npj | Computational Materials

www.nature.com/npjcompumats

ARTICLE OPEN

Identifying Pb-free perovskites for solar cells by machine learning

Jino Im[1], Seongwon Lee[2], Tae-Wook Ko[2], Hyun Woo Kim[1], YunKyong Hyon[2] and Hyunju Chang[1]

Recent advances in computing power have enabled the generation of large datasets for materials, enabling data-driven approaches to problem-solving in materials science, including materials discovery. Machine learning is a primary tool for manipulating such large datasets, predicting unknown material properties and uncovering relationships between structure and property. Among state-of-the-art machine learning algorithms, gradient-boosted regression trees (GBRT) are known to provide highly accurate predictions, as well as interpretable analysis based on the importance of features. Here, in a search for lead-free perovskites for use in solar cells, we applied the GBRT algorithm to a dataset of electronic structures for candidate halide double perovskites to predict heat of formation and bandgap. Statistical analysis of the selected features identifies design guidelines for the discovery of new lead-free perovskites.

npj Computational Materials (2019)5:37 ; https://doi.org/10.1038/s41524-019-0177-0

INTRODUCTION

Identifying optimal materials in applications research is a time-consuming step due to the vast scope of possible materials composed of three-dimensional (3D) networks of elements selected from the periodic table. Data-driven research has recently received attention as a new route to accelerating this step.[1–9] This approach uses a pre-computed materials database and statistical tools that efficiently screen candidates in a search for optimal materials. The availability of open-access databases of material properties,[10–14] along with machine learning (ML) techniques, has rapidly advanced research in this area. Over the last decade, ML has been applied to materials science problems in a variety of directions, such as prediction and classification of crystal structures,[1,15–17] development of interatomic potentials,[18–20] finding of optimal density functionals for density functional theory,[21–23] and building of predictive models of material properties.[24–27]

The use of ML in materials science, however, has been hindered by the accuracy and interpretability of predictive models. Complex interaction among compositions of materials leads to highly nonlinear relationships between material features and target properties. To accurately describe such relationships, nonlinear ML algorithms have been utilized due to their flexible forms. However, lack of interpretability of most nonlinear ML predictive models prevents further mechanistic understanding such as finding key ingredients for target properties. Thus, finding ML algorithms that can achieve both accurate prediction and interpretability is crucial to the further advance of data-driven materials research.

Tree-based learning algorithms can be one candidate due to their advantages in both accuracy and interpretability.[28] Utilizing tree-based algorithms, we here focus on finding optimal candidates for double perovskite solar cells. While recent solar cell technology has been prompted by the development of hybrid lead perovskites having an increase of power conversion efficiency and low-cost manufacturing, the inclusion of lead ion raises environmental and health issues preventing commercialization.[29,30] Alternatively, a new strategy using mixed mono- and tri-valent cations, in the form of the double perovskite $A_2B^{1+}B^{3+}X_6$, has been introduced to replace lead-based perovskite solar cell materials.[31–34] In this approach, sizable combinations of double perovskites can be possible, and thus a combination of high-throughput computations and the ML technique can be a powerful tool to explore the large combinatorial space.

Here, employing the gradient-boosted regression tree (GBRT) algorithm and a dataset of calculated electronic structures of $A_2B^{1+}B^{3+}X_6$, we present an ML-based investigation, which can be ultimately used to identify Pb-free double perovskite solar cell materials. The GBRT method allows us to obtain highly accurate predictive models for the heat of formation (ΔH_F) and bandgap (E_g), with importance scores for each feature of materials. Based on the scores, we extract crucial features to determine the values ΔH_F and E_g of halide double perovskites, enabling an overall understanding of the relationships between features and properties. Finally, we discuss the relevance of extracted features to the chemical and physical aspects of ΔH_F and E_g, and practical approaches of the ML model toward finding optimal candidates of Pb-free halide double perovskites solar cell materials.

RESULTS

Dataset of Pb-free halide double perovskites

For the ML investigation, we first generated a dataset of the electronic structures of halide double perovskites. Figure 1a presents the crystal structure of the double perovskite $A_2B^{1+}B^{3+}X_6$. Compared to the original perovskite, this structure incorporates two different types of cations, B^{1+} and B^{3+}, instead of a single B

[1]Korea Research Institute of Chemical Technology (KRICT), Daejeon 34114, Republic of Korea and [2]National Institute for Mathematical Sciences (NIMS), Daejeon 34047, Republic of Korea
Correspondence: YunKyong Hyon (hyon@nims.re.kr) or Hyunju Chang (hjchang@krict.re.kr)
These authors contributed equally: Jino Im, Seongwon Lee

Received: 4 July 2018 Accepted: 11 March 2019
Published online: 26 March 2019

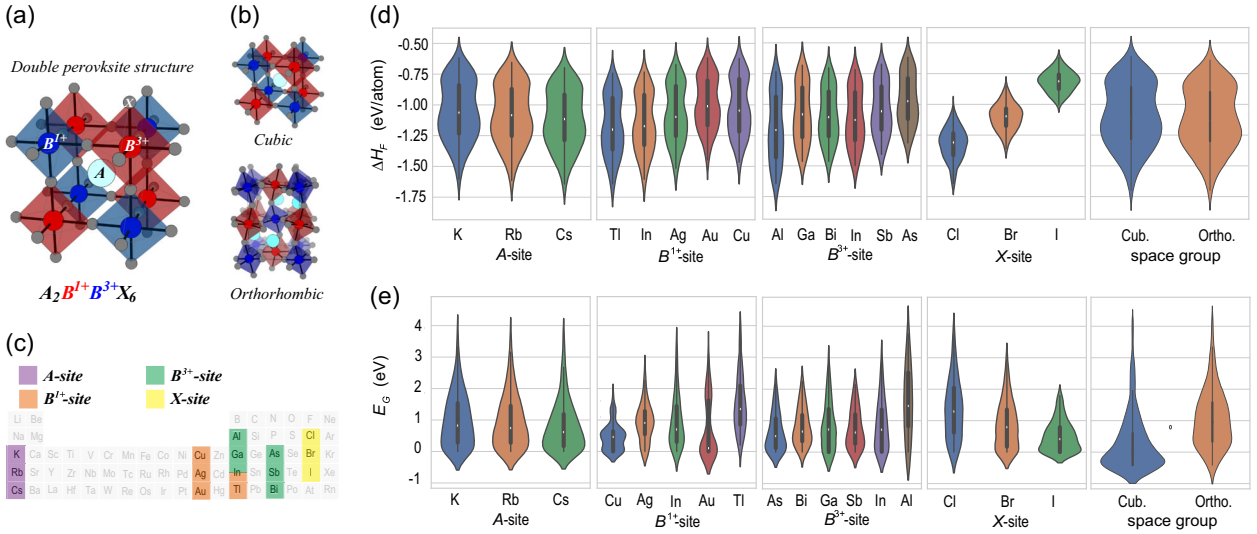

Fig. 1 **a** Crystal structure of double perovskite with *A*, B^{1+}, B^{3+}, and *X*-sites denoted by light blue, blue, red, and gray spheres, respectively. **b** Structural deformation by tilting of octahedral unit presents. **c** List of chemical elements considered in a dataset of halide double perovskites. Distribution of calculated **d** heat of formation (ΔH_F) and **e** bandgap (E_g). In each panel, average values are depicted as a white point

cation. With anion *X*, both B^{1+} and B^{3+} cations form octahedral units. Usually, the perovskite has a structural phase transition upon tilting and rotating of the octahedral unit. As shown in Fig. 1b, two possible crystal structures can be considered: one has a cubic space group; the other has an orthorhombic space group.

In this study, we considered coinage elements and lower group XIII elements for B^{1+}, and upper group XIII and lower group XV elements for B^{3+}. Then, substitutable combinations for di-valent lead ion could span over 30 combinations. Furthermore, we considered a series of alkali metals from K to Cs for the *A*-site (Li and Na were not considered because of their small size). Here, for simplicity of calculation, organic molecules, such as methyl ammonium, were not included. Halogen ions were assigned for the *X*-site. Figure 1c summarizes all combinations of chemical constituents. In sum, along with the two space groups of the crystal structure, 540 hypothetical compounds of $A_2B^{1+}B^{3+}X_6$ were considered.

Using first-principles density functional theory (DFT), we generated a dataset, including values of ΔH_F and E_g for the 540 compounds. ΔH_F indicates the stability of a compound compared to those of the elemental phases of its chemical constituents. Generally, a more negative value of ΔH_F indicates a more stable compound. On the other hand, E_g can represent the capability to absorb solar energy, which is critical to achieving high performance of a solar cell. The optimal value for solar absorption is reported to range from 1.1 eV to 1.8 eV.[35] However, note that E_g is severely underestimated in this work due to the limitation of the standard DFT.[36] Previous studies have shown that a DFT bandgap from 0.3 eV to 0.8 eV can recover to an optimal value of a solar cell material if more accurate computation methods such as hybrid DFT or GW are used.[37,38] Further computational details for these two quantities can be found in the method section.

We can detect a few notable characteristics of ΔH_F and E_g without relying on ML analysis. In the case of ΔH_F, all candidate materials have negative values, indicating that all can be stably synthesized (see Fig. 1d). Another prominent observation about ΔH_F is its dependence on halogen anion, which contrasts to its weak dependence on other elements. As halogen atoms change from iodine to chlorine, ΔH decreases. On the other hand, the relationship between E_g and atomic species and space group is more complicated, two remarkable characteristics of which we summarize in the following. First, in most cases, E_g increases by changing the space group (SG) from cubic to orthorhombic (Fig. 1e). Second, it is found that E_g mostly increases as halogen atoms change from iodine to chlorine. However, in mapping between E_g and materials, no other dependencies are observed.

Machine learning and features

We next apply the machine learning algorithms to the dataset of halide double perovskites. In general, machine learning study requires the appropriate selection of a learning algorithm and an optimized set of input features. In this study, we employed the Gradient-Boosted Regression Tree (GBRT), which is one of the tree-based machine learning algorithms. Decision tree learning is a machine learning method that uses a tree-like diagram, usually a binary tree, to predict a target variable. The goal is to create a tree in which each node represents a split based on one of the input features, and each leaf represents the prediction of the target variable. The prediction can be nonlinear because the partitioning of the input variable space is repeated recursively.[39] Compared to other ML methods, the decision tree is advantageous for its accuracy and speed, although it is prone to overfitting.[40] Using ensemble methods such as bagging and boosting can prevent overfitting, and thus can improve the accuracy.[41–43] GBRT adopts the gradient boosting method, which combines weak learners into one strong learner using the gradient descent algorithm.

Furthermore, the predictive model can be used to record the improvement of prediction results for a specific feature as each node corresponding to a single feature is added to the trees. In this manner, one can measure the feature importance automatically.[44] This feature importance fosters interpretability of predicted results and it leads to the extraction of key-features, which we will show later in the results. In the present work, we adopted a gradient boost method to generate a regression tree ensemble that is implemented in the XGBoost library.[28] See further details in the method section.

Another critical step for achieving good prediction performance is the selection of appropriate input features, referred to as feature engineering. For the dataset of materials, features should clearly describe a single given material and, also, discriminate separate materials. In this study, we selected 32 features, including chemical information of atomic constituents and geometric

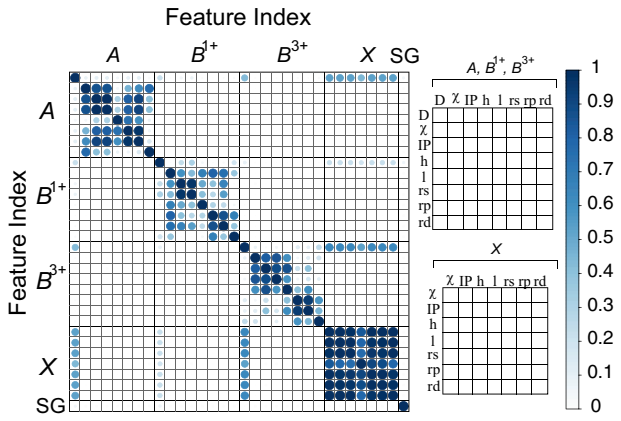

Fig. 2 Square of Pearson correlation coefficient matrix halide double perovskites. Size and color of circles vary with values

information such as bond length and crystal symmetry. The total of 32 features include the following:

Pauling's electronegativity (χ), *ionization potential* (IP), *highest occupied atomic level* (*h*), *lowest unoccupied atomic level* (*l*), and *s-, p-, and d- valence orbital radii* r_s, r_p, and r_d of isolated neutral atoms *A*, B^{1+}, B^{3+}, and *X*; atomic *distance* (*D*) between cations and the nearest halogen atom; *space group of crystal* (SG). Unlike the other features, SG is considered a categorical variable for cubic and orthorhombic symmetry.

Here, we note that the GBRT algorithm cannot appropriately evaluate the importance scores of two strongly correlated features. The reason is that two strongly correlated features cannot be distinguished in the learning process. Thus, reducing the dimensions of the feature space can improve the quality of prediction while simultaneously decreasing the computing cost. Here, we implemented a dimensionality reduction based on the square of the Pearson correlation matrix among features. For each pair of features *x* and *y*, the square of the correlation coefficient R^2_{xy} is defined as

$$R^2_{xy} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2\sum_{i=1}^{n}(y_i-\overline{y})^2}},$$

where $\overline{x}$ and $\overline{y}$ are the sample means of features x_i and y_i of the *i*-th material over a total of n compounds. As shown in Fig. 2, strong correlations are found in several pairs of features: (1) all atomic features of halogen atoms at the *X*-site, (2) r_s and r_p for *A*-, B^{1+}-, or B^{3+}-site atoms, and (3) IE and *h* for *A*-, B^{1+}-, or B^{3+}-site atoms. A lack of atomic variation at the *X*-site (only three atoms: Cl, Br, and I) led to strong correlations among all pairs of atomic features. The same trend was observed at the *A*-site, although this trend was not as strong as it was at the *X*-site. The correlation matrices were used to downselect features of the halide double perovskite dataset. We selected χ as a representative feature of all atomic features of the *X*-site atoms. Furthermore, we selected only r_s and excluded r_p for the *A*-, B^{1+}-, and B^{3+}-site atoms. For IE and *h* of the *A*-, B^{1+}-, or B^{3+}-site atoms, we considered the only *h*. We found that machine learning performance is almost the same under the dimensional reduction from the Pearson correlation matrix among features.

Predictive model and feature importance

We performed regression using GBRT to predict values of E_g and ΔH_F of the halide double perovskites $A_2B^{1+}B^{3+}X_6$. Figure 3a presents the prediction of ΔH_F. The results show that averaged root-mean-square-error (RMSE) of test sets for ΔH_F is 0.021 eV/atom. Even though the number of the current dataset was limited, it is noteworthy that the accuracy of the predictive model of ΔH_F

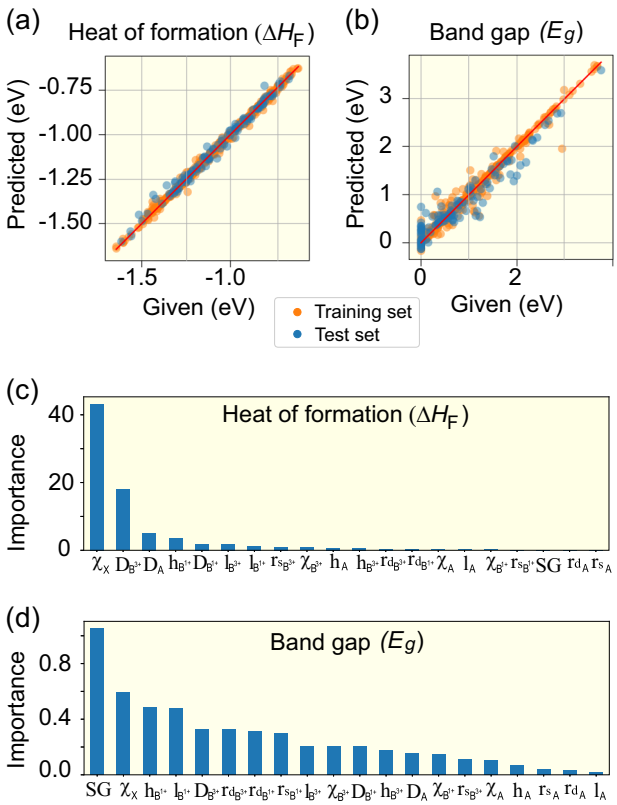

Fig. 3 Prediction of **a** heat of formation and **b** bandgap for halide double perovskites. Orange filled circles correspond to training dataset and blue circles to test dataset. Red solid lines indicate the reference line corresponding to the perfect fit. Feature importance from GBRT for **c** heat of formation and **d** bandgap of halide double perovskite

from GBRT is comparable to the fundamental error of 0.024 eV/atom that results from differences between the experimental and DFT-based values of ΔH_F of ternary oxides.[45] In the case of E_g, the averaged RMSE of the test sets is equal to 0.223 eV (Fig. 3b), which is worse than the error of ΔH_F. However, as the effective range (1.1–1.8 eV) of the bandgap of solar cell material is much larger than the RMSE, such low accuracy of E_g could be acceptable for solar cell applications.

The origin of the high accuracy of the predictive models can be attributed to several things. One is the nonlinear nature of the GBRT algorithm, in contrast to that of most linear algorithms (See Supplementary Information 1). Another reason for the high accuracy could be the structural similarity of the materials considered in the dataset. In the present study, crystal structures of all materials are perovskite. Given the same structure, materials having similar chemical constituents might have similar properties, which allows for more feasible interpolation of properties in the predictive models. Discussing structural similarity is beyond the scope of our study, but it is becoming a significant topic in ML studies of materials science.[46,47]

On top of providing highly accurate predictive models, the GBRT method provides interpretation of the results via feature importance scores, which is the main advantage of this method. Figure 3c, d show the importance score of all features for ΔH_F and E_g, respectively. For ΔH_F, it is revealed that the type of halogen anions, represented by χ_X, is the most important feature (Fig. 3c). Remarkably, the importance score of χ_X is more than two times higher than that of the secondly-ranked feature, $D_{B^{3+}}$. Beyond the first two features, importance scores steeply decrease, indicating that ΔH_F strongly depends on only a few features of materials. On

the other hand, the feature importance used to predict E_g is more dispersed (Fig. 3d). This implies a more complex relationship between E_g and the material features, which is consistent with the tendency observed without ML techniques (see Fig. 1e). We found that although SG is the most important feature, the following features, such as χ_X, $h_{B^{1+}}$, $l_{B^{1+}}$, $D_{B^{3+}}$, $rd_{B^{3+}}$, or $rd_{B^{1+}}$, are not negligible.

Extraction of key-features and application

In this section, we suggest possible utilization of the feature importance scores in an efficient search for target materials. First, we show that the feature importance scores can be utilized to extract key-features determining target properties. To this end, we recursively excluded the least important feature and built a predictive model using only the remaining set of features and learning new decision trees. This process was repeated until the top three features remained for each target property. Figure 4a, b show how the error in the prediction of ΔH_F and E changes under the process of extracting key-features. Remarkably, we found that RMSEs increase almost monotonically in both cases, which indicates that a feature with a higher feature score has stronger predictive power.

Through this process, we selected the most important features for each target properties. For ΔH_F, RMSE abruptly increases when the number of features is smaller than five (Fig. 4a). This means that at least five features are required to predict ΔH_F within RMSE of 0.036 eV. In the case of E_g, seven features comprise minimal set to predict E_g within an RMSE of 0.322 eV (Fig. 4b). The list of selected features includes χ_X, $D_{B^{3+}}$, D_A, $h_{B^{1+}}$, and $D_{B^{1+}}$ for ΔH_F, and SG, χ_X, $h_{B^{1+}}$, $l_{B^{1+}}$, $D_{B^{3+}}$, $rd_{B^{3+}}$, and $rd_{B^{1+}}$ for E_g.

Key-features extracted from the feature importance score of GBRT method can have various implications for a fast search for new material. For instance, selected key-features can be utilized as an optimal set of features in another type of ML algorithms such as classification. We considered a binary classification with class 1 for E_g in the range of 0.3–0.8 eV and class 2 for otherwise, to search for an optimal candidate of solar cell materials. Figure 5 presents the accuracy of classification tasks, as well as corresponding confusion matrices where the classification task was performed with a different number of features. With a smaller number of features, the prediction accuracy is lower than that with all features. However, it is notable that classification with the top seven key-features selected for E_g provides a good approximation to that with the full features. As the classification can be accelerated with the reduced feature space, materials screening before a further investigation can be feasible in a massive dataset.

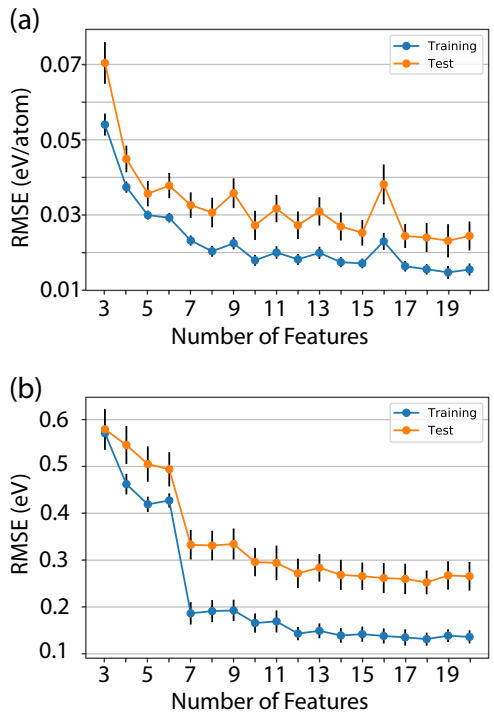

Fig. 4 Root-mean-square-error (RMSE) of **a** heat of formation and the **b** bandgap of halide double perovskite as a function of the number of features. In each panel, the blue curve corresponds to the training set and the orange curve to the test set

The mechanistic understanding of the relationship between features and target properties can provide a practical guide in the search for optimal double perovskite solar cell materials. For example, $Cs_2InAgCl_6$ is one of newly synthesized Pb-free halide double perovskite, but its bandgap energy is too large to be used as a solar cell material.[34] Knowing key-features determining E_g, one may have several plausible remedies to decrease E_g, such as anion-mixing from the electronegativity of halogen anions and partial-substitution/alloying for In and Ag from the highest occupied and lowest unoccupied orbitals of B^{1+} ions, the distance between B^{3+} ion and anion, radii of d-orbitals of B^{1+} and B^{3+}.

However, we note that an explicit relationship between features and properties is still difficult to obtain directly from the feature importance score, and additional steps would be needed. To this end, one may still utilize the process of feature selection described above and fit the properties using basis functions of the reduced features. Such an explicit relationship can be used to further mechanistic understanding of target properties such as revealing interaction between important features, but it is out of the scope of our paper.

DISCUSSION

In the importance-score-based selection process of the GBRT method, scientific knowledge is not reflected. Here, we consider whether the roles of the features selected using the GBRT method in determining the target properties are consistent with previous known scientific knowledge.

First, we investigated the role of the features selected for ΔH_F for the bonding mechanism. The top five selected features for ΔH_F were χ_X, $D_{B^{3+}}$, D_A, $h_{B^{1+}}$, and $D_{B^{1+}}$ (Fig. 4), among which the most important feature was χ_X. Interestingly, it is well-known that differences between electronegativities of bonding partners are good indicators of the bonding character and bonding strength, which strongly affect ΔH_F. Compared to other groups of elements, the halogen group shows relatively large variation in electronegativity. This could be attributed to a strong dependency of ΔH_F on χ_X. Along with electronegativity, the distances between cation and anion (D_A, $D_{B^{1+}}$, and $D_{B^{3+}}$) are also good indicators of bonding strength. Usually, strong bonding is accompanied by a shorter bond length. In the case of $h_{B^{1+}}$, $IP_{B^{1+}}$ is strongly correlated with $h_{B^{1+}}$, (see Fig. 2), and IP is also an important chemical quantity to explain ionic bonding. In this way, the top five selected features for ΔH_F are all relevant to the bonding mechanism, which is important to determine ΔH_F.

In the case of E_g, more complicated theories are required to understand the role of the selected features. For this purpose, we performed DFT analysis of the band structure, explicitly focusing on the selected features. Generally, symmetry-lowering by tilting of the octahedral unit of a halide perovskite increases E_g because of bandwidth shrinkage.[48] In the case of halide double perovskite, a similar transformation of band structure is found. Figure 6a–d show the band structures of cubic and orthorhombic phases of the representative compounds. In all cases, the bandwidths of the conduction and valence bands were reduced by the tilted octahedral units, and this led to increases of the bandgap in

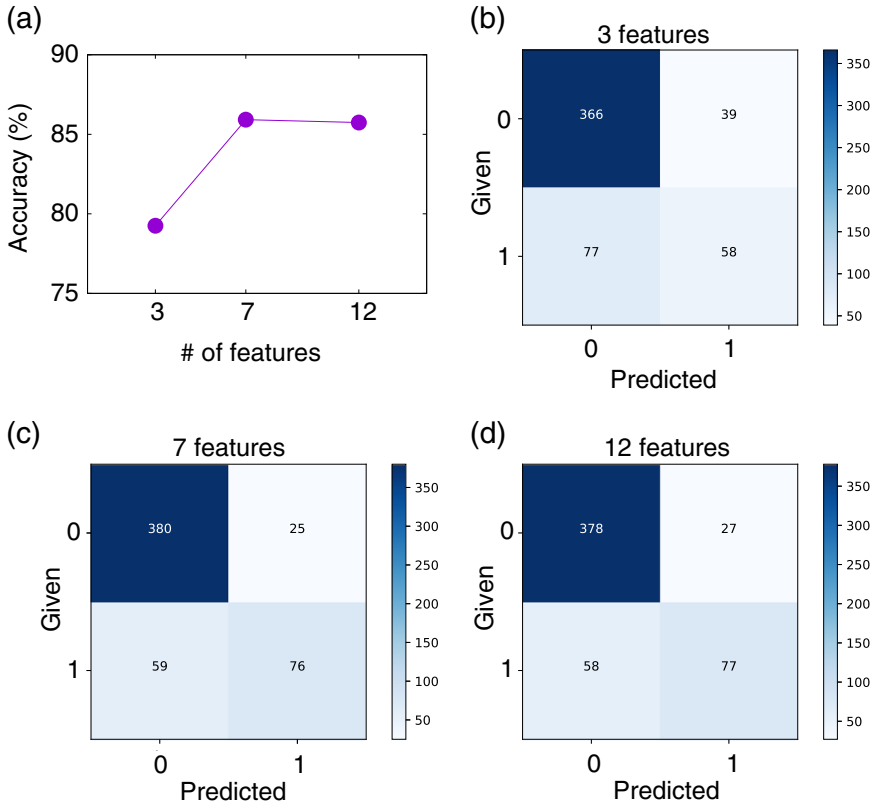

Fig. 5 **a** The accuracy of classification as a function of the number of features. Confusion matrices of classification results for the fivefold test set with **b** 3 features, **c** 7 features, and **d** 12 features. Class "1" means set of materials with a bandgap in a range of [0.3 eV, 0.8 eV] and class "0" of everyone else

orthorhombic phases. For other compounds, the same trend occurs, as shown in Fig. 1e. Thus, SG is relevant in determining E_g.

To check the validity of features χ_X, $h_{B^{1+}}$, $l_{B^{1+}}$, $D_{B^{3+}}$, $rd_{B^{3+}}$, and $rd_{B^{1+}}$, we plotted a schematic diagram of the orbital hybridization between cations and anions in halide double perovskites based on the DFT band structure calculations (see Fig. 6e). The diagram shows that E_g originates from energy differences between two hybridized states, the valence band maximum (VBM), composed of an anti-bonding state involving B^{1+}-site and X-site atoms, and the conduction band minimum (CBM), composed of an anti-bonding state involving B^{3+}-site and X-site atoms. Even though complicated interaction exists, the energy levels of the valence electrons of B^{1+}, B^{3+}, and X^- play important roles in determining the value of E_g of a given compound. The highest occupied and lowest unoccupied atomic levels can be good indicators of the energy levels of the valence electrons. In addition to the atomic levels, the electronegativity can play a crucial role in determining E_g by controlling energy splitting between the bonding and anti-bonding states, denoted by ΔE in Fig. 6e. This splitting indicates the strength of the hybridization among the orbitals. The high electronegativity of the compounds leads to tightly bound electronic distribution around the atoms and reflects strong hybridization via small bonding length, such as $D_{B^{3+}}$.

In this study, we used the GBRT method to investigate the ML prediction of the values ΔH_F and E_g of halide double perovskites. The GBRT method provided outstanding prediction performance for those properties, as well as providing an interpretable feature importance score. Notably, for ΔH_F, GBRT worked accurately, with an error comparable to the fundamental error associated with the difference between experimental and DFT values. The prediction of E_g was also acceptable for use in the search for solar cell materials. Key-features extracted based on the importance score provide a better mechanistic understanding of ΔH_F and E_g. On top of the accurate and interpretable predictive models, we further verified that the key-feature was relevant to scientific knowledge.

METHODS

Density functional theory (DFT)

Structural optimization, total energy, and electronic band structure of 540 halide double perovskites were performed within density functional theory (DFT) formalism. We utilized a plane-wave basis set (cutoff energy = 350 eV), and the projector augmented wave method,[49] implemented in the Vienna Ab-initio simulation package (VASP).[50,51] For the exchange correlational functional, the generalized gradient approximation was adopted within Perdue-Ernzerhof-Burke formalism.[52] A 5 x 5 x 5 regular grid was employed for momentum space sampling. The heat of formation, ΔH_F, was calculated using the following formula:

$$\Delta H_F = E_{tot}\left(A_2B^{1+}B^{3+}X_6\right) - \left(2E_{ref}(A) + E_{rel}\left(B^{1+}\right) + E_{rel}\left(B^{3+}\right) + 6E_{rel}(X)\right),$$

where E_{tot} is the total energy and E_{ref} is the reference energy. For the band structures, spin–orbit interaction was considered.

Atomic features

The highest occupied atomic level (h) and the lowest unoccupied atomic level (l) were taken from the atomic parameters of the VASP pseudopotential.[50,51] We set h as the highest orbital energy of the partially or fully occupied orbitals and l as the lowest orbital energy of the unoccupied ones. If the orbital energies of the unoccupied orbitals were not available in the parameter files, the highest orbital energy was set at l. In determining h and l, we considered degenerated atomic orbitals to be the same orbital.

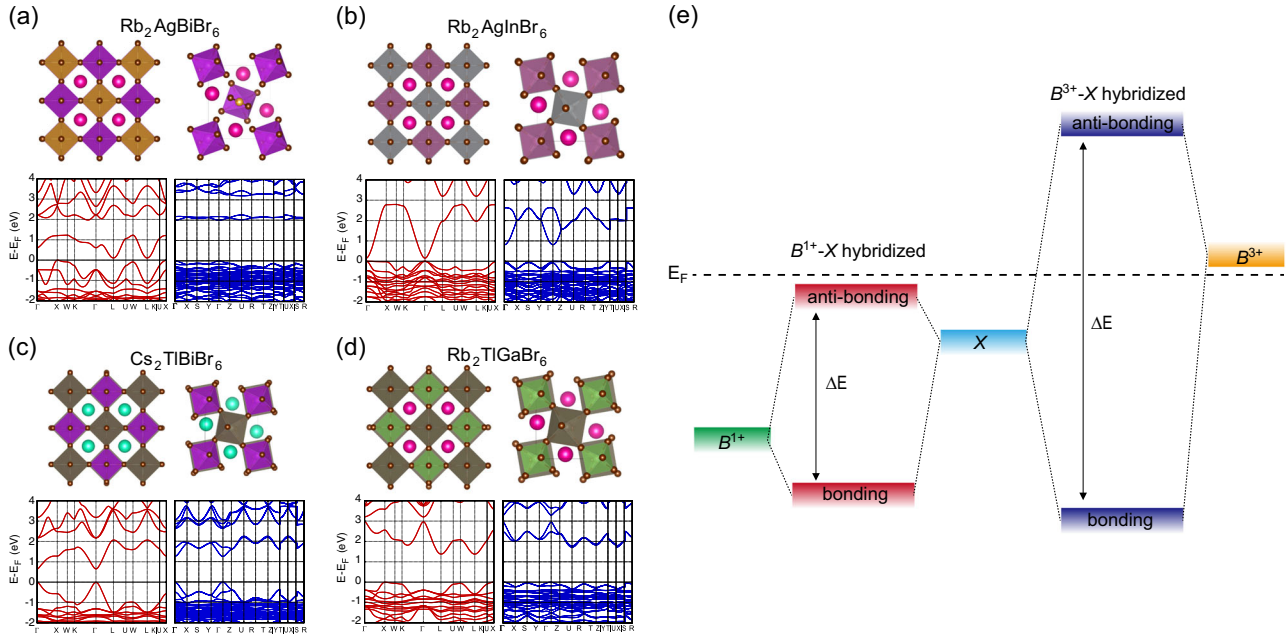

Fig. 6 **a**–**d** Present cubic and orthorhombic crystal structures and corresponding band structures of $Rb_2AgBiBr_6$, $Rb_2AgInBr_6$, $Cs_2TlBiBr_6$, and $Rb_2TlGaBr_6$, respectively. **e** Schematic illustration of a band diagram for halide double perovskite

Gradient-boosted regression tree

The supervised learning model has a loss function to be minimized. In XGBoost the loss function of the model (ensemble of trees f_k) is

$$\mathcal{L} = \sum_i \mathrm{l}(y_i, \hat{y}_i) + \sum_k \Omega(f_k)$$

where l is a function that measures the difference between the prediction and the target and Ω is a regularization term (complexity of the tree) to prevent overfitting. This loss function cannot be optimized using traditional optimization methods: the model is trained in an additive manner by adding a tree at a time that most improves the model (most decreases the loss) to the existing set of trees. Let $\mathbf{x}_i$ be an i-th sample and $\hat{y}_i^{(t-1)}$ be its prediction with the current set of $t-1$ trees. The model needs to add the t-th tree f_t to minimize the loss function

$$\mathcal{L}^{(t)}(f_t) = \sum_{i=1}^{n} \mathrm{l}(y_i, \hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i)) + \Omega(f_t).$$

It is impossible, however, to check all possible tree structures f_t to be added. So, the algorithm starts from a root (single leaf) and greedily adds branches to the tree. For each step, the model finds a leaf to split, and a feature and its value for the split that maximize loss reduction after the split. If the current tree structure is $f_t^{(\mathrm{current})}$ and the structure after the split is $f_t^{(\mathrm{split})}$, then the loss reduction by branching can be calculated by the difference of the loss, $D_{\mathrm{score}}(f_t) = \mathcal{L}^{(t)}\left(f_t^{(\mathrm{current})}\right) - \mathcal{L}^{(t)}\left(f_t^{(\mathrm{split})}\right)$. This means that for each branch of the trees, the model knows, which feature is used for the split and its loss reduction.[28] For each split of the trees during model training, the algorithm finds (approximately) the feature and splitting point that provide the largest cost decrease. It is possible to calculate the number of times each feature is selected for a split in the trees, or the average of the gain of the splits that use each feature. The library offers these numerical values as an F-score of type weight or type gain, respectively. In this study, we used gain.

The hyperparameters of each model were optimized using a cross-validated grid search or a randomized search over the parameter settings. The RMSE of the target variable was used as a cost function for all models. For all ML models and each model's training, randomly chosen 80% samples of the data are used for training; the remaining 20% are used for a test set. We averaged 200 evaluations, 40 sets of fivefold training/test set splitting with random shuffling (which also have 80% of the data as training set) and calculated the importance scores of the figures. In the gradient-boosted regression tree, several hyperparameters exist. A parameter subsample for bagging (subsample ratio of the training instance) and colsample_bytree (subsample ratio of features) for the random forest were optimized to prevent overfitting. The regularization parameters were also optimized. The values of the hyperparameters used here are:

$$\mathrm{max\_depth} = 6, \mathrm{min\_child\_weight} = 1, \mathrm{colsample\_bytree} = 0.5,$$

$$\mathrm{subsample} = 0.7, \mathrm{reg\_alpha} = 0.1, \mathrm{learning\_rate} = 0.03.$$

The regularization parameters were also optimized. The parameter max_depth is for maximum depth of a tree, and min_child_weight is for the minimum number of instances needed in each node. The smaller max_depth and the larger min_child_weight are, the less the training is likely to overfit. The parameter reg_alpha is an L1 regularization term on weights.

Gradient-boosted classification trees

The classifications were performed for halide double perovskite dataset with gradient booted classification trees (GBCT). For these classifications and predictions for all data, the halide double perovskite dataset was separated into five disjoint sub-datasets, and then train data consisted of four sub-datasets and test data of the other sub-dataset. This means that there were five predictions for the full halide double perovskite dataset. In establishing the predictive model, typically given hyperparameters of GBCT were adopted for the classifications. The given parameter values for the classifications used here are:

$$\mathrm{max\_depth} = 4,\ \mathrm{min\_child\_weight} = 4,\ \mathrm{colsample\_bytree} = 0.8,$$

$$\mathrm{subsample} = 0.8,\ \mathrm{reg\_alpha} = 0.1,\ \mathrm{learning\_rate} = 0.1.$$

DATA AVAILABILITY

A dataset on halide double perovskites is provided in the Supplementary Information (See the separate Excel File and corresponding explanation in part 2 of Supplementary Information). Other electronic data in this study are available from the corresponding authors upon reasonable request. A dataset on oxide double perovskites is available via the Computational Materials Repository. [https://cmr.fysik.dtu.dk/low_symmetry_perovskites/low_symmetry_perovskites.html#low-symmetry-perovskites].

ACKNOWLEDGEMENTS

This research was supported by the Nano-Material Technology Development Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Science and ICT (NRF-2016M3A7B4025408 and NRF-2017M3A7B4049366). We especially thank Castelli[53] for data sharing and appreciate S. Lim for valuable discussions.

AUTHOR CONTRIBUTIONS

H.C. and Y.H. designed the project. J.I. performed the high-throughput DFT calculations, and S.L. carried out machine learning calculations. H.W.K. and T.-W.K. developed the features to analyze the data. J.I. and S.L. wrote the manuscript. All authors discussed data and feature analysis and reviewed the manuscript.

ADDITIONAL INFORMATION

Supplementary information accompanies the paper on the *npj Computational Materials* website (https://doi.org/10.1038/s41524-019-0177-0).

Competing interests: The authors declare no competing interests.

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

REFERENCES

1. Fischer, C. C. et al. Predicting crystal structure by merging data mining with quantum mechanics. *Nat. Mater.* **5**, 641 (2006).
2. Armiento, R. et al. Screening for high-performance piezoelectrics using high-throughput density functional theory. *Phys. Rev. B* **84**, 014103 (2011).
3. Hautier, G. et al. Identification and design principles of low hole effective mass p-type transparent conducting oxides. *Nat. Commun.* **4**, 2292 (2013).
4. Carrete, J. et al. Nanograined half-heusler semiconductors as advanced thermoelectrics: An ab initio high-throughput statistical study. *Adv. Funct. Maters.* **24**, 7427 (2014).
5. Yim, K. et al. Novel high-κ dielectrics for next-generation electronic devices screened by automated ab initio calculations. *NPG Asia Mater.* **7**, e190 (2015).
6. Agrawala, A. & Choudhary, A. Perspective: materials informatics and big data: Realization of the "fourth paradigm" of science in materials science. *APL Mater.* **4**, 053208 (2016).
7. Jain, A. et al. New opportunities for materials informatics: Resources and data mining techniques for uncovering hidden relationships. *J. Mater. Res.* **31**, 977 (2016).
8. Kalidindi, S. et al. Role of materials data science and informatics in accelerated materials innovation. *Mrs. Bull.* **41**, 596 (2016).
9. Boyd, P. G., Lee, Y. & Smit, B. Computational development of the nanoporous materials genome. *Nat. Rev. Mater.* **2**, 17037 (2017).
10. Jain, A. et al. The materials project: A materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
11. Saal, J. E., Kirklin, S., Aykol, M., Meredig, B. & Wolverton, C. Materials design and discovery with high-throughput density functional theory: The Open Quantum Materials Database (OQMD). *JOM* **65**, 1501 (2013).
12. Kirklin, S. et al. The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. *npj Comput. Mater.* **1**, 15010 (2015).
13. Draxl, C. & Scheffler, M. NOMAD: The FAIR concept for big data-driven materials science. *Mrs. Bull.* **43**, 676 (2018).
14. Curtarolo, S. et al. AFLOW: An automatic framework for high-throughput materials discovery. *Comp. Mater. Sci.* **58**, 218 (2012).
15. Carr, D. A. et al. Machine learning approach for structure-based zeolite classification. *Micro Meso. Mater.* **117**, 339 (2009).
16. Pilania, G., Gubernatis, J. E. & Lookman, T. Structure classification and melting temperature prediction in octet AB solids via machine learning. *Phys. Rev. B* **91**, 214302 (2015).
17. Yamashita, T. et al. Crystal structure prediction accelerated by Bayesian optimization. *Phys. Rev. Mater.* **2**, 013803 (2018).
18. Hobday, S. et al. Applications of neural networks to fitting interatomic potential functions. *Model. Simul. Mater. Sci. Eng.* **7**, 397 (1999).
19. Handkey, C. M. & Popelier, L. A. Potential energy surfaces fitted by artificial neural networks. *J. Phys. Chem. A* **114**, 3371 (2010).
20. Schneider, E. et al. Stochastic neural network approach for learning high-dimensional free energy surfaces. *Phys. Rev. Lett.* **119**, 150601 (2017).
21. Wellendorff, J. et al. Density functionals for surface science: Exchange-correlation model development with Bayesian error estimation. *Phys. Rev. B* **85**, 235149 (2012).
22. Snyder, J. C. et al. Finding density functionals with machine learning. *Phys. Rev. Lett.* **108**, 253002 (2012).
23. Brockherde, F. et al. Bypassing the Kohn-Sham equations with machine learning. *Nat. Commun.* **8**, 872 (2017).
24. Pilania, G. et al. Accelerating materials property predictions using machine learning. *Sci. Rep.* **3**, 2810 (2013).
25. Lee, J. et al. Prediction model of band gap for inorganic compounds by combination of density functional theory calculations and machine learning techniques. *Phys. Rev. B* **93**, 115104 (2016).
26. Pilania, G. et al. Machine learning bandgaps of double perovskites. *Sci. Rep.* **6**, 19375 (2016).
27. Bartok, A. P. et al. Machine learning unifies the modeling of materials and molecules. *Sci. Adv.* **3**, e170816 (2017).
28. Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. arXiv:1603.02754 (2016).
29. Chen, W. et al. Efficient and stable large-area perovskite solar cells with inorganic charge extraction layers. *Science* **350**, 944–948 (2015).
30. Shin, S. et al. Colloidally prepared La-doped $BaSnO_3$ electrodes for efficient, photostable perovskite solar cells. *Science* **356**, 167–171 (2017).
31. Volonakis, G. et al. Lead-free halide double perovskites via heterovalent substitution of noble metals. *J. Phys. Chem. Lett.* **7**, 1254 (2016).
32. Philip, M. R. et al. Band gaps of the lead-free halide double perovskites $Cs_2BiAgCl_6$ and $Cs_2BiAgBr_6$ from theory and experiment. *J. Phys. Chem. Lett.* **7**, 2579 (2016).
33. Wei, F. et al. Synthesis and properties of a lead-free hybrid double perovskite: $(CH_3NH_3)_2AgBiBr_6$. *Chem. Mater.* **29**, 1089 (2017).
34. Volonakis, G. et al. $Cs_2InAgCl_6$: A new lead-free halide double perovskite with direct band gap. *J. Phys. Chem. Lett.* **8**, 772 (2017).
35. Shockley, W. & Queisser, H. J. Detailed balance limit of efficiency of p-n junction solar cells. *J. Appl. Phys.* **32**, 510–519 (1961).
36. Jin, H., Im, J. & Freeman, A. J. Topological insulator phase in halide perovskite structure. *Phys. Rev. B* **86**, 121102 (2012).
37. Menéndez-Proupin, E., Palacios, P., Wahnón, P. & Conesa, J. C. Self-consistent relativistic band structure of the $CH_3NH_3PbI_3$ perovskite. *Phys. Rev. B* **90**, 045207 (2014).
38. Umari, P., Moscon, E. & Angelis, F. D. Relativistic GW calculations on $CH_3NH_3PbI_3$ and $CH_3NH_3SnI_3$ perovskites for solar cell applications. *Sci. Report.* **4**, 4467 (2014).
39. Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. *Classification and Regression Trees (The Wadsworth statistics/probability series)*. (Chapman and Hall, New York, 1984).
40. Strobl, C., Malley, J. & Tutz, G. An introduction to recursive partitioning: rationale, application and characteristics of classification and regression trees, bagging and random forests. *Psycol. Methods* **14**, 323 (2009).
41. Breiman, L. Bagging predictors. *Mach. Learn.* **24**, 123 (1996).
42. Ho, T. K. The random subspace method for constructing decision forests. *IEEE T. Pattern Anal.* **20**, 832 (1998).
43. Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **29**, 1189 (2001).
44. Hastie, R. & Tibshirani, J. F. The elements of statistical learning: Data mining, inference, and prediction (Springer Series in Statistics, Springer, 2001).
45. Hautier, G. et al. Accuracy of density functional theory in predicting formation energies of ternary oxides from binary oxides and its implication on phase stability. *Phys. Rev. B* **85**, 155208 (2012).
46. Schutt, K. T. et al. How to represent crystal structures for machine learning: Towards fast prediction of electronic properties. *Phys. Rev. B* **89**, 205118 (2014).
47. Ghiringhelli, L. M. et al. Big data of materials science: Critical role of the descriptor. *Phys. Rev. Lett.* **114**, 105503 (2015).
48. Stoumpos, C. C., Malliakas, C. D. & Kanatzidis, M. G. Semiconducting tin and lead iodide perovskites with organic cations: Phase transitions, high mobilities, and near-infrared photoluminescent properties. *Inorg. Chem.* **52**, 9019 (2013).
49. Blöchl, P. E. *Phys. Rev. B* **50**, 17953 (1994).
50. Kresse, G. & Furthmüller Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *J. Comput. Mat. Sci.* **6**, 15 (1996).
51. Kresse, G. & Furthmüller Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *J. Phys. Rev. B* **54**, 11169 (1996).
52. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**(18), 3865 (1996).
53. Castelli, I. E., Thygesen, K. S. & Jacobsen, K. W. Bandgap engineering of double perovskites for one- and two-photon water splitting. *Mater. Res. Soc. Symp. Proc.* **1523**, https://doi.org/10.1557/opl.2013.450 (2013).