Alternative materials for perovskite solar cells from materials informatics

Shohei Kanno, Yutaka Imamura, and Masahiko Hada

Phys. Rev. Materials **3**, 075403 — Published 17 July 2019

DOI: 10.1103/PhysRevMaterials.3.075403

Do Promising Alternative Materials for Perovskite Solar Cells Exist?

An Answer from Materials Informatics

Shohei Kanno, Yutaka Imamura,* Masahiko Hada

Department of Chemistry, Tokyo Metropolitan University, Hachioji, Tokyo, Japan

ABSTRACT

Perovskite solar cells based on lead-halide perovskites have attracted significant attention as prime candidates for next-generation solar cells because of their high-power conversion efficiency. To avoid the toxicity of lead-based perovskites, alternatives such as tin-halide perovskite have been investigated. However, the photovoltaic performance of these alternatives is relatively low, and novel perovskites with low cost, low toxicity, and high performance have not yet been discovered. In this study, to investigate whether promising alternative perovskites exist, a high-throughput material search scheme based on materials informatics was developed and performed for novel perovskite solar cell materials. Using this scheme, over 28 million $AA'BB'X_3X'_3$ double perovskite-like compositions were screened. Among the 24 most promising candidates identified, 5 were well-known organic-inorganic tin-halide perovskites and 17 were novel sodium-, potassium-, and ammonium-based tin-halide perovskites. Interestingly, two novel transition-metal-based perovskites were also identified as promising solar cell materials. The pioneering material search scheme reported is expected to find use in the identification of practically feasible materials for a number of real-world applications.

Introduction

Organic-inorganic hybrid perovskites such as methylammonium lead iodide ($MAPbI_3$) have enormous potential as solar cell materials because of their suitable band gaps for solar light absorption [1], very small exciton binding energies [2], and long carrier diffusion lengths [3]. The power conversion efficiency (PCE) of perovskite solar cells (PSCs) skyrocketed from 3.8% in 2009 to 23.3% in 2018 [4,5]. Therefore, PSCs are prime candidates for next-generation solar cells and are expected to provide a solution to the energy problem. However, the toxicity of lead-based hybrid perovskites is a serious obstacle to their practical application [6]. To avoid the toxicity of lead, lead-free hybrid perovskites in which other ions are substituted for lead have been examined both experimentally and using computational simulations [7–15]. However, the PCEs of lead-free PSCs based on $CH_3NH_3SnI_3$, which is widely used as an alternative to lead-based perovskites, are significantly lower than the PCEs of lead-based PSCs [5,7]. Additionally, SnI_2, which is the main degradation product of tin-based perovskites, may present even greater toxicity concerns than lead-based perovskites [16]. Therefore, the development of novel perovskites with high photovoltaic performance is required.

Recently, data-driven machine learning and materials informatics have succeeded in the discovery of novel materials such as solid-state electrolytes [17], organic light-emitting diodes [18], shape memory alloys [19], piezoelectrics [20], and polymers for organic photovoltaics [21]. These approaches have also been employed for crystal structure prediction [22–25], physical property

prediction [26–31], and high-throughput computational screening [32,33]. Accordingly, massive and efficient material design based on machine learning and materials informatics has attracted significant attention in the field of materials science.

In this study, we developed a massive and highly efficient material search scheme based on materials informatics and applied it to the screening of over 28 million $AA'BB'X_3X'_3$ double perovskite candidates. In this material search scheme, in addition to the semiconductor properties of the candidates such as the band gap and carrier effective mass, the synthetic feasibility, toxicity, and cost, which were rarely considered in previous studies, were systematically analyzed using an informatics strategy based on a combination of experimental and the theoretical databases newly built from our calculations. To date, many lead-free and tin-free perovskites have been reported from material searches based on computational simulations [13–15]. However, there have been no successful experimental reports regarding the alternative perovskites proposed from these computational simulations, possibly because these previous reports used only theoretical databases from first-principles calculations. In particular, it is difficult to calculate the band gaps of hybrid perovskites because of the electron correlation and the strong spin-orbit coupling [34]. In this study, we estimate band gaps based on the experimental databases to guide novel material searches for PSCs. Through the screening of 28 million candidates, we identified alternative perovskites with suitable semiconductor properties, stable cubic or pseudo cubic structures, low toxicity, and low cost for use in PSCs.

68

Methods

Machine learning models for band gap prediction

To train the machine learning model for band gap prediction, a training dataset of high quality is required. Past studies have employed band gap datasets evaluated by first-principles calculations [35,36]. However, studies indicated that electron correlation and relativistic spin-orbit coupling (SOC) play important roles in the band gap calculations of PSC materials [34]. To determine the efficacy of the band gap learning model, we used an experimental band gap dataset of 282 perovskites. The details are summarized in Table S1 and Figure S1 (see Supplemental Material) [37].

To define the feature vectors for each $AA'BB'X_3X'_3$ double perovskite, we used atomic and ionic features of the six constituent ions (A, A', B, B', X, and X') of the perovskite. For each ion, we used nine elemental features (*viz.* first ionization potential, electron affinity, Mulliken electronegativity, ionic radius, group number, Pettifor's Mendeleev number [38], ionic HOMO level, ionic LUMO level, and ionic HOMO-LUMO gap). Therefore, an $AA'BB'X_3X'_3$ double perovskite is described by a 54-dimensional feature vector in the target chemical space. To determine the first ionization potentials and electron affinities for organic molecules and ionic HOMO/LUMO levels for all ions, density functional theory (DFT) calculations were carried out. The computational details and features for each element and organic molecule are summarized in Table S2 (see Supplemental Material).

Using the band gap data set and the feature vectors, we validated machine learning models for band gap prediction. Figure 1 illustrates the validation process. In this process, the band gap dataset was randomly split into a training set and a test set in ratios of 80% and 20%. Ten pairs of a fitting set and a validation set with ratios of 60% and 40% were generated from random splits in the training set. Using the fitting sets, validation sets, and test sets, the accuracy of the band gap predictor was validated. As the preprocess for regression coefficient fitting and cross-validation, the features were scaled by normalization, and the dimension of the feature vector was reduced from 54 dimensions to 15 dimensions by principal component analysis (PCA). After the preprocessing, machine learning models were trained via fitting and cross-validation using the 10 pairs of the fitting set and the validation set. Next, the trained machine learning models were tested using the test set. In order to investigate the generality of the machine learning models, this process was iterated for 50 cycles with different random seeds to split the band gap data set, and the accuracies of the machine learning models were assessed by averaging R^2 values for the training sets (R^2_{train}) and test sets (R^2_{test}).

The averaged R^2_{train}, averaged R^2_{test}, and root mean squared errors (RMSEs) in the test set for each regression model are listed in Table 1. In this assessment, we employed multiple linear regression (MLR), Ridge regression, Lasso regression, support vector machine regression (SVR) with a linear kernel or Gaussian kernel, and Gaussian process regression (GPR) with a Gaussian kernel as regression models. These are implemented in the scikit-learn library [39]. In addition,

ensemble learning models such as random forest and neural network are also known as powerful regression models. However, in this study, the number of data sets was insufficient to perform these regressions, and they were not reasonable for this study. For linear-regression-based prediction models (MLR, Ridge regression, Lasso regression, and SVR with a linear kernel), the averaged R^2_{train} and averaged R^2_{test} are very low; hence, these models cannot predict the band gaps of perovskites. On the other hand, for nonlinear regression-based prediction models (SVR and GPR with a Gaussian kernel), the prediction accuracies are dramatically improved. This result implies that nonlinear correlation between the band gap and the features is important for band gap prediction, and similar perovskites show similar band gaps. In particular, SVR with a Gaussian kernel shows the best prediction accuracy in our examination, and the averaged R^2_{train}, averaged R^2_{test}, and RMSE in the test set are 0.89, 0.65, and 0.81 eV, respectively. We employed this SVR with a Gaussian kernel as the band gap predictor. To improve the prediction accuracy, additional band gap data are required. We believe that our machine learning model will be improved by an expanded band gap dataset in the future.

Computational details for first-principles calculations

First-principles calculations were carried out to perform structural optimization and determine the band gap, electron and hole effective mass, and exciton binding energy. For structural optimization, the ion positions and cell parameters in a 2 × 1 × 1 cubic-based unit cell (Figure S2)

were optimized with the convergence threshold for a change of 10^{-3} eV in the total energy. After the structural optimizations, the direct band gap, indirect band gap, electron and hole effective mass, and exciton binding energy were calculated using the optimized structure. These calculations were performed using the Perdew-Burke-Ernzerhof (PBE) exchange correlation functional with the DFT-D3 method as a van der Waals correction [40,41], projector augmented wave (PAW) pseudopotentials, 700-eV plane-wave cutoff, and 10 × 10 × 10 sampling k-point grid in Vienna ab initio simulation package (VASP) code [42,43].

For the most promising novel perovskites, more accurate DFT calculations were carried out to evaluate the formation enthalpies, light absorption coefficients, and levels of conduction band minimum (CBM) and valence band maximum (VBM). A computationally inexpensive theoretical approach based on PBE + *U* calculations with fitted elemental-phase reference energies (FERE) was used for reoptimizations of the structure and estimations of the formation enthalpy [44]. Here, the values of the effective on-site Coulomb interactions (*U*) were taken from [44], and the most stable crystal structures of each pure element were taken from the Materials Project [45]. The other computational details are similar to the above calculations. Furthermore, a Heyd-Scuseria-Ernzerhof (HSE06) screened hybrid functional with a 2 × 4 × 4 sampling k-point grid was used to evaluate the light absorption coefficients and the levels of CBM and VBM [46]. The light absorption coefficients were estimated based on the calculations of the imaginary part of the frequency-dependent dielectric

matrix implemented in VASP code [47]. The levels of CBM and VBM were determined using empirical equations [48,49]:

$$E_{\mathrm{C}} = \left(\chi_{\mathrm{A}}\chi_{\mathrm{A'}}\chi_{\mathrm{B}}\chi_{\mathrm{B'}}\chi_{\mathrm{X}}^{3}\chi_{\mathrm{X'}}^{3}\right)^{\frac{1}{10}} + \frac{1}{2}E_{\mathrm{g}} \quad (1),$$

$$E_{\mathrm{V}} = E_{\mathrm{C}} - E_{\mathrm{g}} \quad (2).$$

where E_{C} and E_{V} are the CBM level and VBM level relative to the vacuum level, respectively. χ_{A}, $\chi_{\mathrm{A'}}$, χ_{B}, $\chi_{\mathrm{B'}}$, χ_{X}, and $\chi_{\mathrm{X'}}$ denote the absolute electronegativities of A, A', B, B', X, and X' atoms in $AA'BB'X_3X'_3$ double perovskite, respectively [50], and E_{g} is the band gap calculated by an HSE06 functional.

Results and discussion

We developed a high-throughput material search for novel PSC materials. Figure 2 depicts the novel $AA'BB'X_3X'_3$ double perovskite search scheme. In this material search scheme, not only the feasibility of the perovskite structure and its band gap but also its toxicity and cost were systematically considered. First, $AA'BB'X_3X'_3$ compositions were generated from a library of ions. For the A and A′ cations, 18 cations including alkali metal, alkali earth metal, group-3 metal, and organic cations were employed. For the B and B′ cations, 85 cations including transition metals and p-block metals were employed. For the X and X′ anions, nine anions including chalcogens and halogens were employed. The specific ions are listed in Figure 2 (a). From the ion library, 28,125,225 $AA'BB'X_3X'_3$ compositions were generated. These were screened in a stepwise manner according to the material search scheme.

In the first screening step of the material search scheme, the ability of the generated $AA'BB'X_3X'_3$ compositions to form a perovskite was evaluated using the general properties of the constituent ions (*viz.* ionic valence, valence electron number, and ionic radius). A composition was considered to be synthetically feasible if it met the following seven criteria: (1) its charge was neutral, (2) it had an even number of electrons, (3) its tolerance factor *TF* fell between 0.8 and 1.1 [51], (4) its octahedral factor *OF* was greater than 0.4 [51], (5) the ionic radii difference and ratio of its A and A′ cations met the criteria in eq. (3); (6) the ionic radii difference and ratio of its B and B′ cations met

the criteria in eq. (4); and (7) its X and X′ anions had the same group number. *TF* and *OF* are defined as

$$TF = \frac{r_{\bar{A}} + r_{\bar{X}}}{\sqrt{2}(r_{\bar{B}} + r_{\bar{X}})} \quad (3)$$

$$OF = \frac{r_{\bar{B}}}{r_{\bar{X}}} \quad (4).$$

where $r_{\bar{A}}$ is the average ionic radius of the A and A′ cations, $r_{\bar{B}}$ is the average ionic radius of the B and B′ cations, and $r_{\bar{X}}$ is the average ionic radius of the X and X′ anions. Shannon's ionic radii and effective radii were used for atomic ions and molecular ions, respectively [52,53]. However, *TF* and *OF* were defined for ABX_3 single perovskites, and these factors do not consider the differences between the ionic radii of A and A′, B and B′, or X and X′ in $AA'BB'X_3X'_3$ double perovskites. Recently, Bartel *et al.* reported a new tolerance factor to predict the stability of $A_2BB'X_6$-type double perovskites [54]. Even with Bartel's tolerance factor, it may be impossible to predict the stabilities of $AA'BB'X_3X'_3$ double perovskites because Bartel's tolerance factor cannot consider the differences between A and A′ or X and X′. Therefore, the conventional *TF* and *OF* parameters and Bartel's new tolerance factor cannot simply be applied to $AA'BB'X_3X'_3$ double perovskites; additional conditions are required. The 5th and 6th conditions consider the differences in the ionic radii of A and A′ and B and B′, respectively, by applying the following rules:

$$0.73 \leq \frac{r_A}{r_{A'}} \leq 1.37, \qquad |r_A - r_{A'}| \leq 0.45 \text{ Å} \quad (5)$$

$$0.50 \leq \frac{r_B}{r_{B'}} \leq 2.00, \qquad |r_B - r_{B'}| \leq 2.00 \text{ Å} \quad (6).$$

where r_{A}, $r_{\mathrm{A}'}$, r_{B}, and $r_{\mathrm{B}'}$ are the ionic radii of the A, A′, B, and B′ cations, respectively. The 5th, 6th, and 7th conditions were determined from our dataset of 282 experimental perovskite compounds (see Table S1 in the Supplemental Material). $AA'BB'X_3X'_3$ compositions that satisfy the seven criteria are expected to be able to form a perovskite structure; 128,357 such $AA'BB'X_3X'_3$ were identified in this step, and only these combinations proceeded to the next screening step.

In the second screening step, the band gaps of the 128,357 $AA'BB'X_3X'_3$ double perovskites were predicted using a machine learning model. Support vector machine regression (SVR) with a Gaussian kernel trained on the experimental band gap data of 282 perovskite compounds was employed as the machine learning model. The band gaps and compositions of 282 perovskites are listed in Table S1 (see Supplemental Material). The feature vectors for the 128,357 $AA'BB'X_3X'_3$ double perovskites were generated from the atomic and ionic information of the constituent elements. The details are summarized in the Supplemental Material. To identify novel double perovskites with suitable band gaps, we set a band gap criterion of 1.4 ± 0.8 eV. The values of 1.4 and 0.8 eV correspond to the ideal band gap for p-n-junction-based solar cell materials according to the Shockley-Queisser limit and the error bar of our SVR, respectively [55]. Through this screening, 10,918 $AA'BB'X_3X'_3$ double perovskites with suitable band gaps were identified. The 10,918 perovskites included conventional hybrid perovskites such as $MA_2Pb_2I_6$ (=$MAPbI_3$), for which the band gaps determined experimentally and by our SVR were 1.48 and 1.69 eV, respectively [56].

In many material search studies, the development of such a database of 10,918 $AA'BB'X_3X'_3$ double perovskites might be the end goal. However, to target materials with the potential for practical application, we considered two additional criteria (toxicity and cost) in this work. These criteria were selected because the toxicity of the widely used Pb-based hybrid perovskites is a serious problem [6], and low manufacturing cost is a merit of PSCs.

In the third screening step, the 10,918 identified perovskites were evaluated in terms of their toxicity and cost. The *Guideline for Elemental Impurities Q3D* was employed for the toxicity estimation [57]. These guidelines classify elements into four classes: highly toxic Class-1 elements, moderately toxic Class-2 elements, low-toxicity Class-3 elements, and low or nontoxic other elements. Pb, Hg, As, and Cd are Class-1 elements, and Co, V, Ni, Tl, Au, Pd, Ir, Os, Rh, Ru, Se, Ag, and Pt are Class-2 elements. In this evaluation, $AA'BB'X_3X'_3$ perovskites including a Class-1 or -2 element, such as $MA_2Pb_2I_6$ and other Pb-based perovskites, were rejected, and 2,146 low-toxicity double perovskites consisting of only Class-3 and other elements were found. After the toxicity evaluation, the cost of the remaining 2,146 double perovskites was estimated using the Chemicool database [58]. This database lists the price of each element in pure and bulk form. Using this database, the cost of each of the 2,146 double perovskites was estimated in US dollars (USD) per mole, and the 500 $AA'BB'X_3X'3$ double perovskites with the lowest cost were selected. Note that the prices of each element were collected for these estimations in February 2018 and may change over time. The latest prices can be seen in the Chemicool database [58]. Here, the toxicities and costs of

perovskites were considered in the screening scheme. However, we would suggest that this screening step is not necessarily required in general material studies since the toxicities and costs are not physical properties and relatively evaluated.

In the final screening step, first-principles calculations were performed to calculate the structures, band gaps, carrier effective masses, and exciton binding energies of the 500 $AA'BB'X_3X'_3$ double perovskites. The band gaps predicted by machine learning, estimated costs, optimized structures, and semiconductor properties as calculated by density functional theory (DFT) for the 500 candidates are listed in Table S4 (see Supplemental Material), and the computational details are summarized in the Supplemental Material. From among the 500 candidates, the 24 most promising $AA'BB'X_3X'_3$ double perovskites were identified and are listed in Table 2 along with their band gaps, costs, structural properties, hole and electron effective masses, exciton binding energies, and theoretical PCEs. The 24 perovskites identified meet the following criteria: (1) a cubic or pseudo cubic optimized structure with Bravais lattice vector length ratios of $1.90 \leq |\mathbf{a}|/|\mathbf{b}| \leq 2.10$, $1.90 \leq |\mathbf{a}|/|\mathbf{c}| \leq 2.10$, and $0.95 \leq |\mathbf{b}|/|\mathbf{c}| \leq 1.05$, and angles between the Bravais lattice vectors of $89.0° \leq \alpha, \beta, \gamma \leq 91.0°$; (2) a hole and electron effective mass of less than 1.00 a.u.; and (3) equal direct band gap and indirect band gap values. Of the 24 perovskites identified, five are organic tin-halide hybrid perovskites such as methylammonium tin iodide ($MASnI_3$) and formamidinium tin iodide ($FASnI_3$). These organic tin-halide hybrid perovskites are well-known and are already employed as alternative perovskites in PSCs. Their identification reproduces the results of alternative

experimental perovskite searches in recent years, and suggests that our material search scheme is a very reliable method. In addition, sodium-, potassium-, and ammonium-based multi-A-cation tin-halide perovskites that have not been reported in previous studies, namely, $KMASn_2Br_6$, $KMASn_2Br_3I_3$, $KMASn_2I_6$, $KNH_4Sn_2Br_6$, $KNH_4Sn_2Br_3I_3$, $KNH_4Sn_2I_6$, and $NaNH_4Sn_2I_6$, were identified. Furthermore, 10 inorganic tin-halide perovskites, namely, $KSnBr_3$, $K_2Sn_2Br_3I_3$, $K_2Sn_2Cl_3I_3$, $KSnI_3$, $NaKSn_2Br_6$, $NaKSn_2Br_3I_3$, $NaKSn_2I_6$, $Na_2Sn_2Br_3I_3$, $Na_2Sn_2Cl_3I_3$, and $NaSnI_3$, are first reported in this study. These perovskites contain the well-known tin-halide framework. Therefore, the existence of these perovskites is easy to imagine. In addition to the tin-halide perovskites, two novel inorganic perovskites were identified. One is a copper-halide-based perovskite, $CaSrCu_2I_6$, and the other is a sulfide perovskite, $CaBaMnNbS_6$. These perovskites include divalent A- and A′-cations and transition metals as the B- and B′-cations. The compositions are significantly different from those conventionally used in the PSC field, and they have not been investigated as PSC materials. However, we expect that $CaSrCu_2I_6$, $CaBaMnNbS_6$, and similar perovskites will show appropriate properties for use as solar cell materials.

For the novel organic-inorganic tin-halide, inorganic tin-halide, $CaSrCu_2I_6$, and $CaBaMnNbS_6$ perovskites, more detailed examinations are needed to investigate the thermochemical stabilities, light absorbance, and levels of CBM and VBM. Table 3 lists the formation enthalpies calculated by PBE + *U* with FERE, band gaps calculated by the HSE06 functional, gap types, and levels of CBM and VBM relative to the vacuum level for the novel perovskite candidates. In addition,

Figure S3 illustrates the imaginary parts of the frequency-dependent dielectric function corresponding to the light absorption spectra for each novel perovskite. Notably, the HSE06 functional might seriously underestimate the band gaps of $CaSrCu_2I_6$ and $CaBaMnNbS_6$ because of the strong electron correlations of Cu and Mn atoms. The formation enthalpies are exothermic for all novel perovskites, and the candidates discovered in this study can be thermochemically stable as perovskite structures. All novel organic-inorganic tin-halide and inorganic tin-halide perovskites show direct band gaps and are expected to have relatively large light absorption coefficients in the visible light region (see Figure S3). However, the band gaps calculated by HSE06 are smaller than those predicted by machine learning. Therefore, the HSE06 calculations might underestimate the band gaps of the novel organic-inorganic tin-halide and inorganic tin-halide perovskites. By contrast, for the $CaSrCu_2I_6$ and $CaBaMnNbS_6$ perovskites, the gap types are indirect band gaps, and the light absorption coefficients are expected to be relatively small (see Figure S3). However, $CaSrCu_2I_6$ and $CaBaMnNbS_6$ perovskites show thermochemical stability and small effective masses of electrons and holes. These characteristics are related to high carrier mobility. Therefore, $CaSrCu_2I_6$, $CaBaMnNbS_6$, and similar perovskites can be expected as potential candidates for photovoltaic materials, and we suggest that experimental studies are needed for these novel perovskites.

Conclusion

In this study, we examined if promising alternative perovskites with low cost, low toxicity, and high photovoltaic performance exist. To investigate this efficiently, a high-throughput material search scheme based on materials informatics was developed and applied to the screening of 28,125,225 $AA'BB'X_3X'_3$ double perovskite candidates. This scheme systematically considered not only the semiconductor properties of the candidates (such as the band gap and carrier effective mass) but also the feasibility of their synthesis, toxicity, and cost, which have rarely been considered in previous studies. This study used a combination of informatics strategies based on experimental databases and a newly built theoretical database.

To accelerate the material search, the synthetic feasibility, toxicity, and cost were estimated from elemental and atomic information. Furthermore, band gaps were predicted by an SVR machine learning model with a Gaussian kernel. The machine learning model was trained on the experimental band gap data of 282 perovskites. We believe that this dataset will be useful in follow-on studies in material research. Our high-throughput material search scheme can systematically consider the physical properties, toxicity, and cost, and can be modified for use in other material searches. For example, it could be extended to search for novel perovskite materials for the water-splitting photocatalytic reaction by simply changing the band gap screening criteria. Our work represents a pioneering material search method based on materials informatics that can consider various criteria with the aim of identifying materials for practical applications.

Through the materials search scheme, 24 promising candidates were discovered from 28,125,225 $AA'BB'X_3X'_3$-type compositions. Among the 24 discovered perovskites, 22 candidates were tin-halide perovskites, five of which are already well-known and employed as alternative materials for PSCs. Their identification is consistent with the results of recent experimental studies [7,8,59–61] and confirmed the reliability of our material search scheme. Novel sodium, potassium, and ammonium-based tin-halide perovskites were also identified. Therefore, we propose that not only MA-, FA-, and caesium-based perovskites but also sodium-, potassium-, and ammonium-based perovskites represent promising alternative PSC materials. In addition to the tin-halide perovskites, two novel transition-metal-based perovskites, $CaSrCu_2I_6$ and $CaBaMnNbS_6$, were identified. Therefore, the answer to the question, "Do promising alternative perovskites other than tin-halide perovskites exist?" is "yes." We report that there are alternative perovskites other than tin-halide perovskites that show low toxicity, low cost, and high performance as PSCs from the standpoint of materials informatics. This result represents valuable information to guide experimental alternative perovskite searches.

Acknowledgements

Calculations were performed at the Research Center for Computational Science, Okazaki, Japan. This study was supported in part by competitive funding for team-based basic research through the “Creation of Innovative Functions of Intelligent Materials on the Basis of the Element Strategy” from CREST, the Japan Science and Technology (JST) Agency; a Grant-in-Aid for Scientific Research in Innovative Areas of “Coordination Asymmetry” KAKENHI JP17H05380 from the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan; and a Grant-in-Aid from the Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Number JP18J12426. We would like to thank Editage (www.editage.jp) for English language editing.

Conflicts of interest

There are no conflicts to declare.

References

[1] S. De Wolf, J. Holovsky, S.-J. Moon, P. Löper, B. Niesen, M. Ledinsky, F.-J. Haug, J.-H. Yum and C. Ballif, *J. Phys. Chem. Lett.*, 2014, **5**, 1035-1039.

[2] Q. Lin, A. Armin, R. C. R. Nagiri, P. L. Burn and P. Meredith, *Nat. Photonics*, 2015, **9**, 106-112.

[3] G. Xing, N. Mathews, Sun, S. S. Lim, Y. M. Lam, M. Grätzel, S. Mhaisalkar and T. C. Sum, *Science*, 2013, **342**, 344-347.

[4] A. Kojima, K. Teshima, Y. Shirai and T. Miyasaka, *J. Am. Chem. Soc.*, 2009, **131**, 6050-6051.

[5] NREL efficiency chart, https://www.nrel.gov/pv/assets/images/efficiency-chart.png, (accessed September 2018).

[6] A. Babayigit, A. Ethirajan, M. Muller and B. Conings, *Nat. Mater.*, 2016, **15**, 247-251.

[7] N. K. Noel, S. D. Stranks, A. Abate, C. Wehrenfennig, S. Guarnera, A.-A. Haghighirad, A. Sadhanala, G. E. Eperon, S. K. Pathak, M. B. Johnston, A. Petrozza, L. M. Herz and H. J. Snaith, *Energy Environ. Sci.*, 2014, **7**, 3061-3068.

[8] Z. Zhao, F. Gu, Y. Li, W. Sun, S. Ye, H. Rao, Z. Liu, Z. Bian and C. Huang, *Adv. Sci.*, 2017, **4**, 1700204.

[9] A. H. Slavney, T. Hu, A. M. Lindenberg and H. I. Karunadasa, *J. Am. Chem. Soc.*, 2016, **138**, 2138-2141.

[10] M. R. Filip, S. Hillman, A. A. Haghighirad, H. J. Snaith and F. Giustino, *J. Phys. Chem. Lett.*, 2016, **7**, 2579-2585.

[11] G. Volonakis, A. A. Haghighirad, R. L. Milot, W. H. Sio, M. R. Filip, B. Wenger, M. B. Johnston, L. M. Herz, H. J. Snaith and F. Giustino, *J. Phys. Chem. Lett.*, 2017, **8**, 772-778.

[12] C. C. Stoumpos, L. Frazer, D. J. Clark, Y. S. Kim, S. H. Rhim, A. J. Freeman, J. B. Ketterson, J. I. Jang and M. G. Kanatzidis, *J. Am. Chem. Soc.*, 2015, **137**, 6804-6819.

[13] T. Nakajima and K. Sawada, *J. Phys. Chem. Lett.*, 2017, **8**, 4826-4831.

[14] M. R. Filip and F. Giustino, *J. Phys. Chem. C*, 2016, **120**, 166-173.

[15] K. Takahashi, L. Takahashi, I. Miyazato and Y. Tanaka, *ACS Photonics*, 2018, **5**, 771-775.

[16] A. Babayigit, D. D. Thanh, A. Ethirajan, J. Manca, M. Muller, H.-G. Boyen and B. Conings, *Sci. Rep.*, 2016, **6**, 18721.

[17] M. S. Beal, B. E. Hayden, T. L. Gall, C. E. Lee, X. Lu, M. Mirsaneh, C. Mormiche, D. Pasero, D. C. A. Smith, A. Weld, C. Yada and S. Yokoishi, *ACS Comb. Sci.*, 2011, **13**, 375-381.

[18] R. Gómez-Bombarelli, J. Aguilera-Iparraguirre, T. D. Hirzel, D. Duvenaud, D. Maclaurin, M. A. Blood-Forsythe, H. S. Chae, M. Einzinger, D.-G. Ha, T. Wu, G. Markopoulos, S. Jeon, H. Kang, H. Miyazaki, M. Numata, S. Kim, W. Huang, S. I. Hong, M. Baldo, R. P. Adams and A. Aspuru-Guzik, *Nat. Mater.*, 2016, **15**, 1120-1127.

[19] D. Xue, P. V. Balachandran, J. Hogden, J. Theiler, D. Xue and T. Lookman, *Nat. Commun.*, 2016, **7**, 11241.

[20] D. Xue, P. V. Balachandran, R. Yuan, T. Hu, X. Qian, E. R. Dougherty and T. Lookman, *Proc. Natl. Acad. Sci. USA*, 2016, **113**, 13301-13306.

[21] S. Nagasawa, E. Al-Naamani and A. Saeki, *J. Phys. Chem. Lett.*, 2018, **9**, 2639-2646.

[22] S. Curtarolo, D. Morgan, K. Persson, J. Rodgers and G. Ceder, *Phys. Rev. Lett.*, 2003, **91**, 135503.

[23] C. C. Fischer, K. J. Tibbetts, D. Morgan and G. Ceder, *Nat. Mater.*, 2006, **5**, 641-646.

[24] G. Pilania, P. V. Balachandran, C. Kim and T. Lookman, *Front. Mater.;*, 2016, **3**, 19.

[25] F. A. Faber, A. Lindmaa, O. A. von Lilienfeld and R. Armiento, *Phys. Rev. Lett.*, 2016, **117**, 135502.

[26] M. Fernandez, P. G. Boyd, T. D. Daff, M. Z. Aghaji and T. K. Woo, *J. Phys. Chem. Lett.*, 2014, **5**, 3056-3060.

[27] A. Seko, A. Togo, H. Hayashi, K. Tsuda, L. Chaput and I. Tanaka, *Phys. Rev. Lett.*, 2015, **115**, 205901.

[28] G. Pilania, A. Mannodi-Kanakkithodi, B. P. Uberuaga, R. Ramprasad, J. E. Gubernatis and T. Lookman, *Sci. Rep.*, 2016, **6**, 19375.

[29] C. Kim, G. Pilania and R. Ramprasad, *J. Phys. Chem. C*, 2016, **120**, 14575-14580.

[30] M. de Jong, W. Chen, R. Notestine, K. Persson, G. Ceder, A. Jain, M. Asta and A. Gamst, *Sci. Rep.*, 2016, **6**, 34256.

[31] G. Pilania, J. E. Gubernatis and T. Lookman, *Comput. Mater. Sci.*, 2017, **129**, 156-163.

[32] A. D. Sendek, Q. Yang, E. D. Cubuk, K.-A. N. Duerloo, Y. Cuic and E. J. Reed, *Energy Environ. Sci.*, 2017, **10**, 306-320.

[33] A. A. Emery, J. E. Saal, S. Kirklin, V. I. Hegde and C. Wolverton, *Chem. Mater.*, 2016, **28**, 5621-5634.

[34] E. Mosconi, P. Umaribc and F. De Angelis, *J. Mater. Chem. A*, 2015, **3**, 9208-9215.

[35] G. Pilania, A. Mannodi-Kanakkithodi, B. P. Uberuaga, R. Ramprasad, J. E. Gubernatis and T. Lookman, *Sci. Rep.*, 2016, **6**, 19375.

[36] G. Pilania, J. E. Gubernatis and T. Lookman, *Comput. Mater. Sci.*, 2017, **129**, 156-163.

[37] See Supplemental Material at [URL] for more information regarding the dataset and feature vector for machine learning, computational detail of first-principle calculation and power conversion efficiency estimation.

[38] D. G. Pettifor, *J. Phys. C: Solid State Phys.*, 1986, **19**, 285.

[39] 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 É. Duchesnay, *J. Mach. Learn. Res.*, 2011, **12**, 2825-2830.

[40] J. P. Perdew, K. Burke and M. Ernzerhof, *Phys. Rev. Lett.*, 1996, **77**, 3865.

[41] S. Grimme, J. Antony, S. Ehrlich and H. Krieg, *J. Chem. Phys.*, 2010, **132**, 154104.

[42] P. E. Blöchl, *Phys. Rev. B*, 1994, **50**, 17953.

[43] G. Kresse and J. Hafner, *Phys. Rev. B*, 1993, **47**, 558.

[44] V. Stevanović, S. Lany, X. Zhang and A. Zunger, *Phys. Rev. B*, 2012, **85**, 115104.

[45] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D.

398 Skinner, G. Ceder and K. A. Persson, *APL Materials*, 2013, **1**, 011002.

399[46] A. V. Krukau, O. A. Vydrov, A. F. Izmaylov and G. E. Scuseria, *J. Chem. Phys.*, 2006, **125**,

400 224106.

401[47] M. Gajdoš, K. Hummer, G. Kresse, J. Furthmüller and F. Bechstedt, *Phys. Rev. B*, 2006, **73**,

402 045112.

403[48] M. A. Butler and D. S. Ginley, *J. Electrochem. Soc.*, 1978, **125**, 228-232.

404[49] Y. Xu and M. A. A. Schoonen, *Am. Mineral.*, 2000, **85**, 543-556.

405[50] R. G. Pearson, *Inorg. Chem.*, 1988, **27**, 734-740.

406[51] L. M. Feng, L. Q. Jiang, M. Zhu, H. B. Liu, X. Zhou and C. H. Li, *J. Phys. Chem. Solids*, 2008,

407 **69**, 967-974.

408[52] R. D. Shannon, *Acta. Cryst.*, 1976, **32**, 751-757.

409[53] G. Kieslich, S. Suna and A. K. Cheetham, *Chem. Sci.*, 2014, **5**, 4712-4715.

410[54] C. J. Bartel, C. Sutton, B. R. Goldsmith, R. Ouyang, C. B. Musgrave, L. M. Ghiringhelli and M.

411 Scheffler, *Sci. Adv.*, 2019, **5**, eaav0693.

412[55] W. Shockley and H. J. Queisser, *J. Appl. Phys.*, 1961, **32**, 510.

413[56] Y. Dang, Y. Liu, Y. Sun, D. Yuan, X. Liu, W. Lu, G. Liu, H. Xia and X. Tao, *Cryst. Eng. Comm.*,

414 2015, **17**, 665-670.

415[57] Pharmaceuticals and Medical Devices Agency, http://www.pmda.go.jp/files/000197758.pdf,

416 (accessed Feb 2018).

[58] Chemicool Periodic Table, www.chemicool.com, (accessed Feb 2018).

[59] C. C. Stoumpos, C. D. Malliakas and M. G. Kanatzidis, *Inorg. Chem.*, 2013, **52**, 9019-9038.

[60] F. Chiarella, A. Zappettini, F. Licci, I. Borriello, G. Cantele, D. Ninno, A. Cassinese and R. Vaglio, *Phys. Rev. B*, 2008, **77**, 045129.

[61] C. Ferrara, M. Patrini, A. Pisanu, P. Quadrelli, C. Milanese, C. Tealdi and L. Malavasi, *J. Mater. Chem. A*, 2017, **5**, 9391-9395.

Author Information

*Corresponding author

Correspondence and requests for materials should be addressed to yimamura@tmu.ac.jp.

Table 1. Averaged R^2 values for training set R^2_{train} and test set R^2_{test}, and root mean squared error (RMSE) in test set for each band-gap prediction model.

Regression model	R^2_{train}	R^2_{test}	RMSE
MLR	0.46	0.36	1.11
Ridge regression	0.46	0.37	1.10
Lasso regression	0.45	0.36	1.11
SVR with linear kernel	0.43	0.34	1.12
SVR with Gaussian kernel	0.89	0.65	0.81
GPR with Gaussian kernel	0.90	0.58	0.89

Table 2. Band gap predicted by machine learning (E_g), estimated cost, Bravais lattice vector length ratios ($|\mathbf{a}|/|\mathbf{b}|$, $|\mathbf{a}|/|\mathbf{c}|$, and $|\mathbf{b}|/|\mathbf{c}|$), angles between Bravais lattice vectors (α, β, and γ), electron effective mass m_e^*, hole effective mass m_h^*, exciton binding energy E_b, and theoretical PCE of 24 most promising perovskites identified.

Perovskite	E_g/eV	Cost/\$ mol^{-1}	$\|\mathbf{a}\|/\|\mathbf{b}\|$	$\|\mathbf{a}\|/\|\mathbf{c}\|$	$\|\mathbf{b}\|/\|\mathbf{c}\|$	α/degree	β/degree	γ/degree	m_e^*/a.u.	m_h^*/a.u.	E_b/meV	PCE/%
$CaBaMnNbS_6$	2.18	194	1.91	1.91	1.00	90.0	90.0	90.0	0.55	0.62	0	13.6
$CaSrCu_2I_6$	1.81	171	2.00	2.00	1.00	90.0	90.0	90.0	0.36	0.45	10	19.1
$FASnI_3$	1.51	122	1.96	2.00	1.02	90.0	90.0	90.1	0.84	0.10	4	22.2
$KSnBr_3$	1.92	159	2.00	2.00	1.00	90.0	90.0	90.0	0.79	0.09	9	17.4
$K_2Sn_2Br_3I_3$	1.13	179	1.98	1.98	1.00	90.0	90.0	90.0	0.81	0.07	0	22.1
$K_2Sn_2Cl_3I_3$	1.75	167	1.98	1.98	1.00	90.0	90.0	90.0	0.98	0.13	5	19.8
$KSnI_3$	1.14	198	2.00	2.00	1.00	90.0	90.0	90.0	0.74	0.08	2	22.1
$KMASn_2Br_6$	1.99	121	1.99	2.00	1.00	90.5	90.0	90.0	0.94	0.15	33	16.4
$KMASn_2Br_3I_3$	1.17	141	1.98	1.98	1.00	90.2	90.3	89.9	0.97	0.13	14	22.2
$KMASn_2I_6$	1.16	160	2.00	1.99	1.00	90.4	90.0	90.0	0.81	0.11	9	22.2
$KNH_4Sn_2Br_6$	1.97	121	2.00	2.00	1.00	90.2	90.0	90.0	0.83	0.08	1	16.7
$KNH_4Sn_2Br_3I_3$	1.14	140	2.00	1.99	1.00	90.8	90.1	90.0	0.79	0.08	0	22.1
$KNH_4Sn_2I_6$	1.12	160	2.00	2.00	1.00	90.1	90.0	90.0	0.72	0.07	0	22.0
$MAFASn_2I_6$	1.39	122	1.98	1.98	1.00	90.1	89.4	90.0	1.00	0.09	2	23.0
$MA_2Sn_2Br_3I_3$	1.33	103	1.99	1.97	0.99	90.4	90.0	90.1	1.00	0.13	13	22.9
$MA_2Sn_2Cl_3I_3$	1.96	91	1.96	1.96	1.00	89.9	89.8	90.5	0.88	0.12	11	16.8
$MASnI_3$	1.29	122	2.00	1.98	0.99	90.3	90.0	90.0	0.78	0.15	16	22.3
$NaKSn_2Br_6$	1.94	126	2.00	2.00	1.00	90.0	90.0	90.0	0.78	0.07	0	17.2

$NaKSn_2Br_3I_3$	2.12	135	1.99	1.99	1.00	90.0	90.0	90.0	0.86	0.18	13	14.5
$NaKSn_2I_6$	1.54	138	2.00	2.00	1.00	90.0	90.0	90.0	0.73	0.07	0	21.9
$Na_2Sn_2Br_3I_3$	1.48	112	2.00	2.00	1.00	90.0	90.0	90.0	0.81	0.18	22	22.3
$Na_2Sn_2Cl_3I_3$	1.92	100	1.99	1.99	1.00	90.0	90.0	90.0	0.91	0.15	8	17.4
$NaSnI_3$	1.86	147	2.00	2.00	1.00	90.0	90.0	90.0	0.72	0.07	0	18.3
$NaNH_4Sn_2I_6$	1.22	126	2.00	2.00	1.00	90.1	90.0	90.0	1.00	0.07	0	22.3

434

435

436 **Table 3.** Formation enthalpies, direct band gap, indirect band gap, gap type, and levels of CBM and VBM relative to vacuum level of 19 novel

437 perovskites identified.

Perovskite	Formation enthalpy / kJ mol^{-1}	Direct band gap / eV	Indirect band gap / eV	Gap type	CBM level / eV	VBM level / eV
$CaBaMnNbS_6$	-1664	0.55	0.00	Indirect	-4.63	-4.63
$CaSrCu_2I_6$	-1125	0.65	0.08	Indirect	-4.88	-4.97
$KSnBr_3$	-610	0.70	0.70	Direct	-5.04	-5.74
$K_2Sn_2Br_3I_3$	-1053	0.39	0.39	Direct	-5.01	-5.40
$K_2Sn_2Cl_3I_3$	-1161	0.46	0.46	Direct	-5.31	-5.77
$KSnI_3$	-458	0.46	0.46	Direct	-4.80	-5.26
$KMASn_2Br_6$	-1158	0.95	0.95	Direct	-5.53	-6.48
$KMASn_2Br_3I_3$	-1002	0.56	0.56	Direct	-5.51	-6.08
$KMASn_2I_6$	-878	0.54	0.54	Direct	-5.33	-5.87
$KNH_4Sn_2Br_6$	-1117	0.51	0.51	Direct	-5.76	-6.27
$KNH_4Sn_2Br_3I_3$	-960	0.29	0.29	Direct	-5.66	-5.95
$KNH_4Sn_2I_6$	-828	0.32	0.32	Direct	-5.45	-5.77
$NaKSn_2Br_6$	-1075	0.49	0.49	Direct	-5.23	-5.72
$NaKSn_2Br_3I_3$	-919	0.51	0.51	Direct	-5.04	-5.55
$NaKSn_2I_6$	-785	0.37	0.37	Direct	-4.93	-5.29
$Na_2Sn_2Br_3I_3$	-821	0.66	0.66	Direct	-5.05	-5.71
$Na_2Sn_2Cl_3I_3$	-941	0.51	0.51	Direct	-5.27	-5.78
$NaSnI_3$	-327	0.24	0.24	Direct	-5.07	-5.32
$NaNH_4Sn_2I_6$	-699	0.20	0.20	Direct	-5.60	-5.80

438

Figure Caption

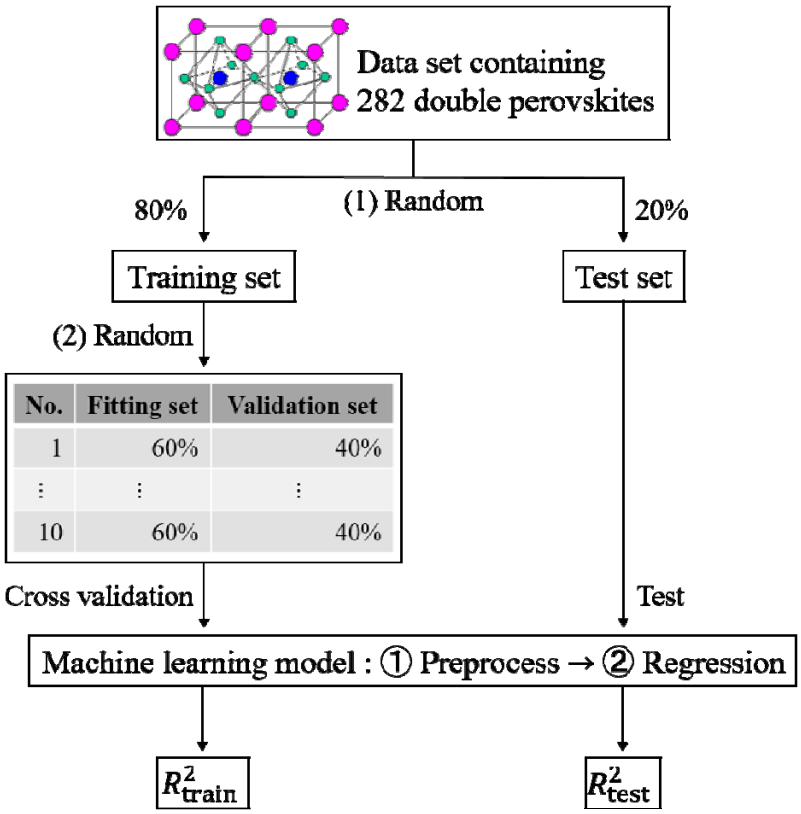
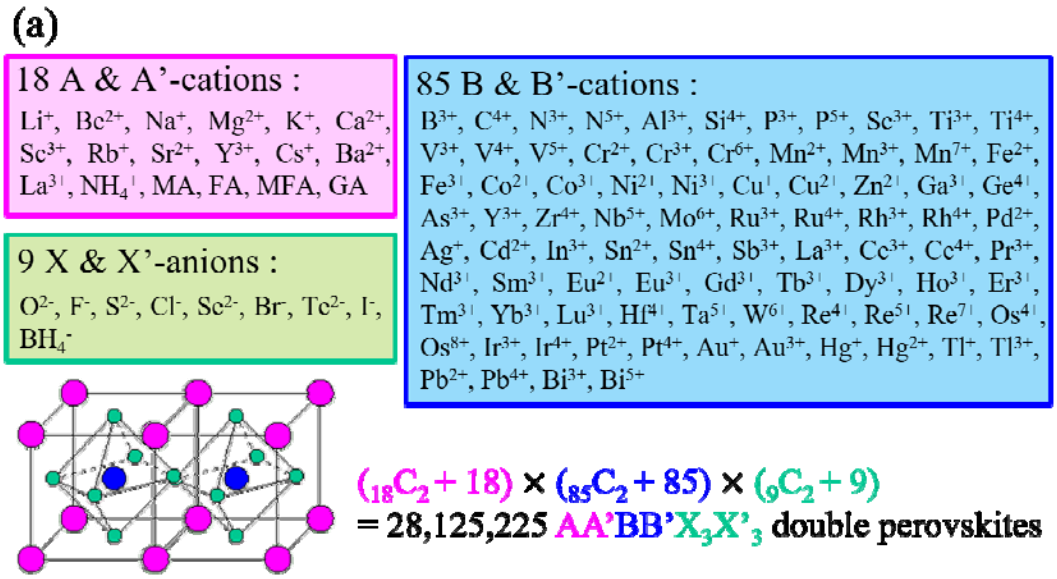
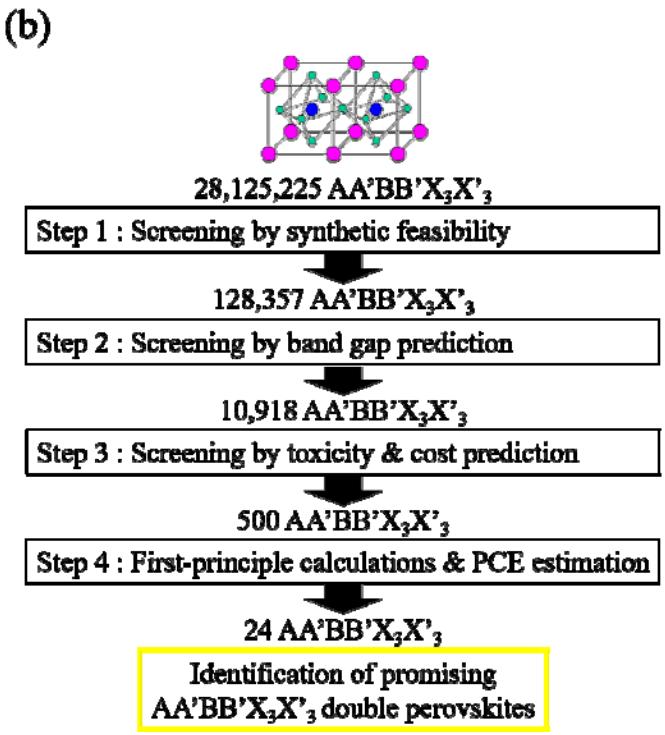


Figure 1. Training and test processes for machine learning models for band gap prediction. Percentages in figure are split ratios for each set. Random seed for (1) Random was changed every time in 50 iterations, and random seed for (2) Random was fixed in the iterations.



445



446

447 **Figure 2.** (a) Ion library for each site, and number of possible $AA'BB'X_3X'_3$ compositions using the

448 library. (b) Diagram of novel $AA'BB'X_3X'_3$ double perovskite search scheme.