Supplementary Information (SI) for Journal of Materials Chemistry A.
This journal is © The Royal Society of Chemistry 2025

Accelerated Screening for High-Efficiency Inorganic Photovoltaic Materials: from Ensemble Voting Regression to Device Performance

Lu-Di Zhang,[1] and Hong-Jian Feng[1,] *

[1]School of Physics, Northwest University, Xi'an 710127, China.

S1 Machine Learning Details

Voronoi tessellation:

The Voronoi tessellation method depends solely on the crystal structure, dividing the crystal into regions associated with the nearest neighbors of central atoms. This approach provides a clear way to describe the local crystal structure of materials. Using the Magpie software package, the local environment characteristics of each atom are first calculated, and then the distribution of these characteristics is statistically analyzed to obtain the overall properties of the crystal.[1,2]

Model comparison:

In this study, the dataset is relatively small in scale with moderate feature dimensions, where the relationships between variables involve complex nonlinear interactions. In such scenarios, tree-based models are particularly well-suited due to their minimal requirements for data preprocessing, robustness to feature scaling, strong capability in modeling nonlinear relationships, and reduced tendency to overfit. The performance of three individual tree-based models—Gradient Boosting Regressor (GBR), Random Forest Regressor (RFR), and Decision Tree Regressor (DTR)—was evaluated in Figure S1 and Table S1, with all achieving R^2 scores exceeding 0.8 on the test set, demonstrating great predictive capabilities. However, analysis of the learning curves revealed potential concerns: both DTR and GBR achieved perfect training scores of 1.0 and exhibited a substantial gap between training and cross-validation performance, indicating a high degree of overfitting. In contrast, RFR also showed signs of overfitting, but to a comparatively lesser extent. These observations are closely related to the inherent learning mechanisms of the models: DTR is a single-tree model with a simple structure and high interpretability, but it relies heavily on training data and is prone to overfitting. GBR employs a boosting strategy that sequentially fits residuals, which makes it particularly susceptible to overfitting to noise in small datasets. In contrast, RFR integrates multiple decision trees via a bagging strategy, effectively reducing model variance and enhancing generalization performance.

Although GBR, RFR, and DTR all belong to the family of tree-based models, their learning mechanisms are significantly different and show strong complementarity. By combining these models, the Voting Regressor can fully leverage their complementary strengths and offset their individual limitations, thereby achieving a more optimal balance in the bias-variance tradeoff and enhancing overall predictive performance and generalization ability.

In addition, to evaluate the impact of different base models on the performance of the ensemble, we explored various combinations of algorithms, as summarized in the table below. When using only Linear Regression (LR) and Kernel Ridge Regression (KRR) as base models, the performance of the Voting Regressor is poor ($R^2 < 0.3$), indicating that linear methods struggle to effectively model the nonlinear structures in the data. However, as GBR, RFR, and DTR are added, the model's R^2 improves significantly. When the voting weights for GBR, RFR, and DTR are set to 90%, the R^2 increases to approximately 0.90, which is close to the performance of our selected VR (GBR + RFR + DTR) model ($R^2 \approx 0.92$). These data demonstrate the critical role of tree-based models in this study and validate the synergistic effect of the three models.

Based on the reviewer's suggestion, a Fully Connected Neural Network (FCNN) and a Stacking Regressor (SR) with GBR, RFR, and DTR as base estimators were constructed. The performance of both models is shown in the Figure S1. The constructed FCNN model consists of five fully connected hidden layers and one output layer, with 10% Dropout applied after each layer to prevent overfitting. The output layer uses the tanh activation function to produce normalized prediction values. The model exhibits nonlinear modeling capabilities, but its performance ($R^2 \approx 0.7$) is significantly lower than that of the tree-based VR model on this dataset, indicating limitations in the neural network's generalization ability and predictive accuracy for this task. Moreover, the FCNN model consumes more resources and operates more slowly (training time for the neural network is 10 minutes, while Stacking Regressor takes 40 seconds and Voting Regressor takes 9.5 seconds). Additionally, the complex structure of the neural network makes its decision-making process difficult to interpret using feature importance or local interpretability methods, limiting its application in material

mechanism studies.

The meta-learner of the Stacking model (using Ridge Regression) learns the optimal combination of the base model predictions (GBR, RFR, DTR) from the training data, with coefficient weights of 34%, 52%, and 14%, respectively. The R^2 of the model is 0.89, which is close to the VR model, but it slightly lags behind in terms of MAE and RMSE. The Stacking model theoretically has stronger nonlinear fitting capabilities by introducing a meta-learner in the secondary layer to re-learn the outputs of the base estimators. However, the VR model has a simpler structure, lower training costs, and is less sensitive to parameters, making it more practical, especially with limited sample sizes in datasets. Additionally, the Voting strategy is essentially a "denoising average" in the form of model fusion, which helps reduce the variance of individual model predictions and improves overall robustness. Therefore, in the condition of the current dataset, the VR model strikes a better balance between performance and complexity.

After thoroughly comparing the Voting, Stacking, and neural network, we ultimately selected the Voting Regressor composed of GBR, RFR, and DTR as the main model. This combination strikes an ideal balance between predictive performance, model interpretability, and computational efficiency, making it especially suitable for research in materials science, where the reliability of results and the interpretability of mechanisms are crucial. We do acknowledge the value of methods such as Stacking and neural networks; however, based on the characteristics of our data and the experimental results, the VR model emerges as a more suitable choice for the current task.

Voting Regressor:

Voting Regressor is an ensemble learning model that leverages the diversity and complementarity of multiple base regressors to enhance overall prediction accuracy and stability. This paper uses three different basic regressors: GBR, RFR and DTR. These basic regressors are based on different algorithmic and provide diverse prediction capabilities. GBR improves predictions by progressively constructing multiple weak learners and optimizing the loss function, making it suitable for handling complex data

relationships. RFR, on the other hand, builds multiple decision trees and uses their average or majority vote as the final prediction, reducing overfitting risk and providing good robustness to noisy data. DTR enhances randomness by randomly selecting features and samples to construct decision trees, thereby improving the model's generalization ability. Each basic regressor independently predicts the given data, and then VR performs a weighted average of these prediction results to improve the prediction ability of the overall model and produce the final prediction value. This ensemble method not only effectively combines the advantages of multiple models but also reduces the bias and variance that a single model might introduce, thereby improving overall prediction accuracy and stability.[3]

To further optimize the performance of the VR model, GridSearchCV was employed to perform systematic hyperparameter tuning for each base regressor, with the optimization objective being the average R^2 score under 10-fold cross-validation. The tuned parameters covered key influencing factors such as learning rate, maximum tree depth, and the number of estimators. The optimal hyperparameters identified were: GBR (learning_rate=0.15, n_estimators=30), RFR (max_depth=8, n_estimators=40), and DTR (max_depth=5, min_samples_split=4). The best_estimator_ from GridSearchCV was used to construct and train the final VR model.

Regarding the dataset, it was divided into a training set and an independent test set in an 8:2 ratio. All model training, parameter tuning, and cross-validation were conducted solely on the training set, while the test set was reserved for evaluating the final model's generalization ability. Additionally, to assess model performance across different data scales, 5-fold cross-validation was used to generate the learning curve. The resulting training and cross-validation scores reflect the model's bias-variance trade-off during training, and the final R^2 score obtained on the test set offers an objective measure of its predictive accuracy on real data.

S2 DFT Calculation Details

In this work, the first-principles calculations have been carried out by using the

QUANTUM-ESPRESSO package.[4,5] Generalized gradient approximation (GGA) in the form of Perdew-Burke-Ernzerhof (PBE) was used to describe the exchange-correlation interactions.[6,7] The periodic boundary condition was applied in all of the plane wave-based electronic structure calculations. The energy cutoffs for the wave functions and charge density are set to 45 Ry and 360 Ry, respectively. The energy convergence threshold of 10^{-6} Ry was used for all the structural optimization. There are totally 250 step ionic movement in the structural relaxation. The criterion of the atomic force during the structural relaxation was 10^{-3} a.u.

To evaluate the dynamic stability of the predicted materials, we performed AIMD simulations for 5000 fs. The time step is 1 fs. 2×2×2 supercell was used for the AIMD simulation. The constant pressure NPT ensemble is adopted, and the temperature was controlled at 300 K. [8]

Optical absorption coefficient was obtained by calculating the real and imaginary parts of the dielectric constants as follows:[9]

$$\epsilon = \epsilon_1 + i\epsilon_2$$

$$\alpha = \frac{\sqrt{2}w}{c}\left(\sqrt{{\epsilon_1}^2 + {\epsilon_2}^2} - \epsilon_1\right)^{1/2}$$

The spectroscopic limited maximum efficiency (SLME) η was calculated as formula:[10]

$$\eta = \frac{P_{max}}{P_{in}} = \frac{max\left\{(J_{sc} - J_0(e^{\frac{eV}{KT}} - 1))V\right\}_V}{\int_0^{\infty} EI_{sun}(E)dE}$$

The decomposition enthalpy (ΔH) was calculated by querying all potential subphases of all candidate materials and using the following formula:

$$\Delta H = E(compounds) - E(competing)$$

H < 0 indicates that the energy of the compound is less than the sum of the energies of its secondary phases, which means that the compound is thermodynamically stable; conversely, H > 0 means that the compound is thermodynamically unstable.

S3 Calculation of SLME potential for specific elements

Assuming there are x materials containing a specific element in the prediction dataset, with y of them having an SLME greater than 30%, the potential of this specific element can be calculated using the following formula: $Potential = \frac{y}{x} \times 100\%$.

S4 Experimental discussion of five candidates

- NaSrP: NaSrP can be synthesized by reacting metallic Na alloy with Sr and red phosphorus under high-temperature sealed conditions in a Nb tube. Dong *et al.* employed vacuum-sealed quartz tubes and carried out heat treatment at 600 °C, eventually obtaining irregular red crystals, indicating a clear synthetic route is experimentally achievable.[11]

- $Mg(InTe_2)_2$: Shahid *et al.* reacted Mg, In, and high-purity Te elements in carbon-coated fused silica tubes, and through heat treatment processes including holding at 1173 K, long-term annealing at 973 K, and programmed cooling, successfully obtained black crystals with a composition close to 1:2:4, exhibiting good elemental homogeneity and stability. This indicates that the Mg-based phosphorus/chalcogenide compound system has good experimental synthesizeability.[12]

- Ba_2PCl and Ba_8P_5Br: Although Ba_2PCl and Ba_8P_5Br currently lack explicit experimental synthesis literature, the reported Ba_2P_7X(X=Cl, Br, and I) provides a solid reference foundation for Ba–P–X system phosphorous halides.[13] Ba_2P_7X is synthesized through Ba, red phosphorus, and BaX_2 in vacuum-sealed silica ampoules, followed by long-term high-temperature annealing to obtain pure crystals. This synthesis process offers a clear pathway reference for the potential preparation and characterization of Ba_2PCl and Ba_8P_5Br. By adjusting the precursor ratio and reaction conditions, especially by reducing the phosphorus/barium ratio or modifying the reaction atmosphere, it is expected to produce structurally stable Ba_2PCl and Ba_8P_5Br materials.

- $Ba_3(InP_2)_2$: There are no experimental synthesis reports available, but its composition and structure are similar to several known indium phosphide materials, indicating good potential for synthesis. The In–P bond structure in this material resembles that of InP and InP_2-type structural units, and there is extensive synthesis

experience with InP materials.[14] Theoretical calculations show that $Ba_3(InP_2)_2$ possesses good thermodynamic stability and kinetic stability, further indicating its potential for stable existence under conventional experimental conditions. Regarding toxicity, Ba is considered a moderately toxic element, and care should be taken to avoid contact with its soluble salts during experiments. In and P elements are commonly used in inorganic synthesis, within vapors or dust potentially posing a respiratory hazard, but the overall toxicity is controllable and can be properly managed through standard laboratory safety measures (such as working in a fume hood and wearing gloves).

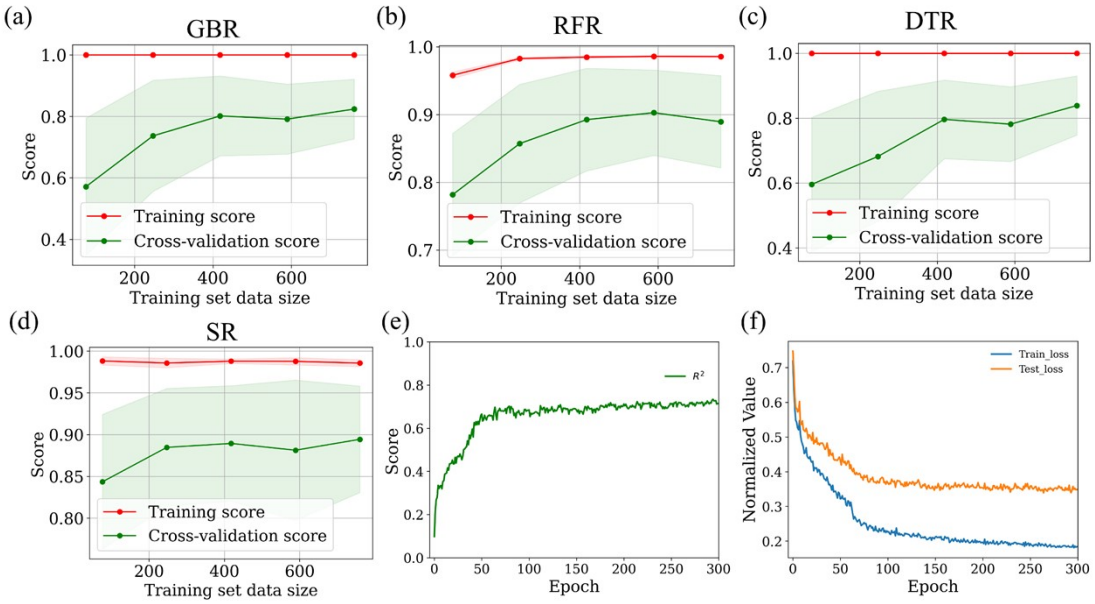

Figure S1. Learning curve of the (a) GBR model, (b) RFR model, (c) DTR model, and (d) SR model. (e) R^2 and (f) losses of FCNN model.

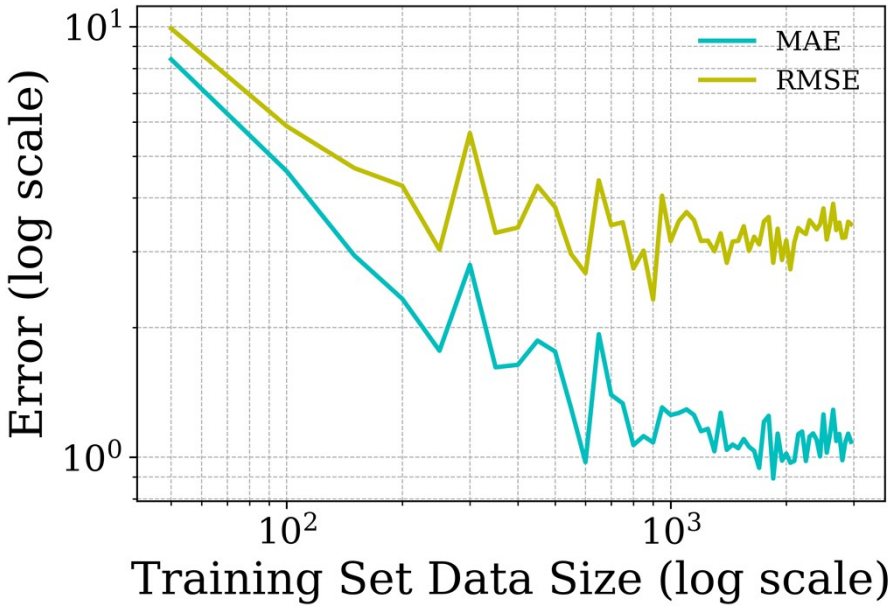

Figure S2. Log–log plot of the error scaling curve.

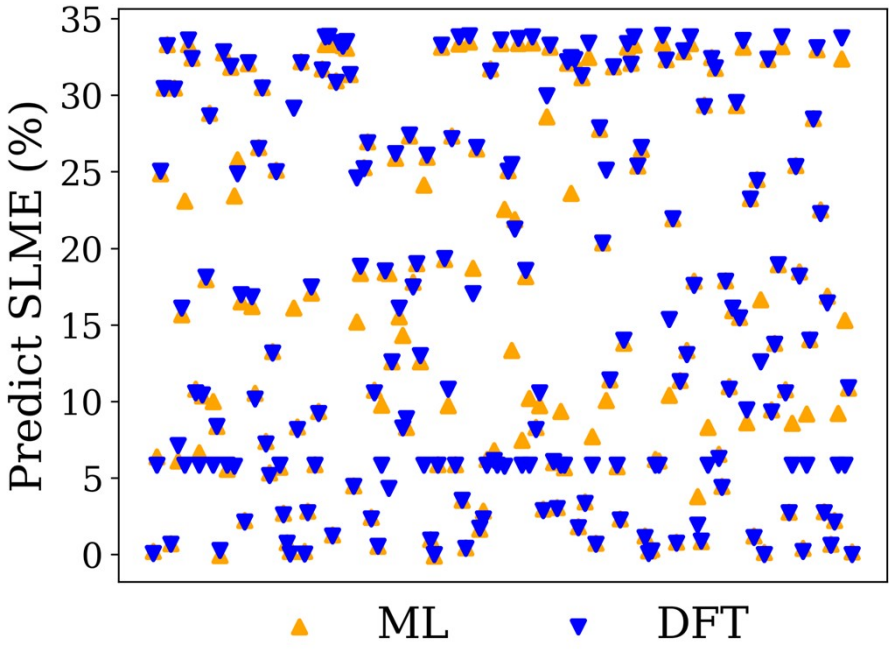

Figure S3. SLME distribution from DFT calculations and VR model predictions.

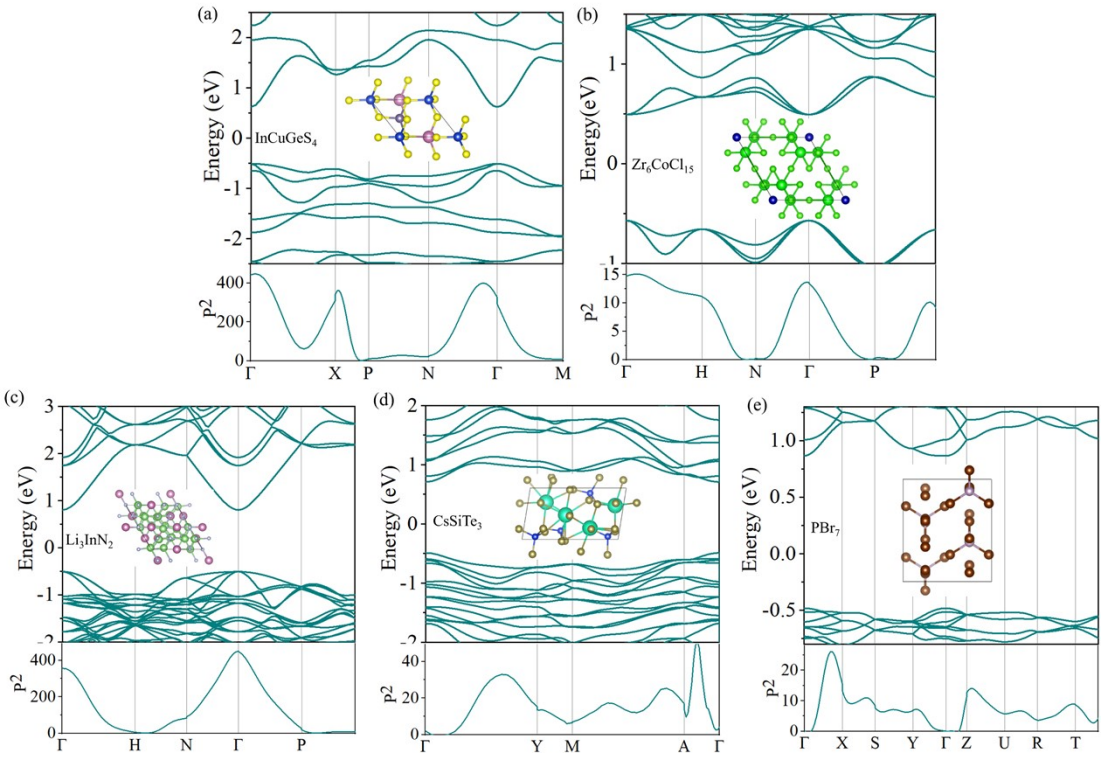

Figure S4. Band structures, transition probabilities at each high symmetry point, and the crystal structures for (a) $InCuGeS_4$, (b) Zr_6CoCl_{15}, (c) Li_3InN_2, (d) $CsSiTe_3$, and (e) PBr_7.

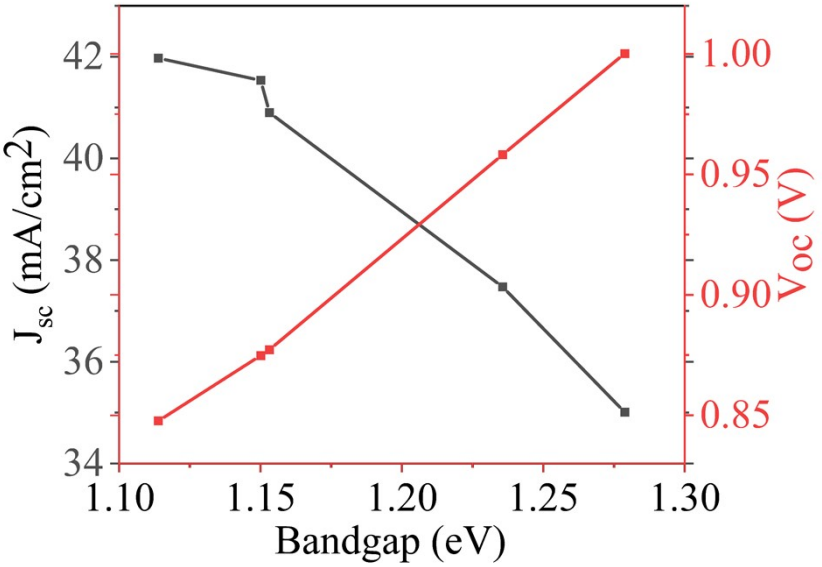

Figure S5. Relationship between J_{sc} and V_{oc} with band gap.

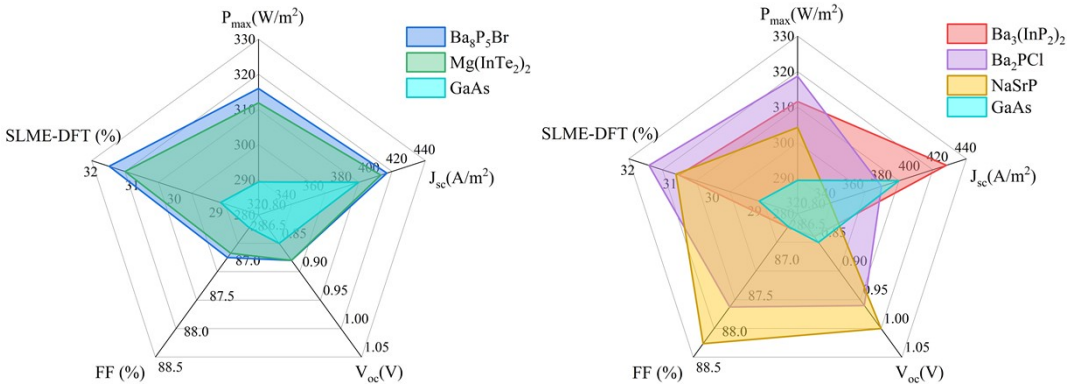

Figure S6. Comparison of device performance.

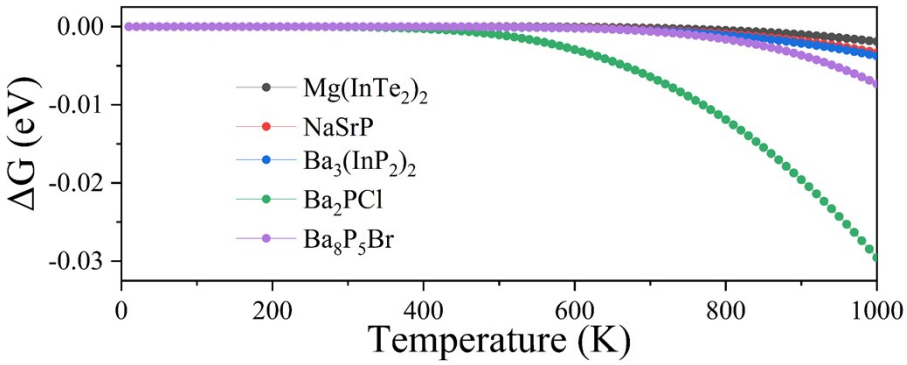

Figure S7. Changes in Gibbs free energy (ΔG) of the candidates at different temperatures.

Table S1. Performance evaluation of different models.

Model	R^2	MAE	RMSE	Weight
GBR	0.82	1.39	5.05	/
RFR	0.87	1.31	4.17	/
DTR	0.81	1.47	5.07	/
VR(LR+KRR)	0.27	8.78	10.10	50%, 50%
VR(LR+KRR+GBR)	0.63	6.22	7.22	33%, 33%,33%
VR(LR+KRR+GBR+RFR)	0.76	4.93	5.84	25%, 25%,25%,25%
VR(LR+KRR+GBR+RFR+DTR)	0.81	2.68	3.92	20%, 20%,20%,20%,20%
VR(LR+KRR+GBR+RFR+DTR)	0.90	1.98	3.80	5%, 5%,30%,30%,30%
VR(GBR+RFR+DTR)	0.92	1.04	3.35	33%, 33%,33%
SR(GBR+RFR+DTR)	0.89	1.28	3.73	34%, 52%, 14%

Table S2. Relative effective mass of electrons and holes for 10 candidate structures.

Formula	Li_3InN_2	$Mg(InTe_2)_2$	Ba_2PCl	$CsSiTe_3$	$InCuGeS_4$
MP id	1029562	1222182	27869	570957	1223929
Space group symbol	$Ia\bar{3}$	$I\bar{4}$	$R\bar{3}m$	Cc	$I\bar{4}$
m* (electron)	0.28	0.16	0.36	0.51	0.25
m* (hole)	2.33	0.64	0.62	0.76	1.16
Band gap (eV)	1.31	1.15	1.23	1.20	1.13

Formula	PBr_7	NaSrP	Zr_6CoCl_{15}	$Ba_3(InP_2)_2$	Ba_8P_5Br
MP id	647343	13275	28734	19913	34034
Space group symbol	Pnma	$P\bar{6}2m$	$Im\bar{3}m$	C2/c	$I\bar{4}$
m* (electron)	0.90	0.13	1.35	0.17	0.23
m* (hole)	3.76	0.12	0.82	0.36	0.18
Band gap (eV)	1.35	1.28	1.06	1.11	1.15

Table S3. Computed parameters for the selected candidate at thickness of 500 nm and temperature of 300 K.

Formula	P_max (W/m^2)	J_{sc} (A/m^2)	V_{oc} (V)	FF (%)	SLME-DFT (%)	SLME-ML (%)
$Ba_3(InP_2)_2$	311.59	425.75	0.84	86.69	30.81	32.36
Ba_8P_5Br	315.96	412.33	0.88	87.10	31.58	33.41
$Mg(InTe_2)_2$	311.89	407.96	0.88	87.04	31.20	32.65
Ba_2PCl	318.70	379.68	0.96	87.80	31.52	33.05
NaSrP	304.28	343.23	1.00	88.31	30.88	33.25
GaAs	289.38	392.09	0.85	86.68	28.91	29.38

Table S4. Representative values for typical photovoltaic materials

Formula	$J_{sc}(A/m^2)$	$V_{oc}(V)$	FF(%)	Efficiency(%)	Ref.
Si	426	0.7434	86.2	27.3	[15]
InP	311.5	0.939	82.6	24.2	[16]
CIGS	395.8	0.734	80.4	23.35	[17]
CIGSSe	395.5	0.6834	75.1	20.3	[18]
CZTSSe	379.0	0.5109	69.5	13.45	[19]
CZTS	217.7	0.7083	65.1	10	[20]
GaInP/GaInAsP//Si	127	3.309	86	36.1	[21]

Table S5. The decomposition paths and the corresponding decomposition enthalpies (ΔH).

Composition	Decomposition Path	ΔH (meV/atom)
NaSrP	(1) SrP + Na → NaSrP	-140
	(2) NaP + Sr → NaSrP	-545
	(3) $\frac{1}{2}SrP_2 + \frac{1}{2}Na_2Sr \rightarrow NaSrP$	-408
	(4) $\frac{1}{3}Sr_3P_2 + \frac{1}{3}Na_3P \rightarrow NaSrP$	-71
	(5) $\frac{1}{4}Sr_3P_4 + \frac{1}{4}Na_2Sr + \frac{1}{2}Na \rightarrow NaSrP$	-262
	(6) $\frac{1}{3}SrP_3 + \frac{1}{2}Na_2Sr + \frac{1}{6}Sr \rightarrow NaSrP$	-517
	(7) $\frac{1}{14}Sr_3P_{14} + \frac{1}{2}Na_2Sr + \frac{4}{14}Sr \rightarrow NaSrP$	-614
Ba_2PCl	(1) $\frac{1}{2}BaCl_2 + \frac{1}{2}Ba_3P_2 \rightarrow Ba_2PCl$	-101
	(2) $\frac{1}{3}Ba_3PCl + \frac{1}{3}Ba_3P_2 \rightarrow Ba_2PCl$	-86
$Ba_3(InP_2)_2$	(1) $2InP + Ba_3P_2 \rightarrow Ba_3(InP_2)_2$	-160
	(2) $InP_3 + Ba_3In + P \rightarrow Ba_3(InP_2)_2$	-685
	(3) $BaIn_2P_2 + 2BaP \rightarrow Ba_3(InP_2)_2$	-280
	(4) $BaP_3 + 2BaIn + P \rightarrow Ba_3(InP_2)_2$	-337
	(5) $BaP_2 + BaIn_2P_2 + Ba \rightarrow Ba_3(InP_2)_2$	-206
	(6) $Ba_3P_4 + 2In \rightarrow Ba_3(InP_2)_2$	-58
	(7) $BaIn_2 + 2BaP_2 \rightarrow Ba_3(InP_2)_2$	-97
$Mg(InTe_2)_2$	(1) $In_2Te_3 + MgTe \rightarrow Mg(InTe_2)_2$	-18
	(2) $MgIn + InTe + 3Te \rightarrow Mg(InTe_2)_2$	-367
	(3) $\frac{1}{2}Mg_2In + \frac{3}{2}InTe + \frac{1}{2}Te \rightarrow Mg(InTe_2)_2$	-326
	(4) $\frac{1}{3}Mg_3In + \frac{5}{3}InTe + \frac{7}{3}Te \rightarrow Mg(InTe_2)_2$	-313
Ba_8P_5Br	(1) $Ba_2PBr + Ba_5P_4 + Ba \rightarrow Ba_8P_5Br$	-40
	(2) $Ba_2PBr + 2BaP_2 + 4Ba \rightarrow Ba_8P_5Br$	-344
	(3) $Ba_2PBr + Ba_3P_4 + 3Ba \rightarrow Ba_8P_5Br$	-214

[1] Logan Ward, Ruoqian Liu, Amar Krishna, Vinay I Hegde, Ankit Agrawal, Alok Choudhary, and Chris Wolverton, Phys. Rev. B 96 (2), 024104 (2017).

[2] Logan Ward, Ankit Agrawal, Alok Choudhary, and Christopher Wolverton, npj Comput. Mater. 2 (1), 1 (2016).

[3] Christopher M Bishop and Nasser M Nasrabadi, Pattern recognition and machine learning. (Springer, 2006).

[4] Paolo Giannozzi, Stefano Baroni, Nicola Bonini, Matteo Calandra, Roberto Car, Carlo Cavazzoni, Davide Ceresoli, Guido L Chiarotti, Matteo Cococcioni, and Ismaila Dabo, J. Phys. Condens. Matter 21 (39), 395502 (2009).

[5] Paolo Giannozzi, Oliviero Andreussi, Thomas Brumme, Oana Bunau, M Buongiorno Nardelli, Matteo Calandra, Roberto Car, Carlo Cavazzoni, Davide Ceresoli, and Matteo Cococcioni, J. Phys. Condens. Matter 29 (46), 465901 (2017).

[6] John P Perdew, Kieron Burke, and Matthias Ernzerhof, Phys. Rev. Lett. 77 (18), 3865 (1996).

[7] John P Perdew, John A Chevary, Sy H Vosko, Koblar A Jackson, Mark R Pederson, Dig J Singh, and Carlos Fiolhais, Phys. Rev. B 46 (11), 6671 (1992).

[8] Weile Jia, Jiyun Fu, Zongyan Cao, Long Wang, Xuebin Chi, Weiguo Gao, and Lin-Wang Wang, J. Comput. Phys. 251, 102 (2013).

[9] Stefano Baroni, Stefano De Gironcoli, Andrea Dal Corso, and Paolo Giannozzi, Rev. Mod. Phys. 73 (2), 515 (2001).

[10] Liping Yu and Alex Zunger, Phys. Rev. Lett. 108 (6), 068701 (2012).

[11] Yongkwan Dong and Francis J DiSalvo, J. Solid State Chem. 180 (2), 432 (2007).

[12] Omair Shahid, Mohd Ishtiyak, Sambit SS Rout, KV Ramanujachary, Manish K Niranjan, and Jai Prakash, Solid State Sci. 152, 107552 (2024).

[13] Juli-Anna Dolyniuk and Kirill Kovnir, Crystals 3 (3), 431 (2013).

[14] Sudarsan Tamang, Christophe Lincheneau, Yannick Hermans, Sohee Jeong, and Peter Reiss, Chem. Mater. 28 (8), 2491 (2016).

[15] LONGi, “LONGi Sets New World-Record for Silicon Solar Cell Efficiency, Launching 2nd Generation Ultra-Efficient BC-Based Module,” 2024. [Online].

Available: https://www.longi.com/en/news/longi-hi-mo9-bc-world-record/

[16] Wanlass M. Systems and Methods for Advanced Ultra-High-Performance InP Solar Cells. US Patent 9,590,131 B2,2017.

[17] Motoshi Nakamura, Koji Yamaguchi, Yoshinori Kimoto, Yusuke Yasaki, Takuya Kato, and Hiroki Sugimoto, IEEE J. Photovoltaics 9 (6), 1863 (2019).

[18] R Diermann, PV Magazine International (2021).

[19] J. Z. Zhou, X. Xu, H. J. Wu, J. L. Wang, L. C. Lou, K. Yin, Y. C. Gong, J. J. Shi, Y. H. Luo, D. M. Li, H. Xin, and Q. B. Meng, Nat. Energy 8 (5), 526 (2023).

[20] Chang Yan, Jialiang Huang, Kaiwen Sun, Steve Johnston, Yuanfang Zhang, Heng Sun, Aobo Pu, Mingrui He, Fangyang Liu, and Katja Eder, Nat. Energy 3 (9), 764 (2018).

[21] Patrick Schygulla, Ralph Müller, Oliver Höhn, Michael Schachtner, David Chojniak, Andrea Cordaro, Stefan Tabernig, Benedikt Bläsi, Albert Polman, and Gerald Siefer, Prog. Photovolt.: Res. Appl. 33 (1), 100 (2025).