Solar Energy 303 (2026) 114112

Contents lists available at ScienceDirect

Solar Energy

journal homepage: www.elsevier.com/locate/solener

Interpretable machine learning on a curated dataset identifies chemical descriptors governing 2D perovskite solar cell performance

Migon Choi [a], Aiden Sable [a], Nicolas Ramos [b], Richard Sheridan [a], Olivier Delaire [a,c], Hai Li [b], David B. Mitzi [a,c], L.Catherine Brinson [a,*]

[a] Department of Mechanical Engineering and Materials Science, Duke University, Durham, NC 27708, USA
[b] Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708, USA
[c] Department of Chemistry, Duke University, Durham, NC 27708, USA

A B S T R A C T

Two-dimensional (2D) hybrid perovskites are promising candidates for stable and efficient solar energy conversion, yet their complex chemistry and device architecture pose challenges for rational design. In this study, we compile a curated dataset—from database and literature sources—of 792 2D perovskite solar cell devices, including detailed chemical descriptors for large organic spacer cations, and investigate how these descriptors relate to photovoltaic performance. Using an interpretable machine learning model, we evaluate the relative importance of chemical and device parameters in predicting key performance metrics. Our analysis reveals that chemical descriptors, particularly the large organic-to-metal ratio and small organic cation type, have a greater influence on device efficiency than stack architecture or deposition method. The large organic-to-metal ratio emerges as the most influential feature in our model, accounting for 19.5 % of the total feature importance. The results highlight the importance of chemically informed data representation and demonstrate how structured datasets can uncover key structure–performance relationships in complex hybrid materials.

1. Introduction

While three-dimensional (3D) perovskites have driven remarkable progress in photovoltaic (PV) applications due to their potential for high efficiency and low-cost energy production [1–3], their long-term stability remains a critical challenge, as solar cells are typically exposed to outdoor conditions for decades [4]. To address this issue, two-dimensional (2D) perovskites have attracted increasing attention. In these layered systems, larger organic cation spacers separate the inorganic layers, while small cations occupy the intralayer spaces, together determining the material's structural and optoelectronic properties (Fig. 1) [5,6]. Introducing organic spacer cations slices the infinite inorganic framework into 2D layers, with varying dimensionality depending on the ratio of small to larger organic cations [7], and forms natural quantum wells, enhancing environmental stability compared to 3D counterparts [8]. Recent advances in 2D perovskites, including developments in spintronics, ferroelectricity, and charge transport, have expanded their applications to LEDs, photodetectors, and solar cells [9–12].

Machine Learning (ML) has emerged as a powerful framework across perovskite-related devices, enabling the prediction of material properties, optimization of synthesis conditions, and forecasting of device performance and stability [13–24]. ML has been successfully applied to solid oxide fuel cells [13], light-emitting diodes [14], photodetectors [15], space applications [16], perovskite dimension classification [17], and solar cells, particularly for performance prediction [18–21], stability modeling [22–24], and synthesis optimization [19]. Machine learning (ML) applications to date have largely focused on 3D perovskites. While early efforts applying ML to 2D perovskite solar cells have made important progress [18,25–27], continued developments, including models predicting photovoltaic behavior from optical spectra [28] and ligand screening for 2D/3D interfaces [29], demonstrate the growing momentum and expanding potential of ML in this field. While previous works have explored different aspects of perovskite design – Mai et al. focused on screening organic spacers at the molecular level [27], Mammeri et al. emphasized device-level stability prediction [23], and Alsulami et al. analyzed stability trends across device architectures [22] – this study bridges the gap between molecular-level chemistry and device-level characteristics in 2D perovskites.

An important factor in 2D perovskite solar cells is how the organic spacer cation governs structural distortion, optical properties, and ultimately device performance. Mahal et al. showed that monovalent and divalent spacers significantly impact the crystal framework [30], while other studies have highlighted the spacer's role in modulating stability,

* Corresponding author.
E-mail address: cate.brinson@duke.edu (L.Catherine Brinson).

https://doi.org/10.1016/j.solener.2025.114112
Received 19 August 2025; Received in revised form 15 October 2025; Accepted 27 October 2025
Available online 31 October 2025
0038-092X/© 2025 International Solar Energy Society. Published by Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

efficiency, and charge transport [31–34]. For example, halogen functionalization of organic spacers, such as fluorine addition, has been shown to enhance power conversion efficiency (*PCE*) [35,36]. The current record *PCE* of 21.6 % was achieved using a glycine-based cation acting as both spacer and additive [37]. While small cations such as formamidinium (FA) can also affect film orientation and device behavior [38,39,40], the primary focus remains on understanding how organic spacer descriptors—such as molecular weight, backbone structure (linear vs. ring) [41], and halogen substitution [42,43]—systematically impact performance. While a growing body of literature on 2D perovskites exists, only a few studies have utilized ML approaches on this data to explore the relationship between these chemical characteristics and 2D perovskite solar cell outcomes to date [27].

Complicating the successful application of ML to perovskites has been the diversity of experimental data and the traditional archival method in text form in individual journal papers. Meftahi et al. utilized 100 entries of experimental data to optimize 2D perovskite solar cell fabrication and enhance performance prediction by applying ML [18], while Dunlap et al. generated 50 data points to develop ML models to predict the stability of 3D perovskite solar cells from their initial dynamics [24]. Despite their significant contributions, these earlier models could be further optimized through application to larger datasets. To overcome the barrier of individual researchers extracting data from the literature, Jacobsson et al. collected more than 40,000 experimental data points from 16,000 perovskite studies to develop a Perovskite Database [44], originally hosted on Materials Zone (https://www.materials.zone/) and now integrated into the NOMAD repository [45]. The original database focused on structural details and key performance metrics but lacked curated chemical descriptors critical for deeper analysis. Subsequently, Zhang et al. effectively utilized this database to perform a statistical analysis evaluating various stability-enhancement strategies encompassing 3D and 2D perovskite systems [46]. However, Saliba et al. pointed out potential biases in the database due to the nature of the experiments [47]. Furthermore, data imbalance can occur as some materials are studied more frequently than others [48]. As a result, applying ML to predict perovskite device performance demands not only more robust and representative datasets but also a cautious approach to avoid overinterpretation or bias.

In this study, we focus on compiling and analyzing a structured dataset of 2D perovskite solar cell devices combining existing data in the Materials Zone database with additional manual literature data extraction, to add chemical descriptors for the 2D systems absent from the original Materials Zone data curation and NOMAD repository. The additional chemical data is added to the Duke Research Data Repository (RDR) to make this expanded dataset available for all researchers (See Data Availability section). Using the combined dataset, we tune and train XGBoost models to predict performance and extract feature importance, aiming to (1) create reliable ML predictors despite data imbalance and (2) identify the most impactful chemical and device descriptors. This work highlights how data-driven insights can inform the design and optimization of next-generation 2D perovskite solar cells.

2. Results and Discussion

2.1. Data acquisition

Data were collected from the Materials Zone database (https://www.materials.zone/, data are now available on NOMAD [45]) and filtered to include only 2D perovskites, excluding entries labeled as 3D, 1D, 0D, or other dimensionalities. As the Materials Zone database does not include chemical information on large cation spacers, we manually collected the data from the original publications or website links (such as PubChem [49]). By identifying 45 unique large cation spacers in the Materials Zone data, we mapped chemical information (molecular weight, number of amine groups, structure, and halide chemical attachment) to the data from Materials Zone (Table S1). Ultimately, the dataset comprised 792 data points—i.e., one of the largest datasets currently employed for ML evaluation of 2D perovskite solar cells. All this data is now available for the community in Duke RDR (See Data Availability Section).

2.2. Dataset description

The input to the ML model in this study consists of a feature vector constructed from perovskite information, device information, and chemical information of the large cation spacer. The predictions of the model include *PCE*, short circuit current (J_{SC}), open circuit voltage (V_{OC}), and fill factor (*FF*), where *PCE* represents a composite of the other parameters and is the ratio of the electrical power output to the incident light power input, indicating the overall efficiency of a solar cell. *Jsc* is the maximum current a solar cell can produce under short-circuit conditions, *Voc* is the maximum voltage available from a solar cell at zero current, and *FF* represents the ratio of the actual maximum obtainable power to the product of *Jsc* and *Voc*, reflecting the cell's quality and performance. The labeled input data and corresponding target data are listed in Table 1.

The dataset was cleaned by correcting minor errors from the Materials Zone database, such as unwanted or non-unified punctuation. Next, the perovskite composition columns were divided into organic and inorganic ion columns, with the organic ion entries further divided into small organic cation and large cation columns. In the "small organic" column, most values involve small organic cations, such as methylammonium (MA) and FA ions, although some studies include the inorganic cesium ion; nevertheless, we will refer to this category as "small organic" for convenience. Finally, the data were preprocessed with Multi-label Binarization (MLB) encoding and normalization methods, such as Min-Max scaling or Standard Scaling. Rows containing missing

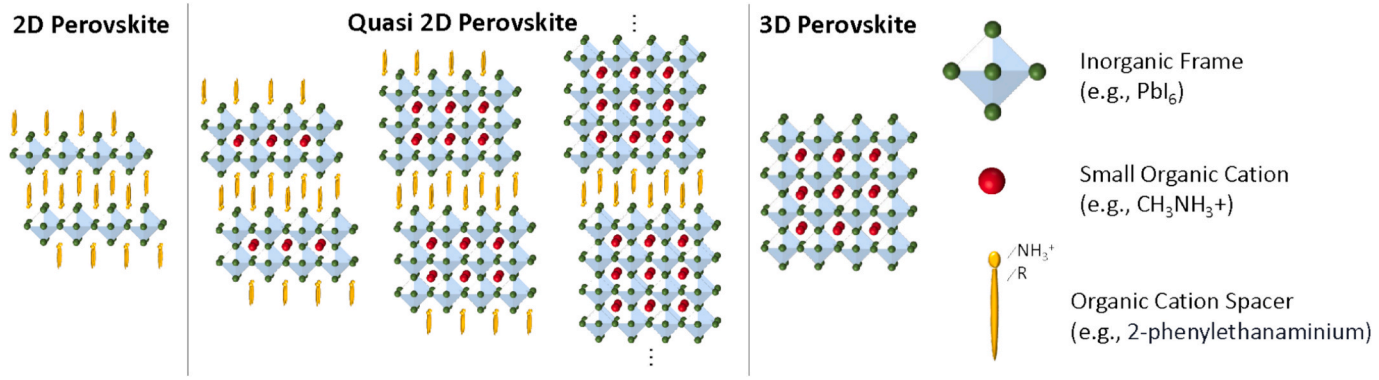


Fig. 1. Illustration of 3D perovskite and 2D hybrid perovskite crystal structures with monovalent ammonium cations. The infinite inorganic layers are sliced into layers characterized by varying degrees of dimensional reduction.

Table 1
Selected features (input) and predictions (output) for ML.

Input	Description	Example(s)
Inorganic	Metal halide ions	Pb, I (lead, iodine)
Small Organic	Small organic cation	MA ($CH_3NH_3^+$, methylammonium)
†Large Organic	Organic Cation Spacer	BA (butylammonium)
Cell architecture	Solar cell architecture	p-i-n
Cell area measured	Solar cell area measured (cm^2)	0.05
ETL stack sequence	Electron transport layer stack sequence	PCBM-60 | LiF
ETL deposition procedure	Electron transport layer deposition procedure	Spin-coating | Evaporation
HTL stack sequence	Hole transport layer stack sequence	PEDOT:PSS
HTL deposition procedure	Hole transport layer deposition procedure	Spin-coating
A ion coefficient	Stoichiometry coefficient of large and small organic cations	2;3
B ion coefficient	Stoichiometry coefficient of metal ions	4
C ion coefficient	Stoichiometry coefficient of halide ions	13
Molecular weight	Molecular weight of fully-protonated large organic cation spacer (g/mol)	74.14
*Number of amines	Number of amine groups for the large organic cation spacer	1
Structure	linear vs. ring structure of large organic cation spacer	linear
Halide chemical	Halide substitution on the large organic cation spacer	None
Large organic and metal ratio	Stoichiometric ratio of the organic cation spacer and metal	0.5
Output	Description	Example
PCE	Power Conversion Efficiency (%)	5.85
J_{sc}	Short Circuit Current (mA/cm^2)	10.84
V_{oc}	Open Circuit Voltage (V)	0.9
FF	Fill Factor	0.6

*In this study, all nitrogen-containing functional groups—including primary, secondary, tertiary, and quaternary amines, as well as imine groups—are collectively categorized under the term "amine groups" for simplicity. This simplification was made to provide a unified framework for investigating nitrogen functionalities. Nitrogen functional groups are typically nitrogen-containing groups bonded to a carbon framework.
†In this study, "cation spacers" are defined operationally as any A-site or additional cation whose incorporation induces interlayer separation and mixed-dimensional stacking, regardless of steric size. Under this functional definition, NH_4^+ was classified as a spacer cation.

data were dropped, ultimately leaving 792 completed samples to use for the modeling, each with 17 features represented (Table 1).

2.3. Dataset analysis

Our first focus was to uncover the chemical composition of the dataset. Here, "chemical composition" refers to the unique combinations of organic spacer cations, small organic cations, and metal halide ions, irrespective of their stoichiometric ratios. The classification aligns with the descriptors defined in Table 1, which include both the types and roles of cations and halide ions in the dataset. We extracted all the unique chemical compositions of samples from the dataset. Then, for each composition, we totaled the number of entries in the dataset within that compositional family, as shown in Fig. 2. From this analysis, we find that the dataset is heavily compositionally imbalanced. The two most prevalent compositions, butylammonium methylammonium lead iodide (BA-MA-Pb-I) and phenethylammonium methylammonium lead iodide (PEA-MA-Pb-I) comprise 34 % of entries in the dataset. At the other extreme, out of the 99 unique chemical compositions in the dataset, 33 are supported by only a single example, leading to extreme sparsity in the remaining features. Though the additional features beyond the

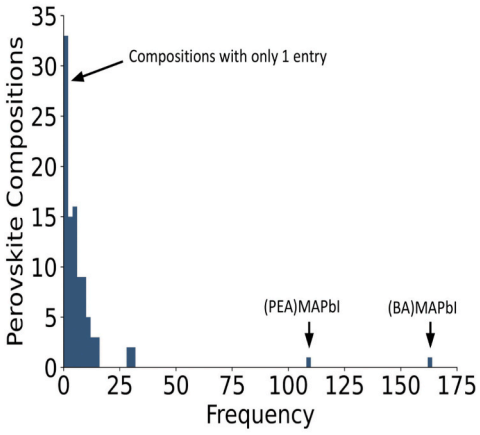


Fig. 2. Perovskite Compositional Analysis: Frequency of Chemical Compositions (x-axis) vs. Types of Perovskite Compositions (y-axis).

chemical composition further diversify entries, the compositional imbalance toward (PEA)- and (BA)-MAPbI derivatives presents a challenge for performing ML and predicting PSC performance across the compositional space.

Another aspect meriting investigation relates to the effect of the measured cell area on the *PCE*. In the literature, it is generally understood that decreasing the cell area leads to an increase in both the achievable *PCE* and PSC stability. This trend is expected due to the increased ease of achieving a uniform and high-quality film and interfaces over smaller areas versus larger areas [50]. Thus, if this effect exists in the 2D PSC data from the database, a normalization procedure for the *PCE* values to the measured cell area is suggested. To investigate this relationship, we considered the intra-compositional variation in *PCE* values for the reported cell areas. However, of the 249 unique stoichiometric compositions, only 24 contained measurements of more than one cell area. This finding largely relates to the fact that most of the compositions contain only one or two entries, though there are also some compounds with many entries that never varied cell areas. From the 24 compositions where the *PCE* vs cell area relationship could be examined, we found no clear evidence of smaller cell areas increasing the *PCE* within the dataset's relatively limited area range of 0.02 cm^2 to 0.4 cm^2 (Fig. S1). The narrow range of reported device areas (0.02—–0.4 cm^2) and the limited number of compositions with multiple measurements (24 out of 249) substantially constrain the statistical robustness of such an analysis. Under these conditions, applying an area-normalization procedure would risk introducing artifacts rather than correcting systematic trends. Accordingly, we did not normalize *PCE* by area in this study; instead, we emphasize the need for broader and more systematic reporting of device areas in future datasets to enable rigorous correction of this effect.

To explore the important impact of the large cation spacer on 2D perovskite solar cell performance, we examined the correlations between the *PCE* and several chemical and structural spacer cation features, which were selected based on prior literature and chemical relevance to perovskite charge transport [27,30,51]. Specifically, we focused on five key descriptors: the cation spacer molecular weight, the number of amine groups present, whether the cationic structure is linear or contains a ring, the type of halide modification, and the stoichiometric ratio of organic cation spacer and metal (Fig. 3). Additionally, we investigate the effect of these descriptors on other performance parameters such as *Jsc*, *Voc*, and *FF* (Figs. S2 – S7).

Considering the impact of molecular weight, the most frequently used cation spacers, specifically butylammonium (74.14 g/mol) and phenethylammonium (122.18 g/mol), appear more densely represented in high-performing devices (Fig. 3(a)). However, this should not be

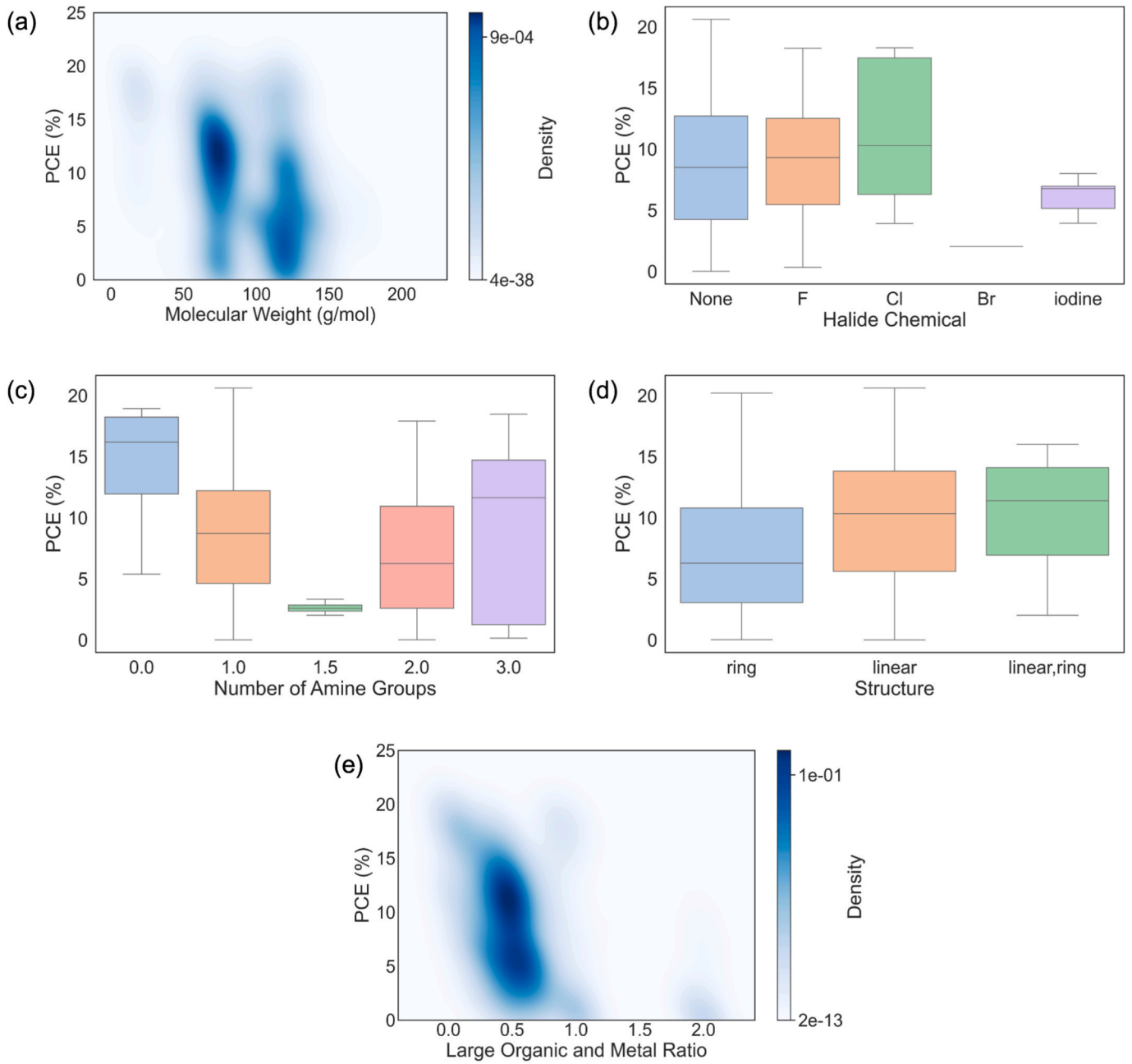


Fig. 3. Relationships between power conversion efficiency (*PCE*) and spacer cation features of (a) molecular weight, (b) halide chemical (since there is only one data point for Br addition, the box plot is reduced to a single line), (c) number of amine groups (Note that fractional values may result from mixtures of cations. A value of 0 indicates a cation that lacks a nitrogen group bonded to a carbon framework), (d) linear vs. ring structure of the organic cation spacer, and (e) stoichiometric ratio of the organic cation spacer and metal—i.e., to represent the number of perovskite layers, n, in the 2D perovskite structures. Kernel density plot is applied for molecular weight and stoichiometric ratio of the organic cation spacer and metal, to visualize the distribution patterns and identify any dominant trends or clusters across the dataset.

interpreted as evidence that these molecular weights inherently yield the highest *PCEs*. Rather, this trend likely reflects historical experimental preference and greater optimization efforts for these specific cations. Importantly, the figure serves as a density summary of existing data and not a prescriptive indicator of optimal values; for instance, sparse data at other molecular weights (e.g., <20 g/mol) also show relatively high *PCEs*. Therefore, caution should be exercised to avoid overgeneralizing from data concentration alone. Regarding halogen addition, cation spacers incorporating chlorine exhibited superior *PCEs*, surpassing those of spacers with other halides or without halide incorporation (Fig. 3(b)). Sharzadi et al. introduced a chlorinated ethylammonium compound, 2-chloroethylammonium, to the MAPbI structure, achieving high efficiency by modulating the bandgap due to distortions in the Pb-I-Pb bonds, stabilizing the perovskite lattice through interactions with grain surfaces, and enhancing thermal stability [52]. Organic cation spacers with zero or three amine groups tend to yield higher *PCE* compared to those with one or two amine groups, as well as those with cation-mixed perovskites (1.5 amine groups) (Fig. 3 (c)). Organic cation spacers with zero or three amine groups may yield higher *PCE* due to either reduced lattice disruption or enhanced structural ordering via multidentate coordination (for tri-amines), consistent with prior findings that spacer–lattice interactions strongly influence crystallinity and charge transport efficiency in low-dimensional perovskites [41].

Further considering spacer cation structure, cation spacers that combine both linear and ring-structured characteristics show a comparable median *PCE* to purely linear spacers and slightly higher than purely ring-based spacers (Fig. 3(d)). While this suggests a potential synergy from combining structural motifs, the apparent performance advantage of the linear and ring group over the linear group is modest (Fig. 3(d)). The combination of linear and ring characteristics arises from cation mixing, such as phenethylammonium (PEA) with butylammonium (BA). Zhou et al. mixed linear and ring structure cations to leverage their structural and electronic properties, with acyclic cations like BA facilitating self-assembly and crystal growth, and ring structure cations like PEA reducing exciton binding energy and improving carrier transport without disturbing the structure [53]. Lastly, the stoichiometric ratio of the organic cation spacer to the metal is investigated for its relationship with *PCE* (Fig. 4(e)). For example, in the quasi-2D perovskite $BA_2MA_3Pb_4I_{13}$, this ratio is 0.5, as there are two BA cations per four Pb ions. A higher spacer-to-metal ratio indicates a greater proportion of organic spacer cations, which leads to fewer inorganic layers (i.e., lower n values) and thus a more two-dimensional structure. Conversely, lower ratios correspond to higher n values and increasingly 3D-like structure. The trend shows that, as the stoichiometric ratio decreases—corresponding to an increase in n—the *PCE* tends to increase. This result is consistent with the findings of Quan et al., which demonstrated improved *PCE* with increasing n. [54]. While higher n values in 2D perovskites improve PCE by approaching 3D-like performance, they reduce the stability benefits of lower-n structures, highlighting the need to balance efficiency and stability.

The above analysis of our dataset shows that the selected spacer cation features may significantly influence device performance. Notably, commonly used cations like butylammonium and phenethylammonium, chlorinated spacers, and mixed linear–ring structures correlate with higher efficiencies.

2.4. XGBoost regressor: Performance analysis

The initial statistical analysis identified general correlations between spacer cation features and device performance, but it is limited in capturing nonlinear interactions or multivariate dependencies. To overcome this limitation, we applied (eXtreme Gradient Boosting) XGBoost, which provides high predictive power and interpretable feature importance metrics. This model facilitates a more comprehensive understanding of how complex combinations of spacer characteristics influence power conversion efficiency.

We used MLB encoding and MinMax scaling to preprocess the features and trained an XGBoost regressor optimized through grid search across eight hyperparameters (details in SI "XGBoost" section and Fig. S8-S10). The optimized model demonstrated strong predictive performance, with average R^2 values close to 0.7 for key photovoltaic metrics (Fig. 4). Each panel in Fig. 4 presents the distribution of R^2 values obtained from repeated 4-fold cross-validation with three random seeds, reflecting model robustness to data partitioning and initialization. The blue histograms represent the empirical distributions, while the overlaid orange curves depict Gaussian fits using the observed means (μ) and standard deviations (σ), as indicated in each subplot. Notably, *PCE*, *Jsc*, and *Voc* (Figs. 4(a) – 4(c)) show consistently high central R^2 values and relatively narrow spreads (σ ≈ 0.04–0.06),

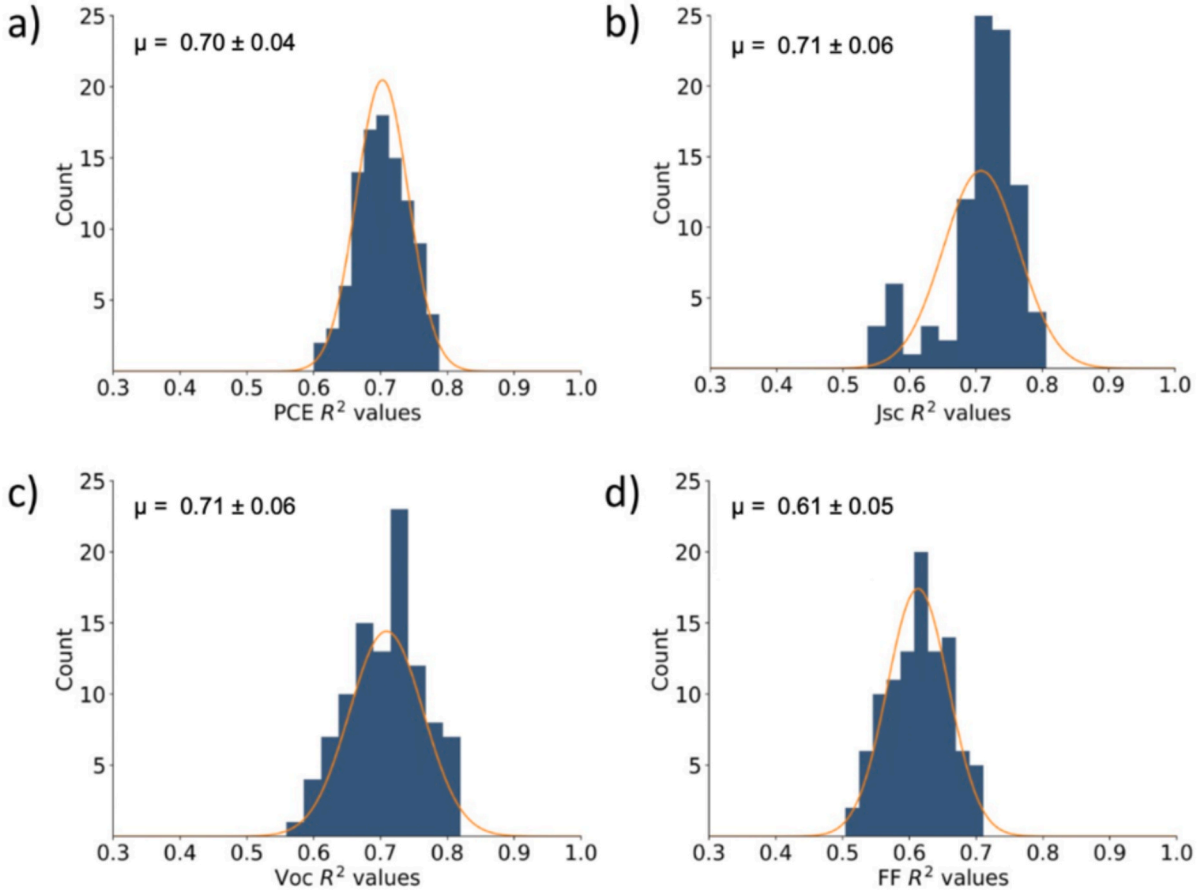


Fig. 4. Performance distribution of optimized XGBoost models, showing R^2 values for predictions of (a) *PCE*, (b) *Jsc*, (c) *Voc*, and (d) *FF*, where μ denotes the mean, and the value after ± represents the standard deviation. The mean and standard deviation are calculated directly from the R^2 values, while the orange curve is a Gaussian distribution plotted using those parameters. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

suggesting stable and accurate modeling across multiple runs. In contrast, *FF* (Fig. 4(d)) exhibits slightly lower predictive accuracy, indicating a more complex or noisier relationship with the input features. This approach of visualizing performance distributions offers a transparent assessment of model generalizability [55].

Based on the XGBoost results, the model performance for *FF* was lower than that for the other photovoltaic parameters (*PCE*, *Jsc*, and *Voc*). This discrepancy likely stems from the fact that, although both device-level and spacer-cation descriptors were included, *FF* is inherently more sensitive to device- and process-dependent factors such as contact resistance and interfacial inhomogeneity [56–58]. To improve *FF* prediction, future work may focus on curating datasets that incorporate additional descriptors, including descriptors for series/shunt resistance (e.g., contact work function mismatch, defect states), morphology-related descriptors (e.g., grain size, orientation), and fabrication variables (e.g., annealing conditions, solvent choice).

We compared the XGBoost model with Random Forest (RF) [59] and Partial Least Squares Regression (PLS) [60]. RF performed comparably to XGBoost, while PLS showed slightly lower but still reasonable predictive performance (Table S2). Conceptually, these three models represent distinct algorithmic families: XGBoost is a boosting-based ensemble that sequentially corrects errors of weak learners, RF is a bagging-based ensemble that aggregates multiple decision trees trained on bootstrapped samples, and PLS is a linear regression method that projects predictors into a lower-dimensional latent space to capture maximum covariance with the target. Although RF yielded similar results to XGBoost with the current dataset, we anticipate that XGBoost may demonstrate superior performance once a larger and more balanced dataset is available, given its boosting-based mechanism that can capture subtle feature–target relationships more effectively than RF.

In our dataset, the main source of compositional imbalance arises from the unequal representation of spacer cations (e.g., BA- and PEA-based systems appear much more frequently than many others). To quantitatively assess the impact of data imbalance and sparsity on model predictions, we further applied weighted and Group K-Fold schemes to XGBoost (Table S2). For the weighted scheme, frequently occurring spacer cations were assigned lower weights, while rarely observed spacer cations were assigned higher weights to balance the training distribution. The results showed that under random K-Fold, weighting had little effect, suggesting that data imbalance itself does not strongly influence model outcomes (R^2 values varied by less than 0.02 across all targets). In contrast, Group K-Fold splitting, where all samples with the same spacer cation are held out from the training set when included in the test set, led to a marked decrease in performance compared to random K-Fold (reduced by roughly 35–60 % for *PCE*, *Jsc*, and *Voc*, and by over 85 % for *FF*, in both the unweighted and weighted models). This result indicates that the model struggles to generalize to unseen spacer cations, highlighting the importance of including a broader range of spacer chemistries in future datasets.

2.5. XGBoost regressor: Feature importance

To interpret model predictions and identify key performance-driving descriptors, we applied three complementary feature attribution methods to the optimized XGBoost model: permutation importance, SHAP (SHapley Additive Explanations), and tree-based feature importance. Among these, we adopted permutation importance as the primary tool, as it directly quantifies each feature's contribution to predictive performance by measuring the reduction in R2 when the feature is randomly shuffled (Fig. 5). *PCE*, *Jsc*, *Voc*, and *FF* are weighted equally. While the derived feature importance rankings are interpretable within the present datasets, their applicability to unseen and more heterogeneous data remains uncertain.

Permutation importance analysis revealed that intrinsic chemical descriptors, most notably the large organic-to-metal ratio and small organic, accounting for 19.5 % and 13.2 % of the total importance, respectively, played a dominant role in determining 2D perovskite solar cell performance, significantly exceeding device-related features such as stack architecture and deposition methods. These trends were further supported by tree-based feature importance results (Fig. S11), though the robustness of tree-based importance is compromised by high cardinality of the large organic category [61]. The small organic category and inorganic presence also emerged as influential contributors.

To complement these results, we conducted a focused SHAP analysis on the large organic-to-metal ratio to further investigate its sample-level influence and directional impact (Fig. S12). The SHAP distribution reveals that lower spacer-to-metal ratios, although only indirectly indicative of the inorganic layer number and thus dimensionality, are correlated with higher predicted *PCE*, consistent with earlier dataset observations. (Fig. 3(e)).

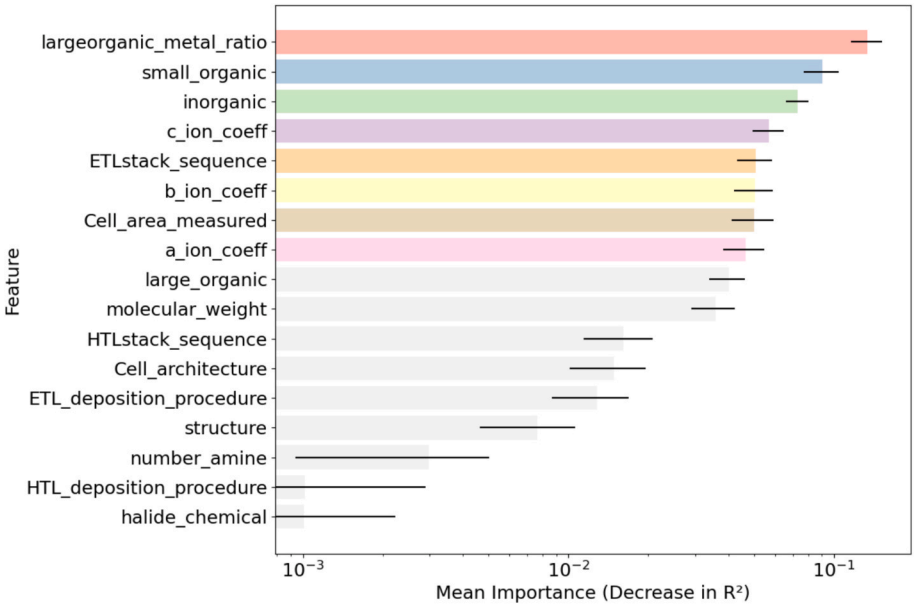


Fig. 5. Permutation importance of input features for the optimized XGBoost model, quantified by the mean decrease in R^2 across 4-fold cross-validation. Note the log scaling, which supports the top feature, large organic-to-metal ratio, as significantly more important than the trailing features.

To interpret why certain chemical descriptors emerged as key predictive factors in our machine learning model, we examined their physical and structural implications. In particular, the large organic-to-metal ratio, an inverse proxy for perovskite dimensionality, plays a critical role in determining optoelectronic properties. Lower dimensionality structures exhibit enhanced exciton binding energy and reduced carrier mobility due to strong quantum and dielectric confinement, whereas structures with increased dimensionality enable better charge transport and higher PCE [62,63]. This trend aligns well with the trend analysis shown in Fig. 3(e), as the large-organic-to-metal ratio decreases, *PCE* increases (Fig. S13). Overall, this correspondence highlights how chemically intuitive descriptors can offer mechanistic insights into the factors governing device efficiency.

Similarly, the small organic cation, which occupies the A-site of the perovskite lattice, has been shown to affect the band edge positions, lattice distortion, and crystallographic phase stability [40,64–66]. Specifically, variations in the composition of FA^+, MA^+, and Cs^+ shift the valence and conduction band positions, influence the formation of photoactive α-phases, and modulate the intermediate crystallization pathway. Cs incorporation has been shown to stabilize the α-phase of FA-based perovskites at room temperature, and triple cation systems have demonstrated improved crystallinity, band alignment, and long-term device performance under ambient conditions. Additionally, we evaluated deep neural networks (DNNs) for performance prediction, as presented in the Supporting Information (Figs. S14–S17). The DNN achieved comparable accuracy to XGBoost models, offering a complementary perspective on model choice.

The data analysis illuminates the bridge between perovskite chemical information and perovskite solar cell performance. However, it is important to consider that the available data for this study does not capture all possible nuances and the relationships between chemical and device information. For example, transitioning from linear to ring compositions likely changes the optimization of device fabrication steps. More extensive and detailed data will be beneficial for elucidating this relationship. To advance ML applications in materials science, we suggest expanding chemical composition coverage to include a wider variety of organic and inorganic cations and their combinations, as this is essential for addressing data imbalance, as shown in Fig. 2. Additionally, documenting detailed processing conditions in the database, including specific steps, environmental conditions, and synthesis methods, will allow ML models to better learn complex relationships. Lastly, gathering comprehensive property measurements, such as continuity and grain size of the perovskite films via scanning electron or optical microscopy, will make the dataset more suitable for ML, especially for structure–property–performance prediction. Addressing these points will make future datasets more robust for ML applications, allowing for extracting more valuable information (See more details in “Enhancing XGBoost Performance by Addressing Data Sparsity and Imbalance” in Supporting Information).

3. Conclusion

A major bottleneck in the development of perovskite solar cells is the overwhelming number of variables to consider – such as device architecture, the choice of materials for each layer, and, in the case of 2D perovskites, the selection of spacer cation molecules for optimal performance. Traditionally, researchers have approached this challenge through trial-and-error, varying one design dimension at a time (e.g., testing different device architectures while keeping other variables constant). While this method is time- and labor-intensive, its biggest limitation is that it only allows comparisons within a single design dimension. It does not reveal which dimension, such as material selection versus device architecture, has the greatest impact on performance. Machine learning-based data analysis enables not only comparison within a single design dimension, but also cross-dimensional insights, revealing which groups of variables have the most significant influence on device performance. Thus, we presented a data-driven investigation into the factors governing the 2D perovskite solar cell performance, using a dataset of 792 curated devices with comprehensive chemical and device descriptors. Through interpretable machine learning analysis, we identify that intrinsic chemical features, most notably the large organic-to-metal ratio and small organic cation, play a more significant role in determining power conversion efficiency than device-related parameters such as stack architecture or deposition method. The crucial role of chemical features is an important finding enabled by our data-driven analysis, which considers in ensemble the results of many previous studies that had focused on either intrinsic chemistry or device related features. The tendency of large organic-to-metal ratio and the small organic cation composition to emerge as critical determinants of perovskite optoelectronic behavior highlights the interplay between dimensionality-driven charge transport and composition-driven phase and band structure stability. The finding suggests that dimensional control through spacer chemistry remains a central lever for performance optimization in 2D perovskites.

Our results highlight the value of incorporating chemically meaningful descriptors in perovskite databases and demonstrate how such structured datasets can reveal key structure–property relationships not easily accessible through experimental trial-and-error. We also contribute a significant expansion of the dataset, adding new chemical diversity and descriptors. The data resource is openly available (See Data Availability section) for the research community to use and further extend, supporting continued progress in data-driven materials design.

4. Limitation

In this study, we utilized hand-curated datasets, including Materials Zone data and manually collected spacer cation data. However, this approach introduces potential sources of human error. One key limitation is the ambiguity in cation spacer acronyms within the Materials Zone dataset. For instance, “HA” can refer to both hexylamine and histamine, leading to potential mismatches when mapping spacer cation data to the Materials Zone dataset. The potential sources of error inherent to a hand-curated dataset may be mitigated in future work through the development of automated data-curation pipelines, for example, systematic parsing of IUPAC nomenclature. Ultimately, such errors could be substantially reduced if researchers in the field were to adopt the practice of consistently providing accurate IUPAC names when publishing manuscripts or depositing data. Additionally, there may be minor inaccuracies in the molecular weight calculations for spacer cations. To maintain consistency across all samples, we assumed a highly acidic environment and assigned the maximum number of protons to each nitrogen-containing organic spacer (i.e., full protonation of all nitrogen functional groups). In reality, the degree of protonation may vary—typically by up to 3 protons, resulting in a deviation in the molecular weight of approximately ± 3 g/mol—depending on the chemical environment, especially in the solid-state perovskite structure. Looking ahead, more rigorous assignments of protonation states could be achieved by integrating cheminformatics pipelines (e.g., SMILES/InChI strings with automated protonation state and pK_a-based rules). In the present study, because reported experimental conditions (e.g., pH, solvent environment) were often under-specified, we assumed full protonation of all nitrogen sites to maintain internal consistency in molecular weight assignments. However, if future reports systematically document the relevant experimental environments (e.g., solution pH, buffer composition, processing conditions), curated databases could leverage such information to assign protonation states more accurately and reduce systematic errors. Furthermore, the Materials Zone dataset includes several entries labeled as 2D perovskites that appear to be misclassified. These discrepancies could introduce unintended noise into the model, possibly influencing its accuracy. Addressing these inconsistencies through enhanced validation methods, improved data preprocessing, and incorporating supplementary data sources could

further strengthen future models. In this study, the dataset is primarily literature-based and therefore lacks direct experimental validation of model predictions. While this limitation does not diminish the value of the current analysis, future work integrating targeted synthesis and device testing of model-predicted candidates will be essential to fully establish and benchmark the predictive power of this approach.

5. Methods

5.1. Data preprocessing

5.1.1. Minimum-maximum (Min-Max) scaling

Data scaling represents an essential step in ML to increase scale invariance and reduce biases toward features with larger magnitudes [67]. Many standardized scaling methods exist in this domain, such as standard scaling, maximum absolute scaling, and Min-Max scaling [68]. Min-Max scaling involves performing a linear transformation on each feature of the original data to constrain the features into a normalized range of 0 to 1. The feature-wise transformation is given by:

$$x_{normalized} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

where $x_{normalized}$ and x are the scaled and original feature vectors, and x_{min} and x_{max} are the minimum and maximum values of the feature, respectively. The implementation of Min-Max scaling used throughout this work comes from the scikit-learn preprocessing package [69].

5.1.2. MultiLabelBinarizer (MLB) Encoding

Since ML algorithms involve performing mathematical operations to learn from the input data, additional preprocessing is required to convert categorical data into a numerical format that can be understood and operated upon. The process is often referred to as data encoding. Like with feature scaling, there are many existing encoding methods with varied levels of complexity and use cases, such as one-hot encoding, label encoding, and Latent Dirichlet Allocation [70]. Selecting an appropriate encoding strategy requires careful consideration of the data's structure and information content as well as the ML algorithm being used. The method used herein is MLB encoding, which transforms list-like multilabel features into a binary representation. The implementation of MultiLabelBinarizer encoding used in this work is from the scikit-learn preprocessing package [69].

6. Machine learning models

6.1. XGBoost regressor

XGBoost or eXtreme Gradient Boosted models are a class of decision-tree-based ensemble models with demonstrated excellent performance over a wide variety of both regression and classification tasks [71–73]. Being an ensemble method, XGBoost consists of many individual decision trees, each of which is trained and performs predictions. Then, the decisions are collectively pooled and used to give the overall model prediction. The utilized XGBoost Regressor model is contained in the open-source XGBoost package [74].

6.2. Grid search model optimization

In AI and ML applications, grid searching is a method employed during model selection to optimize a model's hyperparameters based on some objective function. For example, in the context of this work, the grid search seeks to find the hyperparameters of the XGBoost Regressor that maximize the overall R^2 on the predicted output features. Before performing a grid search, one must first define a list of model-specific hyperparameters and values for each over which to perform the optimization [74]. After defining the model and the parameter grid, various methods exist to find the global optimum within the parameter space [75]. Although it is more computationally expensive, we opted to perform model training and evaluation at every node in the parameter grid to develop a more robust and comprehensive understanding of the model's performance concerning the hyperparameters.

6.3. Random forest (RF)

RF regression is an ensemble method based on bootstrap-aggregated decision trees [59]. Each tree is trained on a randomly sampled subset of the data and features, and the final prediction is obtained by averaging outputs across trees.

6.4. Partial least square regression (PLS)

PLS is applied to model the relationship between input descriptors and target photovoltaic parameters [60]. PLS projects the feature space into a lower-dimensional latent space while maximizing covariance between predictors and response variables.

6.5. Cross validation

Model performance was evaluated using two cross-validation schemes. In random K-fold cross-validation, the dataset was randomly partitioned into K folds, with each fold used once as the test set while the remaining folds were used for training [76]. In Group K-fold cross-validation, samples belonging to the same composition group were restricted to either the training or test set, thereby preventing information leakage across splits [69].

CRediT authorship contribution statement

Migon Choi: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Aiden Sable:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Formal analysis. **Nicolas Ramos:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Formal analysis. **Richard Sheridan:** Writing – review & editing, Supervision. **Olivier Delaire:** Writing – review & editing. **Hai Li:** Writing – review & editing. **David B. Mitzi:** Writing – review & editing. **L.Catherine Brinson:** Writing – review & editing, Supervision.

Declaration of Competing Interest

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

Acknowledgements

M.C. and D.B.M. acknowledge funding from the National Science Foundation under award No. DMR-2323547. A.S. and O.D. acknowledge funding from the U.S. Department of Energy, Office of Science, Basic Energy Sciences, Materials Sciences and Engineering Division, under Award No. DE-SC0019978. N.R. and H.L. acknowledge funding from the Air Force Office of Scientific Research under Award No. FA9550-24-1-0322. M.C., A.S., and N.R. acknowledge funding from the National Science Foundation under grant DGE-2022040. R.S. and L.C.B. acknowledge funding from the National Science Foundation under award CMMI-204067. We also extend our gratitude to Shana Lee McAlexander for the encouragement and guidance throughout the development of our project.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.solener.2025.114112.

Data availability

The datasets generated and/or analyzed during this study are available at the following locations: GitHub: https://github.com/migonchoi/ML-for-2D-Perovskite-Solar-Cell-Performance. Duke Research Data Repository (RDR): https://doi.org/10.7924/r40g3xh1d. The original data were obtained from Materials Zone (https://www.materials.zone/), where the Perovskite Database [44] was originally deposited. The data have now been migrated to NOMAD [45]. We are currently collaborating with NOMAD to incorporate our additional descriptors and data to the existing dataset.

Code Availability

The underlying code for this study is available in GitHub repository and can be accessed via this link https://github.com/migonchoi/ML-for-2D-Perovskite-Solar-Cell-Performance.

References

[1] X.Y. Chin, D. Turkay, J.A. Steele, S. Tabean, S. Eswara, M. Mensi, P. Fiala, C. M. Wolff, A. Paracchino, K. Artuk, D. Jacobs, Q. Guesnay, F. Sahli, G. Andreatta, M. Boccard, Q. Jeangros, C. Ballif, Interface passivation for 31.25%-efficient perovskite/silicon tandem solar cells, Science 381 (6653) (2023) 59–63.

[2] J.J. Berry, J. van de Lagemaat, M.M. Al-Jassim, S. Kurtz, Y. Yan, K. Zhu, Perovskite photovoltaics: The path to a printable terawatt-scale technology, ACS Energy Lett. 2 (11) (2017) 2540–2544.

[3] J. Chen, H. Xiang, J. Wang, R. Wang, Y. Li, Q. Shan, X. Xu, Y. Dong, C. Wei, H. Zeng, Perovskite white light emitting diodes: Progress, challenges, and opportunities, ACS Nano 15 (11) (2021) 17150–17174.

[4] I.C. Smith, E.T. Hoke, D. Solis-Ibarra, M.D. McGehee, H.I. Karunadasa, A layered hybrid perovskite solar-cell absorber with enhanced moisture stability, Angew. Chem. Int. Ed. Engl. 53 (42) (2014) 11232–11235.

[5] G. Yang, Z. Ren, K. Liu, M. Qin, W. Deng, H. Zhang, H. Wang, J. Liang, F. Ye, Q. Liang, H. Yin, Y. Chen, Y. Zhuang, S. Li, B. Gao, J. Wang, T. Shi, X. Wang, X. Lu, H. Wu, J. Hou, D. Lei, S.K. So, Y. Yang, G. Fang, G. Li, Stable and low-photovoltage-loss perovskite solar cells by multifunctional passivation, Nat. Photonics 15 (9) (2021) 681–689.

[6] E.-B. Kim, M.S. Akhtar, H.-S. Shin, S. Ameen, M.K. Nazeeruddin, A review on two-dimensional (2D) and 2D-3D multidimensional perovskite solar cells: Perovskites structures, stability, and photovoltaic performances, J. Photochem Photobiol C: Photochem Rev 48 (2021) 100405.

[7] D.B. Mitzi, Templating and structural engineering in organic–inorganic perovskites, J. Chem. Soc., Dalton Trans. 1 (2001) 1–12.

[8] C. Liang, H. Gu, Y. Xia, Z. Wang, X. Liu, J. Xia, S. Zuo, Y. Hu, X. Gao, W. Hui, L. Chao, T. Niu, M. Fang, H. Lu, H. Dong, H. Yu, S. Chen, X. Ran, L. Song, B. Li, J. Zhang, Y. Peng, G. Shao, J. Wang, Y. Chen, G. Xing, W. Huang, Two-dimensional Ruddlesden–Popper layered perovskite solar cells based on phase-pure thin films, Nat. Energy 6 (1) (2020) 38–45.

[9] X. Yan, R. Cao, R. Zhang, H. Gao, Y. Xiao, Mixed-ligand chiral quasi-2D perovskites for standard and deep blue CP-LEDs, Adv. Funct. Mater. 34 (51) (2024) 2410012.

[10] L. Guo, Y. Qi, Z. Wu, X. Yang, G. Yan, R. Cong, L. Zhao, W. Zhang, S. Wang, C. Pan, Z. Yang, A self-powered UV photodetector with ultrahigh responsivity based on 2D perovskite ferroelectric films with mixed spacer cations, Adv. Mater. 35 (47) (2023) 2301705.

[11] W. Zhang, Z. Liu, L. Zhang, H. Wang, C. Jiang, X. Wu, C. Li, S. Yue, R. Yang, H. Zhang, J. Zhang, X. Liu, Y. Zhang, H. Zhou, Ultrastable and efficient slight-interlayer-displacement 2D Dion-Jacobson perovskite solar cells, Nat. Commun. 15 (1) (2024) 5709.

[12] P. Yan, W. Zhang, C. Wang, Z. Wu, M. Zhu, H. Li, C. Sheng, Acetamidinium based 2D alternating cation perovskite for efficient solar cells, J. Alloy. Compd. 977 (2024) 173298.

[13] R. Jacobs, J. Liu, H. Abernathy, D. Morgan, Machine learning design of perovskite catalytic properties, Adv. Energy Mater. 14 (12) (2024) 2303684.

[14] L. Zhang, F. Lu, G. Tao, M. Li, Z. Yang, A. Wang, W. Zhu, Y. Cao, Y. Jin, L. Zhu, W. Huang, J. Wang, Prediction of operational lifetime of perovskite light emitting diodes by machine learning, Adv. Intell. Syst. 6 (6) (2024) 2300772.

[15] S.V. Pandey, N. Parikh, A. Kalam, D. Prochowicz, S. Satapathi, S. Akin, M. M. Tavakoli, P. Yadav, A machine learning framework for predicting device performance in 2D metal halide perovskite photodetector, Sol. Energy 270 (2024) 112399.

[16] Z. Ma, H.A. Mahmoud, J. Liu, E.M. Awwad, Enhancing perovskite solar cell performance with a novel theoretical method and artificial neural networks, Sol. Energy 267 (2024) 112249.

[17] S. Sun, N.T.P. Hartono, Z.D. Ren, F. Oviedo, A.M. Buscemi, M. Layurova, D. X. Chen, T. Ogunfunmi, J. Thapa, S. Ramasamy, C. Settens, B.L. DeCost, A. G. Kusne, Z. Liu, S.I.P. Tian, I.M. Peters, J.-P. Correa-Baena, T. Buonassisi, Accelerated development of perovskite-inspired materials via high-throughput synthesis and machine-learning diagnosis, Joule 3 (6) (2019) 1437–1451.

[18] N. Meftahi, M.A. Surmiak, S.O. Fürer, K.J. Rietwyk, J. Lu, S.R. Raga, C. Evans, M. Michalska, H. Deng, D.P. McMeekin, T. Alan, D. Vak, A.S.R. Chesman, A. J. Christofferson, D.A. Winkler, U. Bach, S.P. Russo, Machine learning enhanced high-throughput fabrication and optimization of Quasi-2D ruddlesden–popper perovskite solar cells, Adv. Energy Mater. 13 (38) (2023) 2203859.

[19] J. Li, B. Pradhan, S. Gaur, J. Thomas, Predictions and strategies learned from machine learning to develop high-performing perovskite solar cells, Adv. Energy Mater. 9 (46) (2019) 1901891.

[20] Y. Hu, X. Hu, L. Zhang, T. Zheng, J. You, B. Jia, Y. Ma, X. Du, L. Zhang, J. Wang, B. Che, T. Chen, S. Liu, Machine-learning modeling for ultra-stable high-efficiency perovskite solar cells, Adv. Energy Mater. 12 (41) (2022) 2201463.

[21] Y. Liu, W. Yan, S. Han, H. Zhu, Y. Tu, I. Guan, X. Tan, How machine learning predicts and explains the performance of perovskite solar cells, Sol. RRL 6 (2022) 2101100.

[22] B.N.N. Alsulami, T.W. David, A. Essien, S. Kazim, S. Ahmad, T.J. Jacobsson, A. Feeney, J. Kettle, Application of large datasets to assess trends in the stability of perovskite photovoltaics through machine learning, J. Mater. Chem. A 12 (5) (2024) 3122–3132.

[23] M. Mammeri, H. Bencherif, L. Dehimi, A. Hajri, P. Sasikumar, A. Syed, H.A. Al-Shwaiman, Stability forecasting of perovskite solar cells utilizing various machine learning and deep learning techniques, J. Opt. (2024).

[24] W. Dunlap-Shohl, Y. Meng, P. Sunkari, D. Beck, M. Meila, H. Hillhouse, Physiochemical machine learning models predict operational lifetimes of CH3NH3PbI3 perovskite solar, Cells 16 (2022) 9730–9746.

[25] J. Kirman, A. Johnston, D.A. Kuntz, M. Askerka, Y. Gao, P. Todorović, D. Ma, G. G. Privé, E.H. Sargent, Machine-learning-accelerated perovskite crystallization, Matter 2 (4) (2020) 938–947.

[26] Y. Wu, C.-F. Wang, M.-G. Ju, Q. Jia, Q. Zhou, S. Lu, X. Gao, Y. Zhang, J. Wang, Universal machine learning aided synthesis approach of two-dimensional perovskites in a typical laboratory, Nat. Commun. 15 (1) (2024) 138.

[27] Y. Mai, J. Tang, H. Meng, X. Li, M. Liu, Z. Chen, P. Zhang, S. Li, Machine learning-based screening of two-dimensional perovskite organic spacers, Adv. Compos. Hybrid Mater. 7 (3) (2024) 104.

[28] R. Zhang, B. Motes, S. Tan, Y. Lu, M.-C. Shih, Y. Hao, K. Yang, S. Srinivasan, M.G. Bawendi, V. Bulovic, Predicting organic-inorganic halide perovskite photovoltaic performance from optical properties of constituent films through machine learning. arXiv preprint arXiv:2412.09638, (2024).

[29] W. Zhang, J. Zhang, C. Zhang, R. Dong, Y. Xu, Z. Yang, J. Zheng, D. Chen, Q. Liu, W. Yang, M.-H. Shang, Efficiently screening organic ligands by machine learning for stabilizing 2D/3D perovskites, J. Mater. Chem. A (2025).

[30] E. Mahal, S.C. Mandal, B. Pathak, Understanding the role of spacer cation in 2D layered halide perovskites to achieve stable perovskite solar cells, Mater. Adv. 3 (5) (2022) 2464–2474.

[31] A.J. Ramadan, W.H. Jeong, R.D.J. Oliver, J. Jiang, A. Dasgupta, Z. Yuan, J. Smith, J.E. Lee, S.G. Motti, O. Gough, Z. Li, L.M. Herz, M.B. Johnston, H. Choi, J. Even, C. Katan, B.R. Lee, H.J. Snaith, The role of the organic cation in developing efficient green perovskite LEDs based on quasi-2D perovskite heterostructures, Adv. Funct. Mater. 34 (14) (2024).

[32] Y. Shen, L. Luo, Y. Zhang, Y. Meng, Y. Yan, P. Xie, D. Li, Y. Ji, S. Hu, S. Yip, Z. Lai, T.D. Anthopoulos, J.C. Ho, High-performance nanogap photodetectors based on 2D halide perovskites with a novel spacer cation, Adv. Funct. Mater. 34 (41) (2024) 2403746.

[33] C.T. Triggs, R.D. Ross, W. Mihalyi-Koch, C.F.M. Clewett, K.M. Sanders, I.A. Guzei, S. Jin, Spacer cation design motifs for enhanced air stability in lead-free 2D tin halide perovskites, ACS Energy Lett. 9 (4) (2024) 1835–1843.

[34] A.H. Proppe, R. Quintero-Bermudez, H. Tan, O. Voznyy, S.O. Kelley, E.H. Sargent, Synthetic control over quantum well width distribution and carrier migration in low-dimensional perovskite photovoltaics, J. Am. Chem. Soc. 140 (8) (2018) 2890–2896.

[35] Z. Wang, Q. Wei, X. Liu, L. Liu, X. Tang, J. Guo, S. Ren, G. Xing, D. Zhao, Y. Zheng, Spacer cation tuning enables vertically oriented and graded quasi-2D perovskites for efficient solar cells, Adv. Funct. Mater. 31 (5) (2021) 2008404.

[36] T. Zhou, Q. Li, L. Zhou, Fluorinated quasi 2D perovskite solar cells with improved stability and over 19% efficiency, Adv. Energy Mater. 14 (22) (2024) 2400050.

[37] H. Zheng, G. Liu, Y. Wang, F. Chen, X. Dong, C. Wang, C. Wu, L. Yang, X. Ren, L. Yang, X. Pan, Z. Huang, Directional management self-additive spacer cations for stable 2D ruddlesden–popper perovskite solar cells with efficiency over 21%, Adv. Funct. Mater. 34 (32) (2024) 2401546.

[38] M. Shao, T. Bie, L. Yang, Y. Gao, X. Jin, F. He, N. Zheng, Y. Yu, X. Zhang, Over 21% efficiency stable 2D perovskite solar cells, Adv. Mater. 34 (1) (2022) 2107211.

[39] Y. Gao, Z. Song, Q. Fu, Y. Chen, L. Yang, Z. Hu, Y. Chen, Y. Liu, Controlled nucleation and oriented crystallization of methylammonium-free perovskites via in situ generated 2D perovskite phases, Adv. Mater. 36 (33) (2024) 2405921.

[40] S. Ramakrishnan, B. Chen, X. Zhang, Y. Xie, X. Tong, Y. Xu, A.N. Alphenaar, A. Ruthen, A.J. Babatunde, Y. Zhang, M. Cotlet, D.B. Mitzi, Q. Yu, Phase-stabilized 2D/3D hetero-bilayers via lattice matching for efficient and stable inverted solar cells, Joule 9 (6) (2025) 101954.

[41] Q. Cao, P. Li, W. Chen, S. Zang, L. Han, Y. Zhang, Y. Song, Two-dimensional perovskites: Impacts of species, components, and properties of organic spacers on solar cells, Nano Today 43 (2022) 101394.

[42] G. Lv, L. Li, D. Lu, Z. Xu, Y. Dong, Q. Li, Z. Chang, W.-J. Yin, Y. Liu, Multiple-noncovalent-interaction-stabilized layered dion–jacobson perovskite for efficient solar cells, Nano Lett. 21 (13) (2021) 5788–5797.

[43] M.A. Hope, T. Nakamura, P. Ahlawat, A. Mishra, M. Cordova, F. Jahanbakhshi, M. Mladenović, R. Runjhun, L. Merten, A. Hinderhofer, B.I. Carlsen, D.J. Kubicki, R. Gershoni-Poranne, T. Schneeberger, L.C. Carbone, Y. Liu, S.M. Zakeeruddin, J. Lewinski, A. Hagfeldt, F. Schreiber, U. Rothlisberger, M. Grätzel, J.V. Milić, L. Emsley, Nanoscale phase segregation in supramolecular π-Templating for hybrid perovskite photovoltaics from NMR crystallography, J. Am. Chem. Soc. 143 (3) (2021) 1529–1538.
[44] 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, E. Unger, An open-access database and analysis tool for perovskite solar cells based on the FAIR data principles, Nat. Energy 7 (1) (2021) 107–115.
[45] M. Scheidgen, L. Himanen, A.N. Ladines, D. Sikter, M. Nakhaee, Á. Fekete, T. Chang, A. Golparvar, J.A. Márquez, S. Brockhauser, S. Brückner, L. M. Ghiringhelli, F. Dietrich, D. Lehmberg, T. Denell, A. Albino, H. Näsström, S. Shabih, F. Dobener, M. Kühbach, R. Mozumder, J.F. Rudzinski, N. Daelman, J. M. Pizarro, M. Kuban, C. Salazar, P. Ondračka, H.-J. Bungartz, C. Draxl, NOMAD: A distributed web-based platform for managing materials science research data, J. Open Source Software 8 (90) (2023) 5388.
[46] Z. Zhang, H. Wang, T.J. Jacobsson, J. Luo, Big data driven perovskite solar cell stability analysis, Nat. Commun. 13 (1) (2022) 7639.
[47] M. Saliba, E. Unger, L. Etgar, J. Luo, T.J. Jacobsson, A systematic discrepancy between the short circuit current and the integrated quantum efficiency in perovskite solar cells, Nat. Commun. 14 (1) (2023) 5445.
[48] T. Lu, H. Li, M. Li, S. Wang, W. Lu, Predicting experimental formability of hybrid organic-inorganic perovskites via imbalanced learning, J. Phys. Chem. Lett. 13 (13) (2022) 3032–3038.
[49] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B.A. Shoemaker, P. A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang, E.E. Bolton, PubChem 2023 update, Nucleic Acids Res. 51 (D1) (2023) D1373–D1380.
[50] J. Zhao, M. Hou, Y. Wang, R. Wang, J. Zhang, H. Ren, G. Hou, Y. Ding, Y. Zhao, X. Zhang, Strategies for large-scale perovskite solar cells realization, Org. Electron. 122 (2023) 106892.
[51] Z. Lu, X. Xu, Y. Gao, Z. Wu, A. Li, Z. Zhan, Y. Qu, Y. Cai, X. Huang, J. Huang, Z. Zhang, T. Luo, L. Peng, P. Liu, T. Shi, W. Xie, The effect of spacer cations on optoelectronic properties of two-dimensional perovskite based on first-principles calculations, Surf. Interfaces 34 (2022) 102343.
[52] E. Shirzadi, A. Mahata, C. Roldán Carmona, F. De Angelis, P.J. Dyson, M. K. Nazeeruddin, Introduction of a bifunctional cation affords perovskite solar cells stable at temperatures exceeding 80 °C, ACS Energy Lett. 4 (12) (2019) 2989–2994.
[53] N. Zhou, B. Huang, M. Sun, Y. Zhang, L. Li, Y. Lun, X. Wang, J. Hong, Q. Chen, H. Zhou, The spacer cations interplay for efficient and stable layered 2D perovskite solar cells, Adv. Energy Mater. 10 (1) (2020) 1901566.
[54] L.N. Quan, M. Yuan, R. Comin, O. Voznyy, E.M. Beauregard, S. Hoogland, A. Buin, A.R. Kirmani, K. Zhao, A. Amassian, D.H. Kim, E.H. Sargent, Ligand-stabilized reduced-dimensionality perovskites, J. Am. Chem. Soc. 138 (8) (2016) 2649–2655.
[55] S. Varma, R. Simon, Bias in error estimation when using cross-validation for model selection, BMC Bioinf. 7 (1) (2006) 91.
[56] M. Leilaeioun, Z.C. Holman, Accuracy of expressions for the fill factor of a solar cell in terms of open-circuit voltage and ideality factor, J. Appl. Phys. 120 (12) (2016) 123111.
[57] J.-K. Tan, R.-Q. Png, C. Zhao, P.K.H. Ho, Ohmic transition at contacts key to maximizing fill factor and performance of organic solar cells, Nat. Commun. 9 (1) (2018) 3269.
[58] C. Wang, R.C.I. Mackenzie, U. Würfel, D. Neher, T. Kirchartz, C. Deibel, M. Saladina, Transport resistance dominates the fill factor losses in record organic solar cells, Adv. Energy Mater. (2405889) (2025).
[59] L. Breiman, Random Forests, Mach. Learn. 45 (1) (2001) 5–32.
[60] T.K. Dijkstra, Latent Variables and Indices: Herman Wold's Basic Design and Partial Least Squares, Springer, Berlin Heidelberg, 2010, pp. 23–46.
[61] Z. Zhang, T. Zhang, J. Li, Unbiased gradient boosting decision tree with unbiased feature importance. arXiv preprint arXiv:2305.10696, (2023).
[62] R.F. Kahwagi, S.T. Thornton, B. Smith, G.I. Koleilat, Dimensionality engineering of metal halide perovskites, Front. Optoelectron. 13 (3) (2020) 196–224.
[63] L. Etgar, The merit of perovskite's dimensionality; can this replace the 3D halide perovskite? Energ. Environ. Sci. 11 (2) (2018) 234–242.
[64] C. Hu, Y. Bai, S. Xiao, T. Zhang, X. Meng, W.K. Ng, Y. Yang, K.S. Wong, H. Chen, S. Yang, Tuning the A-site cation composition of FA perovskites for efficient and stable NiO-based p–i–n perovskite solar cells, J. Mater. Chem. A 5 (41) (2017) 21858–21865.
[65] W. Jian, R. Jia, H.-X. Zhang, F.-Q. Bai, Arranging strategies for A-site cations: impact on the stability and carrier migration of hybrid perovskite materials, Inorg. Chem. Front. 7 (8) (2020) 1741–1749.
[66] C. Yi, J. Luo, S. Meloni, A. Boziki, N. Ashari-Astani, C. Grätzel, S.M. Zakeeruddin, U. Röthlisberger, M. Grätzel, Entropic stabilization of mixed A-cation ABX3 metal halide perovskites for high performance perovskite solar cells, Energ. Environ. Sci. 9 (2) (2016) 656–662.
[67] L.B.V. De Amorim, G.D.C. Cavalcanti, R.M.O. Cruz, The choice of scaling technique matters for classification performance, Appl. Soft Comput. 133 (2023) 109924.
[68] M. Ahsan, M. Mahmud, P. Saha, K. Gupta, Z. Siddique, Effect of data scaling methods on machine learning algorithms and model performance, Technologies 9 (3) (2021) 52.
[69] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: Machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[70] J.T. Hancock, T.M. Khoshgoftaar, Survey on categorical data for neural networks, Journal of Big Data 7 (1) (2020) 28.
[71] P. Xu, X. Ji, M. Li, W. Lu, Small data machine learning in materials science, npj Comput. Mater. 9 (1) (2023) 42.
[72] Chen, T., Guestrin, C., XGBoost. ACM, pp. 785–794.
[73] M. Zou, W.-G. Jiang, Q. Qin, Y.-C. Liu, M.-L. Li, Optimized XGBoost model with small dataset for predicting relative density of Ti-6Al-4V Parts manufactured by selective laser melting, Materials 15 (2022) 5298.
[74] P.V. Balachandran, B. Kowalski, A. Sehirlioglu, T. Lookman, Experimental search for high-temperature ferroelectric perovskites guided by two-step machine learning, Nat. Commun. 9 (1) (2018) 1668.
[75] D. Belete, D h, m.,, Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results, Int. J. Comput. Appl. 44 (2021) 1–12.
[76] D. Krstajic, L.J. Buturovic, D.E. Leahy, S. Thomas, Cross-validation pitfalls when selecting and assessing regression and classification models, J. Cheminf. 6 (1) (2014) 10.