ARTICLE

DOI: 10.1038/s41467-018-05761-w OPEN

Accelerated discovery of stable lead-free hybrid organic-inorganic perovskites via machine learning

Shuaihua Lu[1], Qionghua Zhou[1], Yixin Ouyang[1], Yilv Guo[1], Qiang Li[1] & Jinlan Wang[1]

Rapidly discovering functional materials remains an open challenge because the traditional trial-and-error methods are usually inefficient especially when thousands of candidates are treated. Here, we develop a target-driven method to predict undiscovered hybrid organic-inorganic perovskites (HOIPs) for photovoltaics. This strategy, combining machine learning techniques and density functional theory calculations, aims to quickly screen the HOIPs based on bandgap and solve the problems of toxicity and poor environmental stability in HOIPs. Successfully, six orthorhombic lead-free HOIPs with proper bandgap for solar cells and room temperature thermal stability are screened out from 5158 unexplored HOIPs and two of them stand out with direct bandgaps in the visible region and excellent environmental stability. Essentially, a close structure-property relationship mapping the HOIPs bandgap is established. Our method can achieve high accuracy in a flash and be applicable to a broad class of functional material design.

[1] School of Physics, Southeast University, Nanjing 211189, China. These authors contributed equally: Shuaihua Lu, Qionghua Zhou. Correspondence and requests for materials should be addressed to J.W. (email: jlwang@seu.edu.cn)

The development of functional materials is the cornerstone of innovations in industry, and discovering materials with targeted property has always been a hotspot in science. The emergence of advanced techniques, such as high-throughput calculations based on density functional theory (DFT) has accelerated the search process at certain level[1–4]. However, the increasing scale of practical problems and complexity of materials require more sophisticated and effective methods for enormous database. Fortunately, the rapid development of material genome project[5] and artificial intelligence technology has brought exciting hope to this dilemma[6–8]. Most recently, machine learning (ML) technology has been made significant progress in the rational material design, such as efficient molecular organic light-emitting diodes[9], low thermal hysteresis shape memory alloys[10], piezoelectrics with large electrostrains[11] and so on. Bypassing complex quantum mechanics, ML technology can not only greatly accelerate materials design with high accuracy, but also learn trends within materials' basic composition from big material data.

While ML technology sheds light on in the field of material design on inorganic perovskites[12–15], the discovery of hybrid organic–inorganic perovskites (HOIPs) has never been reported yet in this way. HOIPs, as one of the most promising photovoltaic materials, have attracted tremendous interest recently. The most distinguished virtues of HOIPs includethe high power conversion efficiency (PCE), low-cost experimental synthesis and tunable bandgaps[16–20]. Since the first successful application of $CH_3NH_3PbX_3$ ($X = $ Cl, Br) with a PCE of 3.8% in 2009 by Kojima et al.[21], great efforts have continually been devoted to improve their PCE. Currently, the PCE of solar cells based on HOIPs has been boosted up to 22.1%[22]. Despite the great progress of HOIP-based solar cells, two key challenges limit the emerging materials for large scale commercial applications. One of the serious issues is toxicity, due to the element of lead (Pb), which contributes to most of the HOIP-based solar cells with high PCEs[23,24]. The other well-known factor is that their environmental stability is particularly poor, even with strict protection procedures. Therefore, it is of paramount importance to find stable Pb-free HOIPs with high PCE and sustainable air stability[25–28]. Unfortunately, due to the complexity and diversity of HOIPs structures (they are composed of organic molecules and inorganic metal frames), DFT-based high throughput calculations are too expensive and time consuming, not to mention experiments.

Here, we develop a target-driven method to discover stable Pb-free HOIPs based on ML technique and DFT calculations. We first train our ML model from 212 reported HOIPs' bandgap values, and predict the bandgaps of 5158 unexplored possible HOIPs. A close structure-property relationship mapping HOIPs' bandgap is concurrently excavated out from ML data, in which the ranges of tolerance factor, octahedral factor, metal electronegativity, and polarizability of organic molecules are defined for ideal HOIP-based solar cells. After further screening, six stable Pb-free HOIPs are selected as promising solar cells materials with proper bandgap.

Results

Design framework. Our multi-stage material design approach is schematically illustrated in Fig. 1, and the prediction engine consists of three integral components: input HOIPs data, ML algorithm, and DFT calculations. As a common ML procedure, an input dataset of HOIPs, each of which is described by features, is built for training and testing ML model. With that, feature engineering is needed in the first place to remove redundant features and establish a structure-property relationship. Once the input feature set is fixed, the best hyper-parameters will be selected using grid searching technique and five-cross-validation procedures (the selection details are given in Methods)[29]. After that, we apply the trained ML model to the prediction dataset. Finally, DFT calculations are performed to study the thermal and environmental stabilities and electronic properties of HOIPs candidates screened out from ML simulations.

Dataset. The input data for this study, containing 346 HOIPs, is obtained from previous high throughput first-principles calculations[30,31]. For data consistency and accuracy of ML predictions, we only select orthorhombic-like crystal structures with bandgap calculated using the Perdew-Burke-Ernzerh (PBE) functional. Therefore, 212 selected HOIP compounds are included in this engine, which completely belong to the family of perovskites of chemical formula ABX_3, as explicitly shown in Fig. 2a. In this structure, the halogen atoms X ($X = $ F, Cl, Br, I) occupy the vertices of regular corner-sharing BX_6 octahedron, while 32 different divalent metal cations B^{2+} sit at the center of the octahedron, and 11 kinds of monovalent cations A^+ fill in the cavity formed by the adjacent octahedrons. In the flow chart, we construct a dataset visualized in the form of plots between tolerance factor and bandgap E_g^{PBE} of HOIPs, in which they are divided into a training dataset (80%) and a test dataset (20%) after one thousand test (see Fig. 2b and Supplementary Fig. 1). As we can see from the data distribution, the input HOIPs dataset is made up of three parts: metals (zero bandgap), semiconductors (bandgap between 0 and 3.5 eV) and insulators (bandgap larger than 3.5 eV).

In fact, there are plenty of choices for the sites A and B in a HOIP. For A-site, we collect other 21 kinds of organic molecular cations A^+, all of which have been considered in the literature[20,32,33] (see Supplementary Fig. 2). Meanwhile, we substitute the B-site with 43 divalent cations across the Periodic Table. Consequently, 5504 different possible HOIP compounds (32 A-site cations, 43 B-site cations and 4 X-site anions) are obtained. Considering that 346 HOIPs have been studied, we explore the rest 5158 potential candidates in this work.

Feature engineering. For any ML method that targets toward a prespecified material property, it usually depends on a certain amount of features (descriptors). The features not only uniquely define each material in input dataset, but also relate to its desired physical and chemical properties. Although there may be many factors that affect the targeted property of materials, the number of features must be reasonable. The best strategy is to choose features that perfectly represent the materials' property and the number of features should be far less than the number of materials in input dataset to avoid the curse of dimensionality[34].

In this work, 30 initial features (the total features are in Supplementary Table. 1 and their sources are given in Supplementary Note. 1) such as ionic radii, tolerance factor and electronegativity are chosen to describe HOIPs in the chemical space collectively. In order to understand the relationship between features and targeted property, we evaluate the initial features via the gradient boosting regression (GBR) algorithm[35]. Furthermore, we incorporate a "last-place elimination" into the GBR algorithm to efficiently exclude the features that have less impact on the bandgap (the computational details are given in Methods). A detailed description of the feature selection procedure is given in and Supplementary Fig. 3 and Supplementary Note. 2. Finally, 14 most important features are sorted out and constitute as an optimal feature set. The new feature set contains structural features (tolerance factor T_f, octahedral factor O_f) as well as the elemental properties of A-, B- and X-site ions (total number of ionic charge IC_B, p orbital electron $X_{p\text{-electron}}$, ionization energy IE_B, electronegativity χ_B, electron affinity EA_B,

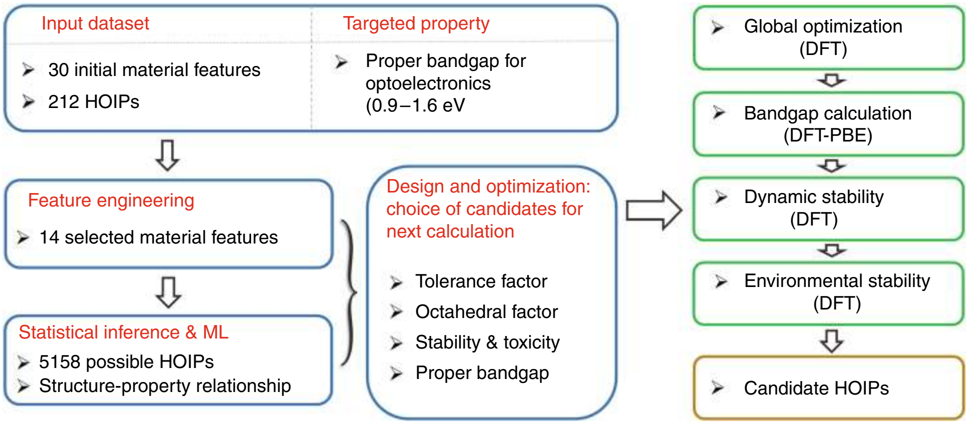
Input dataset
Targeted property
30 initial material features
212 HOIPs
Proper bandgap for optoelectronics (0.9−1.6 eV
Feature engineering
14 selected material features
Statistical inference & ML
5158 possible HOIPs
Structure-property relationship
Design and optimization: choice of candidates for next calculation
Tolerance factor
Octahedral factor
Stability & toxicity
Proper bandgap
Global optimization (DFT)
Bandgap calculation (DFT-PBE)
Dynamic stability (DFT)
Environmental stability (DFT)
Candidate HOIPs

Fig. 1 Lead-free HOIPs design framework. The material design framework combined with ML and DFT to efficiently search for stable Pb-free HOIPs with proper bandgap. The blue boxes represent the material screening process based on the ML algorithm generated from historical HOIP data. Then electronic properties and stability evaluation of these selected candidates are further calculated using DFT, which are shown in the green boxes

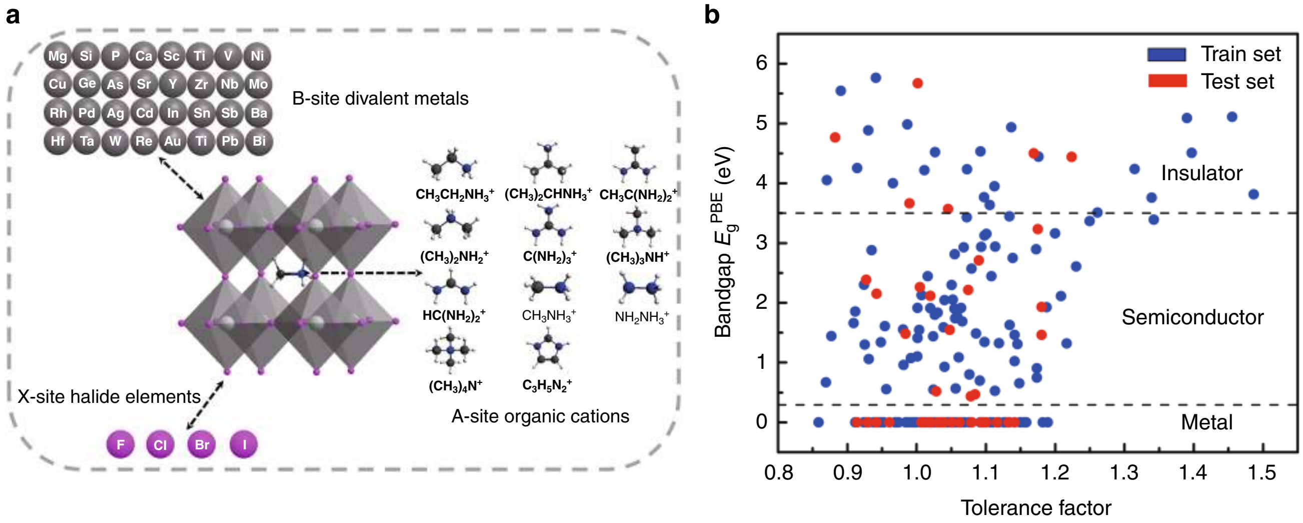

Fig. 2 HOIPs input dataset for training and testing. **a** 212 high throughput HOIP structures. The combination of 11 small organic molecular species (A-site) and 32 divalent metals (B-site) constitutes the input samples of our ML model. X is a typical halide. **b** Data visualization in training (blue dots) and test (red dots) of tolerance factor and bandgap $E_{\mathrm{g}}^{\mathrm{PBE}}$ of HOIPs. The entire dataset consists of metals, semiconductors, and insulators

ionic polarizability (P_{B}, P_{A}), sum of the s and p orbital radii $r_{\mathrm{s+p}\,\mathrm{B}}$, iron radii ($r_{\mathrm{B}}$, r_{A}), the highest occupied molecular orbital and the lowest unoccupied molecular orbital of A site cations ($\mathrm{HOMO_A}$, $\mathrm{LUMO_A}$).

As shown in Fig. 3a, the tolerance factor T_{f} plays the most important role to HOIPs' bandgap, followed by the total number of ionic charge $\mathrm{IC_B}$ and the octahedral factor O_{f}. Interestingly, the properties of B-site elements such as ionization energy $\mathrm{IE_B}$, electronegativity χ_{B} and electron affinity $\mathrm{EA_B}$ show greater influence on the bandgap of HOIPs than those of A- and X-site ions. Pearson correlation coefficient matrices are calculated to identify the positive and negative correlations between pairs of features (Fig. 3b). The low linear correlations for most of features indicate that we have successfully removed redundant and irrelevant features, which will significantly improve the performance of the ML model.

Model inference. In ML method, an appropriate ML algorithm is important. Currently, several supervised ML regression algorithms have been successfully used in material science, such as GBR[32], artificial neural network[36,37], and kernel ridge regression (KRR)[38]. These regression algorithms provide both material property prediction with DFT accuracy and atomic level chemical insights. In this work, we employ six different ML regression algorithms, i.e., GBR, KRR, support vector regression, gaussian process regression, decision trees regression, and multilayer perceptron regression. Each training model is based on a subset of the whole data, known as training data, and the model will be used to predict other new data after training. To evaluate the performance of each ML model, three indexes are chosen to estimate the prediction errors: coefficient of determination (R^2), Pearson coefficient (r), and mean square error (MSE) (the computational details are given in Supplementary Methods). By comparing the three indexes, GBR algorithm reproduces best agreement to the true bandgap values (see details in Supplementary Fig. 4 and 5). Then, we perform a statistical test on R^2 values from 10,000 executions of each model at the 95% confidence level (see details in Supplementary Table. 2 and 3). Evident differences are observed between GBR and other five algorithms. Furthermore, we put standard deviations on the R^2 and MSE values. It is found that GBR algorithm presents more reliable results than other five algorithms (see details in Supplementary Fig. 6 and Supplementary Note. 3). Additionally, the

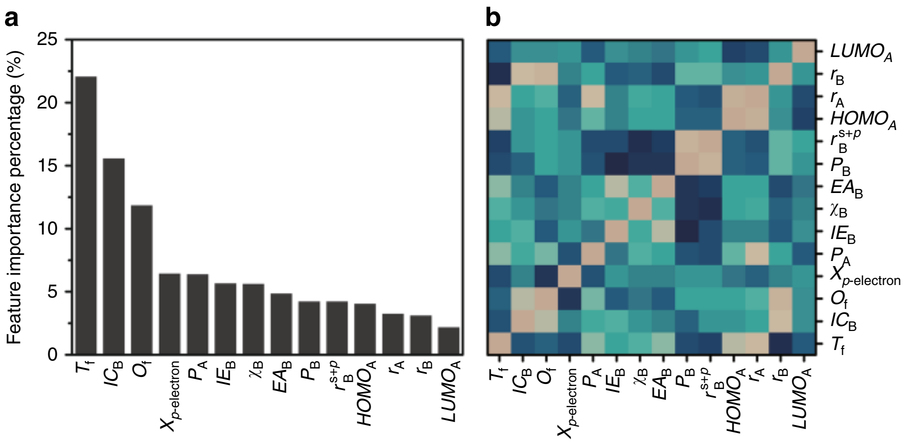

Fig. 3 Importance and correlation of the selected features. **a** The 14 selected features are ranked using GBR algorithm. **b** The heat map of Pearson correlation coefficient matrix among the selected features for HOIPs

GBR algorithm evolves from the combination of boosting methods and regression trees, which makes it suitable for effectively mining features and feature engineering[39]. Therefore, GBR is chosen to establish a nonlinear mapping between the input features and bandgaps and subsequently predicts bandgaps of unexplored HOIPs.

The test results of the GBR model are presented in Fig. 4a. The subplot clearly shows that the training/test set deviance declines gradually with the increase of the boosting iteration numbers. Eventually, R^2, r, and MSE of test data is 97.0%, 98.5%, and 0.086, respectively, witnessing the outstanding performance of our GBR model. Then, the trained GBR model is applied to the 5158 HOIPs to predict their bandgaps, and the prediction results (dark gray dots) are illustrated in Fig. 4b, together with train dataset (blue dots) and test dataset (red dots). We notice that the distribution of post predicted bandgaps is very close to the original input dataset. This proves the reasonability and reliability of our ML model, providing guarantee for further analysis.

Furthermore, data analysis and visualization are performed to unravel hidden trends and periodicities within the HOIPs data. We divide the predicted dataset into four parts according to the X-site elements, i.e., F, Cl, Br, and I in Fig. 4c-f. As seen from the figure, the bandgap of HOIPs tends to increase as the X-site halogen radius reduces and the bandgaps of F- and Cl- HOIPs are too large for photovoltaics applications. Therefore, HOIPs containing Br and I ($ABBr_3$ and ABI_3) with promising prospect for photovoltaic applications are mainly focused in the following discussion. Considering the structural stability for HOIPs, the given features should be restricted in certain ranges. Specifically, tolerance factor T_f, which has been used extensively to predict the stability of the perovskite structure, should be between 0.8 and 1.2 (Fig. 4c). In terms of the octahedral factor O_f, the appropriate range for solar cells is between 0.2 and 0.7 (Fig. 4d). However, the octahedral factor should not be too small for structural stability, so the values are better ranging from 0.4 to 0.7[33,40,41]. Moreover, in order to design HOIPs with proper bandgap, we should select weak polarized organic molecules and the polarization P_A varies between 1×10^{-40} and 5×10^{-40} C m^2 V^{-1} (see Fig. 4e). Finally, the electronegativity of B-site also plays important roles on the bandgap and it needs to be within the range from 1.4 to 2.2 (Fig. 4f). We also analyze the mapping between other eight features with the bandgaps of HOIPs and less obvious correlations are observed (see Supplementary Fig. 7).

Model validation. We have predicted bandgaps E_g for all the possible HOIP structures in the search space via ML technology. To discover stable HOIPs, further screening of the predicted dataset is necessary. 1669 HOIPs are first screened out from the total 5158 HOIPs with ML predicted bandgaps according to the structural stability (T_f between 0.8 and 1.2, O_f between 0.4 and 0.7). These 1669 HOIPs are likely stable and have different potential applications in light of their bandgap values. For examples, HOIPs with small bandgaps (less than 0.9 eV) can be used in infrared sensors[42,43] and large gap HOIPs (larger than 3 eV) may serve as good insulating materials[44,45]. For solar cells, HOIPs with bandgap between 0.9 and 1.6 eV are ideal candidates[46,47]. Therefore, 218 HOIPs with proper bandgap are selected (see full list in Supplementary Table. 4). Since the Br-based HOIPs are more accessible in experiment[19], 22 Br-based HOIPs (i.e., $ABBr_3$) are further selected. Additionally, toxicity of HOIPs will block widespread commercial application and the compounds containing toxic metal elements are excluded as well. To the end, six orthorhombic HOIPs stand out and further thermal and environmental stability evaluation and electronic property exploration are performed by first-principle calculations. The step-by-step screening process are shown in Supplementary Fig. 8 and Supplementary Note 4.

A comparison between ML-predicted and DFT-calculated results of six selected HOIPs is presented in Fig. 5a, with relevant statistics summarized in Table 2. Excellent agreement (ΔE_g no larger than 0.1 eV) is found between the ML predicted and DFT calculated bandgap values, verifying the great superiority of the current ML technology. It takes only a few seconds for all the 5158 HOIPs' bandgaps to be predicted by the ML method. However, if the DFT calculation is adopted, it will take a few days for each HOIP structure. Therefore, we conclude that our current ML scheme provides a possibility of achieving DFT accuracy in a flash and has great priority in complex systems like HOIPs.

Electronic structures of six selected HOIPs. DFT optimizations show that all these six HOIPs hold typical perovskite structures, and corresponding lattice constants are listed in Table 1. The nonbonding electrons of B-site ions lower the coordination symmetry around the cations and lead to the slightly distorted BX_6 octahedral. Sorted by B site metal, the six candidates can be divided into two groups: four $AInBr_3$ and two $ASnBr_3$. Further electronic band structures show that the four indium HOIPs have indirect bandgap between Γ and M/R point in the Brillouin zone (Supplementary Fig. 9 and 10), while the other two tin HOIPs have direct bandgap at Γ point (Fig. 5c). Additionally, the spin-orbital-coupling (SOC) effect is considered for the six selected HOIPs (see details in Supplementary Fig. 11 and Supplementary

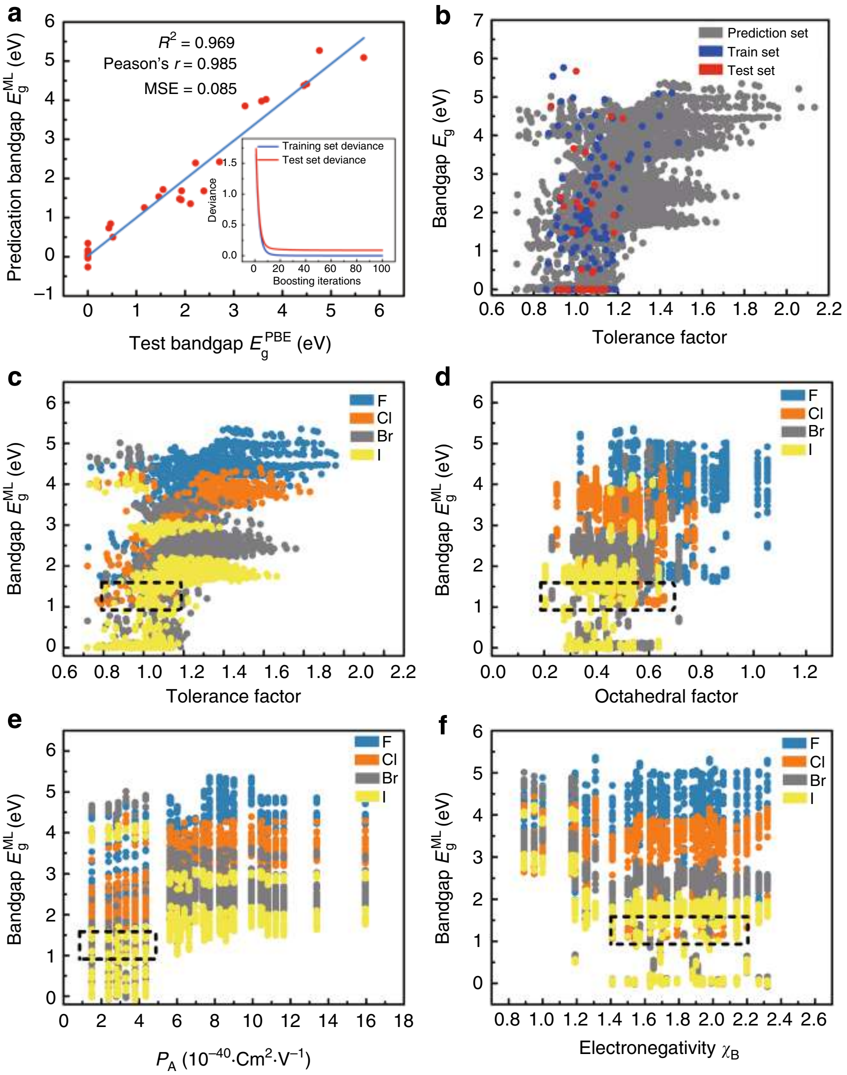

Fig. 4 Results and insights from ML model. **a** The fitting results of test bandgaps E_g^{PBE} and predicted bandgaps E_g^{ML}. Coefficient of determination (R^2), Pearson coefficient (r) and mean squared error (MSE) are computed to estimate the prediction errors. The subplot is the convergence of model accuracy for five cross-validation split of the data. **b** Scatter plots of tolerance factors against the bandgaps for the prediction dataset from trained ML model (blue, red and dark gray plots represent train, test and prediction set, respectively). Data visualization of predicted bandgaps for all possible HOIPs (one color represents a class of halogen perovskites) with (**c**) tolerance factor, (**d**) octahedral factor, (**e**) ionic polarizability for the A-site ions, and (**f**) electronegativity of B-site ions. The dotted box represents the most appropriate range for each feature

Note 5) and the influence to the band structure is not pronounced.

As typical electronic structures of $ASnBr_3$ and $AInBr_3$, the valence band maximums are mainly contributed by the p orbital of the halogen atoms and partly contributed by the s orbital of the metal atom, while the conduction band minimums are dominated by p orbitals of the metal atom (Fig. 5d and Supplementary Fig. 9 and 10). In fact, as a result of the indirect bandgap of $AInBr_3$, the absorption coefficient will be relatively low. Therefore, a relatively thick absorbent layer is required, which definitely increases the material costs, especially with the high cost of rare metal indium. Hence, two $ASnBr_3$ HOIPs are better choices as light-harvesting materials in photovoltaic devices with suitable direct bandgaps and relatively low material costs.

Thermal and environmental stabilities of six selected HOIPs. The thermal and environmental stabilities are important to the practical application of HOIPs, and most reported structures with high PCEs suffer from degradation in ambient. We first perform ab initio molecular dynamics (AIMD) simulations to evaluate the thermal stability of the six selected HOIPs. As shown in Fig. 5e and Supplementary Fig. 9 and 10, the time-dependent evolutions of total energies are oscillating within a very narrow range, indicating that these HOIPs can maintain their structural integrity at room temperature. In the next step of evaluating the environmental stabilities, the adsorption energies (ΔE_{ads}) of molecular water and oxygen on (001) surfaces of these six HOIPs are calculated and listed in Table 2 (the computational details can be found in Supplementary Fig. 12). Comparing with $MAPbI_3$ whose adsorption energies of water and oxygen are −0.48 and −0.15 eV, $C_2H_5OInBr_3$, $C_2H_5OSnBr_3$, and $C_2H_6NSnBr_3$ show better environmental stability against both oxygen and water. Although the degradation mechanism of HIPO is still under discussion, it is widely accepted that water acts as the reactive source for the collapse of the framework. We attribute the significant reduction in adsorption energy of the three systems for water to the weaker polarity of $C_2H_5O^+$ and $C_2H_6N^+$ comparing with MA^+ radical, which present strong interaction with water

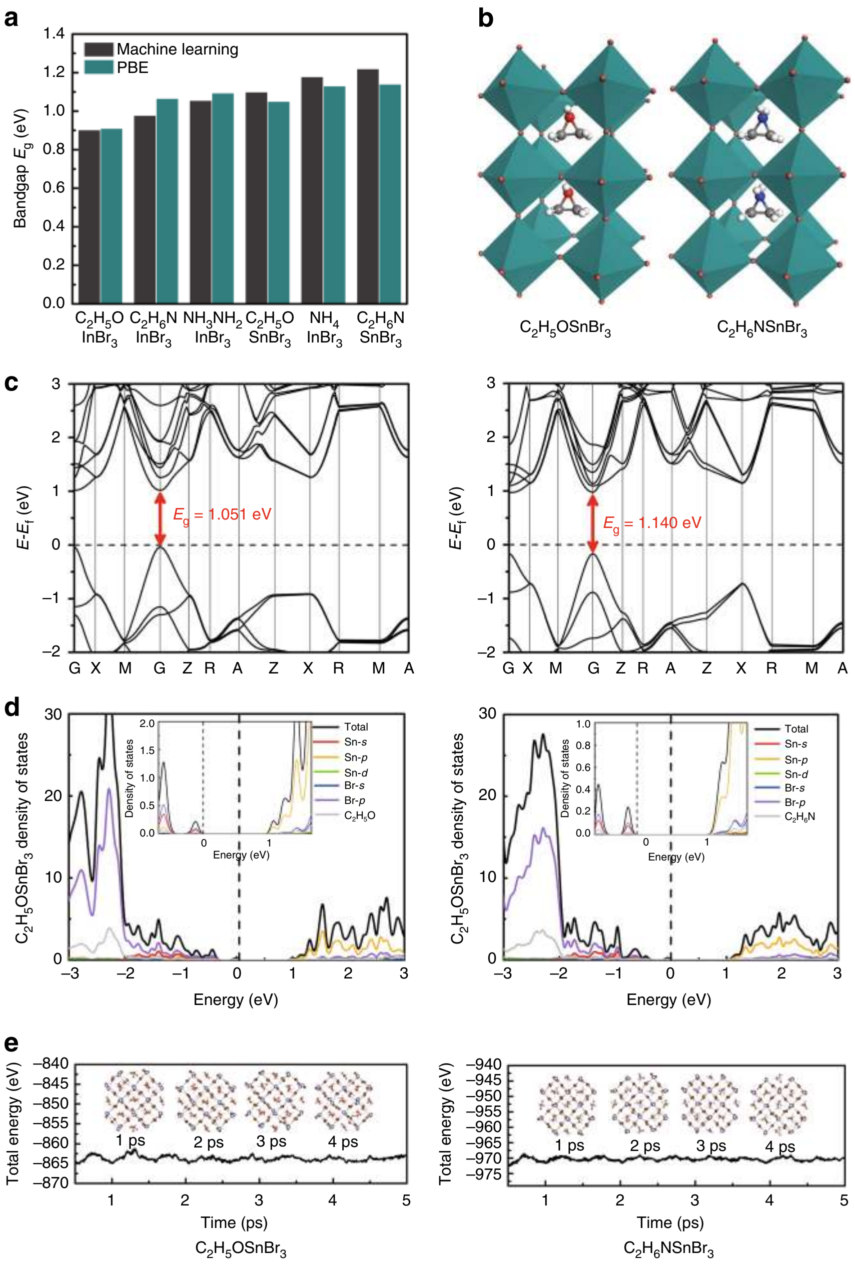

Fig. 5 Comparison with DFT calculations. **a** A comparison between ML-predicted and DFT-calculated results of six selected HOIPs. **b** Optimized structures, **c** band structures, **d** projected density of states (PDOS), **e** total energy during 5 ps AIMD simulations for $C_2H_5OSnBr_3$ and $C_2H_6NSnBr_3$ at 300 K. The subplots in **d** are the PDOS near the Fermi level

Table 1 Lattice constants of six selected HOIPs

HOIPs	a (Å)	b (Å)	c (Å)	α (°)	β (°)	γ (°)
$C_2H_5OInBr_3$	8.44	11.46	8.48	91.63	91.76	91.58
$C_2H_6NInBr_3$	8.50	11.7	8.12	92.17	90.35	88.52
$NH_3NH_2InBr_3$	8.33	11.68	7.67	88.71	89.73	88.69
$C_2H_5OSnBr_3$	8.52	11.60	8.52	91.32	89.74	89.88
NH_4InBr_3	7.92	11.53	7.88	90.42	90.00	90.43
$C_2H_6NSnBr_3$	8.63	11.95	8.26	91.78	89.42	88.58

a, *b*, and *c* are lattice length. *α*, *β*, and *γ* are lattice angle

Table 2 Six selected HOIPs with relevant statistics

HOIPs	T_f	O_f	$E^{ML}{}_g$(eV)	E_g^{PBE} (eV)	$\Delta E_{H_2O\,ads}$ (eV)	$\Delta E_{O_2\,ads}$ (eV)
$C_2H_5OInBr_3$	1.04	0.50	0.90	0.91	−0.301	−0.071
$C_2H_6NInBr_3$	1.04	0.50	0.97	1.07	−0.630	−0.152
$NH_3NH_2InBr_3$	0.99	0.50	1.06	1.09	−0.497	−0.110
$C_2H_5OSnBr_3$	0.99	0.57	1.10	1.05	−0.163	−0.071
NH_4InBr_3	0.82	0.50	1.18	1.13	−0.566	−0.090
$C_2H_6NSnBr_3$	0.99	0.57	1.22	1.14	−0.134	−0.093

T_f and O_f is tolerance factor and octahedral factor respectively. $E_g^{ML}$and E_g^{PBE} are ML-predicted and DFT-calculated results respectively. ΔE_g is the absolute value of the difference between E_g^{ML} and E_g^{PBE}. $\Delta E_{H_2}O$ ads and ΔE_{O_2} ads is the adsorption energy of H_2O and O_2, respectively through hydrogen bond. Therefore, the weak hygroscopicity of $C_2H_5OInBr_3$, $C_2H_5OSnBr_3$, and $C_2H_6NSnBr_3$ suggests that the H_2O molecules are reductant to aggregate on the surface, and increase the energy barrier of the H_2O penetration process in the meantime, preventing the hydration degradation effectively.

Discussion

Combining with ML technology and DFT calculations, we have developed an extremely fast target-driven method to discover HOIPs. Three stable Pb-free HOIPs with proper bandgaps and excellent thermal and environmental stabilities have been successfully selected out from 5158 HIOPs for solar cells. More importantly, a close structure-property map of HOIPs has been well established from the ML predicted big dataset and four stringent conditions in terms of tolerance factor, octahedral factor, electronegativity of metal ions, and polarizability of organic molecules are defined to be ideal HOIP-based solar cells.

Differently from those high-throughput screening methods which the whole chemical space should be searched at DFT level, the current ML and DFT combined scheme only needs to compute the most promising HOIPs at DFT level, which greatly saves the computational resources. Note that the screenings described above are very strict, and in fact, the screening conditions can be adjusted according to targeted goal to find suitable candidates for experimental synthesis. The target-driven method we developed here overcomes a major obstruction in traditional trial-error method. Meanwhile, as this ML technology employs a "last-place elimination" feature selection procedure based on GBR algorithm, it can not only achieve DFT accuracy in a flash (even faster than the popular neural network algorithm), but also works with a small dataset. This means we can achieve accurate prediction with a relatively small training data. Here we only apply this intelligent method to accurately predict the bandgaps of thousands of HOIPs and get over the problem of toxicity and poor environmental stability in HOIPs. In fact, it is applicable to other functional material design and discovery, if the computational or experimental material data are enough to train the ML model.

Methods

Gradient boosted regression. GBR[35,48], a flexible non-parametric statistical machine leaning algorithm in the open-source *scikit-learn* package[49], is implemented to predict bandgaps of undiscovered HOIPs. The learning principle of this method is to improve the accuracy of the final regression results by gradually reducing the algorithm generated by the training process. The final regression algorithm is the weighted sum of several weak regression algorithms obtained by each training, as

$$F_M(x) = \sum_{m=1}^{M} T(x, \theta_m), \tag{1}$$

where m is the times of training, x is the input data, and θ_m is the distribution weight vector. The model is trained M times, and each time it produces a weak regression function T. The loss function of every weak classifier, is defined as

$$\hat{\theta}_m = \arg\min_{\theta_m} \sum_{i=1}^{N} L(y_i, F_{m-1}(x_i) + T(x_i, \theta_m)), \tag{2}$$

where $F_{m\text{-}1}(x_i)$ is the current model, and GBR determines the parameters of the next weak classifier through empirical risk minimization. This work uses ML to analyze a small dataset based on DFT calculation to construct a predictive model.

Hyper-parameters selection. In the ML algorithm, the hyper-parameter is the parameter set before the learning process, rather than that obtained through training model. In general, it is necessary to select a set of optimal hyper-parameters for the learning machine to improve the efficiency and generalization performance of the model. Here, six hyper-parameters in the GBR model are optimized by grid searching method: loss function (least squares), learning rate (0.2), maximum depth of the individual regression estimators (12), the number of features to consider when looking for the best split (0.7), the minimum number of samples required to be at a leaf node (3) and the number of boosting stages to perform (100). The values in parentheses represent the best results for each hyper-parameter.

Last-place elimination feature selection procedure. We employ a "last-place elimination" feature selection procedure into GBR algorithm to optimize the most relevant features. Thirty initial features, which are commonly employed in ML algorithm or structure features of HOIPs, are considered and first ranked by GBR algorithm according to the relative importance. Then we remove the least important feature (i.e., the 30th feature) out of the feature set and the rest 29 features constitute a new dataset for the next step feature selection. After that, we rank the remaining features again and repeat the above step. Finally, we train the ML model using different datasets for twenty-nine times (see workflow in Supplementary Fig. 3). We record the model score (R^2) of each trained model and the ML model performs best when the feature set only includes fourteen features.

Density functional theory. All DFT calculations for selected HOIPs are carried out using the projector-augmented wave method with the generalized gradient approximation, implemented in the Vienna Ab initio Simulation Package package[50]. The exchange-correlation functional is described by PBE[51] functional considering it reproduces more consistent results with the experiments for HOIPs due to fortuitous error–error offset[52,53]. DFT-D3 method is adopted for the van der Waals correction[54]. AIMD simulations are performed at room temperature by using the Nosé-Hoover method[55,56] to verify the thermal stability of selected materials. The environmental stability of selected HOIPs are further evaluated by the adsorption energy calculations. More DFT calculation details can be found in Supplementary Methods.

Data availability. The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Received: 27 April 2018 Accepted: 7 July 2018
Published online: 24 August 2018

References

1. Curtarolo, S. et al. The high-throughput highway to computational materials design. *Nat. Mater.* **12**, 191–201 (2013).
2. Chakraborty, S. et al. Rational design: a high-throughput computational screening and experimental validation methodology for lead-free and emergent hybrid perovskites. *ACS Energy Lett.* **2**, 837–845 (2017).
3. Kuhar, K. et al. Sulfide perovskites for solar energy conversion applications: computational screening and synthesis of the selected compound $LaYS_3$. *Energy Environ. Sci.* **10**, 2579–2593 (2017).
4. Mounet, N. et al. Two-dimensional materials from high-throughput computational exfoliation of experimentally known compounds. *Nat. Nanotech.* **13**, 246–252 (2018).
5. Materials Genome Initiative for Global Competitiveness. https://www.whitehouse.gov/sites/default/files/microsites/ostp/materials_genome_initiativefinal.pdf (2011).
6. Le, T., Epa, V. C., Burden, F. R. & Winkler, D. A. Quantitative structure-property relationship modeling of diverse materials properties. *Chem. Rev.* **112**, 2889–2919 (2012).
7. Tabor, D. P. et al. Accelerating the discovery of materials for clean energy in the era of smart automation. *Nat. Rev. Mater.* **3**, 5–20 (2018).
8. Balachandran, P. V., Kowalski, B., Sehirlioglu, A. & Lookman, T. Experimental search for high-temperature ferroelectric perovskites guided by two-step machine learning. *Nat. Commun.* **9**, 1668 (2018).
9. Gomez-Bombarelli, R. et al. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nat. Mater.* **15**, 1120–1127 (2015).
10. Xue, D. et al. Accelerated search for materials with targeted properties by adaptive design. *Nat. Commun.* **7**, 11241 (2016).
11. Yuan, R. et al. Accelerated discovery of large electrostrains in $BaTiO_3$ -based piezoelectrics using active learning. *Adv. Mater.* **30**, 1702884 (2018).
12. Pilania, G. et al. Machine learning bandgaps of double perovskites. *Sci. Rep.* **6**, 19375 (2016).
13. Geoffroy, H. et al. Finding nature's missing ternary oxide compounds using machine learning and density functional theory. *Chem. Mater.* **22**, 3762–3767 (2010).
14. Seko, A., Hayashi, H., Nakayama, K., Takahashi, A. & Tanaka, I. Representation of compounds for machine-learning prediction of physical properties. *Phys. Rev. B* **95**, 144110 (2017).
15. Schmidt, J. et al. Predicting the thermodynamic stability of solids combining density functional theory and machine learning. *Chem. Mater.* **29**, 5090–5103 (2017).

16. Ball, J. M. & Petrozza, A. Defects in perovskite-halides and their effects in solar cells. *Nat. Energy* **1**, 16149 (2016).
17. Li, F. et al. Ambipolar solution-processed hybrid perovskite phototransistors. *Nat. Commun.* **6**, 8238 (2015).
18. Chen, B., Zheng, X., Bai, Y., Padture, N. P. & Huang, J. Progress in tandem solar cells based on hybrid organic–inorganic perovskites. *Adv. Energy Mater.* **7**, 1602400 (2017).
19. Huang, J., Yuan, Y., Shao, Y. & Yan, Y. Understanding the physical properties of hybrid perovskites for photovoltaic applications. *Nat. Rev. Mater.* **2**, 17042 (2017).
20. Li, W. et al. Chemically diverse and multifunctional hybrid organic–inorganic perovskites. *Nat. Rev. Mater.* **2**, 16099 (2017).
21. Kojima, A. et al. Organometal halide perovskites as visible-light sensitizers for photovoltaic cells. *J. Am. Chem. Soc.* **131**, 6050–6051 (2009).
22. Yang, W. S. et al. Iodide management in formamidinium-lead-halide-based perovskite layers for efficient solar cells. *Science* **356**, 1376–1379 (2017).
23. Noel, N. K. et al. Lead-free organic–inorganic tin halide perovskites for photovoltaic applications. *Energ. Environ. Sci.* **7**, 3061–3068 (2014).
24. Le, T. C. & Winkler, D. A. Discovery and optimization of materials using evolutionary approaches. *Chem. Rev.* **116**, 6107–6132 (2016).
25. Ju, M. G., Dai, J., Ma, L. & Zeng, X. C. Lead-free mixed tin and germanium perovskites for photovoltaic application. *J. Am. Chem. Soc.* **139**, 8038–8043 (2017).
26. Shi, Z. et al. Lead-free organic–inorganic hybrid perovskites for photovoltaic applications: recent advances and perspectives. *Adv. Mater.* **29**, 1605005 (2017).
27. Nie, R. et al. Mixed sulfur and iodide-based lead-free perovskite solar cells. *J. Am. Chem. Soc.* **140**, 872–875 (2018).
28. Zhao, X. G. et al. Design of lead-free inorganic halide perovskites for solar cells via cation-transmutation. *J. Am. Chem. Soc.* **139**, 2630–2638 (2017).
29. Boser, B. E., Guyon, I. M. & Vapnik, V. N. Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory (ACM Press, New York, 1992)
30. Nakajima, T. & Sawada, K. Discovery of Pb-free perovskite solar cells via high-throughput simulation on the K computer. *J. Phys. Chem. Lett.* **8**, 4826–4831 (2017).
31. Kim, C., Huan, T. D., Krishnan, S. & Ramprasad, R. A hybrid organic–inorganic perovskite dataset. *Sci. Data* **4**, 170057 (2017).
32. Kieslich, G., Sun, S. & Cheetham, A. K. An extended tolerance factor approach for organic–inorganic perovskites. *Chem. Sci.* **6**, 3430–3433 (2015).
33. Becker, M., Klüner, T. & Wark, M. Formation of hybrid ABX$_3$ perovskite compounds for solar cell application: first-principles calculations of effective ionic radii and determination of tolerance factors. *Dalton Trans.* **46**, 3500–3509 (2017).
34. Nasrabadi, N. M. *Pattern Recognition and Machine Learning CH.* vol. 14 (Springer Press, New York, 2016).
35. Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Ann. Stat.* **29**, 1189–1232 (2001).
36. Li, Z., Wang, S., Chin, W. S., Achenie, L. E. & Xin, H. High-throughput screening of bimetallic catalysts enabled by machine learning. *J. Mater. Chem. A* **5**, 24131–24138 (2017).
37. Janet, J. P., Chan, L. & Kulik, H. J. Accelerating chemical discovery with machine learning: simulated evolution of spin crossover complexes with an artificial neural network. *J. Phys. Chem. Lett.* **9**, 1064–1071 (2018).
38. Zhuo, Y., Tehrani, A. M. & Brgoch, J. Predicting the band gaps of inorganic solids by machine learning. *J. Phys. Chem. Lett.* **9**, 1668–1673 (2018).
39. Isayev, O. et al. Universal fragment descriptors for predicting properties of inorganic crystals. *Nat. Commun.* **8**, 15679 (2017).
40. Hoefler, S. F., Trimmel, G. & Rath, T. Progress on lead-free metal halide perovskites for photovoltaic applications: a review. *Monash. Chem.* **148**, 795–826 (2017).
41. Travis, W. et al. On the application of the tolerance factor to inorganic and hybrid halide perovskites: a revised system. *Chem. Sci.* **7**, 4548–4556 (2016).
42. Dou, L., Liu, Y., Hong, Z., Li, G. & Yang, Y. Low-Bandgap near-IR conjugated polymers/molecules for organic electronics. *Chem. Rev.* **115**, 12633–12665 (2015).
43. Lin, C., Grassi, R., Low, T. & Helmy, A. S. Multilayer black phosphorus as a versatile mid-infrared electro-optic material. *Nano. Lett.* **16**, 1683–1689 (2016).
44. Liu, Q., Zhang, X., Abdalla, L. B., Fazzio, A. & Zunger, A. Switching a normal insulator into a topological insulator via electric field with application to phosphorene. *Nano. Lett.* **15**, 1222–1228 (2015).
45. Usman, M., Mendiratta, S. & Lu, K. L. Semiconductor metal-organic frameworks: future low-bandgap materials. *Adv. Mater.* **29**, 201605071 (2017).
46. Shockley, W. & Queisser, H. J. Detailed balance limit of efficiency of p–n junction solar cells. *J. Appl. Phys.* **32**, 510–519 (1961).
47. Sun, Q., Wang, J., Yin, W. J. & Yan, Y. Bandgap engineering of stable lead-free oxide double perovskites for photovoltaics. *Adv. Mater.* **30**, 1705901 (2018).
48. Friedman, J. Stochastic gradient boosting. *Comput. Stat. Data. Anal.* **38**, 367–378 (2009).
49. Machine Learning in Python. http://scikit-learn.org/stable/modules/ ensemble.html# regression (2007).
50. Kresse, G. & Furthmüller, J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Comput. Mater. Sci.* **6**, 15–50 (1996).
51. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865 (1996).
52. Motta, C. et al. Revealing the role of organic cations in hybrid halide perovskite CH$_3$NH$_3$PbI$_3$. *Nat. Commun.* **6**, 7026 (2015).
53. Colella, S. et al. MAPbI$_{3-x}$Cl$_x$ mixed halide perovskite for hybrid solar cells: the role of chloride as dopant on the transport and structural properties. *Chem. Mater.* **25**, 4613–4618 (2013).
54. Lee, K., Murray, É. D., Kong, L., Lundqvist, B. I. & Langreth, D. C. Higher-accuracy van der Waals density functional. *Phys. Rev. B* **82**, 081101 (2010).
55. Nose, S. A unified formulation of the constant temperature molecular dynamics methods. *J. Chem. Phys.* **81**, 511–519 (1984).
56. Hoover, W. G. Canonical dynamics: equilibrium phase-space distributions. *Phys. Rev. A* **31**, 1695–1697 (1985).

Acknowledgements

This work is supported by the National Key Research and Development Program of China (2017YFA0204800), Natural Science Foundation of China (21525311, 21773027), Jiangsu 333 project (BRA2016353), and the Fundamental Research Funds for the Central Universities of China. The authors thank the computational resources at the SEU and National Supercomputing Center in Tianjin.

Author contributions

J.W. conceived the project. S.L. and Q.Z. contributed equally to this work. S.L. did ML prediction and Q.Z. carried out DFT calculations. S.L., Q.Z., and J.W co-wrote the paper with all authors contributing to the discussion and preparation of the manuscript.

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-018-05761-w.

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.