Prediction of Aqueous Stable Lead-Free Hybrid Halide Perovskites for Efficient Solar Water Splitting Using Machine Learning and Molecular Dynamics

Mahalaxmi Chandramohan, Madhana Gopal, Tumpa Sadhukhan, Athira Nambiar, Meenal Deo**

Supplementary Information

Note S1: Dataset preparation

The first dataset (D1) being a collection of triple-cation-mixed-halide perovskites with a general structural formula $FA_xMA_yCs_{(1-x-y)}Pb_aSn_{(1-a)}Br_iCl_jI_{(3-i-j)}$, was grouped by Yang et al., and this was originally extracted from 'The Perovskite Database Project' built by Jacobson et al., compiling the experimentally reported values.[1,2] Here, among the 610 entries in D1, 598 were unique, with bandgap values ranging from 1.17 - 3.16 eV, with the majority falling between 1.6 and 1.9 eV. In D1, 33 datapoints correspond to Cs-based systems. These entries were retained with the intention that the compositional diversity present in B- and X-site of Cs-systems can greatly contribute to the structure-property relationship learned by our XGB model, which has a stronger influence in band gap prediction compared to A-site features. However, the limitation to D1 is that the A-site is restricted to mixing of only 3 types of cations. To address these limitations and to introduce numerous variations in A-site substituents, a second database (D2) was included, generated by Kim et al., consisting of 1,346 HOIPs that were optimised through a combined approach of atomistic structure search and DFT.[3] In addition to D1, D2 encompasses 14 additional organic cations in A-site besides MA^+ and FA^+; Pb^{2+}, Sn^{2+} and Ge^{2+} as B-site cations, and all 4 halides, without involving any intra-combinations of A, B, and X sites unlike mixed perovskites. However, the results from D2 should be considered with the proviso that the asymmetrical geometry and bond rotation flexibility of the organic cations used in the A-site can lead to multiple HOIP perovskites with the same elemental composition, which in turn can exhibit different bandgaps.[4] Out of 1346 entries in D2, the bandgaps of such perovskite molecules were averaged, resulting in a final dataset of 192 data points ranging between 1.2 – 4.9 eV. Along with this, 67 manually gathered HOIP points with similar composition to D1 were added, but with greater variety in A-site mixing. Out of this 67, only 27 were unique and are uploaded in the github page (Band-gap-and-band-edge-prediction-of-HOIPs) along with their references.

This approach effectively allows to capture hidden trends in A-site substitution and may enhance the ability of machine learning models to predict the impact of changes in the target property arising from A-site replacement. Further data pre-processing was conducted to merge D1, D2 and extra datapoints into a consolidated table for subsequent ML training, with band gap as the target property.

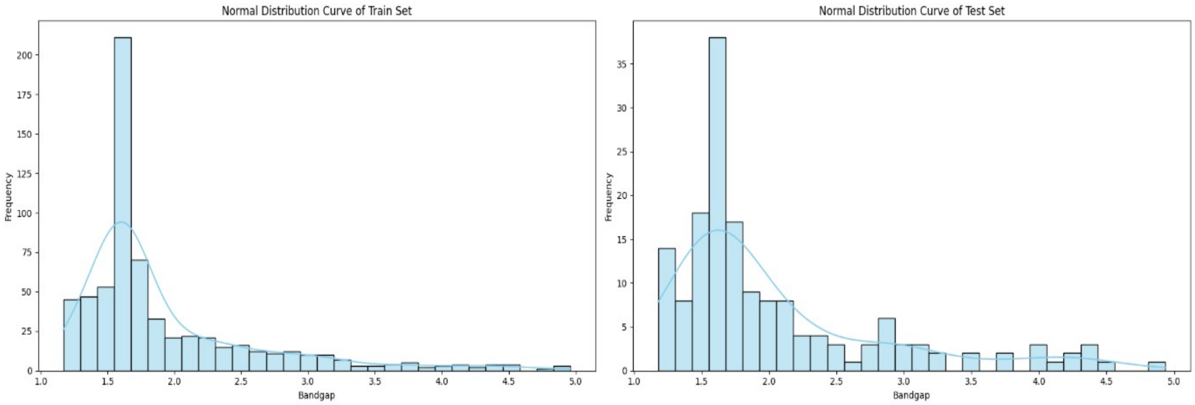

Fig. S1 Normal distribution curves of train and test set for band gap prediction

Table. S1 Initial pool of 82 features generated

S.No.	Features	Corresponding site
1	Ratio	A1, A2, A3, B1, B2, X1, X2 & X3-site
2	Average molecular weight (MW)	A-site
3	Average atomic weight (AW)	B & X-site
4	Atomic number (AN)	B & X-site
5	Average ionic radii (IR)	A, B & X-site
6	Average Mulliken's electronegativity (χ)	A, B & X-site
7	Average ionization energy (IE)	B & X-site
8	Average dipole polarizability (DP)	B & X-site
9	Average electron affinity (EA)	B & X-site
10	Average atomic radii (AR)	B & X-site
11	Average atomic density (AD)	B & X-site

12	Valence electrons (VE)	B1, B2, X1, X2 & X3-site
13	Oxidation state (OS)	A, B & X-site
14	Radii of (s+p) orbital ($r^{(s+p)}$)	B & X-site
15	No. of electrons in s, p, d and f orbitals	B1, B2, X1, X2 & X3-site
16	Feature crossovers from features (2-11) (‘+’, ‘-’, ‘/’)	Respective sites
17	Tolerance factor (Old – t; New - τ)	Perovskite
18	Target property – Band gap (E_g)	Perovskite

The above features are extracted from ‘Mendeleev – A Python package’ except for the calculated electronegativity values of A-site cations and literature sourced values of pseudopotential atomic orbital radii [5].

NOTE S2: Machine Learning methods and Evaluation Metrics

Random Forest Regression (RFR):

RFR is an ensemble learning method that combines the predictions of multiple decision trees to enhance the accuracy and robustness of the model. [6] In this approach, the predicted output $f_{rf}^{N}(X)$ for a given input X is computed as:

$$f_{rf}^{N}(X) = \frac{1}{N}\sum_{i=1}^{N} E_i(X)$$

where $E_i(X)$ represents the outputted energy predicted by the i^{th} tree for input X over the average ensemble of several trees, and N is the total number of trees in the forest. This aggregation of predictions effectively mitigates the risk of overfitting, which can occur in single decision trees, by leveraging the diverse structures and decision boundaries of multiple trees. As a result, RFR not only provides more reliable predictions but also offers improved generalization performance across various datasets. This methodology is particularly advantageous in complex regression tasks where the relationships between features and target variables are nonlinear and multifaceted.

Decision Tree Regression (DTR):

DTR is a predictive modelling approach that utilizes a tree-like structure to partition the input space into smaller subsets.[7] The target property to be predicted is derived by recursively splitting the dataset based on feature values until no further gains can be achieved or the maximum tree depth is reached. At the terminal nodes, the mean predicted output value E_t for a given number of samples n is calculated as:

$$E_t = \frac{1}{n}\sum_{i=1}^{n} E_i$$

where E_i represents the predicted outputs from the samples that fall within that node. This methodology allows DTR to model complex relationships between features and target variables, making it particularly effective in scenarios where the data exhibits nonlinear characteristics. The tree's interpretability further enhances its appeal in various applications, providing clear insights into the decision-making process.

Gradient Boosting Regression (GBR):

GBR is an advanced predictive modelling technique that enhances overall prediction accuracy by combining multiple weak models, typically decision trees.[8] This method iteratively refines its predictions by focusing on minimizing the errors of preceding models, resulting in a more accurate ensemble. The final prediction function $F_N(X)$ can be expressed as

$$F_N(X) = \sum_{n=1}^{N} f(X, w_n)$$

In this equation, N denotes the total number of trees in the ensemble, $f(X, w_n)$ represents the prediction from the n^{th} model, and w_n indicates the weight assigned to this model. By aggregating the outputs of these individual models, GBR effectively captures complex relationships within the data, making it particularly effective for regression tasks that involve nonlinear dependencies and high-dimensional feature spaces.

Extreme Gradient Boosting Regression (XGB):

XGB builds upon the principles of Gradient Boosting Regression (GBR) while introducing several key enhancements that improve performance and efficiency.[9] While it shares the same prediction function $F_N(X) = \sum_{n=1}^{N} f(X, w_n)$, XGB employs advanced optimization techniques, including second-order Taylor approximations, to refine the model updates. Additionally, it incorporates regularization terms to reduce overfitting and enhance generalization. Designed for computational efficiency, XGB also supports parallel processing and has robust mechanisms for handling missing values, making it particularly suitable for large and complex datasets.

Coefficient of Determination (R²): The coefficient of determination, R^2, measures the goodness of fit of the regression model by indicating how well the predicted values align with the actual values. It is calculated as:

$$R^2 = 1 - \frac{\sum (y_{true} - y_{pred})^2}{\sum (y_{true} - \bar{y}_{true})^2}$$

Where y_{true} represents the actual values, y_{pred} represents the predicted values, and $\bar{y}_{true}$ is the mean of the actual values. The closer the R^2 value is to 1, the better the model fits the data.

Root Mean Squared Error (RMSE): RMSE is the square root of MSE and provides a measure of the model's prediction error in the same units as the target variable. RMSE is often preferred because it combines both the magnitude and variance of errors, penalizing large deviations more than MAE. RMSE is defined as:

$$RMSE = \sqrt{\frac{1}{N} \sum (y_{true} - y_{pred})^2}$$

A lower RMSE indicates a better fit and less prediction error.

Pearson's Correlation (r): Pearson's correlation coefficient r measures the linear relationship between two variables, indicating the strength and direction of the relationship. It is defined as:

$$r = \frac{\sum(y_{true} - \bar{y}_{true})(y_{pred} - \bar{y}_{pred})}{\sqrt{\sum(y_{true} - \bar{y}_{true})^2 \sum(y_{pred} - \bar{y}_{pred})^2}}$$

Here, r ranges from -1 to 1, where:

- $r > 0$: Positive correlation.
- $r < 0$: Negative correlation.
- $r = 0$: No correlation. The larger the absolute value of r, the stronger the correlation.

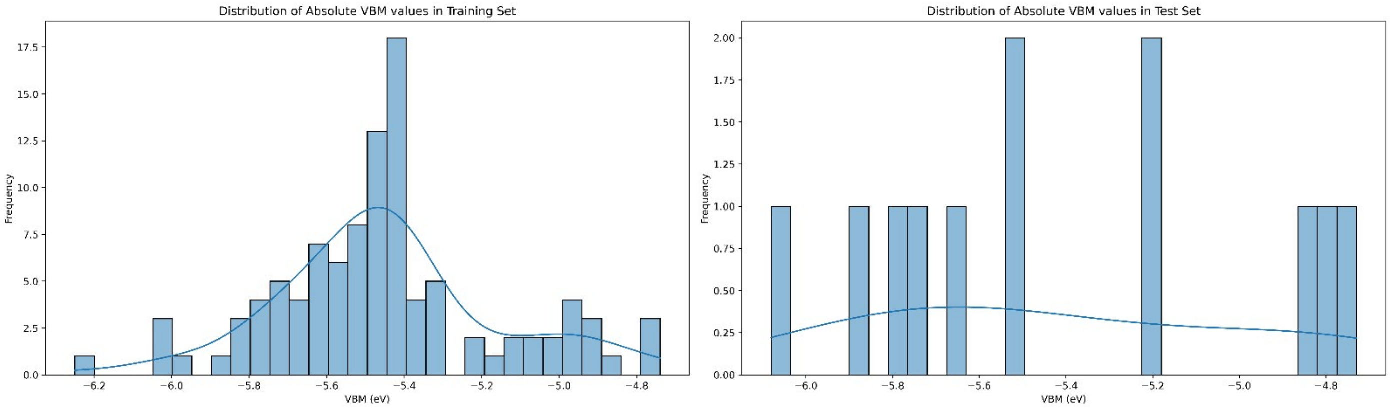

Fig. S2 Normal distribution curve of the train and test set for band edge (VBM) prediction

Table. S2 29 A-site cations and their ionic radii [10–12]

A-site cation	Chemical formulae	Ionic radii (Å)
Ammonium (Am)	$[NH^4]^+$	1.46
Methylammonium (MA)	$[CH_3NH_3]^+$	2.17
Dimethylammonium (DMA)	$[(CH_3)_2NH_3]^+$	2.72
Trimethylammonium (TMA)	$[(CH_3)_3NH_3]^+$	2.53
Tetramethylammonium (TTMA)	$[(CH_3)_4NH_3]^+$	2.92

Ethylammonium (EA)	$[C_2H_5NH_3]^+$	2.74
Propylammonium (PA)	$[C_3H_{10}N]^+$	2.52
Isopropylammonium (IPA)	$[C_3H_{10}N]^+$	2.58
Butylammonium (BA)	$[C_4H_{12}N]^+$	2.86
Hydroxylammonium (HA)	$[NH_3OH]^+$	2.16
Formamidinium (FA)	$[NH_2(CH)NH_2]^+$	2.53
Acetamidinium (Ac)	$[C_2H_7N_2]^+$	2.77
Hydrazinium (Hz)	$[NH_3NH_2]^+$	2.17
Guanidinium (Gua)	$[C(NH_2)_3]^+$	2.78
Imidazolium (Im)	$[C_3N_2H_5]^+$	2.58
Azetidinium (Az)	$[(CH_2)_3NH_2]^+$	2.50
Tropylium (Tr)	$[C_7H_7]^+$	3.33
Methylphosphonium (MP)	$[CH_3PH_3]^+$	2.49
Ethylenediammonium (EDA)	$[C_2H_{10}N]^+$	2.54
Tertiarybutylammonium (TBA)	$[C_4H_{12}N]^+$	2.62
Thiazolium (Th)	$[C_3H_4NS]^+$	3.20
Piperazinium (Pz)	$[C_4H_{11}N_2]^+$	3.22
Dabconium (DABCO)	$[C_6H_{13}N_2]^+$	3.39
Pyrollinium (Py)	$[NC_4H_8]^+$	2.72
Isobutylammonium (IBA)	$[C_4H_{12}N]^+$	3.60
Diethylammonium (DEA)	$[C_4H_{12}N]^+$	3.85
Phenylammonium (PhA)	$[C_6H_8N]^+$	3.88
Pyrollidinium (Pyd)	$[C_4H_6N]^+$	3.22
Protonated Protonated Formamide (Fm)	$[NH_3COH]^+$	1.90

The ionic radii of all the divalent B-site cations used in this study are extracted from Shannon's ionic radii database except for Bi^{2+} with a reported radius of 1.14 Å[13–15].

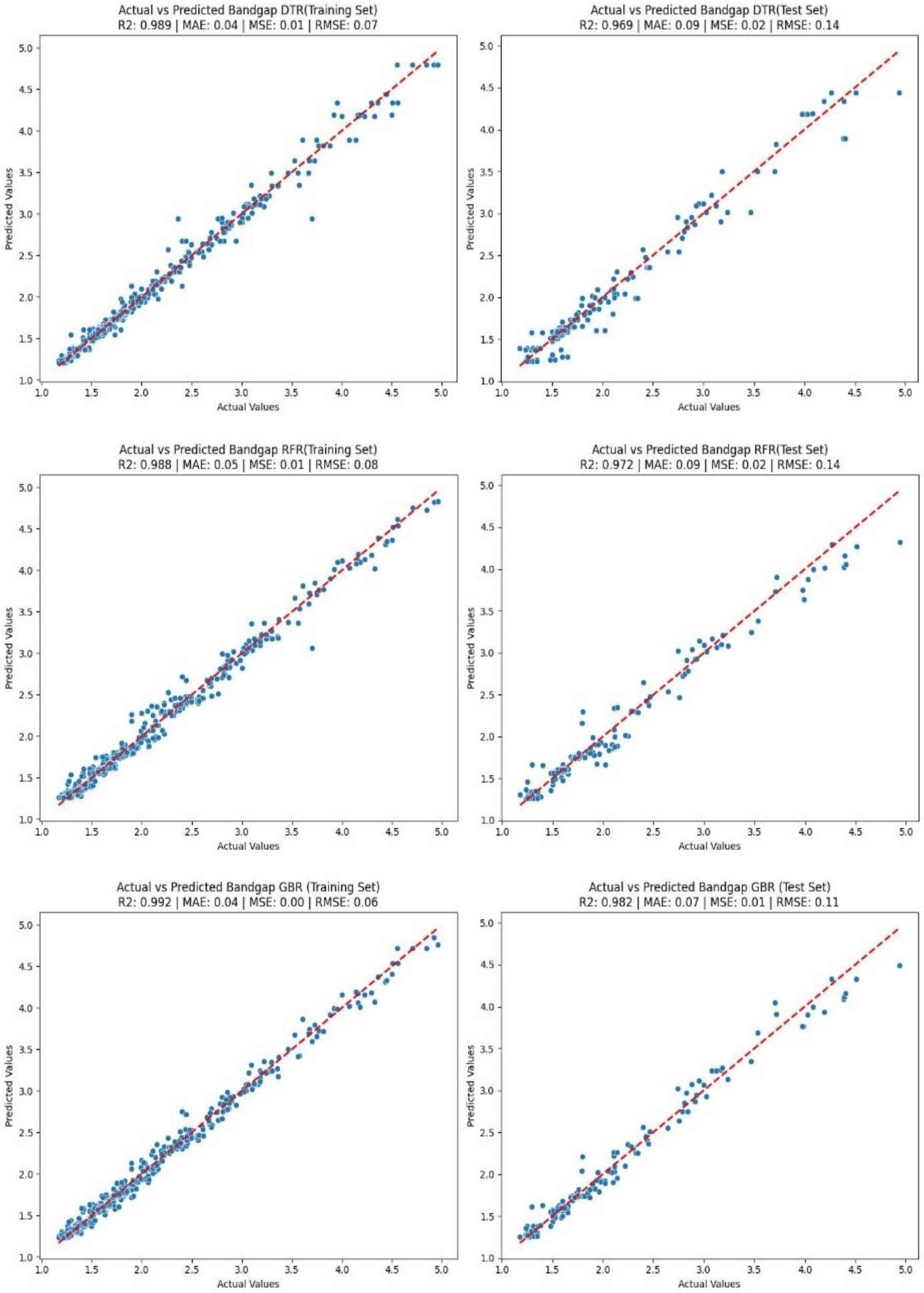

Fig. S3 Regression fit plot for DTR, RFR and GBR algorithms for both train and test sets.

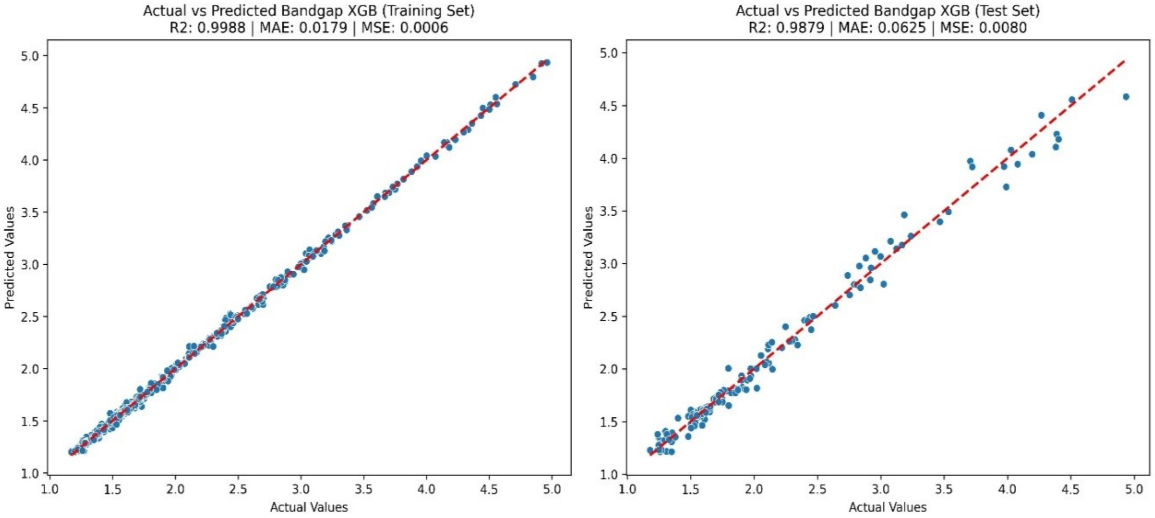

Fig. S4 Regression fit plot for best performed XGB model.

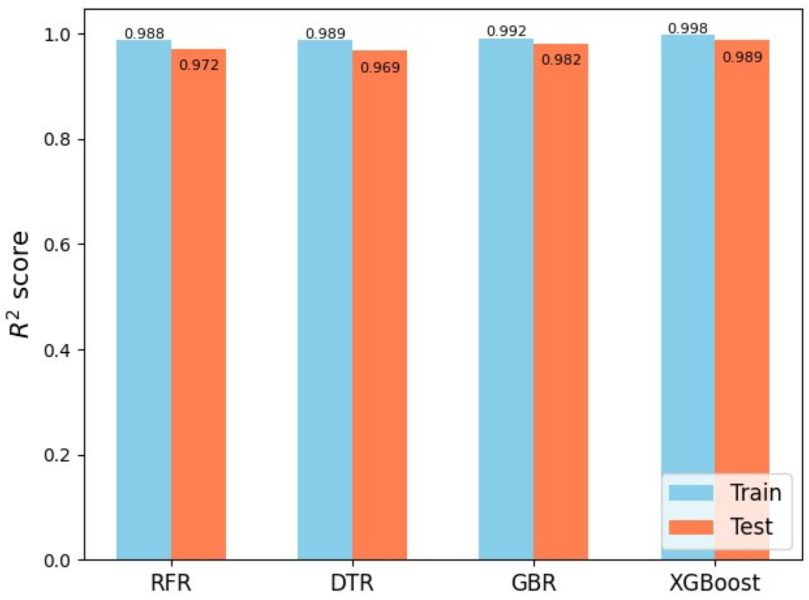

Fig. S5 Comparison plot between R^2 score of different algorithms for band gap prediction

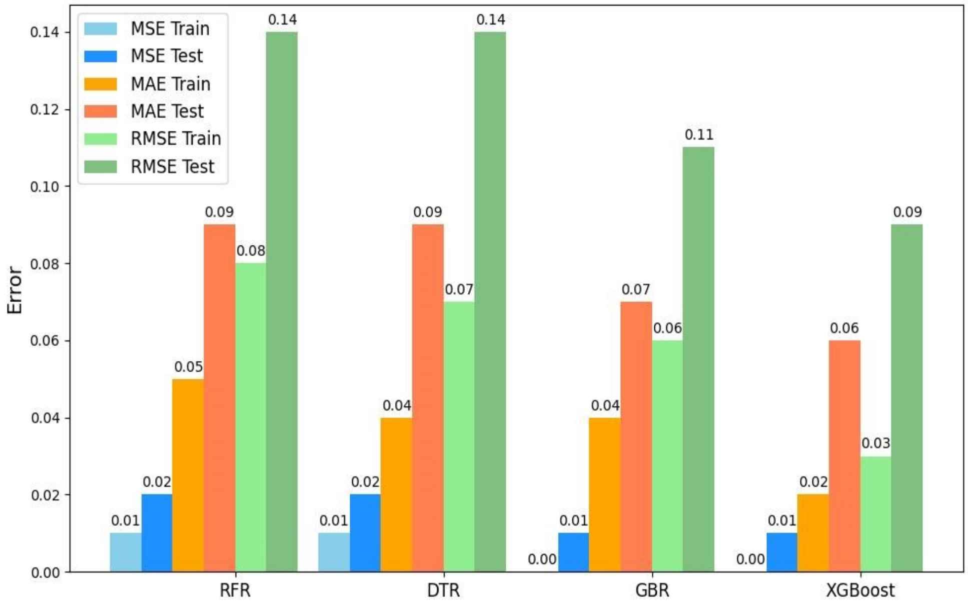

Fig. S6 Comparison plot between loss functions of different algorithm for band gap prediction

Table S3. Comparison between evaluation metrics for bandgap prediction

Evaluation metrics	Machine learning models							
	RFR		DTR		GBR		XGB	
	Train	Test	Train	Test	Train	Test	Train	Test
R^2	0.988	0.972	0.989	0.969	0.992	0.982	0.998	0.989
RMSE	0.08	0.14	0.07	0.14	0.06	0.11	0.03	0.09

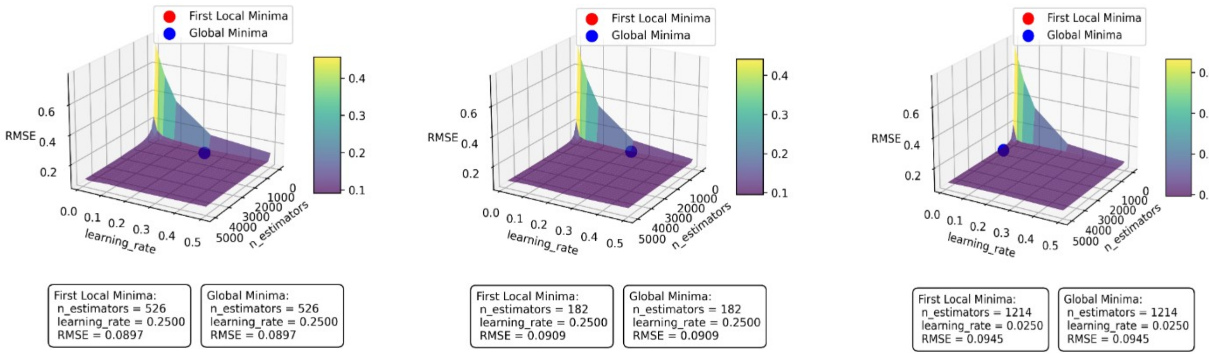

Fig. S7 To identify the optimal combination of hyperparameters, an exhaustive grid search was performed, focusing on two highly influential hyperparameters: n_estimators and learning_rate, while keeping the maximum depth constant at (a) max depth = 2, (b) max depth = 3 and (c) max depth = 4. A 2D surface plot was generated to visualize the relationship between the cost function (RMSE) and these hyperparameters. For instance, a lower RMSE was achieved when the maximum depth was set to 2. The figures shown above provide further insights into how the error landscape evolves with changes in depth and other parameters. In these cases, the simplicity of the error surface resulted in the first local minimum being identical to the global minimum, meaning that the optimization process quickly found the most optimal solution without getting side tracked by other potential minima.

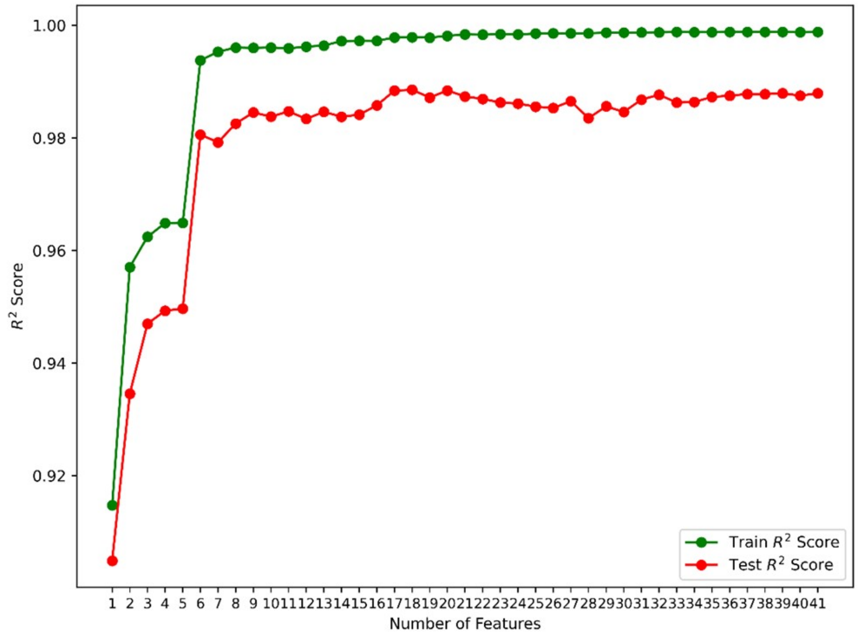

Fig. S8 The above plot shows the visualization of forward selection strategy. The model shows highest R2 score when the number of features is 18, after which there is no significant improvement in the training score and the test score shows fluctuations.

Table S4. Optimized subset of 17 features after feature engineering

Features	Importance score in (%)	Expansion
AD (X)	50.72	Atomic Density of X-site
χ(B+X)	22.60	Mulliken's electronegativity of (B+X)
χ(A+X)	8.28	Mulliken's electronegativity of (A+X)
IR (B+X)	5.55	Ionic Radii of (B+X)
IE (B)	3.26	Ionization energy of B-site
Ratio (X2)	2.30	Proportion of X2-site
Ratio (X1)	1.98	Proportion of X1-site

AW (B)	1.61	Atomic weight of B-site
Ratio (X3)	0.94	Proportion of X3-site
AN (X3)	0.68	Atomic Number of X3-site
AN (X1)	0.60	Atomic number of X1-site
IR (A/X)	0.59	Ionic Radii of (A/X)
χ(A-B)	0.32	Mulliken's electronegativity of (A-B)
χ(A)	0.29	Mulliken's electronegativity of A
Ratio (B1)	0.17	Proportion of B1-site
IR (B/X)	0.06	Ionic Radii of (B/X)
IR (X)	0.04	Ionic radii of X-site

Table S5. Comparison of evaluation metrics with available literature (*PC – Perovskite composition, *c – Manually calculated values which were not available in the corresponding literature)

S. No.	Perovskite description	Area of application	Algorithm	Data Points	Features	R^2	RMSE	Year	Ref
1	HOIP-ABX_3	Solar cells	GBR	212	14	0.969	0.291^{C}	2018	16
2	HOIP-ABX_3	Solar cells	GBR	192	32	0.827	0.448^{C}	2019	10
3	Inorganic-ABX_3	Ferroelectric Photovoltaic solar cells	GBR	564	134	0.923	0.619^{C}	2019	17
4	Mixed HOIP-ABX_3	Solar cells	ANN	69 (Train) + 9 (Test)	PC	-	0.06	2019	18
5	HOIP-ABX_3	Solar cells	GBR	906	32	0.943	0.293	2020	19
6	DHOIP	Solar cells	GBR	525	8	0.97	0.244	2021	20
7	DHOIP	Solar cells	GBR-Model1	196	24	0.908	0.256	2022	21
			GBR-Model2	196	19	0.887	0.284		
			GBR-Model3	11161	24	0.876	0.3		
			GBR-Model4	2113	24	0.769	0.533		
			GBR-Model5	2113	24	0.791	0.216		
8	DHOIP	Solar cells	GBR	4456	95	0.92	0.307	2022	22
9	Mixed HOIP-ABX_3	Solar cells	XGB	204 (Train) + 23 (Test)	PC	-	0.055	2022	23
10	Mixed	Solar cells	GBR	610	27	0.99	0.059	2023	2

	HOIP-ABX_3								
11	Mixed HOIP-ABX_3	Solar cells	GBDT	245	PC	0.93	0.09	2024	[24]
12	**Mixed HOIP-ABX_3**	**Solar water splitting**	**XGB**	**818**	**82**	**0.989**	**0.03**	**2025**	**This work**

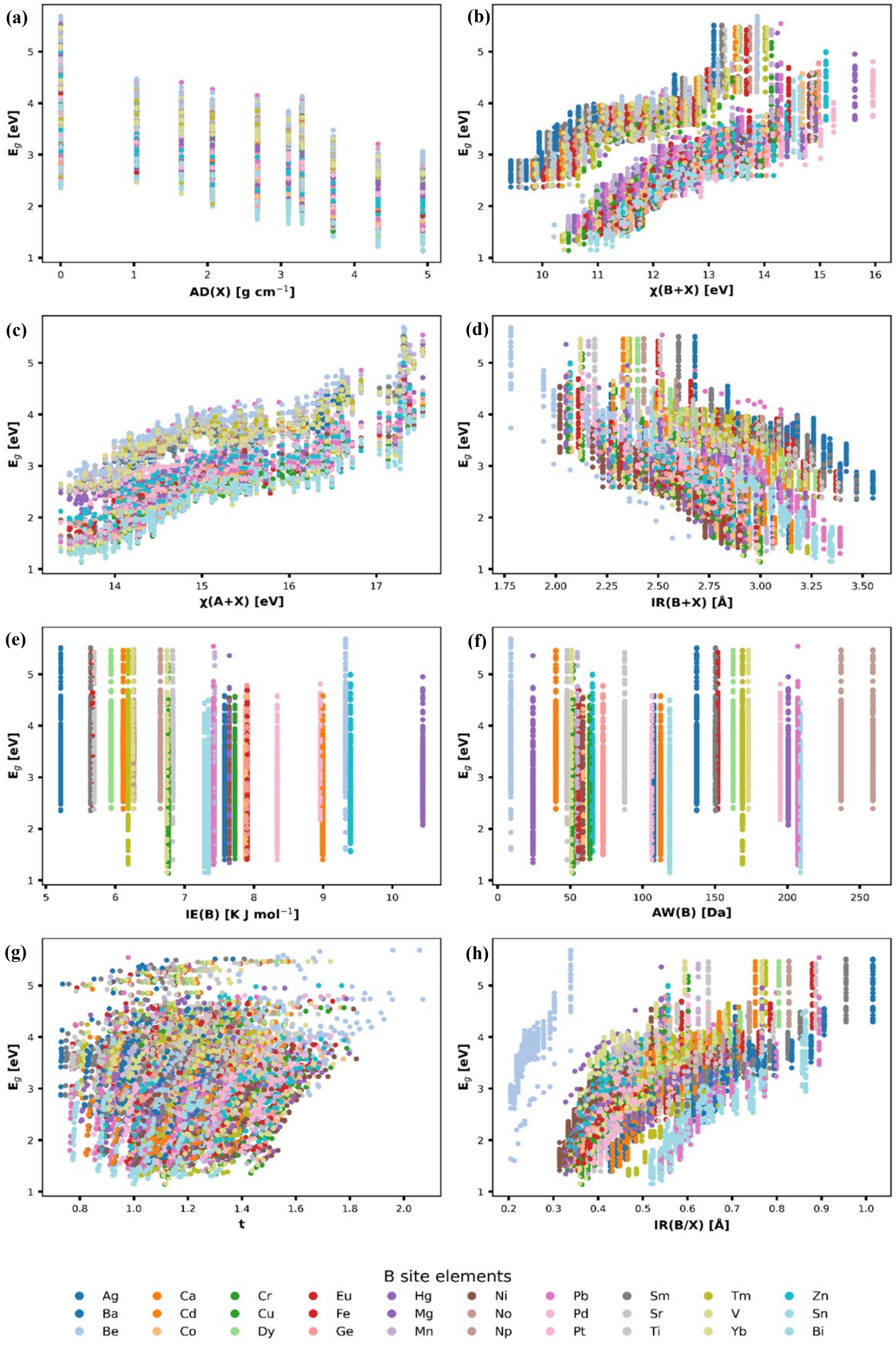

Fig. S9 Scatter plots between predicted bandgap values by XGB and important features

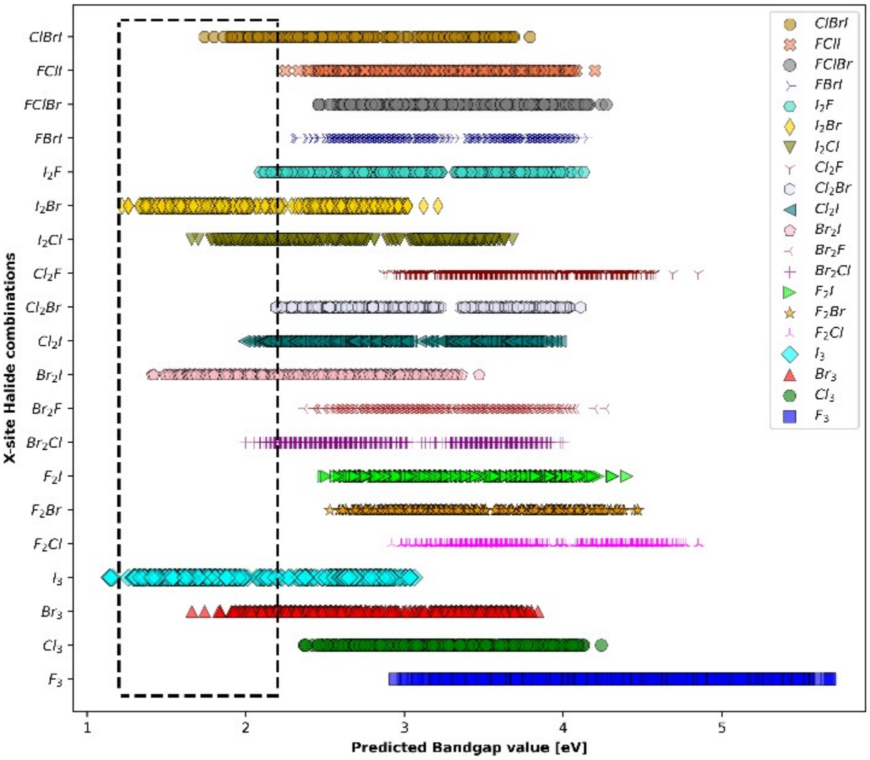

Fig. S10 Impact of halide combination in bandgap tuning

Table S6. Evaluation metrics obtained for band edge prediction

Evaluation metrics	Machine learning models for band edge prediction							
	RFR		DTR		GBR		XGB	
	Train	Test	Train	Test	Train	Test	Train	Test
R^2	0.932	0.756	0.942	0.832	0.974	0.902	0.992	0.812
RMSE	0.08	0.15	0.07	0.12	0.05	0.13	0.03	0.13

Table S7. Other formulae generated using GPSR with their RMSE values.

S. No.	Formula	RMSE
1	$E_{VBM} = \left[0.451 - \left(E_g + t\right)\right] - IR(B + X)$	0.159
2	$E_{VBM} = (0.452 - t) - \left[E_g + IR(B + X)\right]$	0.160
3	$E_{VBM} = -0.513 - \left[E_g + IR(B + X)\right]$	0.161

Table S8. Comparison of VBM & CBM values obtained through GBR, GPSR (SR) and conventional formula (CF) with reported literature (Rep). All the VBM and CBM values taken here are with respect to Absolute Vacuum Scale.

S. No	HOIP	Rep CBM [eV]	Rep VBM [eV]	Rep E_g [eV]	XGB E_g [eV]	VBM - GBR [eV]	Error (GBR)	VBM -SR [eV]	Error (SR)	VBM - CF [eV]	Error (CF)	VBM - Avg. [eV]	Error (Avg.)	Ref
1	$MAPbI_3$	-3.9	-5.45	1.55	1.52	-5.42	0.03	-5.37	0.08	-5.32	0.13	-5.37	0.08	
2	$MAPbI_3$	-4.06	-5.67	1.61	1.52	-5.42	0.25	-5.37	0.30	-5.32	0.35	-5.37	0.30	
3	$MASnBr_3$	-3.39	-5.54	2.15	1.91	-5.48	0.06	-5.50	0.04	-5.67	-0.13	-5.55	-0.01	
4	$FASnI_3$	-4.5	-5.9	1.4	1.15	-5.07	0.83	-5.04	0.86	-5.59	0.31	-5.23	0.67	
5	$MAPbI_3$	-3.85	-5.45	1.6	1.52	-5.42	0.03	-5.37	0.08	-5.32	0.13	-5.37	0.08	
6	$MASnI_3$	-3.48	-4.74	1.26	1.32	-4.99	-0.25	-5.13	-0.39	-5.52	-0.78	-5.21	-0.47	
7	$MAPbI_3$	-3.88	-5.39	1.51	1.52	-5.42	-0.03	-5.37	0.02	-5.32	0.07	-5.37	0.02	
8	$MASnI_3$	-3.63	-4.73	1.1	1.32	-4.99	-0.26	-5.13	-0.40	-5.52	-0.79	-5.21	-0.48	23
9	$FAPbI_3$	-4.23	-5.76	1.53	1.3	-5.04	0.72	-5.22	0.54	-5.41	0.35	-5.22	0.54	
10	$MAPbI_3$	-3.9	-5.4	1.5	1.52	-5.42	-0.02	-5.37	0.03	-5.32	0.08	-5.37	0.03	
11	$MASnBr_3$	-3.42	-5.67	2.25	1.91	-5.48	0.19	-5.50	0.17	-5.67	0.00	-5.55	0.12	
12	$FASnI_3$	-4.12	-5.34	1.22	1.15	-5.07	0.27	-5.04	0.30	-5.59	-0.25	-5.23	0.11	
13	$MASnI_3$	-3.7	-5	1.3	1.32	-4.99	0.01	-5.13	-0.13	-5.52	-0.52	-5.21	-0.21	
14	$MAPbBr_3$	-3.69	-6	2.31	2.16	-5.80	0.20	-5.47	0.53	-5.44	0.56	-5.57	0.43	
15	$MAPbBr_3$	-3.42	-5.73	2.31	2.16	-5.80	-0.07	-5.47	0.26	-5.44	0.29	-5.57	0.16	
16	$FASnI_2Br$	-4.3	-6	1.7	1.22	-5.33	0.67	-5.04	0.96	-5.71	0.29	-5.36	0.64	25
17	$FASnI_3$	-4.26	-5.47	1.21	1.15	-5.07	0.40	-5.04	0.43	-5.59	-0.12	-5.23	0.24	
18	$FASnI_2Br$	-3.96	-5.52	1.56	1.22	-5.33	0.19	-5.04	0.48	-5.71	-0.19	-5.36	0.16	26
19	$FASnIBr_2$	-3.78	-5.53	1.75	1.41	-5.74	-0.21	-5.16	0.37	-5.76	-0.23	-5.55	-0.02	
20	$FASnBr_3$	-3.39	-5.54	2.15	1.66	-5.67	-0.13	-5.34	0.20	-5.78	-0.24	-5.60	-0.06	
21	$DMASnBr_3$	-3.65	-6.5	2.85	2.36	-5.84	0.66	-6.08	0.42	-5.44	1.06	-5.79	0.71	27
22	$MASnI_3$	-4	-5.5	1.5	1.32	-5	0.5	-5.13	0.37	-5.52	-0.02	-5.22	0.28	
23	$MASnI_2Br$	-3.6	-5.4	1.8	1.34	-5.1	0.3	-5.08	0.32	-5.66	-0.26	-5.28	0.12	28
24	$MASnBr_2I$	-3.2	-5.2	2	1.52	-5.47	-0.27	-5.19	0.01	-5.72	-0.52	-5.46	-0.26	
25	$MASnBr_3$	-2.8	-5.1	2.3	1.91	-5.48	-0.38	-5.5	-0.4	-5.67	-0.57	-5.55	-0.45	
26	$MAGeI_3$	-3.2	-5.2	2	1.50	-5.90	-0.70	-5.02	0.18	-5.53	-0.33	-5.48	-0.28	29
27	$FAGeI_3$	-3.2	-5.5	2.3	1.70	-5.90	-0.40	-5.31	0.19	-5.41	0.09	-5.54	-0.04	
28	$DMASnI_3$	-1.42	-4.19	2.77	1.75	-5.95	-1.76	-5.69	-1.50	-5.30	-1.11	-5.65	-1.46	30
29	$DMAGeI_3$	-0.51	-3.59	3.08	1.75	-5.91	-2.32	-5.42	-1.83	-5.39	-1.80	-5.57	-1.98	
30	$FASnCl_3$	-3.83	-7.33	3.5	2.38	-6.00	1.33	-5.92	1.41	-5.78	1.55	-5.90	1.43	
31	$MASnCl_3$	-3.36	-6.85	3.49	2.55	-5.97	0.88	-6.01	0.84	-5.71	1.14	-5.90	0.95	
32	$MAPbCl_3$	-3.77	-6.92	3.15	2.82	-5.95	0.97	-6.30	0.62	-5.46	1.46	-5.90	1.02	31
33	$FAPbCl_3$	-3.98	-6.94	2.96	2.51	-6.00	0.94	-6.08	0.86	-5.60	1.34	-5.89	1.05	
34	$FASnBr_3$	-3.6	-6.23	2.63	1.66	-5.67	0.56	-5.34	0.89	-5.78	0.45	-5.60	0.63	

35	$MAPbBr_3$	-4.25	-6.6	2.35	2.16	-5.80	0.80	-5.48	1.12	-5.44	1.16	-5.57	1.03
36	$FAPbBr_3$	-4.51	-6.7	2.19	1.92	-5.86	0.84	-5.62	1.08	-5.54	1.16	-5.67	1.03
37	$MASnBr_3$	-3.42	-5.67	2.25	1.91	-5.48	0.19	-5.50	0.17	-5.67	0.00	-5.55	0.12
38	$MAPbI_3$	-4.36	-5.93	1.57	1.52	-5.42	0.51	-5.37	0.56	-5.32	0.61	-5.37	0.56
39	$FAPbI_3$	-4.47	-6.24	1.77	1.30	-5.04	1.20	-5.22	1.02	-5.41	0.83	-5.22	1.02
40	$MASnI_3$	-4.07	-5.39	1.32	1.32	-4.99	0.40	-5.13	0.26	-5.52	-0.13	-5.21	0.18
41	$FASnI_3$	-4.12	-5.34	1.22	1.15	-5.07	0.27	-5.04	0.30	-5.59	-0.25	-5.23	0.11

Note S3: Statistical Analysis of Solar to Hydrogen Conversion (STH) Efficiency

The solar to hydrogen conversion efficiency can be determined statistically by the following consecutive equations.

$$\eta_{STH} = \eta_{abs} \cdot \eta_{cu} \tag{1}$$

$$\eta_{abs} = \frac{\int_{E_g}^{\infty} P(h\omega)\, d(h\omega)}{\int_{0}^{\infty} P(h\omega)\, d(h\omega)} \tag{2}$$

$$\eta_{cu} = \frac{\Delta G \int_{E}^{\infty} \frac{P(h\omega)}{h\omega}\, d(h\omega)}{\int_{E_g}^{\infty} P(h\omega)\, d(h\omega)} \tag{3}$$

where, η_{abs} and η_{cu} corresponds to efficiency of light absorption and carrier utilization. $P(h\omega)$ represents the AM1.5 G solar flux as a function of photon energy $(h\omega)$. The XGB predicted band gap values are given as input for the term E_g with $\Delta G = 1.23\ eV$, minimum Gibbs' free energy required for water splitting and E represents minimum photon energy to participate in the redox reactions.

$$E = \begin{cases} E_g, \ \chi(H_2) \geq 0.2, \chi(O_2) \geq 0.6 \\ E_g + 0.2 - \ \chi(H_2), \ \ \chi(H_2) < 0.2, \chi(O_2) \geq 0.6 \\ E_g + 0.6 - \ \chi(O_2), \ \ \chi(H_2) \geq 0.2, \chi(O_2) < 0.6 \\ E_g + 0.8 - \ \chi(H_2) - \chi(O_2), \ \ \chi(H_2) < 0.2, \chi(O_2) < 0.6 \end{cases}$$

Here, $\chi(H_2)$ and $\chi(O_2)$ represents the overpotentials required for HER and OER reaction to occur.

Table S9. Predicted band gap, band edge and STH efficiency values for the final 21 compounds

S.No.	A	B	X1	X2	X3	Compound	ML predicted E_g (eV)	VBM (V)	CBM (V)	η_{STH} (%)
1	Methylphosphonium	Sn	Br	Br	I	MPSnBr$_2$I	1.6	1.17	-0.43	23.14
2	Protonated Formamide	Sn	I	I	Br	FmSnI$_2$Br	1.47	0.98	-0.49	19.79
3	Methylammonium	Ag	I	I	Br	MAAgI$_3$	1.53	1.03	-0.50	19.53
4	Hydrazinium	Sn	I	I	Br	HzSnI$_2$Br	1.52	1.02	-0.50	19.26
5	Methylammonium	Ag	I	I	I	MAAgI$_3$	1.49	0.97	-0.51	18.74
6	Methylammonium	Ag	Br	Br	I	MAAgBr$_2$I	1.63	1.11	-0.53	17.96
7	Hydroxylammonium	Sn	I	I	Br	HASnI$_2$Br	1.57	1.03	-0.54	17.45
8	Hydrazinium	Ag	I	I	Br	HzAgI$_2$Br	1.64	1.07	-0.57	15.97
9	Hydrazinium	Ag	I	I	I	HzAgI$_3$	1.58	1.01	-0.57	15.73
10	Propylammonium	Sn	I	I	Br	PASnI$_2$Br	1.8	1.20	-0.60	14.62
11	Protonated Formamide	Sn	Br	Br	I	FmSnBr$_2$I	1.66	1.06	-0.60	14.39
12	Hydrazinium	Ag	Br	Br	I	HzAgBr$_2$I	1.74	1.13	-0.61	13.94
13	Protonated Formamide	Ag	I	I	Br	FmAgI$_2$Br	1.59	0.96	-0.62	13.49
14	Hydrazinium	Sn	Br	Br	I	HzSnBr$_2$I	1.7	1.07	-0.63	13.26
15	Hydroxylammonium	Ag	I	I	I	HAAgI$_3$	1.68	1.04	-0.64	12.62
16	Azetidinium	Sn	I	I	Cl	AzSnI$_2$Cl	1.89	1.23	-0.66	11.98
17	Methylphosphonium	Sn	I	I	Cl	MPSnI$_2$Cl	1.88	1.22	-0.65	11.98
18	Protonated Formamide	Ag	Br	Br	I	FmAgBr$_2$I	1.68	1.02	-0.67	11.77
19	Hydroxylammonium	Sn	Br	Br	I	HASnBr$_2$I	1.77	1.10	-0.67	11.57
20	Hydroxylammonium	Ag	Br	Br	I	HAAgBr$_2$I	1.84	1.17	-0.67	11.37
21	Hydroxylammonium	Ag	I	I	Br	HAAgI$_2$Br	1.77	1.10	-0.68	11.37

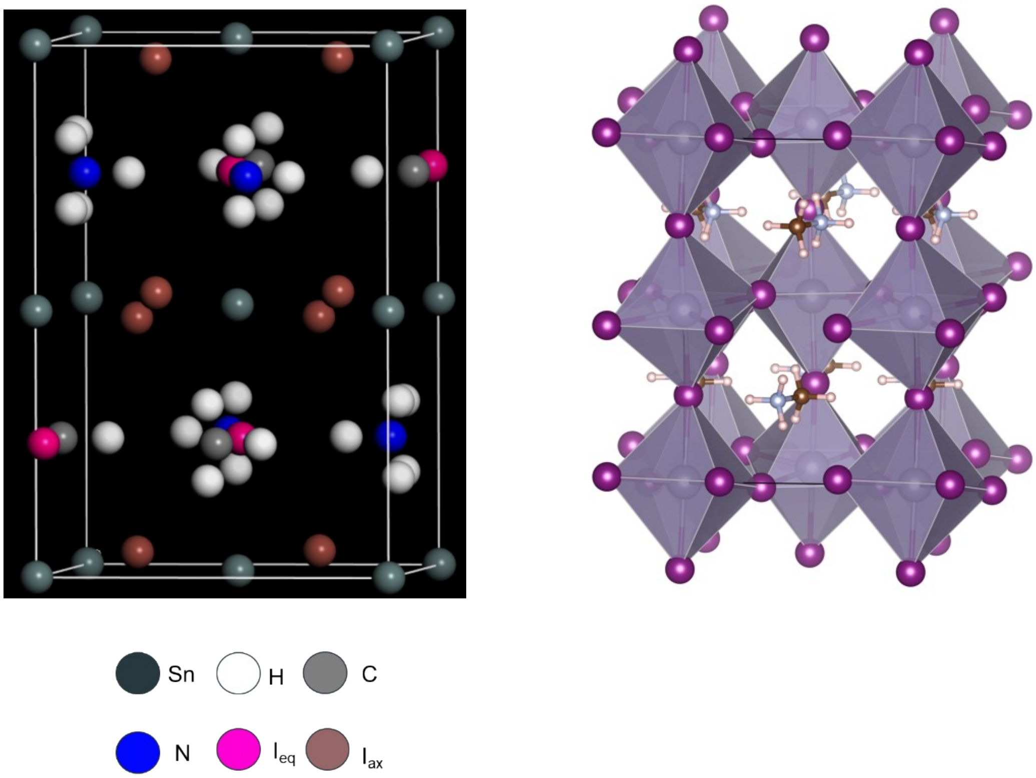

Fig. S11 (a) Orthorhombic MASnI$_3$ unit cell and (b) MASnI$_3$ perovskite structure.

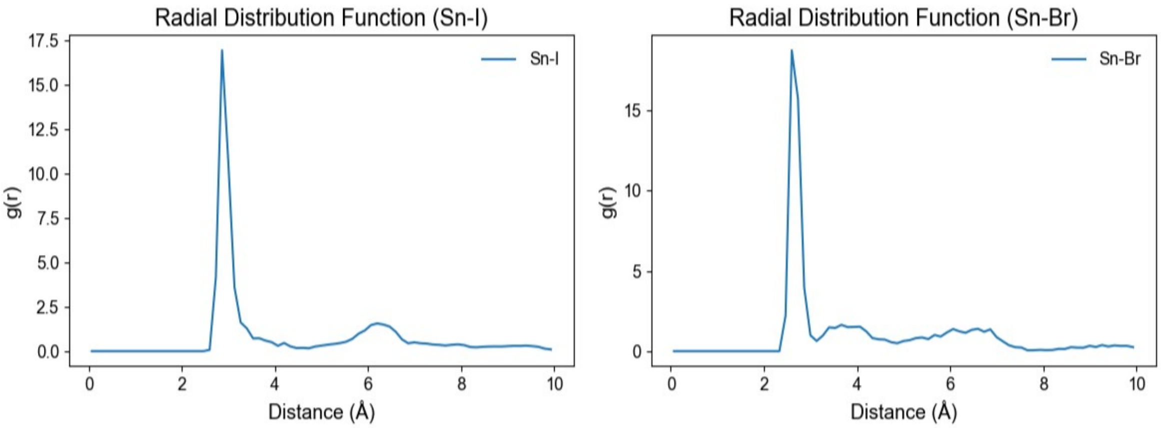

Fig. S12 Radial Distribution Function of Sn-I and Sn-Br for FmSnI2Br

References

1 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, Mahmoud. H. Aldamasy, M. Vasquez-Montoya, M. A. Ruiz-Preciado, M. A. Flatken, M. V. Khenkin, M. Grischek, M. Kedia, M. Saliba, M. Anaya, M. Veldhoen, N. Arora, O. Shargaieva, O. Maus, O. S. Game, O. Yudilevich, P. Fassl, Q. Zhou, R. Betancur, R. Munir, R. Patidar, S. D. Stranks, S. Alam, S. Kar, T. Unold, T. Abzieher, T. Edvinsson, T. W. David, U. W. Paetzold, W. Zia, W. Fu, W. Zuo, V. R. F. Schröder, W. Tress, X. Zhang, Y.-H. Chiang, Z. Iqbal, Z. Xie and E. Unger, *Nat. Energy*, 2021, **7**, 107–115.

2 C. Yang, X. Chong, M. Hu, W. Yu, J. He, Y. Zhang, J. Feng, Y. Zhou and L. W. Wang, *ACS Appl. Mater. Interfaces*, 2023, **15**, 40419–40427.

3 C. Kim, T. D. Huan, S. Krishnan and R. Ramprasad, *Sci. Data*, DOI:10.1038/sdata.2017.57.

4 J. Chen, W. Xu and R. Zhang, *J. Mater. Chem. A Mater.*, 2022, **10**, 1402–1413.

5 K. M. Rabe, J. C. Phillips, P. Villars and I. D. Brown, *Global multinary structural chemistry of stable quasicrystals, high-Tc ferroelectrics, and high-T, superconductors*, 1992, vol. 45.

6 L. Breiman, *Mach. Learn.*, 2001, **45**, 5–32.

7 J. R. Quinlan, *Machine Learning 1986 1:1*, 1986, **1**, 81–106.

8 J. H. Friedman, *https://doi.org/10.1214/aos/1013203451*, 2001, **29**, 1189–1232.

9 T. Chen and C. Guestrin, *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, **13-17-August-2016**, 785–794.

10 T. Wu and J. Wang, *Nano Energy*, DOI:10.1016/j.nanoen.2019.104070.

11 S. Alidoust, F. Jamalinabijan and A. Tekin, *ACS Appl. Energy Mater.*, 2024, **7**, 785–798.

12 G. Kieslich, S. Sun and A. K. Cheetham, *Chem. Sci.*, 2015, **6**, 3430–3433.

13 Shannon Radii, http://abulafia.mt.ic.ac.uk/shannon/ptable.php, (accessed 29 January 2026).

14 M. Peng, J. Lei, L. Li, L. Wondraczek, Q. Zhang and J. Qiu, *J. Mater. Chem. C Mater.*, 2013, **1**, 5303–5308.

15 R. Cao, Y. Cao, T. Fu, S. Jiang, W. Li, Z. Luo and J. Fu, *J. Alloys Compd.*, 2016, **661**, 77–81.

16 S. Lu, Q. Zhou, Y. Ouyang, Y. Guo, Q. Li and J. Wang, *Nat. Commun.*, DOI:10.1038/s41467-018-05761-w.

17 S. Lu, Q. Zhou, L. Ma, Y. Guo and J. Wang, *Small Methods*, DOI:10.1002/smtd.201900360.

18 J. Li, B. Pradhan, S. Gaur and J. Thomas, *Adv. Energy Mater.*, DOI:10.1002/aenm.201901891.

19 T. Wu and J. Wang, *ACS Appl. Mater. Interfaces*, 2020, **12**, 57821–57831.

20 Y. Wu, S. Lu, M. G. Ju, Q. Zhou and J. Wang, *Nanoscale*, 2021, **13**, 12250–12259.

21 J. Chen, W. Xu and R. Zhang, *J. Mater. Chem. A Mater.*, 2022, **10**, 1402–1413.

22 X. Cai, Y. Zhang, Z. Shi, Y. Chen, Y. Xia, A. Yu, Y. Xu, F. Xie, H. Shao, H. Zhu, D. Fu, Y. Zhan and H. Zhang, *Advanced Science*, DOI:10.1002/advs.202103648.

23 Y. Liu, W. Yan, H. Zhu, Y. Tu, L. Guan and X. Tan, *Org. Electron.*, DOI:10.1016/j.orgel.2021.106426.

24 C. Ren, Y. Wu, J. Zou and B. Cai, *Materials*, DOI:10.3390/ma17112686.

25 M. Zhang, M. Lyu, J. H. Yun, M. Noori, X. Zhou, N. A. Cooling, Q. Wang, H. Yu, P. C. Dastoor and L. Wang, *Nano Res.*, 2016, **9**, 1570–1577.

26 F. Hao, C. C. Stoumpos, D. H. Cao, R. P. H. Chang and M. G. Kanatzidis, *Nature Photonics 2014 8:6*, 2014, **8**, 489–494.

27 L. Romani, A. Speltini, F. Ambrosio, E. Mosconi, A. Profumo, M. Marelli, S. Margadonna, A. Milella, F. Fracassi, A. Listorti, F. De Angelis and L. Malavasi, *Angewandte Chemie International Edition*, 2021, **60**, 3611–3618.

28 H. Xu, H. Yuan, J. Duan, Y. Zhao, Z. Jiao and Q. Tang, *Electrochim. Acta*, 2018, **282**, 807–812.

29 T. Krishnamoorthy, H. Ding, C. Yan, W. L. Leong, T. Baikie, Z. Zhang, M. Sherburne, S. Li, M. Asta, N. Mathews and S. G. Mhaisalkar, *J. Mater. Chem. A Mater.*, 2015, **3**, 23829–23832.

30 T. Chutia and D. J. Kalita, *RSC Adv.*, 2022, **12**, 25511–25519.

31 S. Tao, I. Schmidt, G. Brocks, J. Jiang, I. Tranca, K. Meerholz and S. Olthof, *Nature Communications 2019 10:1*, 2019, **10**, 1–10.