughput property screening. An integrated active-learning loop facilitates iterative retraining on high-uncertainty candidates to maximize chemical space exploration. Note: Candidate screening operates conditionally following cross-validation and independent test-set evaluation to ensure predictive fidelity.

7. Simulation of Perovskite Solar Cells at the Device Level

While atomistic calculations are useful tools to gain knowledge of the intrinsic properties of perovskite materials, device-level modelling is essential to connect material properties and measurable photovoltaic properties. Numerical device simulations can link material and interface parameters to important solar-cell performance parameters, such as open-circuit voltage (Voc), short-circuit current density (J_{sc}), fill factor (FF), and power conversion efficiency (PCE). This modelling is especially helpful for the case of perovskite solar cells where the combined effects of charge transport, energetic alignment, recombination, defects and interfaces on device performance are analyzed (Tessler and Vaynzof, 2020; Le Corre et al., 2020). The device modelling and analysis can give insights into the effects of ions, dielectric properties, density of states, energetic alignment and the spatial distribution of the recombination losses in perovskite devices, The typical structure of perovskite solar cells includes a transparent conducting electrode, an electron-transport layer (ETL), a perovskite absorber, a hole-transport layer (HTL), and a metallic contact. Numerical device models are used to solve coupled equations for the electrostatic potential and electron and hole population in a device, under given illumination and electrical-bias conditions, to describe the electrostatic and carrier transport behavior. Such simulations allow for systematic study of the various parameters such as absorber thickness, carrier concentrations, defect densities, interface properties and energy-level alignment without the need to experimentally fabricate each and every parameter combination (Srivastava et al., 2022; Bhavsar and Lapsiwala, 2021). In particular, SCAPS-1D has been extensively used for numerical simulation of perovskite solar-cell designs and optimization of its parameters.

At the device level, simulations have also shown that defects and interfaces have a significant impact on the performance of a PV. The simulations show that the Shockley-Read-Hall recombination becomes significant as the density of the absorber defects increases, resulting in a significant reduction of the open-circuit voltage (Voc) while Interface defect states enhance recombination losses at the heterojunction which further reduces the fill factor (FF) and the power conversion efficiency (Jamal et al., 2019). The results show that the performance of such a perovskite solar cell is not solely determined by the material properties. Rather, a combination of the absorber properties, transport layers, interfaces, defect population and device architecture are all factors that need to be taken into account when evaluating the performance of a photovoltaic.

A key benefit of numerical device modelling is that it can be used to conduct systematic sweeps of various parameters of the device and materials, and determine combinations of these parameters which might help to enhance the simulated performance of a photovoltaic device. The simulated device characteristics are studied for the impact of absorber thickness, bandgap, doping concentration, defect density, transport layer, and operating temperature by various mechanisms in SCAPS-1D studies (Bhavsar and Lapsiwala, 2021; Srivastava et al., 2022). Simulated efficiencies should however be viewed as model dependent predictions and not as actual efficiencies. The results are highly sensitive to the physical parameters provided to the simulator and to assumptions regarding defect distributions, recombination, interfaces, transport and boundary conditions (Tessler and Vaynzof, 2020; Abbas et al., 2025).

While convenient, one-dimensional device simulations involve a number of simplifications of the actual operation of perovskite photovoltaics. SCAPS-1D cannot explicitly simulate the complexity of a real device in three dimensions nor can it explicitly solve all problems involving spatial inhomogeneity, transient ion migration, grain boundaries, degradation path, or coupled

multidimensional transport. As a result, the confidence level of the simulated PCE is highly dependent on the accuracy and consistency of the input parameters. Such constraints highlight the need for linking the device modelling with reliable descriptor information of materials from first-principles calculations and well validated machine-learning models. A more systematic pathway from the material level properties to the predicted photovoltaic performance can be created through such cross-scale integration (Tessler and Vaynzof, 2020; Le Corre et al., 2020).

Table 6. Examples of computational tools used for perovskite solar-cell device modelling and analysis.

Tool / Approach	**Dimensionality**	**Representative Capability**
SCAPS-1D	1D	Drift–diffusion device modelling, defect and recombination analysis, J–V and quantum-efficiency simulations
SILVACO ATLAS	2D/3D	Semiconductor device simulation, advanced transport modelling, tunnelling and complex interface analysis
wxAMPS	1D	Heterojunction device modelling and parameter-sweep studies
Custom MATLAB models	Model-dependent	Numerical solution of transport equations, parameter sweeps and optimization studies
WENDI	Model-dependent	Energy-level alignment and device-structure analysis

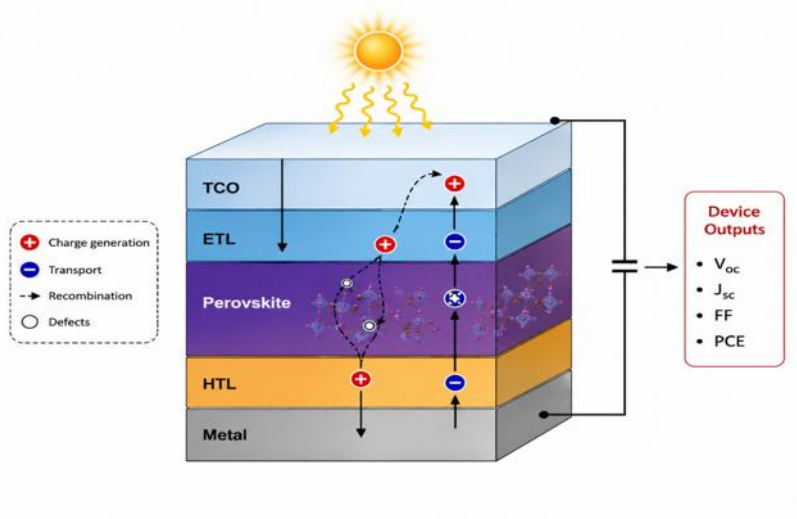


Figure 3. Schematic architecture and operational mechanism of a planar heterojunction perovskite solar cell in drift–diffusion device modeling. The diagram illustrates photogenerated carrier transport, bulk and interfacial defect-assisted Shockley–Read–Hall (SRH) recombination, and the extraction of macroscopic photovoltaic metrics (V_{oc}, J_{sc}, FF, and PCE).

8. Machine Learning with Density Functional Theory

Intersection of machine learning (ML) and Density Functional Theory (DFT) has become a paradigm shift in the study of all-perovskite materials where the huge space of compositions can be effectively explored and the physical functionality is still understandable. DFT and ML are complementary in their own strengths: the former, an atomistic description based on quantum mechanics, and the latter

has scalability and speed in even large-scale screening. Their unification can solve the basic tradeoff of computational cost that restricts purely first-principles methods (methodologically demonstrated for oxides in Pilania et al., 2016; Butler et al., 2018).

It is frequently the case that in integrated DFT-ML models a large amount of training information is obtained by performing DFT calculations. This includes the calculation of a representative set of perovskite compounds' bandgap, density of states, formation energy, dielectric constant, defect energetics and others. These models are then trained using these datasets generated from DFT predictions at very low computational cost to make predictions for the same properties of unknown compositions. This way of doing it ensures physical fidelity and enables heavy materials checking (Pilania et al., 2016).

Successful integration has to be based on descriptor design. Descriptors of DFT elements based on physics, such as the orbital contribution at the band edges, and the effective masses, tolerance and formation enthalpies, provide higher accuracy and transferability compared to composition-based features. The researchers can minimize the chances of spurious correlation and enhance the generalization of Timely processed models in developing various families of perovskites by incorporating domain knowledge in the ML pipeline (Schmidt et al., 2019). Recent work also indicates that feature importance analysis can provide physically significant structure-property relationships which can provide information other than black-box prediction.

Besides forward prediction, Iterative and Active Learning models are favored when integrating DFT-ML. The models under such schemes discover candidate materials with large predicted performance or huge uncertainty, and selectively examined with DFT to improve training data. This closed loop technique that provides the optimum material solution in a short time (Lookman et al., 2019). Fully integrated workflows can be applied to optimize end-to-end between atomic scale chemistry and PV efficiency measures, thanks to the integration with device level simulations.
There are still some challenges to be resolved, despite the potential, including the lack of consistency in the data across codes of DFT, the sensitivity of the exchange-correlation functional, and not all available data for defects and interfaces. Incorporated with integrated DFT-ML systems, however, a scalable route to efficiency, stability, and sustainability optimization can provide an effective base of data-driven discovery of next-generation perovskite materials.

Table 7. Roles of DFT and Machine Learning in an Integrated Perovskite Materials-Design Workflow.

Aspect	**Role of DFT**	**Role of Machine Learning**
Data generation	Provides high-fidelity atomistic and electronic-structure calculations	Learns from DFT and/or experimental datasets to enable rapid prediction
Property prediction	Provides physically grounded predictions but can be computationally expensive	Provides fast and scalable predictions within the validated training domain
Descriptor development	Provides physics-based descriptors such as bandgap, band-edge positions, effective masses, dielectric response, and formation energies	Selects, combines, and weights informative descriptors for predictive modelling
Materials screening	Evaluates selected candidate materials with high-fidelity calculations	Enables large-scale screening of compositions and structures

Model validation	Provides high-fidelity reference data for model training and validation	Requires independent validation, uncertainty estimation, and assessment of extrapolation limits

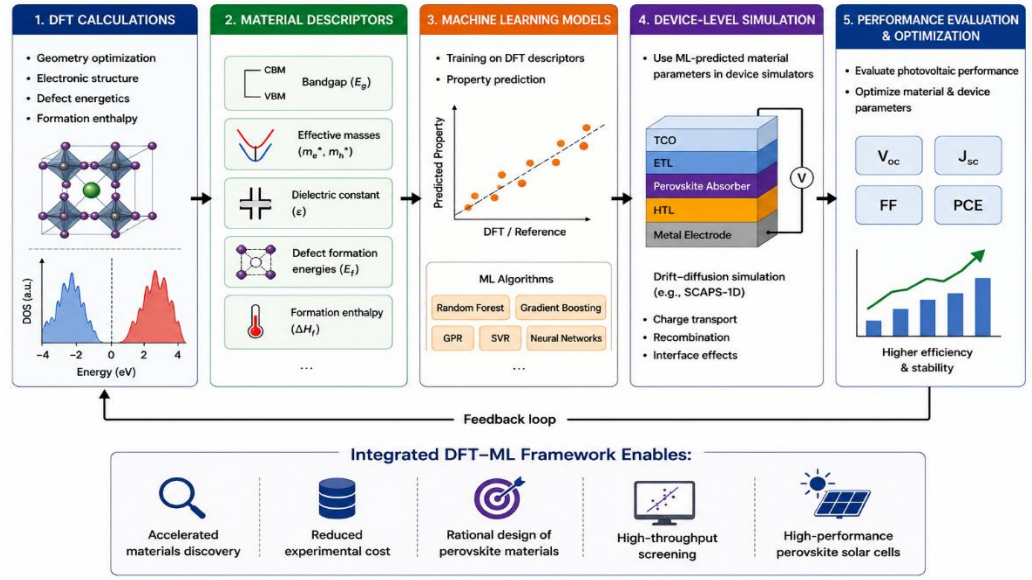


Figure 4. Integrated DFT-ML-SCAPS framework: Materials design and optimization for photovoltaic devices based on perovskite materials. Note: Conceptual performance trend (rightmost panel) shows trajectory of target optimization during the iteration.

A conceptual DFT–ML–SCAPS workflow can be created to link material-level descriptors with device-level PV performance by combining first principles calculations, machine-learning-assisted screening, and numerical device simulation. In the first stage, DFT calculations are performed for a representative set of perovskite compositions to obtain physically meaningful descriptors such as lattice parameters and bandgap, Band-edge positions, dielectric response, effective masses, formation energies, and defect-related energetics. The selection of the electronic structure method is especially crucial for halide perovskites, as the calculation of electronic properties can be significantly affected by the inclusion of spin–orbit coupling and the consideration of exchange–correlation effects. Thus, it is necessary to develop descriptors by DFT with a specific computational protocol and test them against suitable reference data before their use in machine-learning models (Tao et al., 2021; Wang et al., 2023).

This DFT-derived data can then be used to train supervised ML models that can be used to predict selected material properties in unexplored regions of compositional space. Depending on the number of data points, the properties of the descriptors, and the prediction task, a number of algorithms can be considered: random forests, support-vector methods, gradient-boosting models, and Gaussian-process approaches. With proper training and validation, such models can dramatically reduce the time required for screening of candidate compositions from typical first-principles calculations of each composition. Previous studies have showcased the use of ML to explore perovskite chemical spaces, such as double perovskites and mixed-composition systems, and recent reviews have focused on the critical role of descriptor selection, validation and chemical domain coverage to guarantee reliable predictions (Tao et al., 2021; Wang et al., 2023; Pandey et al., 2026).

The second step is to identify potentially good materials for evaluation at the device level. A key difference needs to be drawn between the material descriptors predicted by the ML and the physical parameters needed by a device simulator. The predicted bandgap can be used directly as (or can be

used to provide) the absorber-layer parameter in a device model. However, quantities like carrier mobility, density of traps, capture cross sections, interface recombination parameters, and defect distributions are in general not determined uniquely from one ML prediction of the effective mass or defect formation energy. Therefore, extra physical models, experimental data or parameterization procedures are needed before these quantities may be taken into account for a drift-diffusion device simulation. This separation is crucial for ensuring physical consistency in the linking of atomistic calculations with the device modelling level.

Selected material parameters that are suitable for the specific device-simulation stage can be used in the numerical device model (SCAPS-1D) to calculate device parameters like (Voc, Jsc, FF, and PCE).

It has recently been shown that SCAPS-1D and ML can be used together for device optimization. In one such application, Hasanzadeh Azar et al. used the SCAPS-1D software to simulate device performance and then fed these simulated data into a machine learning model to predict the performance of photovoltaics (PV) with various combinations of absorber and hole-transport-layer materials (Hasanzadeh Azar et al., 2023). More recently, data sets generated from SCAPS simulations of thousands of device configurations have been used in combination with ML and SHAP analysis to determine which parameters are important in affecting PCE (Novoselov et al., 2025). The above-mentioned studies indicate that device simulation is a viable method for coupling with ML, but do not need to be seen as proof that any DFT-derived quantities can be directly applied in SCAPS without going through the phase of intermediate physical device parameterization.

Accordingly, the workflow can be expressed as a series of DFT descriptor generation, ML prediction of physical properties, candidate selection, physical parameter mapping, SCAPS/device simulation, and photovoltaic-performance evaluation. It enables the labor- and time-intensive first-principles stage to be performed on representative or strategically chosen compositions, with faster exploration of the wider chemical space achieved using ML, and the implications of chosen material properties explored at the device level through device simulation. Recent studies have shown how this method can be combined with SCAPS-1D to predict new device configurations that have never been explored before and then subsequently verified by performing full simulation of the device using the SCAPS-1D software (Mustafa et al., 2025)

In summary, the DFT–ML–SCAPS approach offers a conceptual approach to linking atomic-scale chemistry with PV macroscopic metrics. The main strength of this is not that all material parameters can be transferred automatically across the computational levels, but rather that each method can solve a different part of the materials-to-device problem: DFT can supply physically grounded descriptors, ML can supply scalable prediction and screening, and device simulation can give insights into the consequences of selected material and interface parameters for PV performance. The ability of the predictions to be reliable, however, is contingent on the consistent use of DFT methodology, the representation of training data, the careful validation of the ML, the physically valid choices of parameter maps, and the realistic device-model assumptions.

8.1 Quantitative Benchmark Case Study: DFT–ML–SCAPS-1D Workflow for Lead-Free $Cs_2AgBiBr_6$

We show a concrete procedure for lead-free $Cs_2AgBiBr_6$ double perovskite as an example of parameter mapping across different scales, benchmarked against recent multiscale findings (Pandey, 2026c). The optical bandgap, electron affinity, and the relative permittivity obtained, using first principles calculations within the framework of Quantum ESPRESSO with the GGA-PBE, are 2.00 eV, 3.90 eV, and 10.0, respectively. In the case of $FTO/SnO_2/Cs_2AgBiBr_6/NiO_x/Au$, the ground

state quantities are mapped into SCAPS-1D, with $N_c = 2.2 \times 10^{18}$ cm^{-3}and $N_v = 1.8 \times 10^{19}$ cm^{-3} .

Device sensitivity sweeps show that reducing bulk defect density from 10^{18} cm^{-3} to 10^{12} cm^{-3} and optimizing absorber thickness to 1200 nm suppresses non-radiative Shockley-Read-Hall recombination, boosting simulated efficiency from a 4.82% baseline to 11.46% with $V_{oc} = 0.93$ V, $J_{sc} = 14.32$ mA cm^{-2}, and FF = 85.8%. A Random Forest surrogate model trained on this simulation space predicts device efficiency with high fidelity ($R^2 = 0.96$, MAE = 0.12%), where feature importance confirms absorber thickness (81%) and bulk defect density (19%) as dominant governing parameters. This benchmarks the transition from quantum ground states to rapid machine learning surrogate screening.

Level	**Method**	**Parameter**	**Quantified Value**	**Role in Device Modeling**
Microscopic	DFT GGA-PBE	E_g, χ, ε_r	2.00 eV, 3.90 eV, 10.0	Band alignment and dielectric response
Continuum	SCAPS-1D	N_c, N_v, t, N_t, PCE	1200 nm, 10^{12} cm^{-3}, 11.46%	Drift-diffusion carrier transport and SRH loss
Surrogate	Random Forest	R^2, MAE, Feature Weight	0.96, 0.12%, t (81%) vs N_t (19%)	Fast parameter prediction without numerical solvers

Table 8. Multiscale Parameter Bridging in $Cs_2AgBiBr_6$ Photovoltaics

9. Limitations of Integrated DFT–ML–SCAPS Workflows

While significant advances have been made, there are still major challenges to overcome in providing scalable computational tools and workflows that can be used reliably in the study of perovskite photovoltaics. Consistency and transferability of calculated descriptors may also be dependent on the exchange–correlation functional, pseudopotentials, spin–orbit coupling, structural configuration, and supercell size for DFT. These methodological decisions can result in systematic differences for calculated bandgap, formation energy, defect formation energy, and electronic defect level. Therefore, compatibility of the multiple datasets collected from different computational protocols, if not done using a common normalization or validation procedure, can lead to inconsistencies in further

machine-learning approaches. It is therefore crucial to develop a standard procedure for computing and well-documented calculation parameters to enable generation of reliable DFT derived datasets.

Machine-learning models pose different challenges. If experimental data is used together with computational data that is obtained under different conditions, the models can learn correlations with data provenance, but not true structure–property relationships. In addition to this limited chemical diversity, the transferability of models may be further limited if predictions are made outside the compositional and structural range of the training data set. These problems are especially relevant in the case of mixed-cation, mixed-halide and lead-free perovskites, because of the presence of chemical substitution which can induce structural and electronic properties that are not well described in the current dataset. To determine if an ML model can successfully extrapolate to unexplored materials, therefore, feature selection, chemically meaningful descriptors, appropriate train/test partitioning, uncertainty estimation, and external validation are required (Tao et al., 2021; Wang et al., 2023).

Another constraint is that many of the traditional ML models describe material properties as static parameters, while the perovskite materials and devices have highly dynamic behaviour. Device composition, electronic structure, interfaces and device performance can change over time due to ion migration, light exposure, thermal exposure, and processes that depend on bias. Nowadays, ion migration is known to impact the operational behaviour and stability of perovskite solar cells (PSC) (Ren et al., 2021; Zhao et al., 2021b; Roose et al., 2021). Thus, static ML predictions, which rely primarily on equilibrium material descriptors, might not adequately describe time-dependent processes that are important for operational device performance.

One-dimensional models like SCAPS-1D are useful to analyze multilayer PV structures, but the results generated from these models rely on the assumptions and input parameters for the model. Simplifications can be: one-dimensional geometry, fixed defect distributions, steady-state transport, simplified interface description, and parameterized recombination mechanisms. These assumptions do not capture the spatial variations, transient response, migration of ions, degradation processes, and complicated interfacial behavior that are usually observed in real perovskite devices. In particular, recent analyses have singled out the issue of over-optimistic or overly high efficiencies predicted by SCAPS-1D models and demonstrated that using improper input parameters such as defect densities, recombination coefficients, and doping concentrations, could yield efficiencies significantly different from experimental results (Saidarsan et al., 2025).

Thus, it is important to note that the photovoltaic efficiencies obtained from SCAPS are only model dependent numerical values and are not predictive of experimental or long-term operation. The reliability of a simulated PCE is highly dependent on the physical realism of the input parameters, and on the experimental justification of those parameters. This distinction is especially significant if SCAPS is used in conjunction with ML, as unrealistic assumptions in the training data can be reproduced or amplified by an ML model. The predictions should then be tested with experimentally realistic ranges of parameters, sensitivity analysis, estimates of uncertainty, and if possible, comparison with experimental measurements (Saidarsan et al., 2025).

To overcome these limitations, greater consistency in the generation of DFT data is needed, along with the availability of uncertainty-aware and chemically diverse ML datasets, active learning strategies to target acquisition of high-fidelity data and device models that represent time-dependent

processes. Especially, it would be desirable to incorporate ion migration, transient carrier dynamics, interfacial changes, and degradation mechanisms into suitable devices-modelling frameworks to better link the static material descriptors with the operational PV behavior. This is crucial for advancing towards more accurate and physically consistent multiscale design strategies for PVs, which can be then validated in experiments (Ren et al., 2021; Zhao et al., 2021b; Roose et al., 2021).

10. Challenges and Future Directions in Integrated Computational Perovskite Design

While the first-principles modelling, machine learning and numerical device simulation of a perovskite PV are progressing quickly, a predictive multiscale computational framework has not yet been adopted in this field. The uniformity of material properties over computer levels is one challenge. The DFT-derived quantities like the bandgaps, energetics of defect states, and dielectric property can vary from one type of exchange–correlation functional, pseudopotentials, treatment of spin–orbit coupling, structural configuration to supercell size. Methodological variations may be introduced when datasets are collected using different computational protocols, without proper validation and normalization, and can lead to systematic variations in the training data for ML, which consequently affect the transferability of the ML across different perovskite chemistries (Meggiolaro et al., 2018; Xue et al., 2022; Tao et al., 2021).

The second crucial issue is the lack of coverage of the current computational datasets. While many existing datasets focus on bulk equilibrium structures and relatively easily accessible material properties, other quantities like interface energetics, defect kinetics, ion-migration barriers, metastable phases, and degradation-related processes are more difficult to compute and are less well represented. This imbalance may prevent learning of phenomena that become relevant in real-world operations from existing datasets, making it difficult for the resulting ML models to capture them. Such high-fidelity calculations of defects, interfaces, non-equilibrium structures, and ion-migration processes should thus be increasingly included in future computational databases in addition to the bulk descriptors already defined (Ren et al., 2021; Zhao et al., 2021b; Roose et al., 2021).

Another crucial path is to enhance the extrapolative ability and physical interpretation of ML models. The use of physics-informed ML architectures by adding physics-meaningful constraints, symmetry information, conservation principles, and chemically relevant descriptors might help control for spurious correlations and enhance models' generalization. Simultaneously, active-learning strategies can facilitate iterative building of more informative datasets through the creation of data from more high-fidelity calculations that can be used when needed to learn about compositions or structures of which the model is uncertain. This can decrease the number of calculations and gradually increase the chemical space that is covered by the training set (Tao et al., 2021; Wang et al., 2023).

Additional problems occur at the level of devices and simulations. One-dimensional steady-state models such as SCAPS-1D offer powerful tools for analyzing multilayer solar cell architectures but are not able to account for all aspects of spatial inhomogeneity, transient ion migration, changes induced by illumination and degradation processes in long term operation. Therefore, the time-dependent transport, ion migration, interfacial evolution, and degradation kinetics in the devices will increasingly be taken into account in future device-modelling frameworks, if these effects are relevant to research questions. Standardized methods for mapping DFT-generated and ML-derived descriptors to the device-model parameters to achieve higher reproducibility and enable comparisons across studies are also required (Tessler and Vaynzof, 2020; Saidarsan et al., 2025).

One of the key long-term goals is to create autonomous or closed-loop computational materials discovery frameworks. In this workflow, DFT calculations at high fidelity would produce physically realistic data in order to validate the properties of unexplored chemical space in the ML model, followed by device-level simulations to identify promising candidates, and ultimately feedback to the ML model for further DFT calculations. Such strategies of active and closed learning could be used to decrease the number of costly calculations needed for materials exploration while continually enhancing the predictive model (Tao et al., 2021; Srivastava et al., 2021).

More community-level standardization will be necessary in order to make this vision a reality. Reproducibility could benefit from consistency in data formats, clear reporting of computational parameters, benchmarking datasets, clear validation protocols, uncertainty quantification, as well as from open computational tools. This would aid the development of a more robust link between atomistic simulations, data-driven materials discovery, and device-level modelling of photovoltaic performance for perovskites, enabling their computational design in a more systematic way.

11. Future Perspectives and New Trends

The future of perovskite materials discovery will increasingly depend on autonomous and adaptive computational workflows, that integrate first-principles calculations, machine learning, device modelling and experiments. Active-learning and closed-loop optimization methods are on the other hand very promising, as they can employ uncertainty or information gain in the model to prioritize the next set of calculations or experiments. Recent studies have shown active-learning strategies for perovskite materials and the potential of such strategies to decrease the amount of high fidelity data needed for materials exploration. Uncertainty-aware active learning has also recently been utilized to determine regions with unreliable model predictions for lead-free halide-perovskite screening, demonstrating its potential and the need to identify regions where model predictions are unreliable (Min et al., 2020; Lourenço et al., 2023; Yu, 2026).

A key research trajectory is the field of physics-informed and explainable machine learning methods. Using physically meaningful descriptors, symmetry relationships, conservation principles, or other domain constraints when constructing the ML model could enhance the interpretability and minimize the chance of learning spurious correlations. Related to the above, in perovskite research, explainable ML has already been applied to discover the compositional and elemental descriptors that influence the bandgaps of ABX_3 materials. Ensemble ML models trained with DFT-derived data showed their ability to predict the bandgap of perovskite materials and yield insightful information regarding elemental descriptor importance (Obada et al., 2023). More generally, physics-informed ML methods are being developed to fuse quantum-mechanical descriptors with data-driven models that are physically constrained, which could offer the potential for more interpretable and physically consistent prediction frameworks (Khatua et al., 2024).

Another challenge lies in the creation of multimodal datasets and models that integrate information across various points along the materials-to-device pathway. Apart from the numerical composition or processing descriptors, which are the most commonly used in conventional ML models, the experimental perovskite research also provides valuable information, such as microscopy images, spectral measurements, processing histories, device architectures, degradation data, etc. Recent studies have shown that multimodal ML models have been developed using microstructural data from SEM and composition and processing parameters to predict perovskite-device performance. This demonstrates the potential of synthesizing complementary data modalities to capture the relations that could not be obtained from composition-only datasets (Liu et al., 2025; Rahman et al., 2025).

An integration of computational and experimental datasets may prove to be especially useful in the quest to understand long-term stability and degradation that cannot be well captured by static equilibrium descriptors. Taking into account the DFT-generated properties along with the experimental characterization, processing conditions, microstructural information, and time-dependent measurements of the devices could enable the use of ML models to predict quantities related to device stability, including efficiency retention, and degradation behavior. This is aligned with the recent advances in ML research for perovskite development, which have started to take into account various aspects of the materials synthesis, processing, characterization, and device operation (Mao and Xiang, 2025; Li and Yi, 2025).

Another research direction is the development of computational approaches for lead-free, low-toxicity perovskites. Beyond lead-based perovskites, lead-free halide double perovskites and related mixed-composition materials offer available chemical space for computational screening, but can have electronic structures, defect chemistry, and photovoltaic properties that vary significantly from that of lead-based perovskites. The use of uncertainty-aware ML and active learning to guide screening of lead-free halide-perovskites is especially intriguing as it can assist in differentiating predictions from regions of chemical space where the data are missing. (Yu, 2026)

Last but not least, deeper connections with open databases and standardized data formats, reproducible computational workflows and digital materials platforms will be vital to enhancing collaboration and reproducibility. It would be helpful to have a common format for reporting DFT computational parameters, ML descriptors, training/test procedures, uncertainty estimates, and device-simulation inputs so as to facilitate comparison of the results produced by different research groups. Such infrastructure, alongside open-source computational tools and automated data-generation workflows, could enable more autonomous materials-discovery systems where DFT calculations generate high fidelity reference data, the ML models can indicate useful candidate materials with quantified uncertainty, and experimental or simulation work on devices can inform the ML models to improve the subsequent prediction. Active-learning strategies can also be applied to minimize the number of DFT calculations that need to be performed throughout model development, as has been done for data-efficient ML dataset generation for perovskite alloys recently.

12. Conclusions

The complementary use of density functional theory (DFT), machine learning (ML) and numerical device simulation in computational investigation and design of perovskite photovoltaic materials has been inspected. However, these methods do not stand alone but they have been discussed here with regard to how DFT descriptors of structure, electronic properties, thermodynamic and defect properties can feed into ML models with a physically informed approach, and how the large compositional spaces explored by ML can speed up the discovery process. The candidate materials obtained are then incorporated in the device level modelling frameworks to explore their possible effects on PV parameters like Voc, Jsc, FF, and PCE (Tao et al., 2021; Srivastava et al., 2021; Wang et al., 2023).

The literature reviewed shows complementary aspects of the materials to the device problem being tackled by DFT, ML and device simulation. The physically grounded basis of DFT, combined with the efficiency of ML for prediction and screening over large chemical spaces, and device simulation of the relation of selected material and interface parameters to the macroscopic photovoltaic performance make it an ideal complement. Thus, the conceptual DFT–ML–device parameter-mapping workflow outlined in this review is helpful in connecting various computational length and time scales. The transfer of parameters between these levels, however, will not necessarily occur

automatically and needs to be done properly on the basis of suitable physical assumptions, calibration, uncertainty assessment and validation. Specifically, some quantities, like defect formation energies, effective masses, carrier mobilities, or trap densities, should not be directly compared between atomistic and ML models and device models.

The review also points to some obstacles to the construction of reliable multi-scale computational workflows. They encompass methodological inconsistencies in DFT data, as well as small chemical and physical range of ML training data, poor extrapolation out of the ML training domain, insufficient representation of defects and interfaces, and simplified description of dynamic processes in conventional device simulations. Solving these problems will rely on the use of computational protocols (standardized), uncertainty-aware datasets (curated), physically informed ML models, and the treatment of time-dependent processes like ion migration and degradation (Ren et al., 2021; Zhao et al., 2021b; Saidarsan et al., 2025).

Going forward, it is envisioned that the materials discovery process could become even more computational and closed loop, with computational processes increasingly accurate and providing reference data to a machine-learning model that would provide guidance on promising materials with uncertainty quantification, and with device-level simulations providing a quantitative assessment of photovoltaic device performance. This information could then be utilized in the active-learning strategies to decide on what further calculations/experiments would best improve the models. This would not only decrease the waste of calculations but would also gradually improve the accuracy of materials screening, as proposed by Tao et al. (2021) and Srivastava et al. (2021).

As a whole, the coupling of DFT-based methods with ML and device simulation holds great promise for computational research in perovskite photovoltaics. But the current state of the field is more of a developing end-to-end predictive pipeline than a validated end-to-end cross-scale framework. Future progress in data standardization, model validation, physically consistent parameter mapping, uncertainty quantification and experimental benchmarking will be critical in bridging the conceptual framework to reliable computational tools to discover and optimize efficient and stable perovskite photovoltaic materials.

Declarations

Funding

No funding was received for conducting this study.

Data Availability

Not applicable. No datasets were generated or analyzed during the current study as this is a review article.

Ethical Approval

Not applicable. This study does not involve human participants or animal subjects.

Consent to Participate

Not applicable. This review does not involve human subjects.

Consent to Publish

Not applicable.

Competing Interests

The author declares that there are no competing financial or non-financial interests.

Author Contributions

Sameer Pandey is the sole author of this work. S.P. conceptualized the review, performed the literature curation and comparative analysis, developed the multi-scale mapping framework, authored the initial manuscript, and revised the final text.

References :

1. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O., & Walsh, A. (2018). Machine learning for molecular and materials science. Nature, 559, 547–555. DOI: 10.1038/s41586-018-0337-2

2. Kojima, A., Teshima, K., Shirai, Y., & Miyasaka, T. (2009). Organometal halide perovskites as visible-light sensitizers for photovoltaic cells. Journal of the American Chemical Society, 131, 6050–6051. DOI: 10.1021/ja809598r

3. Snaith, H. J. (2013). Perovskites: The emergence of a new era for low-cost, high-efficiency solar cells. The Journal of Physical Chemistry Letters, 4, 3623–3630. DOI: 10.1021/jz4020162

4. Yin, W.-J., Shi, T., & Yan, Y. (2014). Unusual defect physics in $CH_3NH_3PbI_3$ perovskite solar cell absorber. Applied Physics Letters, 104, 063903. DOI: 10.1063/1.4864778

5. Jeong, M., Choi, I. W., Go, E. M., et al. (2020). Stable perovskite solar cells with efficiency exceeding 24.8% and 0.3-V voltage loss. Science, 369, 1615–1620. DOI: 10.1126/science.abb7167

6. Gao, W., Chen, C., Ran, C., et al. (2020). A-site cation engineering of metal halide perovskites: Version 3.0 of efficient tin-based lead-free perovskite solar cells. Advanced Functional Materials, 30, 2000794. DOI: 10.1002/adfm.202000794

7. Huang, Y., et al. (2019). A-site cation engineering for highly efficient $MAPbI_3$ single-crystal X-ray detector. Angewandte Chemie International Edition. DOI: 10.1002/anie.201911281

8. Fedorovskiy, A. E., Drigo, N. A., & Nazeeruddin, M. K. (2020). The role of Goldschmidt's tolerance factor in the formation of A_2BX_6 double halide perovskites and its optimal range. Small Methods, 4, 1900426. DOI: 10.1002/smtd.201900426

9. Meyer, E., Mutukwa, D., Zingwe, N., & Taziwa, R. (2018). Lead-free halide double perovskites: A review of the structural, optical, and stability properties as well as their viability to replace lead halide perovskites. Metals, 8, 667. DOI: 10.3390/met8090667

10. Bibi, A., Lee, I., Nah, Y., et al. (2021). Lead-free halide double perovskites: Toward stable and sustainable optoelectronic devices. Materials Today, 49, 123–144. DOI: 10.1016/j.mattod.2020.11.026

11. Wolf, N. R., Connor, B. A., Slavney, A. H., & Karunadasa, H. I. (2021). Doubling the stakes: The promise of halide double perovskites. Angewandte Chemie International Edition, 60, 16264–16278. DOI: 10.1002/anie.202016185

12. Ghasemi, M., Hao, M., Xiao, M., et al. (2021). Lead-free metal-halide double perovskites: From optoelectronic properties to applications. Nanophotonics, 10, 2181–2219. DOI: 10.1515/nanoph-2020-0548

13. Ji, G., Han, C., Hu, S., et al. (2021). B-site columnar-ordered halide double perovskites: Theoretical design and experimental verification. Journal of the American Chemical Society, 143, 10275–10281. DOI: 10.1021/jacs.1c03825

14. Huang, Y., Yin, W.-J., & He, Y. (2018). Intrinsic point defects in inorganic cesium lead iodide perovskite $CsPbI_3$. The Journal of Physical Chemistry C, 122, 1345–1350. DOI: 10.1021/acs.jpcc.7b10045

15. Meggiolaro, D., & De Angelis, F. (2018). First-principles modeling of defects in lead halide perovskites: Best practices and open issues. ACS Energy Letters, 3, 2206–2222. DOI: 10.1021/acsenergylett.8b01212

16. Xue, H., Brocks, G., & Tao, S. (2022). Intrinsic defects in primary halide perovskites: A first-principles study of the thermodynamic trends. Physical Review Materials, 6, 055402. DOI: 10.1103/PhysRevMaterials.6.055402

17. Pilania, G., Mannodi-Kanakkithodi, A., Uberuaga, B. P., et al. (2016). Machine learning bandgaps of double perovskites. Scientific Reports, 6, 19375. DOI: 10.1038/srep19375

18. Schmidt, J., Marques, M. R. G., Botti, S., & Marques, M. A. L. (2019). Recent advances and applications of machine learning in solid-state materials science. npj Computational Materials, 5, 83. DOI: 10.1038/s41524-019-0221-0

19. Tao, Q., Xu, P., Li, M., & Lu, W. (2021). Machine learning for perovskite materials design and discovery. npj Computational Materials, 7, 23. DOI: 10.1038/s41524-021-00495-8

20. Lookman, T., Balachandran, P. V., Xue, D., & Yuan, R. (2019). Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. npj Computational Materials, 5, 21. DOI: 10.1038/s41524-019-0153-8

21. Yang, Z., et al. (2021). Machine learning accelerates the discovery of light-absorbing materials for double perovskite solar cells. The Journal of Physical Chemistry C, 125, 22483–22492. DOI: 10.1021/acs.jpcc.1c07262

22. Filipoiu, N., Mitran, T. L., Anghel, D. V., et al. (2021). Investigation of opto-electronic properties and stability of mixed-cation mixed-halide perovskite materials with machine-learning implementation. Energies, 14, 5431. DOI: 10.3390/en14175431

23. Srivastava, M., Howard, J. M., Gong, T., et al. (2021). Machine learning roadmap for perovskite photovoltaics. The Journal of Physical Chemistry Letters, 12, 7866–7877. DOI: 10.1021/acs.jpclett.1c01961

24. Wang, J., Xu, P., Ji, X., Li, M., & Lu, W. (2023). Feature selection in machine learning for perovskite materials design and discovery. Materials, 16, 3134. DOI: 10.3390/ma16083134

25. Yılmaz, B., Odabaşı, Ç., & Yıldırım, R. (2022). Efficiency and stability analysis of 2D/3D perovskite solar cells using machine learning. Energy Technology, 10, 2100948. DOI: 10.1002/ente.202100948

26. Tessler, N., & Vaynzof, Y. (2020). Insights from device modeling of perovskite solar cells. ACS Energy Letters, 5, 1260–1270. DOI: 10.1021/acsenergylett.0c00172

27. Le Corre, V. M., Wang, Z., Koster, L. J. A., & Tress, W. (2020). Device modeling of perovskite solar cells: Insights and outlooks. In Soft-Matter Thin Film Solar Cells: Physical Processes and Device Simulation, 109–140. American Institute of Physics. DOI: 10.1063/9780735422414_004

28. Islam, M. S., Sobayel, K., Al-Kahtani, A., et al. (2021). Defect study and modelling of SnX_3-based perovskite solar cells with SCAPS-1D. Nanomaterials, 11, 1218. DOI: 10.3390/nano11051218

29. Bhavsar, K., & Lapsiwala, P. B. (2021). Numerical simulation of perovskite solar cell with different material as electron transport layer using SCAPS-1D software. Semiconductor Physics, Quantum Electronics & Optoelectronics, 24, 341–347. DOI: 10.15407/spqeo24.03.341

30. Srivastava, P., Sadanand, Rai, S., et al. (2022). Theoretical study of perovskite solar cell for enhancement of device performance using SCAPS-1D. Physica Scripta, 97, 125004. DOI: 10.1088/1402-4896/ac9dc5

31. Jamal, M. S., Shahahmadi, S. A., Chelvanathan, P., et al. (2019). Effect of defect density and energy level mismatch on the performance of perovskite solar cells by numerical simulation. Optik, 182, 1204–1210. DOI: 10.1016/j.ijleo.2018.12.163

32. Hasanzadeh Azar, M., Aynehband, S., Abdollahi, H., et al. (2023). SCAPS empowered machine learning modelling of perovskite solar cells: Predictive design of active layer and hole transport materials. Photonics, 10, 271. DOI: 10.3390/photonics10030271

33. Ren, X., Zhang, L., Yuan, Y., & Ding, L. (2021). Ion migration in perovskite solar cells. Journal of Semiconductors, 42, 010201. DOI: 10.1088/1674-4926/42/1/010201

34. Zhao, Y., Zhou, W., Han, Z., Yu, D., & Zhao, Q. (2021b). Effects of ion migration and improvement strategies for the operational stability of perovskite solar cells. Physical Chemistry Chemical Physics, 23, 94–106. DOI: 10.1039/D0CP04418K

35. Roose, B., et al. (2021). Ion migration drives self-passivation in perovskite solar cells and is enhanced by light soaking. RSC Advances, 11, 12095–12101. DOI: 10.1039/D1RA01166A

36. Saidarsan, A., Guruprasad, S., Malik, A., Basumatary, P., & Ghosh, D. S. (2025). A critical review of unrealistic results in SCAPS-1D simulations: Causes, practical solutions and roadmap ahead. Solar Energy Materials and Solar Cells, 279, 113230. DOI: 10.1016/j.solmat.2024.113230

37. Bare, Z. J. L., Morelock, R. J., & Musgrave, C. B. (2023). Dataset of theoretical multinary perovskite oxides for data-driven materials design. Scientific Data, 10, 244. DOI: 10.1038/s41597-023-02127-w

38. Mayr, F., & Gagliardi, A. (2021). Global property prediction: A benchmark study on open-source, perovskite-like datasets. ACS Omega, 6, 12722–12732. DOI: 10.1021/acsomega.1c00991

39. Lu, F., Liang, Y., Wang, N., et al. (2024). Machine learning for perovskite optoelectronics: A review. Advanced Photonics, 6, 054001. DOI: 10.1117/1.AP.6.5.054001

40. Roberts, N., Jones, D., Schuy, A., Hsu, S.-C., & Lin, L. Y. (2024). Machine Learning for Perovskite Solar Cells: An Open-Source Pipeline. Advanced Physics Research, 3(11), 2400060. DOI: 10.1002/apxr.202400060

41. Pandey, S., Shukla, N., & Sharma, V. K. (2026a). Machine Learning-Assisted First-Principles Investigation of Structural, Electronic, and Photovoltaic Properties of Perovskite Materials. Applied Nanoscience. https://doi.org/10.1007/s13204-026-03184-4

42. Pandey, S., Shukla, N., Sharma, V. K., & Tiwari, S. (2026b). From first-principles to device performance: a computational perspective on perovskite solar cell design. Russian Physics Journal. https://doi.org/10.1007/s11182-026-03763-6

43. Pandey, S. (2026c). Computational Investigation and Defect-Related Optimization of Lead-Free $Cs_2AgBiBr_6$ Double-Perovskite Solar Cells Using DFT, SCAPS-1D, and Machine Learning. *Research Square* (Preprint). https://doi.org/10.21203/rs.3.rs-10810357/v1

44. Abbas, S. R., Mir, B. A., Ryu, J., & Lee, S. W. (2025). Toward sustainable solar energy: Predicting recombination losses in perovskite solar cells with deep learning. Sustainability, 17, 5287. DOI: 10.3390/su17125287

45. Min, K., & Cho, E. (2020). Accelerated discovery of potential ferroelectric perovskite via active learning. Journal of Materials Chemistry C, 8, 7866–7872. DOI: 10.1039/D0TC00985G

46. Lourenço, M. P., Tchagang, A., Shankar, K., Thangadurai, V., & Salahub, D. R. (2023). Active learning for optimum experimental design—insight into perovskite oxides. Canadian Journal of Chemistry, 101, 734–744. DOI: 10.1139/cjc-2022-0198

47. Obada, D. O., Okafor, E., Abolade, S. A., et al. (2023). Explainable machine learning for predicting the band gaps of ABX_3 perovskites. Materials Science in Semiconductor Processing, 161, 107427. DOI: 10.1016/j.mssp.2023.107427

48. Khatua, R., Das, B., & Mondal, A. (2024). Physics-informed machine learning with data-driven equations for predicting organic solar cell performance. ACS Applied Materials & Interfaces, 16, 57467–57480. DOI: 10.1021/acsami.4c10868

49. Liu, H., Yang, A., Zhong, C., et al. (2025). Integration of microstructural image data into machine learning models for advancing high-performance perovskite solar cell design. ACS Energy Letters, 10, 1884–1891. DOI: 10.1021/acsenergylett.5c00626

50. Rahman, W., Zhong, C., Liu, H., et al. (2025). Enhancing perovskite solar cell efficiency and stability: A multimodal prediction approach integrating microstructure, composition, and processing technology. Nanoscale, 17, 15935–15949. DOI: 10.1039/D5NR00364D

51. Mao, L., & Xiang, C. (2025). A comprehensive review of machine learning applications in perovskite solar cells: Materials discovery, device performance, process optimization and systems integration. Materials Today Energy, 47, 101742. DOI: 10.1016/j.mtener.2024.101742

52. Li, A., & Yi, C. (2025). From data-driven to intelligence-driven: Innovating research paradigms of perovskite solar cells. Solar RRL, 9, e202500681. DOI: 10.1002/solr.202500681

53. Yu, X. (2026). Uncertainty-aware active learning reveals reliability limits in lead-free halide perovskite screening. Digital Discovery, 5, 2342–2352. DOI: 10.1039/D6DD00153J

54. Merchant, A., Batzner, S., Schoenholz, S. S., Aykol, M., Cheon, G., & Cubuk, E. D. (2023). Scaling deep learning for materials discovery. Nature, 624, 80–85. DOI: 10.1038/s41586-023-06735-9

55. Bansal, N. K., Mishra, S., Dixit, H., Porwal, S., Singh, P., & Singh, T. (2023). Machine learning in perovskite solar cells: Recent developments and future perspectives. Energy Technology, 11, 2300735. DOI: 10.1002/ente.202300735

56. Liu, T., Wang, S., Shi, Y., Wu, L., Zhu, R., Wang, Y., Zhou, J., & Choy, W. C. H. (2023). Machine-learning accelerating the development of perovskite photovoltaics. Solar RRL, 7, 2300650. DOI: 10.1002/solr.202300650

57. Chen, M., Yin, Z., Shan, Z., Zheng, X., Liu, L., Dai, Z., Zhang, J., Liu, S. Z., & Xu, Z. (2024). Application of machine learning in perovskite materials and devices: A review. Journal of Energy Chemistry, 94, 254–272. DOI: 10.1016/j.jechem.2024.02.035

58. Zhao, Y., Zhang, J., Xu, Z., et al. (2021a). Discovery of temperature-induced stability reversal in perovskites using high-throughput robotic learning. Nature Communications, 12, 2191. DOI: 10.1038/s41467-021-22472-x

59. Novoselov, I. E., Gvozdev, A. M., Smirnov, A. A., Zhidkov, I. S., et al. (2025). Dataset of SCAPS-1D simulated halide perovskite solar cells with SHAP and machine learning-based PCE optimization. Data in Brief, 60, 111653. DOI: 10.1016/j.dib.2025.111653

60. Mustafa, U. G., Wu, W., Wang, M., Hashibon, A., & Anwar, H. (2025). Machine learning-assisted optimization of $CsPbI_3$-based all-inorganic perovskite solar cells: A combined SCAPS-1D and XGBoost approach. Energy and AI, 21, 100559. DOI: 10.1016/j.egyai.2025.100559