ESI Preprints **Not Peer-reviewed**

Computational Evaluation of Perovskite Materials through Machine Learning-Driven Prediction and Screening of Photovoltaic Cell Candidates

Muyi Chen
Palak Kundalia
Research in Science and Engineering (RISE),
Boston University, Boston, USA

Approved: 09 November 2026
Posted: 11 November 2026

Copyright 2026 Author(s)
Under Creative Commons CC-BY 4.0
OPEN ACCESS

Cite As:
Chen, M., & Kundalia, P. (2026). *Computational Evaluation of Perovskite Materials through Machine Learning-Driven Prediction and Screening of Photovoltaic Cell Candidates.* ESI Preprints. https://doi.org/10.19044/esipreprint.11.2026.p31

Abstract

While current solar technology mainly uses silicon, perovskites — crystalline compounds with a chemical formula of ABX3 — are emerging technologies due to their high energy conversion efficiency. However, experimentally discovering new perovskite compositions is both costly and time-intensive. This study introduces Random Forest models to predict photovoltaic-relevant properties of perovskites and identify promising candidates for solar cell applications. A dataset from the Open Quantum Materials Database was reconstructed due to inconsistencies and used to predict the band gap, stability energy, and formation energy of the compounds. Further, ablation studies and Shapley-value analyses indicated greater dependence on crystal-structure inputs. Finally, a Photovoltaic Suitability Score ranked candidate materials based on these three properties relevant to solar-cell performance, revealing that rare-earth vanadate compounds carry strong prospects for future investigations. Based on these findings, future efforts in engineering photovoltaic materials can consider crystallographic features alongside elemental properties. Broadly, this analysis of band gap, stability energy, and formation energy serves as a preliminary step to identify candidates for photovoltaic cell engineering and inverse design pipelines.

Keywords: OQMD, Random Forest, Photovoltaic Suitability Score, Band Gap, Stability Energy

Introduction

As renewable energy advocacy grows, scientists attempt to improve solar energy production. One tactic is to improve the sunlight conversion efficiency. Increasingly, researchers are observing perovskites, crystalline compounds in the chemical blueprint of ABX3, shown in Figure 1.

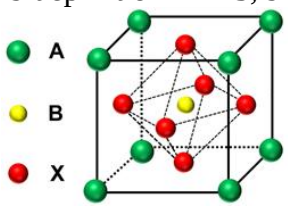


Figure 1: The general structure of a perovskite (Zhang et al., 2020)

Perovskites are efficient energy converters, highlighting them as promising candidates for solar cells. Existing research mainly references tandem solar cells, multi-layer constructions of perovskite and silicon subcells (U.S. Department of Energy, n.d.). In this technology, a top perovskite layer captures higher energy photons, while the silicon sublayer captures lower energy particles. As these photovoltaic systems develop, they continue to break conversion efficiency records. For example, the current tandem solar cell efficiency record is 34.6%, achieved by LONGi in June 2024 (Sheppard, n.d.). With ongoing development in tandem solar cell technology, there is foresight of its commercial production in the next decade.

While perovskites carry prospects for photovoltaic technology, discovery and testing is both costly and time consuming. Thousands of perovskite compounds can be computationally generated, but their energy capabilities and chemical stability remain unpredictable to an extent. To enhance ongoing research, fast-computing machine learning (ML) models can be designed to identify the strongest perovskite candidates for lab exploration based on the ideal ranges of energy performance parameters.

Some factors that determine material suitability for photovoltaic use include band gap, stability energy, and formation energy. While these certainly do not cover all photovoltaic application ideals, they remain relevant for efficiency and durability concerns.

During the process of solar conversion, a light particle can hit a solar cell with enough energy, and an electron can be knocked free. Band gap (EG) is the energy range where the electron has sufficient energy to jump from the valence band of the cell to the conduction band, creating a current. This process is illustrated in Figure 2.

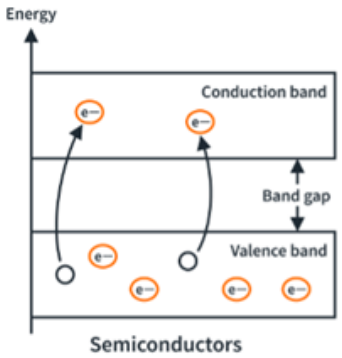


Figure 2: After electrons cross the band gap, they are free to move in the conduction band (Bharathi D., n.d.)

With band gaps that are too large or small, energy is either wasted as heat or cannot be absorbed. For this reason, most electronic components rely on semiconductors due to their controllable band gap. The ideal band gap for solar cells specifically is 1.34 eV, revealing how semiconductors such as silicon are prevalent in existing photovoltaic technologies (Rühle, 2016).

Stability energy (ES) explains a material's resistance to decomposition from heat, moisture, and other external effects. Since solar cells operate outdoors for commercial use, it is imperative that the material does not degrade under harsh conditions. Within the dataset, more favorable stability energy values are associated with greater thermodynamic stability, making ES an important factor in material screening (Qu & Li, 2023).

Formation energy (EF) is the energy change after a material is formed from its constituent elements and measures how favorable it is to synthesize the material. A negative formation energy indicates that formation is thermodynamically favorable relative to the constituent reference states. While formation energy does not imply significant information about the performance of the solar cell, it is a key insight into the chemical likelihood of formation.

For this study, solar cell candidacy will be determined primarily from these three output values. To evaluate promising perovskites, their respective outputs will be filtered through a photovoltaic scoring system (PSS) designed around the ideal ranges of Band Gap, Stability Energy, and Formation Energy.

Methods

Dataset

A publicly available dataset from the Open Quantum Materials Database (OQMD), containing Density Functional Theory (DFT) calculated thermodynamic and structural properties of numerous compounds, was used to train the models (Chenebuah, 2022). The dataset outlined theoretical

perovskite compositions and respective crystal structure and elemental composition properties. The dataset includes 16,323 perovskite entries with elemental, structural, and target-property fields. Sections of the dataset had been constructed via elemental computation, resulting in mean and standard deviation outputs of those values. Upon further investigation of the dataset, inconsistencies were discovered and consequently, the dataset was reconstructed to replace flawed values with recalculated inputs from A-, B-, and X-site properties from Ptable (Ptable, n.d.). As shown in Figure 3, one-hot encoding was utilized to convert the A-, B-, and X-site elemental identities into binary input features, allowing the models to associate specific elemental compositions with the target properties rather than treating each chemical formula as an independent label.

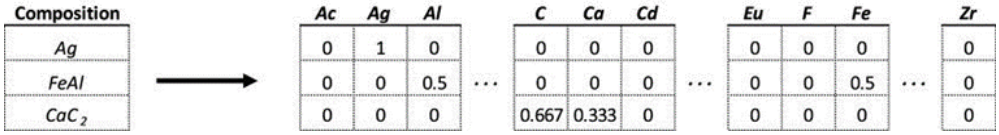


Figure 3: One-hot encoding process (converting compounds into binary) for chemical compounds (Peterson & Brgoch, 2021)

The revised dataset contains the theoretical perovskite compounds, but certain descriptors were removed as they did not affect the accuracy of the model's performance, as determined by a preliminary random forest model. The final features are depicted in Figure 4.

Dataset reconstruction and sample accounting

A direct row-by-row comparison of the publicly released oqmd_data.csv and the reconstructed oqmd_REVISEDdata.csv showed 16,323 entries in both files (3,501 distinct chemical formulas), in identical row order. The original file has 66 columns; the reconstructed version retains 41, removing 25 fields, including entry identifiers, crystallographic labels, and elemental descriptors not retained for modeling. The 41 retained fields comprise a formula identifier; the mean and standard deviation of 17 elemental properties (z, grp, row, av_ionrad, atom_rad, x, val, ea, ie, atom_mass, bp, rho, mp, k, heat_fus, heat_vap, and spec_heat); the gtf and of descriptors; unit-cell volume (vol); and the three target properties (Eg, Es, and Ef). Unit-cell volume and all three target properties are identical across the two files for every matched record.

The 34 elemental-property descriptors were reconstructed from A-, B-, and X-site reference values using ABX3 stoichiometric weights of 1:1:3. For an elemental property p available at all sites, the weighted mean is mu = (p_A + p_B + 3p_X)/5, and the weighted population standard deviation is sigma = sqrt{[(p_A - mu)^2 + (p_B - mu)^2 + 3(p_X - mu)^2]/5}. When a required elemental value was absent (except in the group-number case

described below), its reconstructed mean and standard deviation were left missing rather than imputed. For example, the source KAuBr3 record has z_mean = 25.8, whereas the reconstructed value is 40.6, consistent with (19 + 79 + 3 x 35)/5. Other observed changes include CaInBr3 val_mean (4.0 to 5.2) and KAuBr3 rho_mean (2.5352 to 5.9044).

The original and revised CSVs were compared using an absolute and relative tolerance of 10^-6, with missing versus numeric values counted as different. Across the 34 shared elemental-feature columns, 361,002 of 554,982 aligned values differed; examples include 89 z_mean values, 16,042 val_mean values, and 16,315 rho_mean values. The additional gtf and of fields differed in 13,357 of 32,646 aligned values, while vol, Eg, Es, and Ef remained unchanged. These counts document the extent of reconstruction, but differences alone cannot establish that each original value was incorrect; changes may also reflect differences in elemental reference values or descriptor definitions.

The 34 reconstructed elemental-feature columns were reproduced from the corrected customized periodic-table reference table within a numerical tolerance of 10^-6, with two documented handling rules. For periodic-table group (grp), unavailable group assignments were excluded and the remaining site weights renormalized (6,498 records). For average ionic radius (av_ionrad), the initial reference table assigned hydrogen a value of 0.00, while the 163 hydrogen-containing records in the reconstructed dataset consistently reflected 1.54. The hydrogen entry in the reference table was therefore corrected to 1.54 to reproduce the reconstructed means and standard deviations. This reconciliation establishes internal numerical consistency, not independent validation of the scientific basis for the 1.54 value; the research authors should verify its original source and applicability before submission.

Missing values were confined to six descriptor columns: ea_mean and ea_std (2,044 missing entries each), heat_fus_mean and heat_fus_std (339 each), and spec_heat_mean and spec_heat_std (983 each). Because some rows are missing more than one property, these correspond to 2,373 distinct incomplete material records, not the sum of the column-level missing counts. Complete-case filtering before model training excluded these 2,373 records, leaving 13,950 material entries for the machine-learning evaluation. All 16,323 records had nonmissing Eg, Es, and Ef targets. Therefore, the 16,323-record reconstructed screening dataset and the 13,950-record model-evaluation subset serve different purposes and should not be reported as the same sample count. In contrast, the original CSV contains no missing values in the retained numeric fields; the missingness in the reconstructed features corresponds to unavailable values in the supplied elemental-property reference table.

Features	Abbreviation	Unit
Atomic Number	*Z*	-
Group Number	*grp*	-
Row Number	*row*	-
Average Ionic Radius	*av_ionrad*	angstrom
Atomic Radius	*atom_rad*	angstrom
Valence Electrons	*val*	-
Atomic Mass	*atom_mass*	g/mol
Elemental Solid Density	*rho*	g/cc
Pauling Electronegativity	*x*	pauling
Electron Affinity	*ea*	eV
First Ionization Energy	*ie*	eV
Boiling Point	*bp*	kelvin
Melting Point	*mp*	kelvin
Thermal Conductivity	*k*	W/m.k
Heat of Fusion	*heat_fus*	eV
Heat of Vaporization	*haet_vap*	eV
Specific Heat	*spec_heat*	KJ/kg.k
Goldschmidt Tolerance Factor	*gtf*	-
Octahedral Factor	*Of*	-
Unit Cell Volume	*vol*	cubic-Angstrom
Stability Energy	*Es*	eV/atom
Formation Energy	*Ef*	eV/atom
Energy Bandgap	*Eg*	eV

Figure 4: Revised dataset features (E. Chenebuah & D. Chenebuah, 2023)

Modeling

Since the dataset included numerous input features used to map three target properties (EG, ES, EF), multiple regression models were evaluated to determine the most effective predictive model. Linear regression was tested but performed poorly due to its inability to capture the complex, non-linear relationships between the input descriptors and thermodynamic output features. Tree-based methods, including Extreme Gradient Boosting (XGBoost), Random Forest, and Extremely Randomized Trees (Extra Trees), were investigated because of their ability to recognize nonlinear relationships while remaining resistant to overfitting. Simultaneously, crystal structure input was strategically removed to provide an insight into feature importance. Additionally, some models were designed to have individual outputs (EG, ES, and EF were output through three separate models) while others had one multi-output prediction. Each model's performance was evaluated using the coefficient of determination (R2), mean absolute error (MAE), and root mean squared error (RMSE), with the highest-performing model selected for further analysis.

The Random Forest (RF), XGBoost, and Extra Trees models were trained using a 85-15 train-test split. The Random Forest model constructed multiple decision trees using random subsets of the training data and while the trees would independently predict the desired target values, the final prediction was outputted as an average of all the trees. XGBoost took the simple RF slightly further; trees were designed to correct prediction inaccuracies of previous trees using gradient optimization. Finally, Extra Trees were tested which introduced additional randomness during the tree construction.

Complex neural network architectures (such as CNNS) were not prioritized for this study due to the desired interpretability. With 13,950 complete-case samples, each with 202 encoded input features, tree-based models allowed for an efficient approach to capture nonlinear relationships between inputs and the target properties. Therefore, Random Forest, XGBoost, and Extra Trees were compared for their nonlinear modeling capabilities while enabling feature importance and SHAP interpretations.

Results

Figure 5 portrays how RF and XGBoost performed similarly, but a multi-output model was prioritized to satisfy the goal of a photovoltaic scoring system curated from all the outputs. Ultimately, the multi-output RF model without crystal inputs, but with volume, was selected, which had similar ES and EF accuracy to the best-performing values but with a slight advantage in EG. The model's performance is shown in Table 1, with calculation methods in Figure 6.

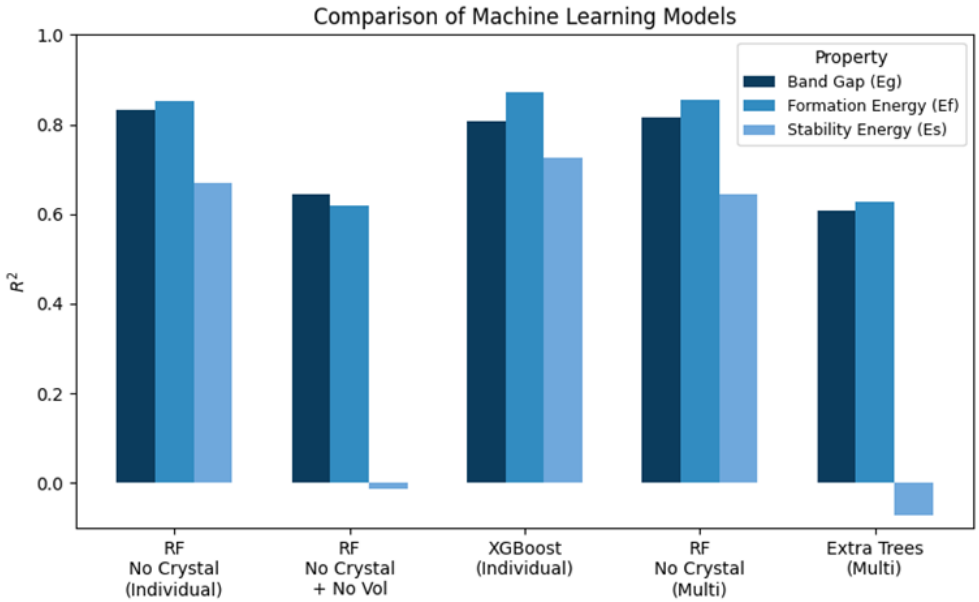


Figure 5: Performance evaluations of various ML model approaches

Table 1: Model performance of the RF multi-output model

Evaluation Metric	Value
EG R^2	0.815
EF R^2	0.855
ES R^2	0.643
MSE	0.185
RMSE	0.430
MAE	0.208
R^2	0.809

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

Figure 6: Formula for the coefficient of determination (Camp, n.d.)

The comparison of ML models also revealed that volume was an essential input for the model to have a stronger predictive accuracy. Because of the inherent cubic lattice-nested octahedral layered structure, perovskites' thermodynamic capabilities are highly dependent on each crystal's volume. Optimal volume values can stabilize the crystal structure, which in turn boosts the stability energy of the material as well (Mortazavi et al., 2018). When crystal structure features and unit cell volume were removed, the R2 of ES significantly dropped to a value below zero, indicating the model performed worse than a baseline predicting the mean target value for each sample. For Extra Trees, it is likely that the model's randomness reduced the predictive accuracy for ES.

Model validation

For reproducibility, the supplied reconstructed CSV contained 16,323 material records. Complete-case filtering excluded 2,373 records with missing feature or target values, leaving 13,950 samples corresponding to 3,002 distinct chemical formulas for model evaluation. Element identities at the A, B, and X sites were one-hot encoded, giving 202 model input features. The previously reported Random Forest results were reproduced with an 85/15 random train-test split (random_state = 42), 250 trees, and maximum tree depth of 25 (R^2 = 0.815, 0.855, and 0.643 for Eg, Ef, and Es, respectively).

The random split permitted formulas to occur in both partitions: 1,958 of the 2,093 test records (93.5%) shared their chemical formula with at least one training record. To assess generalization to unseen compositions more stringently, an additional formula-grouped split was performed with GroupShuffleSplit (test_size = 0.15; random_state = 42). It assigned 11,946 records from 2,551 unique formulas to training and 2,004 records from 451 different formulas to the held-out test set; no chemical formula occurred in both sets.

Hyperparameters were selected using only the grouped training partition. A three-fold GroupKFold search compared n_estimators = 120 or 250 and max_depth = 15 or 25, with selection based on the mean of the three target R^2 scores across folds. The selected configuration (250 trees; maximum depth 25) achieved a mean cross-validated R^2 of 0.740 in this

search. It was then fitted to all grouped training records and evaluated once on the previously untouched grouped test set. No held-out test labels were used in hyperparameter selection.

Table 2*:* Supplementary validation results (Random Forest)

Property	Original random R^2	Grouped holdout R^2	Grouped holdout MAE	Grouped holdout RMSE	5-fold grouped CV R^2 (mean ± SD)
Eg	0.815	0.751	0.400	0.715	0.680 ± 0.046
Ef	0.855	0.892	0.208	0.313	0.865 ± 0.016
Es	0.643	0.786	0.162	0.270	0.731 ± 0.024

As a complementary stability check, five-fold GroupKFold evaluation was performed within the grouped training partition using the selected Random Forest configuration; the mean ± standard deviation R^2 values are shown in the last column. The grouped holdout R^2 for band gap was lower than under random splitting, whereas the two energy targets were higher on this particular held-out set. Because the evaluation partitions differ, these results should not be interpreted as evidence that grouping improves model performance. Rather, they show that predictive accuracy remains substantial when chemical formulas are held out, while band-gap generalization is more challenging. These experiments evaluate compositions with the existing structural volume descriptor available and do not establish prediction from chemical formula alone.

Model evaluation and interpretation

Beyond evaluating the predictive accuracy of the model, the model's decision making process was extracted through ablation study and SHapley Additive exPlanations (SHAP) analysis. These examinations underlie the explainable AI (XAI) framework that measures the global feature importance and individual feature contributions used in the model's development (Bhalodia, n.d.).

The ablation study, shown in Figure 7, removed features categorized into descriptor groups to determine their feature contributions, a metric of how the input affects the model's prediction (Zyl et al., 2024). The dataset inputs were split into five groups: Crystal Structure (rho, vol), Electronic (x, val ie, ea, of), Atomic Size (av_ionrad, atom_rad), Periodic (grp, row), and Thermodynamic (bp, mp, heat_fus, heat_vap, spec_heat). The model's change in performance for each target output is highlighted in Table 3.

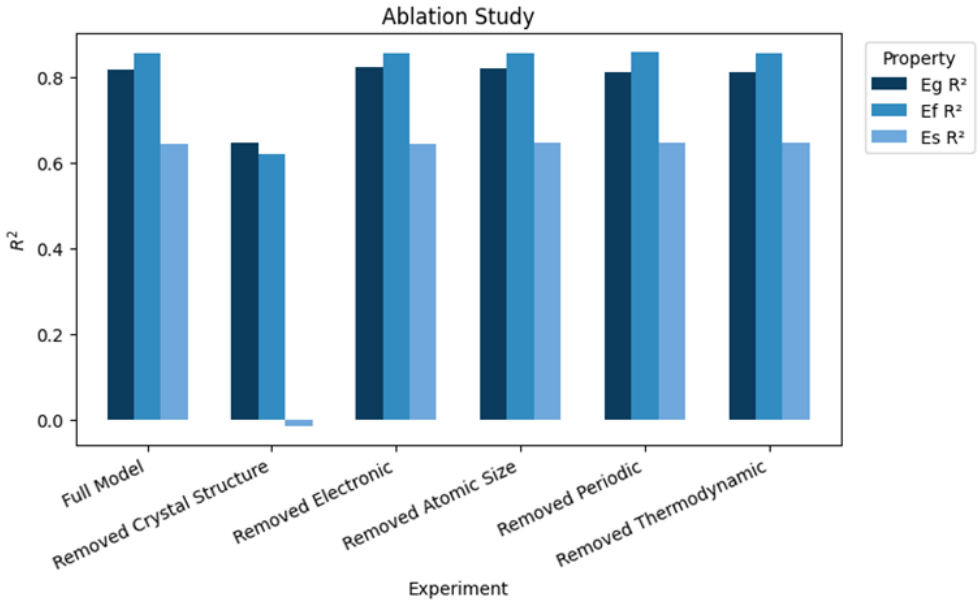


Figure 7: Ablation study outlining the change in model performance with removed feature groups

Table 3*:* Ablation study results

Experiment	EG	EF	Es
Full model	0.816	0.855	0.644
Removed crystal structure	0.646	0.619	−0.015
Removed electronic	0.823	0.853	0.644
Removed atomic size	0.821	0.855	0.647
Removed periodic	0.811	0.857	0.646
Removed thermodynamic	0.812	0.856	0.645

SHAP Analysis is an XAI method quantifying the contribution of each input feature to the model's prediction and providing an insight into the model's decision-making process. Unlike simpler feature importance methods, SHAP includes both the magnitude and direction of feature's contribution which allows for both individual prediction and global trend revelation. Because the selected model was a random forest, the implemented SHAP model was TreeExplainer (GeeksForGeeks, n.d.). The individual output results are summarized in Figure 8.

These beeswarm plots display the ranked importance of features in the model and their relationships with the predicted outcome. Each point is a perovskite compound, with horizontal position indicating to which degree the feature affects the predicted property. The color gradient indicates the relative numeric value of the feature. Together, these plots visualize the strongest influences on material properties and how variations in them affect the model's predictions (Chugani, n.d.). According to the beeswarm plot,

band gap is highly dependent on volume and periodic table (PTE) placement (period), revealing how period-dependent features like valence electrons and electronegativity, which relating to orbital distance and ionicity, affect band gap.

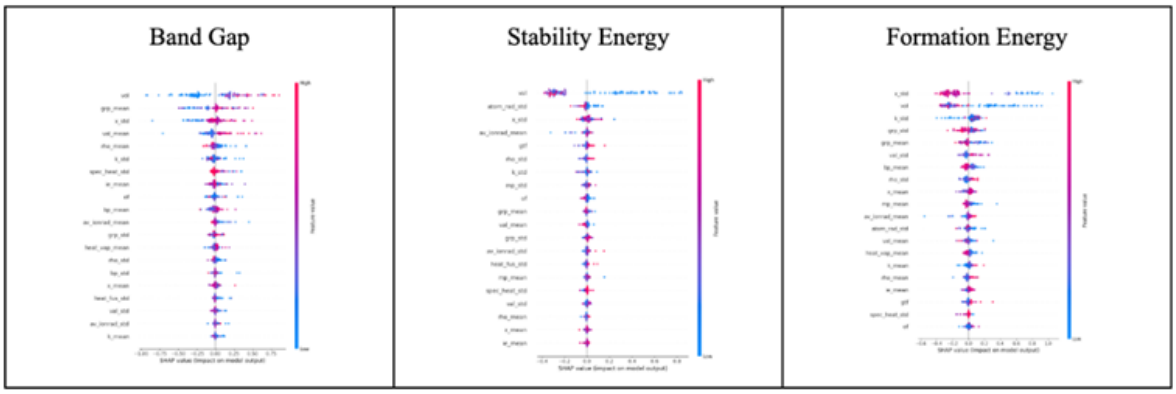


Figure 8: SHAP beeswarm plots showing feature contributions for Eg, Es, and Ef predictions

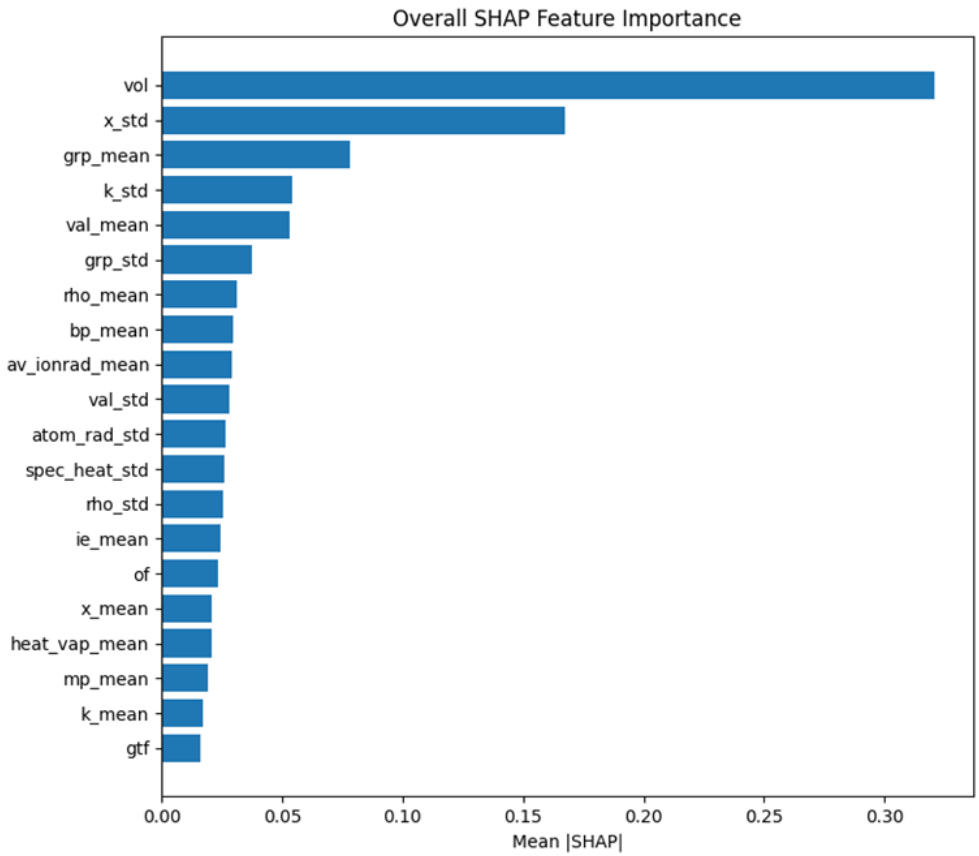


Figure 9: Highest ranking overall feature contributions based on mean SHAP values

While the feature importance was analyzed for each target output, they displayed similar trends in the most influential inputs. Figure 9 reiterates how crystal features had a significant impact on the model's summative predictive accuracy, while thermal features are not as significant to the model's performance.

Scoring system

The OQMD dataset contained thousands of theoretical perovskites with ranges of inputs for numerous electronic and energy uses. To identify the top perovskites for solar cell efficiency, a Photovoltaic Scoring System (PSS) was designed that outputted the highest scoring candidates in regards to the optimal ranges of EG, ES and EF.

Ideally, for photovoltaic applications, the Shockley-Queisser limit for single-junction solar cells (one semiconductor layer) is 1.34 eV under standard solar illumination (Zanatta, 2022). Band gap values with a smaller absolute distance to this value perform better and to reflect this behavior, the EG for PSS utilized a Gaussian function with greater penalizations as values diverged. Both stability energy and formation energy favor negative values for solar cells. However, cells with slightly positive ES values also perform well and thus, these values were not heavily penalized (Saliba et al., 2018). For formation energy, negative values with greater magnitudes were further rewarded (Guo et al., 2024). To identify the top perovskites for photovoltaic use, these three properties were combined into a scoring system with a 40-40-20 split. Band gap and stability energy are outputs detailing long-lasting energy conversion efficiency and thus they were more heavily weighted than formation energy, which explained chemical favorability in the material's creation. The top-scoring compounds are identified in Table 5.

PSS definition and weight sensitivity analysis

The PSS combines three property-specific subscores on a nominal 0-100 scale. The band-gap subscore is S_Eg = 100 exp[-(Eg - 1.34)^2/(2 x 0.20^2)]. The stability subscore is S_Es = 100 for Es <= 0; S_Es = 100 - 300 Es for 0 < Es <= 0.05; and S_Es = 85 exp[-18(Es - 0.05)] for Es > 0.05. The formation-energy subscore is S_Ef = min[100, max(0, -25 Ef)]. The combined score is PSS = w_Eg S_Eg + w_Es S_Es + w_Ef S_Ef, where the nonnegative weights sum to one. Band gap (40%) and stability energy (40%) were assigned equal emphasis as proxies for photovoltaic light-harvesting suitability and thermodynamic stability. Formation energy (20%) was weighted less because it adds auxiliary information about thermodynamic formation favorability, rather than directly measuring conversion efficiency or device lifetime. These percentages are scientifically motivated screening priorities, not experimentally optimized or uniquely correct constants.

For the sensitivity analysis, the same scoring functions and input records were retained while the three weights were varied. Scores were calculated for all 16,323 records before ranking; to avoid counting repeated crystal-structure records as distinct chemical candidates, each chemical formula was represented by its highest-scoring entry under each weighting scheme (3,501 distinct formulas). Six alternative 10-percentage-point combinations were compared with the 40/40/20 baseline. As shown in Table 4, the top five candidate formulas remained identical for four alternatives, while two alternatives retained four of the baseline top five. The highest-ranked candidate, HoVO3, was unchanged in every tested alternative.

Table 4: Alternative PSS weighting schemes and shortlist stability

Weights Eg/Es/Ef (%)	Top-5 overlap (of 5)	Exact Top-5 order?	Top-10 overlap (of 10)
40/40/20 (baseline)	5	Yes	10
30/50/20	5	Yes	9
50/30/20	5	Yes	8
40/30/30	5	Yes	9
30/40/30	5	No	9
40/50/10	4	No	9
50/40/10	4	No	7

A more systematic grid examined 91 weight combinations with w_Eg and w_Es each between 0.30 and 0.50 (increments of 0.02), w_Ef between 0.10 and 0.30, and w_Eg + w_Es + w_Ef = 1. Across these 91 scenarios, HoVO3 ranked first in all 91; the original set of five candidate formulas was fully retained in 82 scenarios (90.1%), with four of five retained in the remaining nine. The exact order of the first five was preserved in 69 of 91 scenarios (75.8%). These results support robustness of the leading candidate and broad stability of the top-five shortlist over the tested weight ranges, while showing that the precise ranking and fifth-place membership depend on the heuristic weighting scheme. Results do not establish an empirically optimal choice of weights or measured photovoltaic performance.

Table 5: Top-scoring perovskite compounds

Highest-ranking candidates	Photovoltaic Suitability Score (PSS)
$HoVO_3$	96.23
$PrVO_3$	95.88
$TmVO_3$	95.59
$LaVO_3$	95.30
$NdVO_3$	94.62

Discussion

The model predicted that unit cell volume has the highest feature importance. This is backed by prior research as lattice volume affects the orbital overlap of atoms; a smaller volume means smaller interatomic

distance, which in turn affects the band gap (Lee et al., 2008). The agreement between the ablation and SHAP analyses further supports the importance of crystallographic descriptors. Removing crystal structure descriptors, particularly volume, reduced predictive performance, while SHAP identified volume as one of the strongest individual contributors across the target properties. In addition, rare-earth vanadates (RVO3) have the highest PSS values among all perovskites within our dataset, identifying them as promising targets for future experimental research.

RF and XGBoost are the strongest performing models. Since RF was a multi-output model while XGBoost used three models to calculate every output value separately, RF was chosen for the purpose of photovoltaic screening. However, future studies optimizing individual outputs can emphasize XGBoost models, especially on larger datasets that are not limited within perovskites.

While most chemical properties can be generated from individual element descriptors, crystal structure features such as unit cell volume vary between samples and require deep quantum mechanics calculations, like Density Functional Theory (DFT), to compute. Because of an inability to generate structural features, the complete inverse design pipeline could not be executed without significant error. For future research, the model can be expanded with a volume predictor that, once fully trained, can allow for the inverse design of photovoltaic cells; in this manner, thousands of theoretical perovskites can be narrowed down for in-lab testing.

Conclusions

The study identified unit-cell volume as an important predictor of the examined material properties and demonstrated the use of a multi-output Random Forest for preliminary screening. The additional composition-grouped validation and PSS weight sensitivity analysis clarify the reliability and limitations of these results. Rare-earth vanadates remain leading screening candidates, although further computational and experimental evaluation is needed before their photovoltaic suitability can be established.

Acknowledgments

The authors would like to thank Dr. Indrajit Kalita and Dr. Eugene Pinsky from Boston University for their instruction and mentorship in data science. They would also like to thank Boston University and the RISE program for hosting and providing an engaging environment for scientific research.

Conflict of Interest: The authors reported no conflict of interest.

Data Availability: All data are included in the content of the paper.

Funding Statement: This work was supported by the Boston University Research in Science & Engineering (RISE) Program.

References:

1. Bhalodia, S. (n.d.). Explainable AI (XAI): How to interpret your machine learning model's decisions with LIME and SHAP. Medium. https://medium.com/@shradhdha.bhalodia/explainable-ai-xai-how-to-interpret-your-machine-learning-models-decisions-with-lime-and-shap-0e7dff696df
2. Bharathi D., U. (n.d.). What are energy bands, and how are they formed? Circuit Bread. https://www.circuitbread.com/ee-faq/what-are-energy-bands-and-how-are-they-formed
3. Camp, M. (n.d.). How to calculate R^2 in Tensorflow [Online forum post]. Stack Overflow. https://stackoverflow.com/questions/42351184/how-to-calculate-r2-in-tensorflow
4. Chenebuah, E. (2022, February 9). ML_abx3_dataset [Data set]. GitHub. https://github.com/chenebuah/ML_abx3_dataset/blob/main/oqmd_data.csv
5. Chenebuah, E., & Chenebuah, D. (2023). An inorganic ABX3 perovskite materials dataset for target property prediction and classification using machine learning [Preprint]. arXiv. https://arxiv.org/abs/2312.11335
6. Chugani, V. (n.d.). A gentle introduction to SHAP for tree-based models. Machine Learning Mastery.
7. GeeksForGeeks. (n.d.). SHAP: A comprehensive guide to SHapley Additive exPlanations.
8. Guo, Z., Yuan, M., Chen, G., Liu, F., Lu, R., & Yin, W. (2024). Understanding defects in perovskite solar cells through computation: Current knowledge and future challenge. Advanced Science, 11(20). https://doi.org/10.1002/advs.202305799
9. Lee, S., Woodford, W., & Randall, C. (2008). Crystal and defect chemistry influences on band gap trends in alkaline earth perovskites. Applied Physics Letters, 92(20). https://doi.org/10.1063/1.2936091
10. Mortazavi, M., Brandenburg, J., Maurer, R., & Tkatchenko, A. (2018). Structure and stability of molecular crystals with many-body dispersion-inclusive density functional tight binding. The Journal of Physical Chemistry Letters, 9(2), 399–405. https://doi.org/10.1021/acs.jpclett.7b03234

11. Peterson, G., & Brgoch, J. (2021). Materials discovery through machine learning formation energy. JPhys Energy, 3(2). https://doi.org/10.1088/2515-7655/abe425
12. Ptable. (n.d.). Properties. https://ptable.com
13. Qu, H., & Li, Y. (2023). Giant enhancement of exciton radiative lifetime by ferroelectric polarization: The case of monolayer TiOCl2. Physical Review B, 107(23), 235407. https://doi.org/10.1103/PhysRevB.107.235407
14. Rühle, S. (2016). Tabulated values of the Shockley–Queisser limit for single junction solar cells. Solar Energy, 130, 139–147. https://doi.org/10.1016/j.solener.2016.02.015
15. Saliba, M., Stolterfoht, M., Wolff, C., Neher, D., & Abate, A. (2018). Measuring aging stability of perovskite solar cells. Joule, 2(6). https://doi.org/10.1016/j.joule.2018.05.005
16. Sheppard, L. (n.d.). Perovskite solar cells: Progress continues in efficiency, durability, and commercialization. The American Ceramic Society. https://ceramics.org/ceramic-tech-today/perovskite-solar-cells-progress-2025/
17. U.S. Department of Energy. (n.d.). Perovskite solar cells. https://www.energy.gov/cmei/systems/perovskite-solar-cells
18. Zanatta, A. (2022). The Shockley–Queisser limit and the conversion efficiency of silicon-based solar cells. Results in Optics, 9, 100320. https://doi.org/10.1016/j.rio.2022.100320
19. Zhang, X., Dengbao, H., Chen, X., Yu, C., Chang, S., & Zhong, H. (2020). Effects of solvent coordination on perovskite crystallization. Acta Physico-Chimica Sinica. https://doi.org/10.3866/PKU.WHXB202008055
20. Zyl, C., Ye, X., & Naidoo, R. (2024). Harnessing eXplainable artificial intelligence for feature selection in time series energy forecasting: A comparative analysis of Grad-CAM and SHAP. Applied Energy, 353, 122079. https://doi.org/10.1016/j.apenergy.2023.122079