Supplemental Information

Accelerated design of promising mixed lead-free double halide organic-inorganic perovskites for photovoltaics using machine learning

Yilei Wu[†], Shuaihua Lu[†], Minggang Ju[*], Qionghua Zhou[*]& Jinlan Wang[*]

School of Physics, Southeast University, Nanjing 211189, China.

These authors contributed equally: Yilei Wu, Shuaihua Lu.

Correspondence and requests for materials should be addressed to M. J., Q.Z. and J.W. (email: juming@seu.edu.cn; qh.zhou@seu.edu.cn; jlwang@seu.edu.cn)

Table of Contents

Supplementary Methods

Model selection. The accuracy of machine learning (ML) models relies on appropriate algorithm, thus selecting out the best model from a series of available ML models is necessary. For ML classification, six classifiers are considered: gradient boosting classifier (GBC), support vector machine (SVM), AdaBoost classifier, random forest classifier (RFC), stochastic gradient descent classifier (SGDC), and decision trees classifier (DTC). For ML regression, six regressors are considered as well: gradient boosting regressor (GBR), kernel ridge regressor (KRR), bagging regressor, random forest regressor (RFR), kernel neighbors' regressor (KNR) and decision trees regressor (DTR).[1] The same training & test sets are applied to train each classification (regression) model. The same model evaluation indexes are utilized to evaluate fairly the performance of classification (regression) model. The GBC model shows the best performance among these six classification models, and the most appropriate regression model is GBR model.

Model evaluation for classification and regression algorithms. Model evaluation indexes are essential for measuring the performance of ML models. ML classification and regression models correspond to different model evaluation indexes. Four evaluation indexes (area under curve (AUC), accuracy, precision and recall) are applied for classification models and three different evaluation indexes (coefficient of determination (R^2), mean square error (MSE) and mean absolute error (MAE)) are used for regression models.

The classification models produce the prediction probability for samples. The classification threshold is set to 0.5 in this work, and the prediction probability results of samples are compared to the pre-defined threshold. The prediction probability results correspond to the probabilities that samples belong to positive class (i.e. perovskite and perovskite with bandgaps smaller than 0.2 eV) or negative class (i.e. non-perovskite and perovskite with bandgaps larger than 0.2 eV). Therefore, samples with prediction probability results larger than 0.5 are classified into positive class, and

samples with prediction probability results less than 0.5 are classified into negative class.

According to the classification results, the count of positive samples predicted correctly is defined as true positive (TP), and the count of positive samples predicted falsely is defined as false positive (FP). The count of negative samples predicted correctly is defined as true negative (TN), and the count of negative samples predicted falsely is defined as false negative (FN). By calculating the values of TP, FP, TN and FN, we obtained the confusion matrix, which represents the counts of the predicted classes versus the true classes of test set. The confusion matrix is shown as follows:

	Predicted positive	Predicted negative
True positive	TP	FN
True negative	FP	TN

The true positive rate (TPR) and false positive rate (FPR) are calculated based on TP, TN, FP and FN.

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \quad (1)$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{TN+FP}} \quad (2)$$

The receiver operating characteristic (ROC) curve can be drawn using TPR and FPR as coordinates, and is often used to measure the performance of classification models. For ROC curves in this work, the probability of positive prediction is the number of positive prediction times derived by total number of times after 100 executions. When comparing the performance of different classification models, if the ROC curve of one model is completely below the ROC curve of the other, it means that the performance of the latter model is better than the former. If the ROC curves of two classification models intersect, the comparison is difficult. Thus the more appropriate model evaluation index is the AUC value. The higher AUC value correspond to the better performance of classification model. The AUC value of a classification model without learning algorithm is equal to 0.5, and the AUC value of the perfect classification model is equal to 1.

Accuracy is the proportion of correctly classified samples among all samples. In

general, the better classification model has the higher accuracy. Accuracy is defined as follows:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \tag{3}$$

Precision is the proportion of true positive samples in predicted positive class. Precision is defined as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{4}$$

Recall represents the ability for identifying positive samples of classification models. Recall is defined as follows:

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{5}$$

The above indexes are used to evaluate the performance of classification models, and three different indexes are chosen for evaluating the performance of regression models.

R^2 is the proportion of the variance in the dependent variable that is predictable form the independent variable. R^2 is defined as follows:

$$R^2 = 1 - \frac{\sum_i \left(y_i^{\text{true}} - y_i^{\text{pred}}\right)^2}{\sum_i \left(y_i^{\text{true}} - \bar{y}_i^{\text{true}}\right)^2} \tag{6}$$

Where y_i^{true} are the true values, and y_i^{pred} are the predicted values. The predicted values of the perfect regression model are equal to true values, thus the value of R^2 is equal to 1.

Mean square error (MSE) represents the average squared difference between the predicted values and true values. MSE is defined as follows:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left(y_i^{\text{true}} - y_i^{\text{pred}}\right)^2 \tag{7}$$

Mean absolute error (MAE) represents the arithmetic average of the absolute errors between predicted values and true values. MAE is defined as follows:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|y_i^{\text{true}} - y_i^{\text{pred}}\right| \tag{8}$$

Hyper-parameters selection. After model selection, hyper-parameters are optimized by applying a global search algorithm based on the simulated annealing algorithm.[2] Hyper-parameters are optimized before training ML models, and optimized hyper-parameters can improve the performance of ML models. We selected 7 initial hyper-parameters for GBC and GBR models, respectively. For GBC model, max_depth is in the range of 2~30, n_estimators is in the range of 40~200, learning_rate is in the range of 0.005~0.5, subsample is in the range of 0.5~1, min_child_weight is in the range of 0.5~10, the random_state_seed is set to 42, and the loss function is set to logistic regression. For GBR model, max_depth is in the range of 1~200, n_estimators is in the range of 100~500, learning_rate is in the range of 0.001~0.5, min_sample_leaf is in the range of 1~50, max_features is in the range of 0.01~1, the random_state_seed is set to 42, the range of loss function is set to [least squares (ls), least absolute deviation (lad), the combination of ls and lad (huber), quantile regression (quantile)]. During the optimization process, ML models with the best hyper-parameters can achieve the maximum accuracy.

Last-place elimination feature selection procedure. To search the most relevant features, a "last-place elimination" feature selection procedure[2, 3] is introduced to GBC algorithm and GBR algorithm. In this work, the initial feature set is consisted of 87 features relating to polarizability, ionic radii and electronegativity. In the first step, 87 features are ranked according to the relative importance after training model. Then the feature at the last position is removed, and the remaining 86 features consist a new feature set. Next, the performance of model with new feature set is evaluated, and the above step is repeated until only two features left. Finally, the accuracy of the model at each step is analyzed, and the feature set corresponds to the inflection point is selected as the optimized feature set.

Supplementary Notes

Note S1. The ionic polarizability of A-site ions in Table S1 is obtained from

Amsterdam Density Functional program package (ADF2013).[3, 4] The ionic radii of A-, B-, B'- and X-site ions and electronegativity of B-, B'- and X-site ions are obtained from the python Mendeleev package 0.5.1.[5]

Note S2. Among all DHOIPs data with the chemical formula $A_2BB'X_6$ in literature, we only selected compositions that satisfy the charge neutrality condition and Pauling's valence rule, resulting in the number of DHOIPs in training & test sets reduced to 2274. Then all of them are labeled 'perovskite' or 'non-perovskite'. The criterion is based on (i) $\theta_{\mathrm{X-B-X}} > 160°$; (ii) $R_{\mathrm{B-X}}^{\min}/R_{\mathrm{B-X}}^{\max} > 2/3$. Where $\theta_{\mathrm{X-B-X}}$ represents the angle of the X-B-X bonds, $R_{\mathrm{B-X}}^{\min}$ and $R_{\mathrm{B-X}}^{\max}$ represent the minimum length and the maximum length of B-X bonds, respectively.

Note S3. To obtain geometric boundaries for double perovskite structure, the geometric limits using rigid sphere model are derived. The generalized Goldschmidt's parameters are introduced as follows: (i) the average octahedron factor: $\bar{\mu} = (IR_{\mathrm{B}} + IR_{\mathrm{B'}})/2IR_{\mathrm{X}}$, (ii) the octahedron mismatch: $\Delta\mu = (IR_{\mathrm{B}} - IR_{\mathrm{B'}})/2IR_{\mathrm{X}}$, (iii) the generalized tolerance factor: $\mathrm{t} = \frac{IR_{\mathrm{A}}+IR_{\mathrm{X}}}{\sqrt{2}\{[(IR_{\mathrm{B}}+IR_{\mathrm{B'}})/2+IR_{\mathrm{X}}]^2+(IR_{\mathrm{B}}-IR_{\mathrm{B'}})^2/4\}^{1/2}}$.[6] Where IR_{A}, IR_{B}, $IR_{\mathrm{B'}}$ and IR_{X} represent the ionic radii of A-, B-, B'- and X-site ions, respectively.

When IR_{B} is equal to $IR_{\mathrm{B'}}$, X-site ions sit at the midway between B- and B'-site ions (Figure S3a). When IR_{B} is different from $IR_{\mathrm{B'}}$, X-site ions shift the distance of $1/2\left|IR_{\mathrm{B}} - IR_{\mathrm{B'}}\right|$ from midway O toward large ions between B- and B'-site ions. This offset helps to relieve the lattice strain caused by size mismatch and to reach the overall electrostatic energy of double perovskite (Figure S3b).[7] The larger difference between IR_{B} and $IR_{\mathrm{B'}}$, the larger the distance of offset. The octahedron limit corresponds to the extremal situation wherein two adjacent X-site ions in the same octahedron are tangent to each other (Figure S3c). In this situation the distance between centers of B'-site ions and X-site ions satisfies the condition $IR_{\mathrm{B'}} = (\sqrt{2}-1)IR_{\mathrm{X}}$. Then both sides of equation are divided by $\mathrm{R_X}$ and $\bar{\mu} - \Delta\mu = (\sqrt{2}-1)$

is obtained. Therefore, the ionic radii must satisfy the condition $\bar{\mu} - \Delta\mu \geq (\sqrt{2} - 1)$. If $\bar{\mu} - \Delta\mu < (\sqrt{2} - 1)$, the B/B'-site ions cannot touch six adjacent X-site ions, bringing the instability from reduced coordination number of B/B'-site ions.

The stretch limit is also considered, which corresponds to the extremal situation wherein A-site ion is so large that it is tangent to all twelve X-site ions around the octahedron cavity (Figure S3d). The distance between A-site ions and midway O is $\frac{\sqrt{2}}{2}(IR_{\mathrm{B}'} + 2IR_{\mathrm{X}} + IR_{\mathrm{B}})$, and the distance between X-site ions and midway O is $\frac{1}{2}(IR_{\mathrm{B}} - IR_{\mathrm{B}'})$. According to Pitagoras' theorem to the triangle, the boundary condition $(IR_{\mathrm{A}} + IR_{\mathrm{X}})^2 = \frac{1}{4}(IR_{\mathrm{B}} - IR_{\mathrm{B}'})^2 + \frac{1}{2}(IR_{\mathrm{B}'} + 2IR_{\mathrm{X}} + IR_{\mathrm{B}})^2$ is obtained. After combining this condition to the generalized tolerance factor, the geometric boundary t = 1 is obtained. For $t > 1$, A-site ions are too large to maintain three-dimensional perovskite structure, thus perovskites are likely to form low dimensional structures.

Note S4. To validate the generalization ability of the bandgap regression model, we randomly divided the training & test sets (525 DHOIPs) into two subsets based on combinations of B and B'-site cations. In detail, each DHOIPs corresponding to the same combination appears in the same subset. Then one subset (498 DHOIPs) is used to train the regression model, while another subset (27 DHOIPs) contains 6 combinations, and MDHOIPs corresponding to these combinations are labeled as out-of-sample systems, e.g. Ag&In, In&Sb, Ag&Sb, Au&In, Sn&Pb, and As&In. Subsequently, two models are utilized to predict the bandgap values of these out-of-sample systems: ML model trained by 525 DHOIPs (model-1), and ML model trained by 498 DHOIPs (model-2). The comparison between ML-predicted and DFT-calculated results is listed in Table R1. For model-2, the maximum error between ML-predicted and DFT-calculated bandgap values is 0.228 eV, and most of MDHOIPs have errors within 0.15 eV, which is slightly higher than that of ML model-1. Overall, our ML model shows well generalization ability on these

out-of-sample systems.

Note S5. In this work, the decomposition energy (ΔH) of $MA_2GeSnI_4Br_2$ is calculated through the pathways as follows:

$$MA_2GeSnI_4Br_2 \rightarrow \frac{2}{3}(2MAI+SnI_2+GeI_2)+\frac{1}{3}(2MABr+SnBr_2+GeBr_2)$$

$$MA_2GeSnI_4Br_2 \rightarrow \frac{2}{3}(MASnI_3+MAI+GeI_2)+\frac{1}{3}(MASnBr_3+MABr+GeBr_2)$$

$$MA_2GeSnI_4Br_2 \rightarrow \frac{2}{3}(MAGeI_3+MAI+SnI_2)+\frac{1}{3}(MAGeBr_3+MABr+SnBr_2)$$

The ΔH of $A_2InB^{3+}X_4X'_2$ (i.e., $MA_2InBiI_2Br_4$ and $FA_2InSbBr_2Cl_4$) is calculated through the pathways as follows:

$$A_2InB^{3+}X_4X'_2 \rightarrow \frac{2}{3}(2AX+InX+BX_3)+\frac{1}{3}(2AX'+InX'+BX'_3)$$

$$A_2InB^{3+}X_4X'_2 \rightarrow \frac{2}{3}(2AX+InX_3+\frac{1}{3}BX_3+\frac{2}{3}B)+\frac{1}{3}(2AX'+InX'_3+\frac{1}{3}BX'_3+\frac{2}{3}B)$$

$$A_2InB^{3+}X_4X'_2 \rightarrow \frac{2}{3}(\frac{1}{2}AX+\frac{1}{2}A_3B_2X_9+InX)+\frac{1}{3}(\frac{1}{2}AX'+\frac{1}{2}A_3B_2X'_9+InX')$$

$$A_2InB^{3+}X_4X'_2 \rightarrow \frac{2}{3}(\frac{3}{2}AX+\frac{1}{6}A_3B_2X_9+InX_3+\frac{2}{3}B)+\frac{1}{3}(\frac{3}{2}AX'+\frac{1}{6}A_3B_2X'_9+InX'_3+\frac{2}{3}B)$$

$$A_2InB^{3+}X_4X'_2 \rightarrow \frac{2}{3}(\frac{1}{2}A_3In_2X_9+\frac{1}{2}AX+\frac{1}{3}BX_3+\frac{2}{3}B)+\frac{1}{3}(\frac{1}{2}A_3In_2X'_9+\frac{1}{2}AX'+\frac{1}{3}BX'_3+\frac{2}{3}B)$$

$$A_2InB^{3+}X_4X'_2 \rightarrow \frac{2}{3}(\frac{1}{2}A_3In_2X_9+\frac{1}{6}A_3B_2X_9+\frac{2}{3}B)+\frac{1}{3}(\frac{1}{2}A_3In_2X'_9+\frac{1}{6}A_3B_2X'_9+\frac{2}{3}B)$$

Supplementary Figures

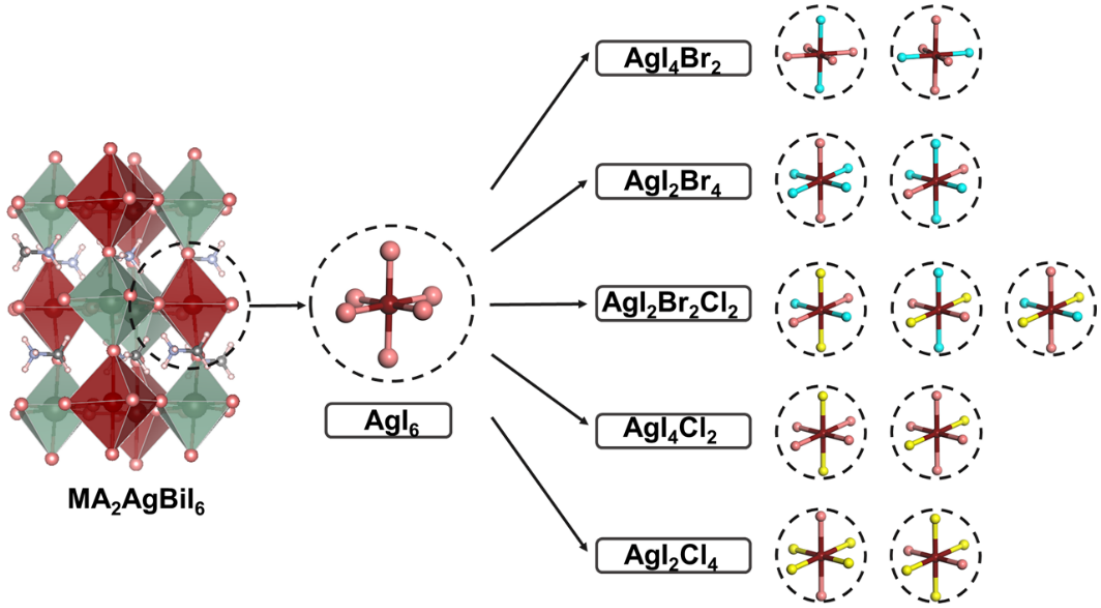

Figure S1. 11 candidates in the prediction set are obtained based on MA_2AgBiI_6 in training&test sets.

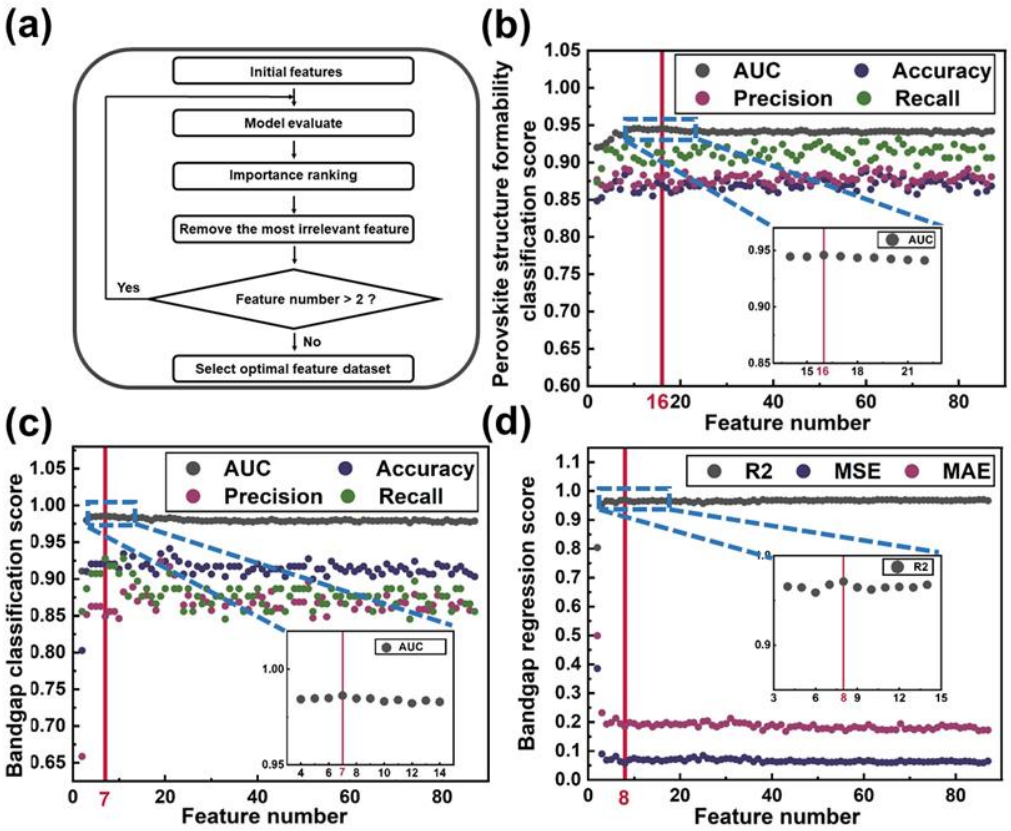

Figure S2. Flowchart and results of last-place elimination feature selection procedure. (a) Feature engineering framework combined with "last-place elimination" method. Optimized feature set of (b) perovskite structure formability classification, (c) bandgap classification, and (d) bandgap regression.

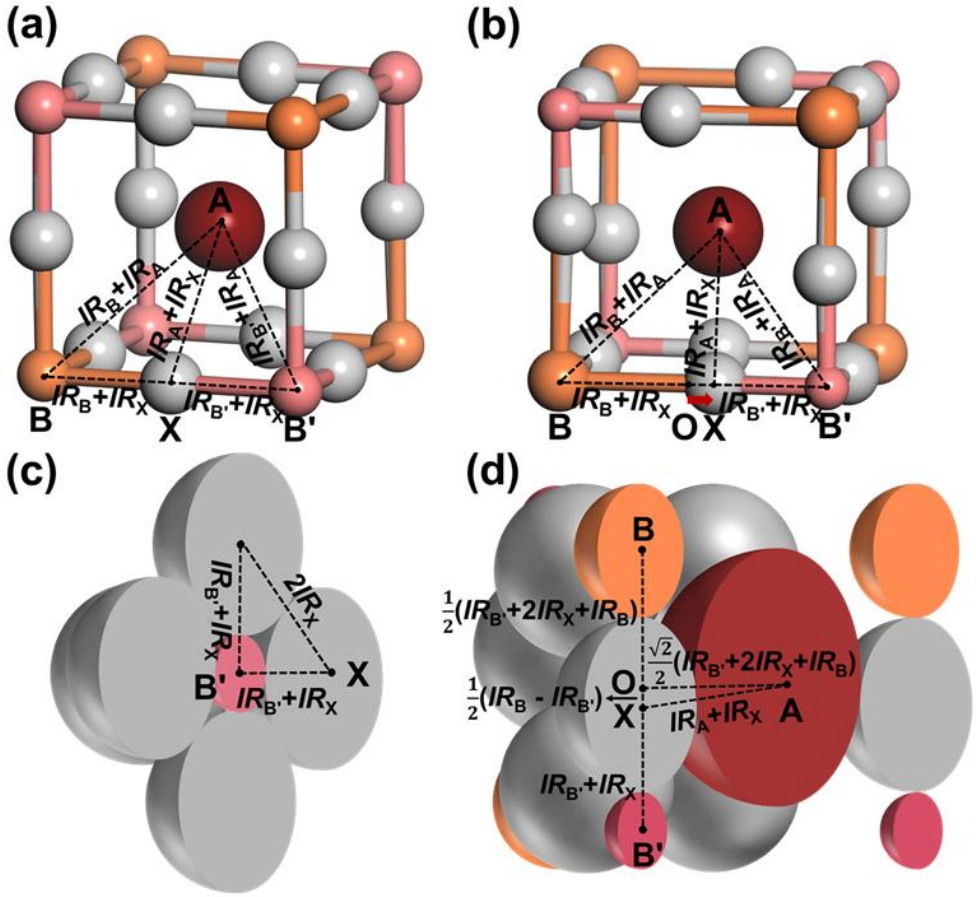

Figure S3. Representation for conventional cell of double perovskites. Red, orange, pink and grey spheres represent A-site ions, spheres represent B-site ions, B'-site ions and X-site ions, respectively. Schematic diagram corresponds to the situation wherein B- and B'-site ions possess (a) same ionic radii and (b) different ionic radii. Schematic representations for (c) the octahedron limit and (d) the stretch limit of double perovskites.

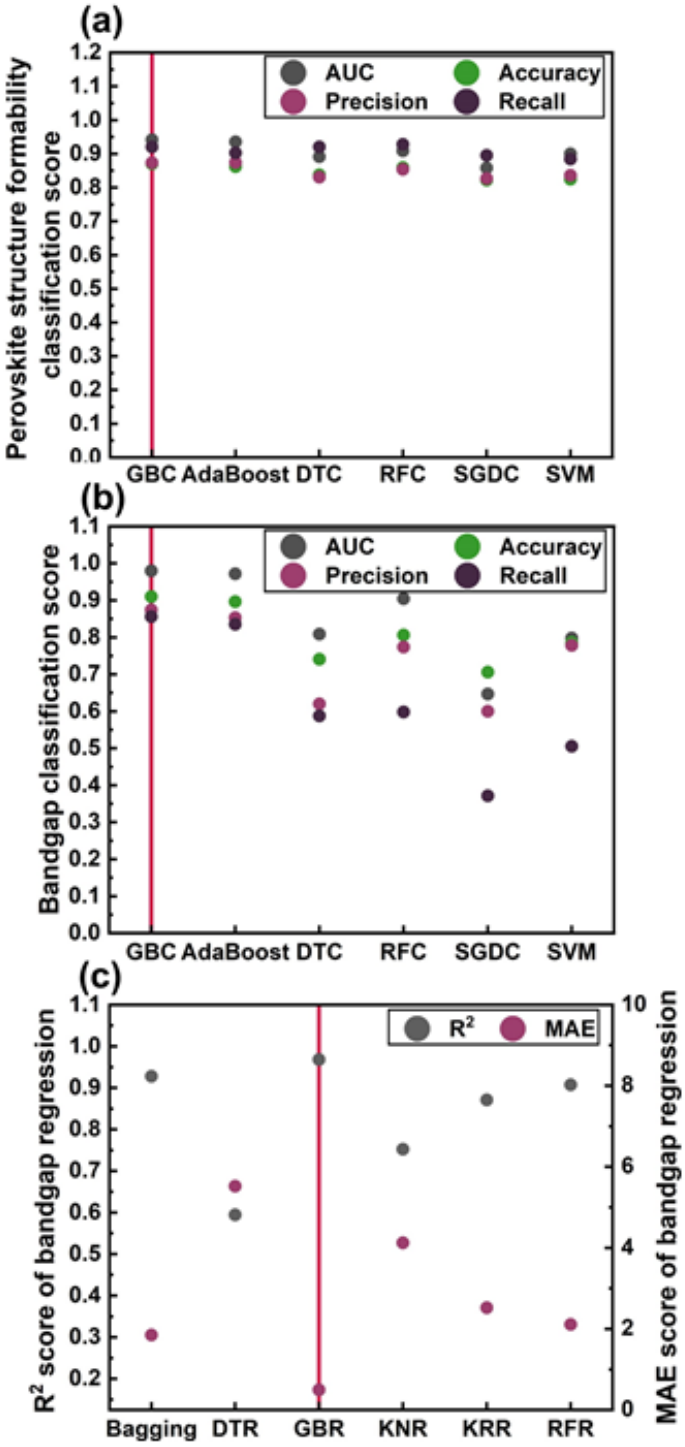

Figure S4. Classification model score of (a) perovskite structure formability classification and (b) bandgap classification. (c) Regression model score of bandgap regression.

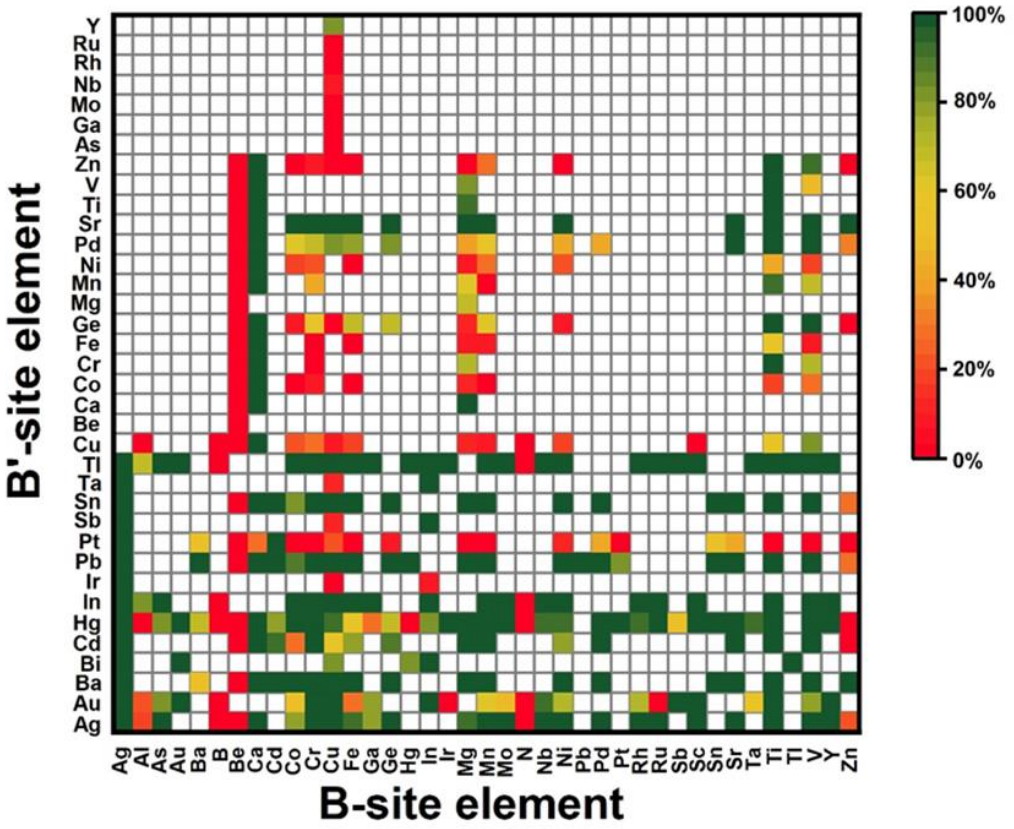

Figure S5. Proportion of perovskites in candidates based on different combinations of B- and B'-site ions.

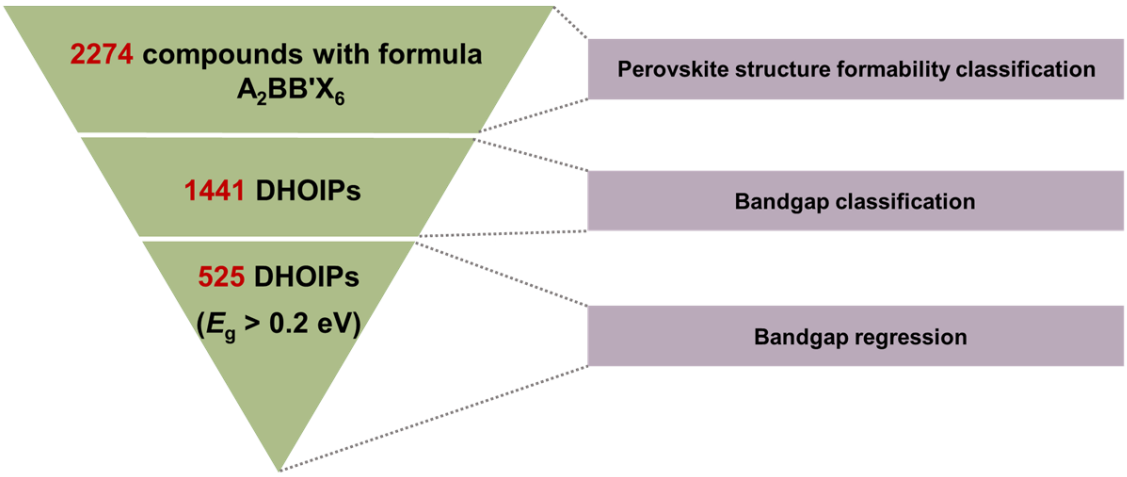

Figure S6. Selection of training & test sets corresponds to three ML models with different target properties. Preparing for model training and test, the training set and test set for each model are divided according to the proportion of 80% and 20%.

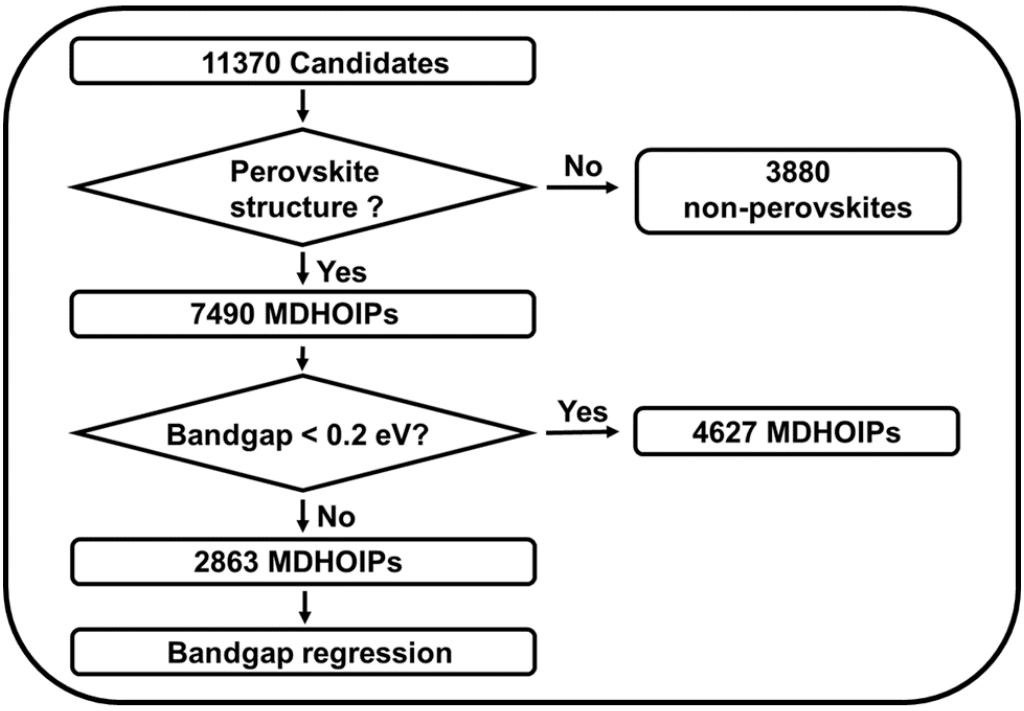

Figure S7. Flowchart for predicting candidates using trained ML models.

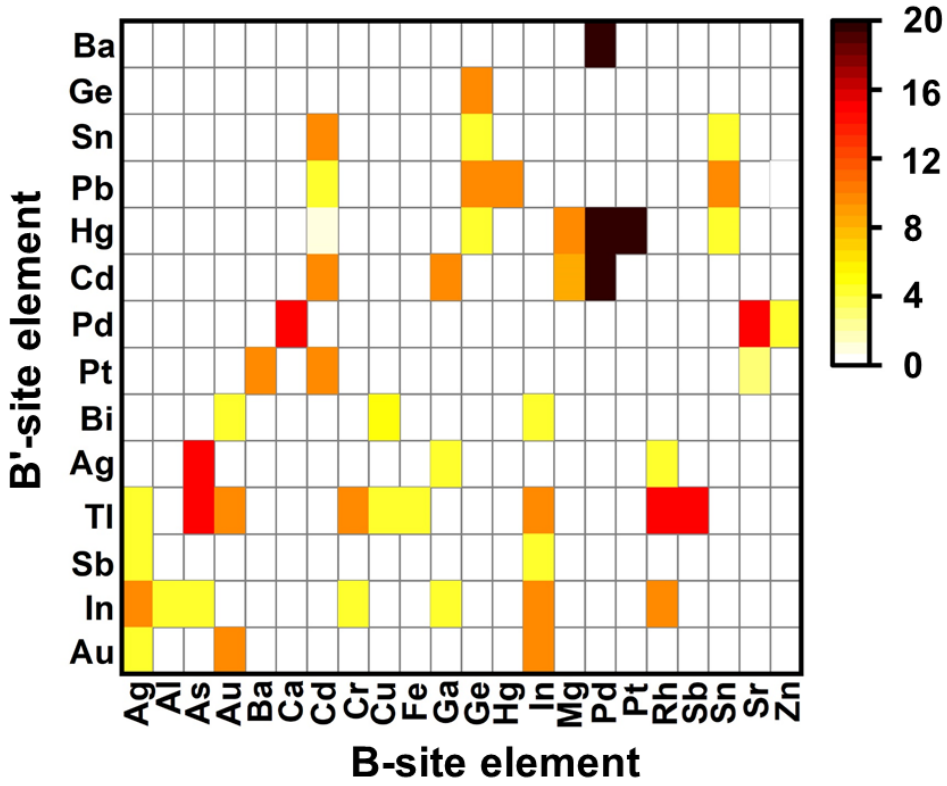

Figure S8. Number of optimal MDHOIPs based on different combinations of B- and B'-site ions.

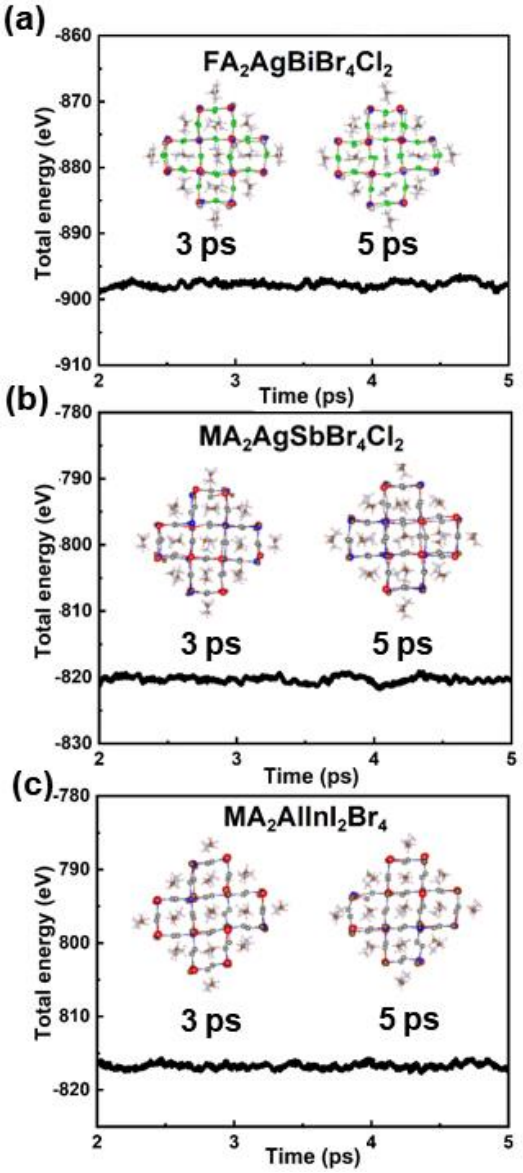

Figure S9. Total energy during 5 ps ab initio molecular dynamics (AIMD) simulations for (a) $FA_2AgBiBr_4Cl_2$, (b) $MA_2AgSbBr_4Cl_2$, and (c) $MA_2AlInI_2Br_4$.

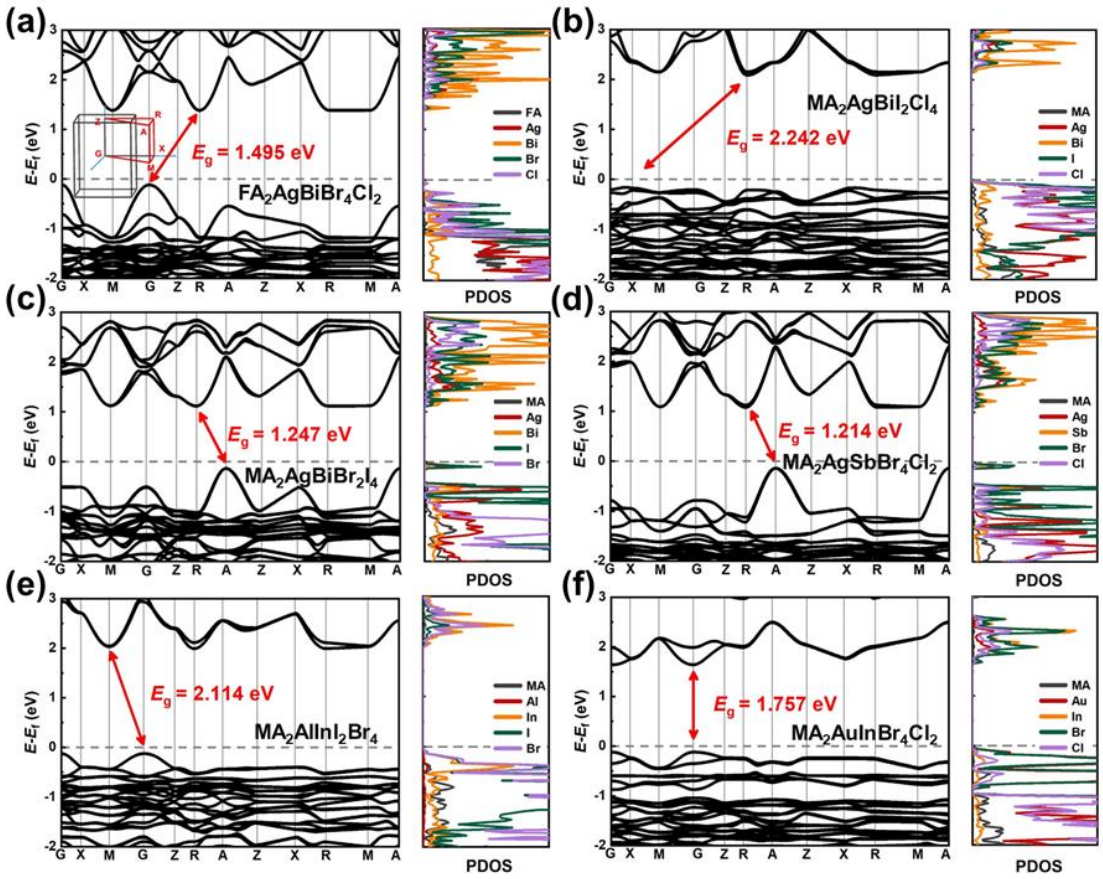

Figure S10. Calculated band structures and PDOS of (a) $FA_2AgBiBr_4Cl_2$, (b) $MA_2AgBiI_2Cl_4$, (c) $MA_2AgBiBr_2I_4$, (d) $MA_2AgSbBr_4Cl_2$, (e) $MA_2AlInI_2Br_4$, and (f) $MA_2AuInBr_4Cl_2$. Subgraph represents the Brillouin zone for tetragonal lattice.

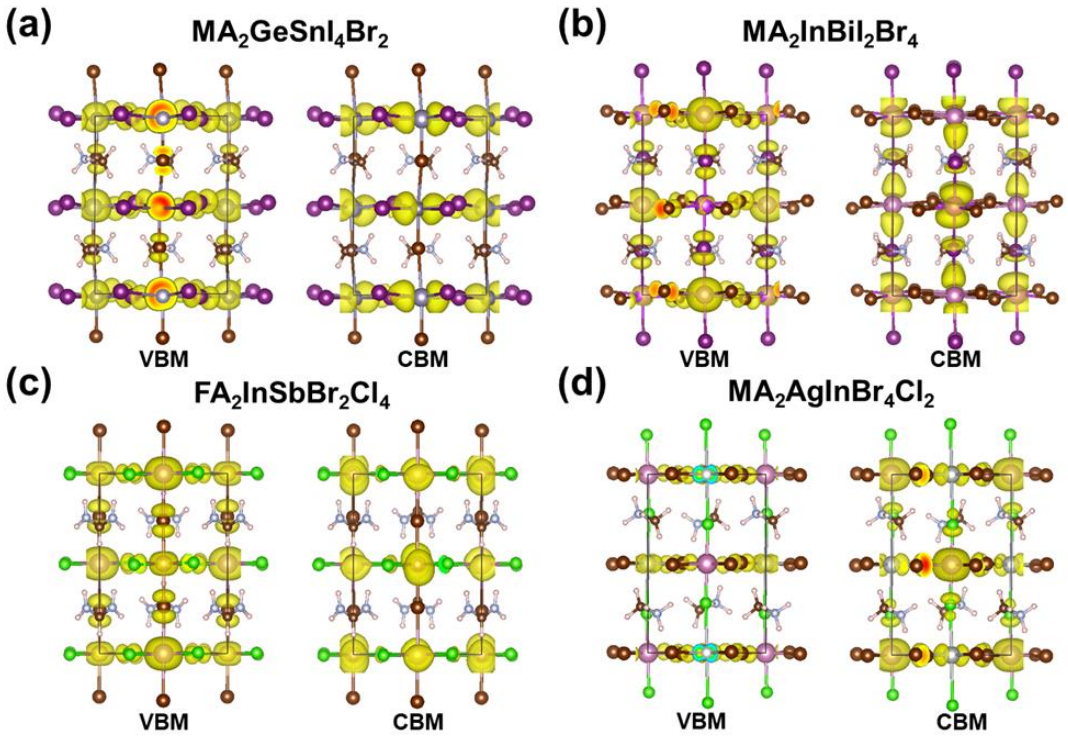

Figure S11. Band decomposed charge density of (a) $MA_2GeSnI_4Br_2$, (b) $MA_2InBiI_2Br_4$, (c) $FA_2InSbBr_2Cl_4$, and (d) $MA_2AgInBr_4Cl_2$.

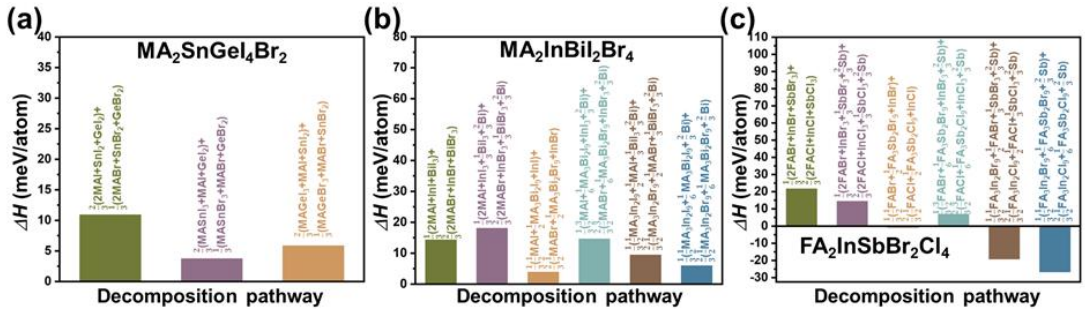

Figure S12. DFT-calculated decomposition energies of (a) $MA_2GeSnI_4Br_2$, (b) $MA_2InBiI_2Br_4$, and (c) $FA_2InSbBr_2Cl_4$.

Supplementary Tables

Table S1. Different elements with common valence states.

Valence	elements
+1	Ag, Au, Cu, Hg, In, Tl
+2	Ag, Ba, Be, Ca, Cd, Co, Cr, Cu, Fe, Ge, Hg, Mg, Mn, Ni, Pd, Pb, Pt, Sn, Sr, Ti, V, Zn
+3	Al, As, Au, B, Bi, Co, Cr, Fe, Ga, In, Ir, Mn, Mo, N, Nb, Ni, Rh, Ru, Sb, Sc, Ta, Ti, V, Y

Table S2. Eighty-seven initial features with description.

Feature	Description
P_A	Ionic polarizability of the A-site cations
χ_B, $\chi_{B'}$, χ_{X1}, χ_{X2} and χ_{X3}	Electronegativity of the B-, B'-, X_1, X_2 and X_3-site ions
IR_A, IR_B, $IR_{B'}$, IR_{X1}, IR_{X2} and IR_{X3}	Ionic radii of the A-, B-, B'-, X_1, X_2 and X_3-site ions
IR_{i+j}	Sum of two ionic radii
IR_{i-j}	Difference between two ionic radii
$IR_{i/j}$	Ratio between two ionic radii
χ_{i+j}	Sum of electronegativity of two ions
χ_{i-j}	Difference between electronegativity of two ions
$\chi_{i/j}$	Ratio between electronegativity of two ions

Table S3. Comparison between DFT-calculated and ML-predicted results.

System	E_g^{DFT} (eV) [a]	E_g^{ML1} (eV) [a]	E_g^{ML2} (eV) [a]
$FA_2InSbBr_2Cl_4$	0.985	1.013	1.011
$MA_2AgInBr_4Cl_2$	0.949	0.860	1.177
$MA_2AgSbBr_4Cl_2$	1.214	1.316	1.096
$MA_2AuInBr_4Cl_2$	1.757	1.609	1.613
$FA_2SnPbBr_2Cl_4$	1.507	1.442	1.401
$MA_2AsInCl_4I_2$	0.712	0.786	0.826

[a] E_g^{DFT} represents DFT-calculated bandgap values, and E_g^{ML1} and E_g^{ML2} represent ML-predicted bandgap values obtained by model-1 and model-2, respectively. In which model-1 is trained by all 525 DHOIPs, and model-2 is trained by 498 DHOIPs.

Table S4. Comparison between bandgaps from the database and our DFT results.

System	Database (eV)	Our PBE (eV)
$MAPbI_3$	1.74	1.74
MA_2AgBiI_6	1.34	1.23
$Cs_2AgBiCl_6$	1.80	1.82
MA_2AgSbI_6	1.04	0.92
$Cs_2AgInCl_6$	1.01	1.00

Supplementary Reference

1 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, and E. Duchesnay, *J. Mach. Learn. Res.*, 2011, **12,** 2825−2830.

2 S.; Lu, Q.; Zhou, L.; Ma, Y.; Guo, and J. Wang, *Small Methods*, 2019, **3,** 1900360.

3 S. Lu, Q. Zhou, Y. Ouyang, Y. Guo, Q. Li, and J. Wang, *Nat. Commun.*, 2018, **9,** 3405.

4 G. Te Velde, F. M. Bickelhaupt, E. J. Baerends, C. Fonseca Guerra, S. J. A. van Gisbergen, J. G. Snijders, and T. Ziegler, *J. Comput. Chem.*, 2001, **22,** 931-967.

5 L. M. Mentel, https://bitbucket.org/lukaszmentel/mendeleev, 2014.

6 M. R. Filip and F. Giustino, *Proc. Natl. Acad. Sci. U. S. A.*, 2018, **115,** 5397-5402.

7 G. Pilania, A. Ghosh, S. T. Hartman, R. Mishra, C. R. Stanek, and B. P. Uberuaga, *npj Comput. Mater.*, 2020, **6,** 71.