npj | computational materials

Article

Published in partnership with the Shanghai Institute of Ceramics of the Chinese Academy of Sciences

https://doi.org/10.1038/s41524-024-01270-1

Machine learning-enabled chemical space exploration of all-inorganic perovskites for photovoltaics

Check for updates

Jin-Soo Kim, Juhwan Noh & Jino Im

The vast compositional and configurational spaces of multi-element metal halide perovskites (MHPs) result in significant challenges when designing MHPs with promising stability and optoelectronic properties. In this paper, we propose a framework for the design of B-site-alloyed ABX_3 MHPs by combining density functional theory (DFT) and machine learning (ML). We performed generalized gradient approximation with Perdew–Burke–Ernzerhof functional for solids (PBEsol) on 3,159 B-site-alloyed perovskite structures using a compositional step of 1/4. Crystal graph convolution neural networks (CGCNNs) were trained on the 3159 DFT datasets to predict the decomposition energy, bandgap, and types of bandgaps. The trained CGCNN models were used to explore the compositional and configurational spaces of 41,400 B-site-alloyed ABX_3 MHPs with a compositional step of 1/16, by accessing all possible configurations for each composition. The electronic band structures of the selected compounds were calculated using the hybrid functional (PBE0). Then, we calculated the optical absorption spectra and spectroscopic limited maximum efficiency of the selected compounds. Based on the DFT/ML-combined screening, 10 promising compounds with optimal bandgaps were selected, and from among these 10 compounds, $CsGe_{0.3125}Sn_{0.6875}I_3$ and $CsGe_{0.0625}Pb_{0.3125}Sn_{0.625}Br_3$ were suggested as photon absorbers for single-junction and tandem solar cells, respectively. The design framework presented herein is a good starting point for the design of mixed MHPs for optoelectronic applications.

Metal halide perovskites (MHPs) are a class of materials with the chemical formula of ABX_3, where A, B, and X represent monovalent organic/inorganic cations, divalent metal cations, and monovalent halide anions, respectively[1] MHPs have received considerable research efforts[1] owing to their outstanding optoelectronic properties, low cost, and easy fabrication, which make them promising materials for photovoltaics[2], light-emitting diodes[3], lasers[4], and photodetectors[5]. Among the various MHPs, compounds containing methylammonium (MA) or formamidinium (FA) at the A-site and Pb at the B-site have been extensively studied and have demonstrated the best optoelectronic properties[5]. However, commercialization of the MHPs is still challenging because of several obstacles, such as the toxicity of Pb and instability under adverse conditions such as high temperature and humidity, which primarily originate from organic cations[6]. Therefore, to facilitate their commercialization, MHPs with better stability while containing less Pb should be developed without compromising their optoelectronic properties[6].

To overcome these challenges, several researchers have suggested substitutional alloying (i.e., doping or mixing with different elements or chemicals at each site in an MHP)[7] as a promising approach for tuning the stability and optoelectronic properties of MHPs[8], and thus, the Pb-derived toxicity of MHPs can be effectively mitigated by partially replacing Pb with other elements. In the field of metallurgy, the concept of high-entropy alloys has recently demonstrated that mixing multiple elements can improve thermodynamic stability by exploiting the large configurational entropy[9]. This concept has also been applied in MHPs[10]. For example, the A-site mixing entropy improved the stability of perovskite solar cells[11,12] and light-emitting diodes[13]. Moreover, high-entropy alloying of the B-site of $MAPbBr_3$ improved the colloidal stability and optical performance while reducing the Pb content[14]. In this study, we primarily focused on B-site alloying to achieve both improved stability and lower Pb content.

However, because the compositional space of MHPs with multiple alloying elements is virtually infinite, utilizing an experimental design

Chemical Data-driven Research Center, Korea Research Institute of Chemical Technology, Daejeon 34114, Republic of Korea.
e-mail: js.kim@krict.re.kr; jhwan@krict.re.kr; imjino@krict.re.kr

approach is not practical for exploring the chemical space of element-mixed MHPs. Even the use of high-throughput experiments only achieved the synthesis of a few tens of binary and ternary alloys of MHPs[15–17]. In this regard, several computational approaches utilizing either density functional theory (DFT) calculations[18–21] or a combination of DFT with data-driven machine learning (ML) models[22–24] have been proposed for the rapid screening of the compositional space of MHPs. For example, Wang et al.[18] investigated the effect of entropy on the thermodynamic stability and optoelectronic properties of $Cs_2B^+B^{3+}Cl_6$ double-perovskite alloys using DFT calculations. They used the special quasi-random structure (SQS)[25,26] approach to model disordered double-perovskite alloys, showing entropy-driven stabilization. Moreover, Yang et al.[19] presented generalized gradient approximation and hybrid-level computations on a DFT dataset of 495 ABX_3 perovskite alloys, performed modeling mixed perovskites using the SQS approach, and suggested 32 compounds showing promising stability and photovoltaic efficiency from the screening.

For exploration using ML, Choubisa et al.[23] developed a "crystal site feature embedding" (CSFE) representation, which achieves low errors when predicting DFT energies and bandgaps of mixed MHPs. They used CSFE to explore the chemical space of MHPs and discovered how a small amount of Cd doping (i.e., ~1.5%) can change the bandgap to 1.1 eV, which is desirable for the active layers of solar cells[23]. Moreover, Mannodi-Kanakkithodi et al.[24] proposed an ML-driven high-throughput screening framework based on the stability, bandgap, and defect tolerance of mixed MHPs, where the tabulated elemental properties of each species of the MHPs were used as inputs for ML (i.e., neural networks). Using the proposed screening framework, 392 out of 17955 MHPs were identified as promising candidates for photon absorbers[24].

Despite the practical applicability of the aforementioned computational design framework, most previous studies do not guarantee the most stable atomic configuration for each composition. That is, previous works[18–21,23,24] utilized random sampling or SQS methods to estimate the properties of random alloys and did not explore all possible configurations exhaustively. Understanding the properties of the ground-state atomic configuration is important because the atomic configuration of an alloy is likely to be the most stable configuration, and the properties of alloys, such as the bandgap, can be significantly altered by configuration changes, even at the same composition[27,28]. In this regard, Yamamoto et al.[22] utilized the cluster expansion approach to identify the ground state of B-site mixed iodide perovskites (i.e., $AB_xB'_{1-x}I_3$); however, the authors only investigated the effect of B-site mixing on thermodynamic stability.

In this work, we propose a DFT/ML-combined framework (Fig. 1) for the design of B-site-alloyed MHPs with improved stability and optoelectronic properties by accessing all possible atomic configurations. To achieve this, we employed a crystal graph convolution neural network (CGCNN)[29] as a surrogate model for predicting the stability and the electronic band structure computed at the Perdew–Burke–Ernzerhof functional revised for solids (PBEsol) level[30]. For the construction of the DFT-calculated database used for training the CGCNN, a 20-atom unit cell was used, in which B-site alloyed MHPs were modeled with a compositional step of 1/4. The trained CGCNN is then used to exhaustively explore the enlarged chemical space of B-site alloyed MHPs with the increased compositional resolution by four times (i.e., 1/16 compositional step) remaining the A- and X-sites as unary (i.e., A: Cs, K, and Rb; X: Br, Cl, and I). Through ML-assisted screening, 110 compounds were predicted to be thermodynamically stable and exhibited promising electronic band structures, and validations on the

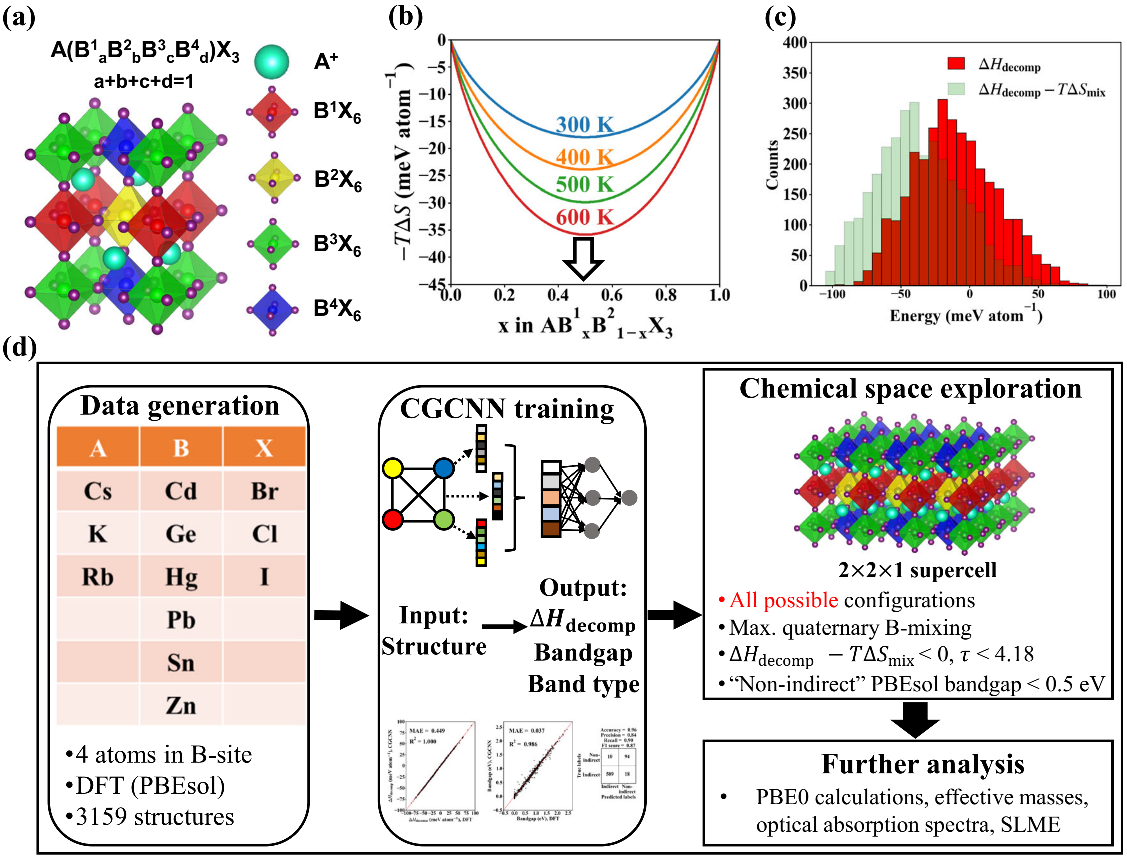


Fig. 1 | Overview of the present study. **a** Schematic representation of crystal structure for ABX_3 metal halide perovskites (MHPs). B^1, B^2, B^3, and B^4 represent elements positioned in the B-site of MHPs. **b** Mixing of entropic contribution ($-T\Delta S_{mix}$) in $AB^1_xB^2_{1-x}X_3$ system at varying temperature. **c** Histogram of decomposition enthalpy (ΔH_{decomp}) with and without the mixed entropic contribution (at $T = 298$ K) in our Perdew–Berke–Ernzerhof functional revised for solids (PBEsol) dataset. **d** Schematic workflow of the present study.

thermodynamic stability and electronic band structures were further conducted using DFT calculations with the semi-local PBEsol and hybrid PBE0[31] functionals, respectively. From this, we identified 10 promising candidates for solar-cell application after examining the PBE0-calculated bandgaps. To this end, the main contributions of our work can be summarized as follows.

- We explored the whole possible atomic configurational space using the trained CGCNN model to identify the most stable atomic configurations within the given B-site mixed MHP composition which is not explicitly identified in the previous work as mentioned earlier.
- We expanded the explored B-site mixed MHP domain up to the quaternary phase including the mixing entropy effect in terms of both thermodynamic stability and the electronic properties compared to the previous works which explored up to the ternary phase for B-site mixing without the mixing entropy effect[23,24].
- We proposed to use PBE0[31] combined with the spin-orbit coupling (SOC) correction[32] term as a more accurate approach for estimating experimental bandgaps of MHPs exhibiting lower prediction error compared to the other hybrid functionals such as HSE06[33] and B3LYP[34].

Results

Design strategy

Figure 1d illustrates the overall workflow of this study. The generated PBEsol data were used to train the CGCNN models (see details in **Training data generation** in **METHODS** section). Specifically, three CGCNN models were independently trained to predict three target properties (see details in **Crystal Graph Convolutional Neural Networks training** in **METHODS** section): regression of the decomposition enthalpy ($\Delta H_{\mathrm{decomp}}$) and bandgap ($E_{\mathrm{gap}}$) and binary classification for band type (see also **Thermodynamic stability and bandgap of B-site mixed metal halide perovskites** in **METHODS** section for details on calculating each target property). For the band type classification, the non-indirect and indirect band types were labeled as positive and negative classes, respectively.

To explore the chemical space of multi-element MHPs, we used a four-fold enlarged structure compared to the training data, containing 16 B-sites (i.e., 80 atoms per unit cell). This allows a finer resolution of the compositional step (i.e., 6.25 at.%) and we considered B-site alloying up to the quaternary system. For each composition of the MHP, CGCNN-predicted $\Delta H_{\mathrm{decomp}}$ (i.e., $\Delta H^{\mathrm{CGCNN}}_{\mathrm{decomp}}$) was used to identify the most stable atomic configurations, and the mixing entropy term ($-T\Delta S_{\mathrm{mix}}$) was added to $\Delta H^{\mathrm{CGCNN}}_{\mathrm{decomp}}$ at a temperature of 298 K. The entropic contribution, $-T\Delta S_{\mathrm{mix}}$, is proportional to temperature, as can be seen in Fig. 1b. Therefore, the entropy-driven stabilization is more effective at higher temperatures. Even at 298 K, the distribution of $\Delta H_{\mathrm{decomp}} - T\Delta S_{\mathrm{mix}}$ in our training data shifted negatively when compared to that of $\Delta H_{\mathrm{decomp}}$, as shown in Fig. 1c.

Next, we considered Bartel's tolerance factor[35], τ, which is a new data-driven tolerance factor, to classify whether an arbitrary compound is a perovskite (see details in **Tolerance factor of B-site mixed metal halide perovskites** in **METHODS** section). τ exhibited better classification accuracy than Goldschmidt's tolerance factor[36]. We also compared the classification accuracy of τ with that of Filip's geometric limits for the formability of perovskites[37]. After a comparison of classification accuracy for eight materials that were experimentally confirmed to exist as perovskites, we concluded that τ predicts closer to the experimental results (see details in Supplementary Note 2).

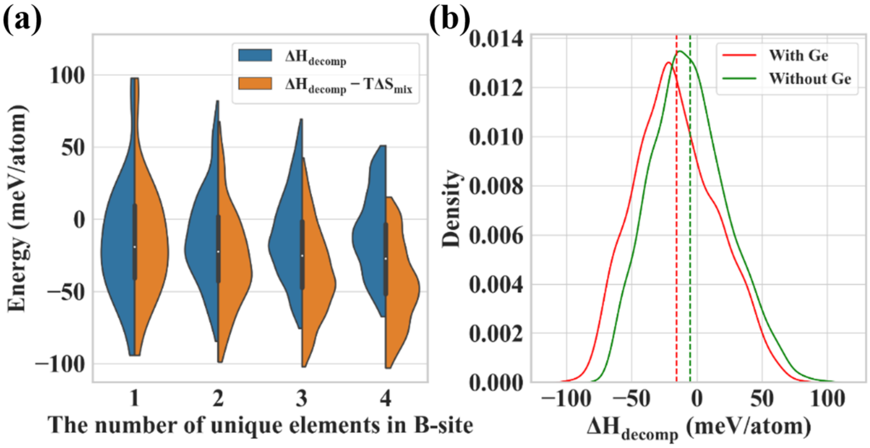
For solar-cell applications, CGCNN predicts the band type and bandgap ($E^{\mathrm{CGCNN}}_{\mathrm{gap}}$, hereafter) of the atomic configuration with the lowest $\Delta H_{\mathrm{decomp}}$ at each composition. Because the CGCNN model is trained to predict PBEsol-calculated bandgap ($E^{\mathrm{PBEsol}}_{\mathrm{gap}}$), we selected the compositions showing direct bandgaps with the condition $E^{\mathrm{CGCNN}}_{\mathrm{gap}} < 0.5$ eV as promising compounds for photovoltaics. Considering that PBEsol underestimates bandgap by approximately 1.0–1.5 eV when compared to PBE0 (refer to Supplementary Table 5), we expect that the candidates satisfying the latter condition ($E^{\mathrm{CGCNN}}_{\mathrm{gap}} < 0.5$ eV) will show a PBE0-calculated bandgap ($E^{\mathrm{PBE0}}_{\mathrm{gap}}$) of approximately 1.0–2.0 eV, which is close to an ideal bandgap range of approximately 1.2–1.4 eV for single-junction photovoltaics according to the Shockley–Queisser limit[38]. Moreover, this bandgap range ($1.0 < E^{\mathrm{PBE0}}_{\mathrm{gap}} < 2.0$ eV) is close to the ideal bandgap for a top cell in a tandem solar cell (1.72 eV), with silicon as the bottom cell, which has a bandgap of 1.1 eV[17].

The compounds selected based on the stability and band structure criteria were further validated for their stability and band structure through PBEsol and PBE0 calculations, respectively, to examine their potential application in photovoltaics. We optimized the geometry of the selected compounds using PBEsol and then calculated the band structure using PBE0. We calculated the carrier effective mass, optical absorption spectra, and spectroscopic limited maximum efficiency (SLME). The SLME was calculated according to a previous study by Yu et al.[39] using the open-source SL3ME code (https://github.com/ldwillia/SL3ME). In the SLME calculations, the optical absorption spectra and bandgap were used as inputs, assuming an air mass 1.5 global solar spectrum (https://www.nrel.gov/grid/solar-resource/spectra-am1.5.html).

Analysis of the training dataset

The distribution of the thermodynamic stability of the 3,159 DFT (PBESol)-calculated training data was first analyzed, as shown in Fig. 2. As shown in Fig. 2a, we observed that $\Delta H_{\mathrm{decomp}}$ distribution moves toward a lower energy regime (i.e., orange color in Fig. 2a) by employing the mixing entropy term (i.e., $-T\Delta S_{\mathrm{mix}}$). As expected, a greater shift to the stable region can be obtained by mixing more elements. In fact, without considering the mixing entropy, $CsGeBr_3$ showed the lowest $\Delta H_{\mathrm{decomp}}$ value among the 3159 DFT-calculated data; however, we obtained more stable data for the composition

Fig. 2 | Analysis of the training dataset.
a Decomposition enthalpy ($\Delta H_{\mathrm{decomp}}$) distribution of 3159 density functional theory (DFT)-calculated data with respect to the number of unique elements in the B-site before (blue) and after (orange) including the mixing entropy term (at $T = 298$ K), respectively. **b** Difference in distribution of the decomposition energy with the existence of Ge element in the training dataset. Dashed lines indicate the mean value of each decomposition energy distribution.



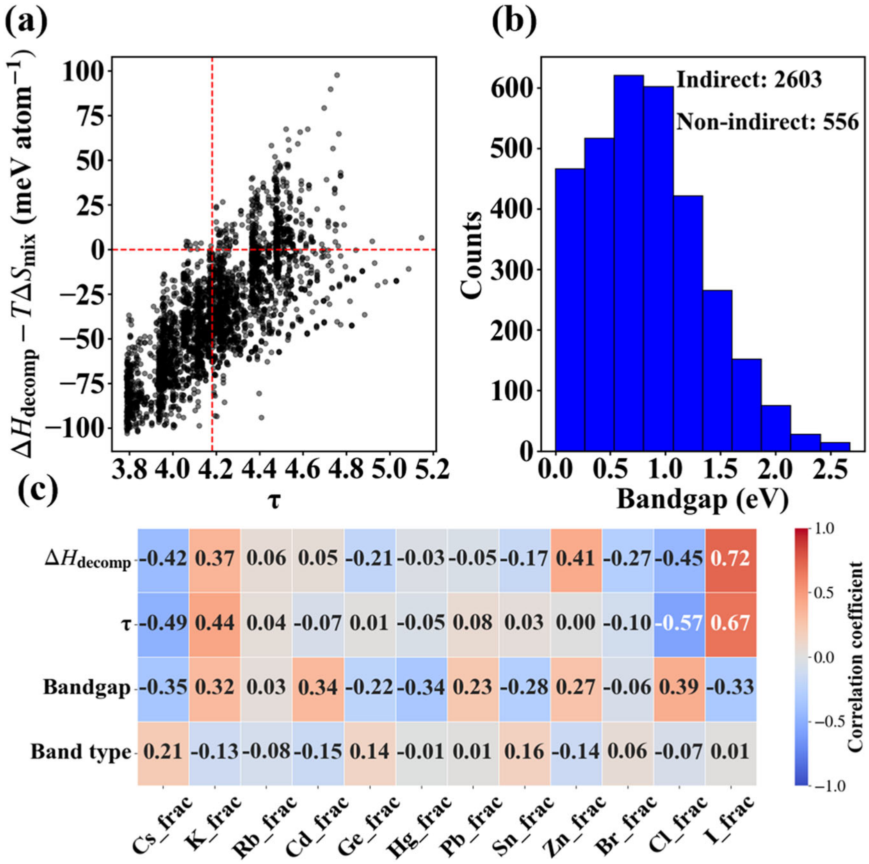


Fig. 3 | Visualization of the training data set. **a** Decomposition enthalpy with mixing entropy term ($\Delta H_{decomp} - T\Delta S_{mix}$) versus Bartel's tolerance factor (τ) and **b** histogram of bandgap with the number of data for indirect and non-indirect bandgap displayed. "Non-indirect" includes direct, metallic, and semi-metallic materials. **c** Pearson correlation coefficient between the four output properties and fractions of elements.

of $CsGe_{0.25}Hg_{0.25}Pb_{0.25}Sn_{0.25}Cl_3$ corresponding to the lowest $\Delta H_{decomp} - T\Delta S_{mix}$ value among the entire 3159 data. Additionally, we note that alloys with Ge lead to ΔH_{decomp} distribution in the lower energy range when compared to that of alloys without Ge (Fig. 2b), which is consistent with the results of previous computational studies[19,24] (refer to Supplementary Fig. 6 for the effects of including other elements at the B-site). As shown in Supplementary Fig. 6, alloys with Zn result in ΔH_{decomp} distribution in a higher energy range when compared to alloys without Zn. Consequently, ΔH_{decomp} becomes lowest (highest) when the B-site contains only Ge (Zn). Therefore, when the B-site contains a single element, the ΔH_{decomp} distribution is widest, while the ΔH_{decomp} distribution narrows as the number of elements in the B-site increases.

We analyzed the correlation between perovskite formability as described by Bartel's tolerance factor (τ) and thermodynamic stability (Fig. 3a). Fig. 3a shows a positive correlation between τ and $\Delta H_{decomp} - T\Delta S_{mix}$, as both values indicate that compounds with lower values are more stable. However, a region showing inconsistency between the two metrics exists at $\tau > 4.18$, which might be due to the imperfectness of τ. The dataset used to derive τ does not contain B-site alloyed perovskites[35], thereby resulting in an inaccurate prediction for B-site alloyed perovskites. Moreover, τ ignores the mixing entropy-driven stabilization (Eq. (3)); thus, when we compare ΔH_{decomp} with τ, less data exists in the inconsistent region ($\tau > 4.18$ and $\Delta H_{decomp} < 0$), as shown in Supplementary Fig. 7. Specifically, the data deviating from the positive correlation (i.e., $\tau > 4.18$ and ΔH_{decomp}-$T\Delta S_{mix} < -85$ meV atom^{-1}) are all Ge-containing compounds, namely, $CsGeBr_3$, $CsGeCl_3$, $CsCd_{0.25}Ge_{0.75}Br_3$, and $CsGe_{0.75}Hg_{0.25}Br_3$; however, experimental results have confirmed that $CsGeBr_3$ and $CsGeCl_3$ exist as perovskite structures[40]. Despite this imperfection of τ, we can safely apply the condition, $\tau < 4.18$, to screen stable compounds if combined with the condition of maintaining the lowest possible value of $\Delta H_{decomp} - T\Delta S_{mix}$, because the two metrics show a positive correlation. For electronic band-related properties, as shown in Fig. 3b, the amount of indirect-bandgap data was 4.68 times larger than that of the non-indirect-bandgap data. Moreover, of the 905 data with a bandgap of less than 0.5 eV, 654 data were indirect.

Next, we calculated the Pearson correlation coefficient between the elemental fractions and the four output properties, ΔH_{decomp}, τ, bandgap, and band type, and the results are illustrated in Fig. 3c. ΔH_{decomp} and τ tend to be lower (more stable) when the A-site is occupied by Cs, whereas I in the X-site increases ΔH_{decomp} and τ (more unstable). An increase in the fraction of Ge tends to decrease ΔH_{decomp}, whereas increasing the fraction of Zn tends to increase ΔH_{decomp}. The bandgap tends to decrease when the A- and X-sites are occupied by Cs and I, respectively. Notably, the correlation trend between ΔH_{decomp} and bandgap is opposite in the X-site: I (Cl) tends to increase (decrease) ΔH_{decomp} but decrease (increase) the bandgap. Increasing the Ge, Hg, and Sn fractions decreases the bandgap, whereas increasing the Cd, Pb, and Zn fractions increases it. In this correlation analysis of the band type, we represented the indirect band type as zero and the non-indirect band type as one. Thus, a positive correlation implies that increasing the fraction of certain elements tends to change the compounds into non-indirect-bandgap materials and vice versa. The Cs fraction showed the highest positive correlation (0.21), whereas the fractions of all other elements exhibited a relatively low correlation with the band type.

Validation of the trained CGCNN models

To validate the prediction performance of the trained CGCNN models, we plotted the test set prediction results, as shown in Fig. 4. When predicting the thermodynamic stability (Fig. 4a), the trained CGCNN model showed promising prediction accuracy with a mean absolute error (MAE) of 0.45 meV atom^{-1}, indicating that the model could effectively capture the quantitative relation between the crystal structure and thermodynamic stability.

The trained models also showed highly promising prediction performance for the prediction of the bandgap (Fig. 4b) and band type (Fig. 4c). Specifically, the prediction performance of the trained CGCNN model for

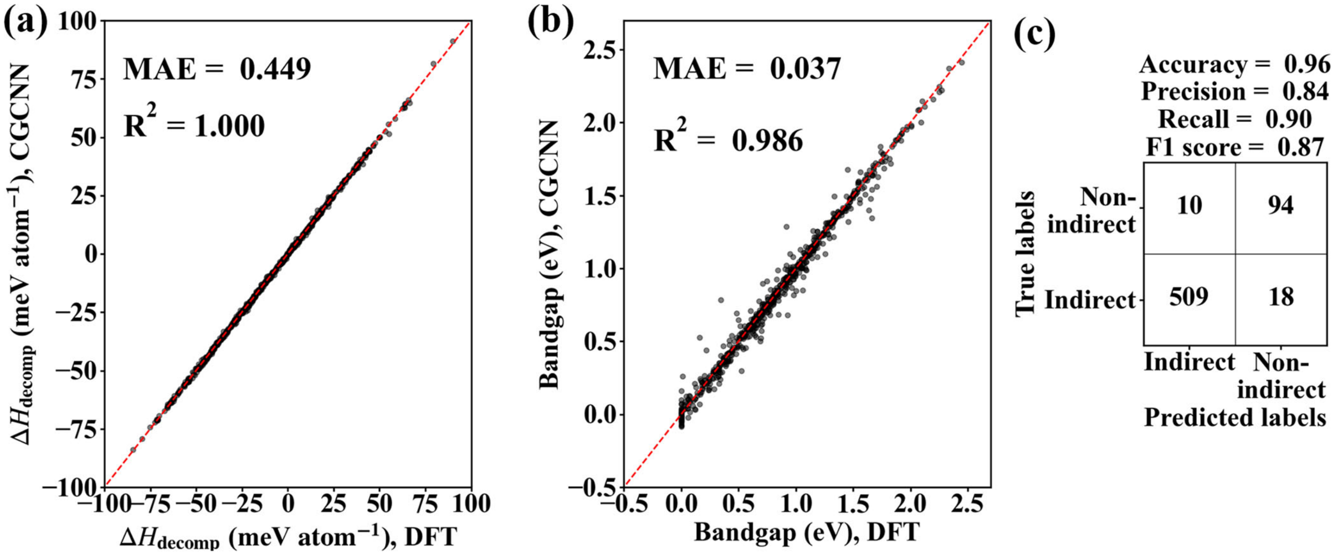


Fig. 4 | Validation of trained models. Parity plot between crystal graph convolution neural network (CGCNN)-predicted and DFT (PBESol)-calculated (**a**) $\Delta H_{\mathrm{decomp}}$ and **b** bandgap on the test set. **c**: Confusion matrix of classification test. “Non-indirect” includes direct, metallic, and semi-metallic materials.

bandgap prediction (Fig. 4b, MAE of 0.037 eV and root mean squared error of 0.061 eV) was superior to that of a previous study[24].

In the classification of the band type (Fig. 4c), our model showed promising classification performance with a low inclusion of misclassified data (i.e., false-negative and false-positive data). Only 3 and 10% of the negative and positive classes, respectively, were misclassified with a high accuracy of 0.96 and recall value of 0.90. In other words, our model could effectively recover the true-positive distribution with a relatively low inclusion of false-positive data from the negative class. Here, we noted that the relatively low precision value of 0.84 was due to the class distribution imbalance of the negative data (Fig. 3b). Despite the large class imbalance between the positive and negative data, the model shows a similar amount of misclassified positive and negative data, as shown in Fig. 4c. Therefore, we expect no significant reduction in sampling efficiency due to the inclusion of false-positive data.

Design of B-site mixed MHPs for solar-cell application

As described previously, we applied the trained CGCNN models to explore the chemical space of the B-site-alloyed MHPs with a four-fold enlarged unit cell consisting of 80 atoms. For A- and X-sites, we considered the chemical elements of {Cs, K, Rb} and {Br, Cl, I}, respectively. For B-site alloying, we utilized the chemical elements of {Ge, Sn, Pb, Zn, Cd, and Hg} under binary, ternary, and quaternary alloying conditions. The compositional ratio of the alloying was varied from zero to one in 0.0625 steps. Then the search space consisted of 2025 chemical compositions for binary alloying and 18,900 for ternary alloying. For quaternary alloying, we considered only Cs for the A-site to reduce the computational cost of the search procedure. Through the exploration of binary and ternary alloying systems, we determined that $\Delta H_{\mathrm{decomp}}$ becomes lower when Cs is in the A-site. Thus, the search space for quaternary alloying was reduced to 20,475 compositions. For each chemical composition, the CGCNN model was applied to all possible atomic configurations to determine the lowest $\Delta H_{\mathrm{decomp}} - T\Delta S_{\mathrm{mix}}$. In total, 41,400 compositions and approximately 5.6 × 10^{11} atomic configurations were explored for alloyed MHPs using the CGCNN model during the search process. The details of computing 5.6 × 10^{11} atomic configurations are described in the Supplementary Note 3.

Figure 5 shows the distributions of $\Delta H^{\mathrm{CGCNN}}_{\mathrm{decomp}} - T\Delta S_{\mathrm{mix}}$ of $\mathrm{CsGe_xSn_{1\text{-}x}Br_3}$ and $\mathrm{CsGe_xHg_ySn_{1\text{-}x\text{-}y}Cl_3}$ systems as case examples, showing the lowest $\Delta H^{\mathrm{CGCNN}}_{\mathrm{decomp}} - T\Delta S_{\mathrm{mix}}$ in binary and ternary systems, respectively. As mentioned previously, the CGCNN covers the compositions beyond the chemical space of the training data because the unit cell size is four times larger than that of the training data. We also observed large variations in the $\Delta H^{\mathrm{CGCNN}}_{\mathrm{decomp}} - T\Delta S_{\mathrm{mix}}$ for compositions with a high ΔS_{mix} (i.e., close to equiatomic ratio). In particular, at the fraction of 0.5 in Fig. 5a, three inequivalent atomic configurations exhibit different $\Delta H_{\mathrm{decomp}}$ in the training data, and the energy deviation lies on the prediction range of our CGCNN models. The lowest $\Delta H^{\mathrm{CGCNN}}_{\mathrm{decomp}} - T\Delta S_{\mathrm{mix}}$ values predicted by the CGCNN model (x = 0.75 for Fig. 5a and x = 0.5625, y = 0.25 for Fig. 5b) were validated once again using DFT calculations, as shown in Fig. 5. The error between the CGCNN predictions and the DFT data (80 atoms) was below 10 meV atom^{-1}.

With the most stable atomic configurations for each composition, we applied the following four screening criteria to discover potentially promising candidates for solar-cell application: (1) non-indirect bandgap, (2) $E^{\mathrm{CGCNN}}_{\mathrm{gap}} < 0.5$ eV, (3) τ value that is as low as possible (i.e., $\tau < 4.18$), and (4) $\Delta H^{\mathrm{CGCNN}}_{\mathrm{decomp}} - T\Delta S_{\mathrm{mix}}$ that is as low as possible (i.e., $\Delta H^{\mathrm{CGCNN}}_{\mathrm{decomp}} - T\Delta S_{\mathrm{mix}} < 0$). However, increasing the fraction of Ge tends to decrease $\Delta H_{\mathrm{decomp}}$, as shown in Fig. 2b; therefore, we selected candidates with the top three low values of $\Delta H^{\mathrm{CGCNN}}_{\mathrm{decomp}} - T\Delta S_{\mathrm{mix}}$ within each region divided into five intervals for τ lower than 4.18. We additionally selected candidates containing at least 50% Sn or Pb, because Sn- and Pb-containing MHPs are the most promising materials for solar-cell applications[41]. Throughout the screening procedure, 110 compounds were selected.

Subsequently, the 110 selected compounds were validated using DFT calculations with the PBEsol functional. We determined that the error between $\Delta H^{\mathrm{CGCNN}}_{\mathrm{decomp}} - T\Delta S_{\mathrm{mix}}$ and $\Delta H^{\mathrm{PBEsol}}_{\mathrm{decomp}} - T\Delta S_{\mathrm{mix}}$ did not exceed 6 meV atom^{-1} (view Supplementary Fig. 8a for the error distribution). However, in the case of the bandgap and band type, the predictions from the CGCNN were significantly inaccurate. That is, the CGCNN predicted that all 110 compounds would have a non-indirect bandgap; however, 79 compounds exhibited an indirect bandgap. Moreover, the MAE for bandgap predictions was 0.038 for systems with 20 atoms, whereas it increased to 0.14 for systems with 80 atoms (Supplementary Fig. 8b). We assume that these results were obtained as the band structure-related properties do not change continuously with compositional variations when compared to the thermodynamic stability, causing high prediction uncertainty for unseen compositional space. The detailed analysis of increased bandgap prediction error can be found in the Supplementary Note 4, Supplementary Table 1, Supplementary Fig. 2, and Supplementary Fig. 3. One practically applicable strategy to enhance the prediction reliability could be adding training data randomly selected from the target domain after the DFT calculations as similarly done in the active learning strategy. A detailed description of the

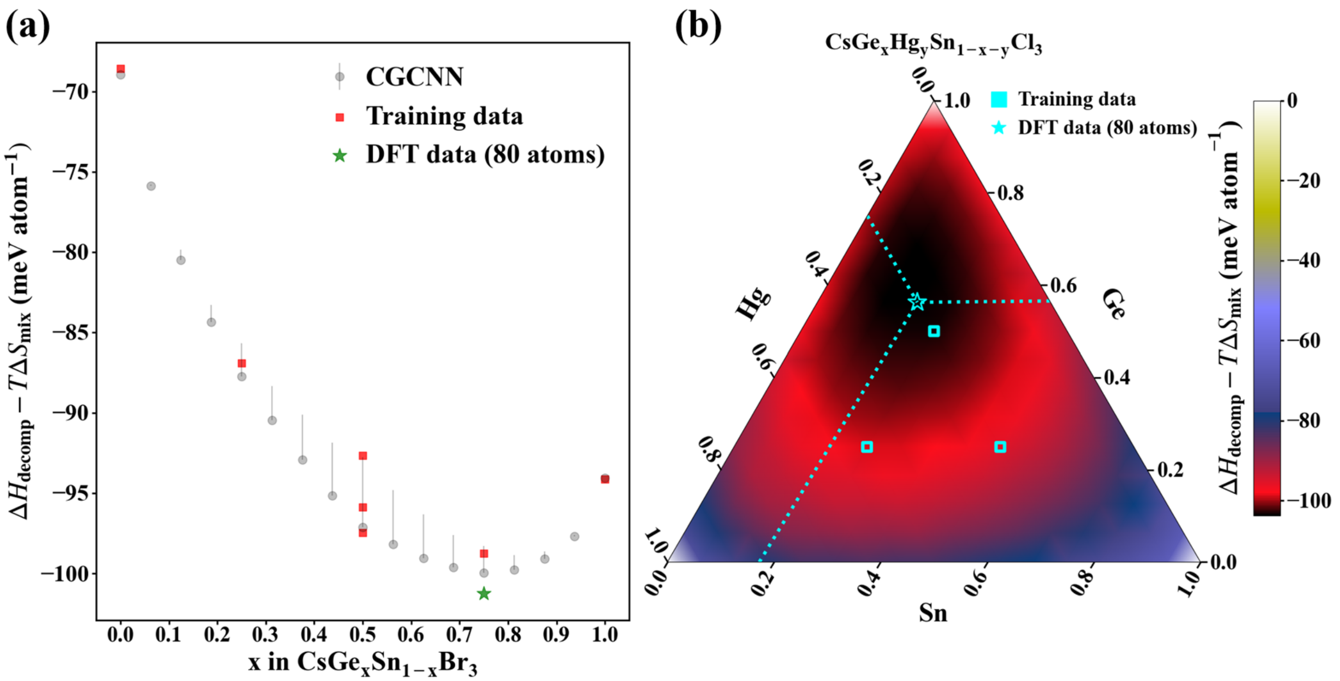


Fig. 5 | ML-enabled search of alloyed perovskite systems. **a** CGCNN-predicted $\Delta H_{\mathrm{decomp}} - T\Delta S_{\mathrm{mix}}$ of **a** $CsGe_xSn_{1-x}Br_3$ and **b** $CsGe_xHg_ySn_{1-x-y}Cl_3$ systems in comparison with the training and DFT data (80 atoms). The variance of CGCNN data is represented as a line at each composition. In **b**, only the lowest $\Delta H_{\mathrm{decomp}} - T\Delta S_{\mathrm{mix}}$ value is plotted at each composition for clearance. In **b**, the validation and CGCNN data at the same composition (Ge: Hg: Sn = 0.5625: 0.25: 0.1875) are −107.60 and −103.79 meV atom^{-1}, respectively.

mitigation strategy can also be found in the Supplementary Note 5 and Supplementary Fig. 4.

We calculated the PBE0 bandgap for only those compounds that were calculated to have a direct bandgap using PBEsol, yielding 31 candidates (refer to Supplementary Table 5 for the full list of the selected 31 candidates) with $E_{\mathrm{gap}}^{\mathrm{PBE0}}$ close to the Shockley–Queisser limit[38] (1.2–1.4 eV) or the optimal bandgap of the top cell of a tandem solar cell[17] (1.73 eV). From among these 31 compounds, 10 compounds showing an optimal bandgap were selected for each system, as summarized in Table 1. All selected compounds have direct bandgaps, as shown in Supplementary Fig. 9.

Previous experimental results have been reported for compositions similar to those listed in Table 1. We compared experimental and PBE0-calculated bandgaps of those compounds in Table 2. The difference between experimentally reported band gaps and DFT-calculated band gaps for 19 reported perovskite compounds is summarized in Supplementary Table 4. Previous experimental studies have reported $CsGe_xSn_{(1-x)}Br_3$[42], $CsGe_{0.5}Sn_{0.5}I_3$[43], $CsPb_xSn_{(1-x)}Br_3$[44], and $CsGe_xPb_{(1-x)}I_3$[45] as light absorbers for perovskite solar cells. Combining Table 2 and Supplementary Table 4, the root mean squared error (RMSE) between the experimentally reported bandgaps and our PBE0-calculated bandgaps is approximately 0.30 eV (see also Supplementary Fig. 5). Specifically, in the $CsGe_xSn_{(1-x)}Br_3$ system, the PBE0 band gap underestimates the experimental band gap by approximately 0.5 ~ 0.6 eV, while for other systems, it underestimates approximately 0.1 eV.

The effective mass of the carriers listed in Table 1 does not exceed 1 m_e (mass of free electrons), except for $CsGe_{0.5625}Hg_{0.3125}Sn_{0.125}Cl_3$ and $CsCd_{0.125}Ge_{0.5625}Hg_{0.1875}Sn_{0.125}Cl_3$. The carrier effective mass is known to directly affect the mobility of electrons in the electron transport layer, which in turn affects the power conversion efficiency of solar cells. The optical absorption spectra and SLME versus the sample thickness are shown in Supplementary Fig. 10. Note that the SLME in Table 1 is a convenient metric for comparing the suitability of the compounds for photon absorbers; however, it does not represent a realistic power conversion efficiency. Considering that our $E_{\mathrm{gap}}^{\mathrm{PBE0}}$ value underestimates $E_{\mathrm{gap}}^{\mathrm{exp}}$ by approximately 0.5–0.6 eV in the $CsGe_xSn_{1-x}Br_3$ system, the SLME of $CsGe_{0.5625}Sn_{0.4375}Br_3$ and $CsGe_{0.4375}Pb_{0.0625}Sn_{0.5}Br_3$ will be lower than that of $CsCd_{0.125}Ge_{0.5625}Hg_{0.1875}Sn_{0.125}Cl_3$. Thus, considering both the stability and SLME, we suggest that $CsGe_{0.3125}Sn_{0.6875}I_3$ and $CsGe_{0.0625}Pb_{0.3125}Sn_{0.625}Br_3$ are promising single-junction and tandem perovskite solar cells, respectively.

Discussion

Although the DFT/ML design framework presented here shows several promising aspects, substantial challenges still exist, which must be overcome for the experimental achievement of this framework. A discrepancy exists between the $E_{\mathrm{gap}}^{\mathrm{PBE0}}$ and $E_{\mathrm{gap}}^{\mathrm{exp}}$ values. Owing to the significant discrepancies between CGCNN predictions and DFT results regarding the band type, certain compounds that were predicted as indirect-bandgap materials by the CGCNN may be direct bandgap materials in reality. The mitigation strategy for bandgap prediction is described in the Supplementary Note 5. Furthermore, our metric of thermodynamic stability, $\Delta H_{\mathrm{decomp}} - T\Delta S_{\mathrm{mix}}$, ignores the fact that Ge and Sn readily oxidize to a + 4 state during the fabrication procedure, resulting in a significant loss of open-circuit voltage[46,47]. Further extension should be made to consider other important properties such as defect formation energy, electronic levels of such defects, charge carrier transport properties, and stability of surfaces and interfaces of perovskites for photovoltaics and other optoelectronic applications. Finally, A- and X-site alloying should also be considered in future studies.

This study explored the chemical space (compositional and configurational) of B-site-mixed all-inorganic perovskites by combining DFT and ML. A total of 3,159 PBEsol data points were generated for the thermodynamic stability, bandgap, and bandgap type. The CGCNN models were trained on the PBEsol data using the crystalline structures as inputs to predict the aforementioned properties. The trained CGCNN models screened 110 of the 41,400 compounds with desirable properties. Of the 110 compounds, 31 were selected after examining their band types and PBE0-calculated bandgaps. Furthermore, the carrier effective masses, optical absorption spectra, and SLME of 10 of the 31 compounds were calculated. Among the 10 compounds, $CsGe_{0.3125}Sn_{0.6875}I_3$ and $CsGe_{0.0625}Pb_{0.3125}Sn_{0.625}Br_3$ were suggested as promising single-junction and tandem perovskite solar cells, respectively, considering their stability and SLME. We have also presented the limitations of this study and perspectives for future research. The DFT/ML design framework proposed herein forms the basis for designing perovskite alloys with desirable properties.

Methods

Training data generation

Training data were generated through DFT calculations using PBEsol[30]. As shown in Fig. 1a, we employed a crystal structure consisting of four formula units of ABX_3 as a computational unit cell (i.e., the number of atoms in the unit cell = 20), with a focus on multi-element mixing at the B-site. For the A- and X-sites in ABX_3, we used one of the three monovalent alkali cations (Cs, K, and Rb) and one of the three halogen anions (Br, Cl, and I), respectively. We excluded organic cations such as methylammonium (MA) and formamidinium (FA) from the A-site for two reasons: (1) instability issues of perovskite devices that primarily arise from organic components[6] and (2) higher positional degrees of freedom than alkali elements, which cause additional energy contributions from various molecular configurations. For the B-site, we considered multi-element mixing of six metal cations (Cd, Ge, Hg, Pb, Sn, and Zn) up to the quaternary system because four available atomic positions of the B-site exist within the considered computational unit cell, giving 126 (= ${}_6H_4$, combination with repetition) compositions. From this, 1134 (= 3 × 126 × 3) unique ABX_3 compositions were obtained, and by considering all possible atomic configurations within each composition, 3159 unique B-site-mixed ABX_3 structures were obtained for the training data. The details of obtaining 3159 unique structures are described in Supplementary Note 1 and Supplementary Fig. 1. The aforementioned six metal elements for the B-site were chosen to minimize any significant perturbation of the electronic structure of the pristine Pb-based halide perovskite, which is good for applications in solar cells. To achieve this, we intentionally chose cations that exhibit electronic similarities to Pb^{2+} ions. This selection was based on two key conditions: 1) maintaining the same charged state (2+) and 2) ensuring filled d-electron orbitals. The idea behind these conditions originated from the understanding that considering a cation with a different charge state can result in charge imbalances, potentially leading to the formation of metallic materials. Additionally, the presence of unfilled d-orbitals tends to introduce complex interactions with anions, leading to a markedly different electronic structure.

Crystal Graph Convolutional Neural Networks training

For all CGCNN models, an unrelaxed input geometry was used as the input to predict the target properties of the DFT-relaxed structures, which allowed us to predict the target properties of the relaxed structures without the costly DFT structural relaxations. Here, the experimentally well-known structure of $CsPbI_3$ was used as the input for the CGCNN models; moreover, we present the crystal structure of $CsPbI_3$ in the CIF format in Supplementary Note 6. The atomic features of the original CGCNN paper[29] and default hyperparameters of the CGCNN code distributed in GitHub (https://github.com/txie-93/cgcnn) were used. The hyperparameters used in this study are listed in Supplementary Table 3. A total of 3159 PBEsol data points were randomly divided in a ratio of 7:1:2 as training, validation, and test data. The validation data were used to determine the model parameters using an early-stopping approach.

Table 1 | Selected 10 perovskite compounds

Formula	$\Delta H_{decomp} - T\Delta S_{mix}$ (meV atom^{-1})		τ	E_{gap}(eV)		E_{gap}^{PBE0} (eV)	$m_e^*(m_e)$		$m_h^*(m_e)$		SLME (%)@ 1μm
	CGCNN	PBESol		CGCNN	PBESol		min	max	min	max	
$CsGe_{0.5625}Sn_{0.4375}Br_3$	−98.18	−98.58	4.00	0.07	0.11	1.39	0.06	0.11	0.04	0.05	36.26
$CsGe_{0.3125}Sn_{0.6875}I_3$	−46.58	−47.46	4.17	0.17	0.20	1.34	0.07	0.10	0.05	0.06	36.40
$CsPb_{0.3125}Sn_{0.6875}Br_3$	−78.19	−78.36	4.05	0.41	0.41	1.73	0.13	0.16	0.01	0.11	28.53
$CsGe_{0.375}Pb_{0.625}I_3$	−29.73	−30.94	4.17	0.49	0.53	1.77	0.11	0.14	0.11	0.14	29.33
$CsGe_{0.0625}Pb_{0.3125}Sn_{0.625}Br_3$	−84.17	−83.95	4.01	0.38	0.39	1.71	0.12	0.16	0.09	0.10	28.96
$CsGe_{0.4375}Pb_{0.0625}Sn_{0.5}Br_3$	−98.20	−98.23	3.95	0.10	0.18	1.48	0.08	0.10	0.05	0.06	34.58
$CsGe_{0.25}Pb_{0.5625}Sn_{0.1875}I_3$	−42.97	−44.17	4.17	0.46	0.50	1.71	0.12	0.15	0.11	0.14	30.20
$CsGe_{0.5625}Hg_{0.3125}Sn_{0.125}Cl_3$	−103.41	−108.21	3.88	0.12	0.33	1.75	0.27	1.86	0.26	0.59	21.18
$CsCd_{0.0625}Ge_{0.25}Pb_{0.5}Sn_{0.1875}I_3$	−46.57	−44.17	4.17	0.48	0.55	1.75	0.12	0.20	0.12	0.18	28.82
$CsCd_{0.125}Ge_{0.5625}Hg_{0.1875}Sn_{0.125}Cl_3$	−106.60	−110.31	3.89	0.31	0.44	1.87	0.36	1.91	0.22	0.46	20.83

m_e^* and m_h^* represent the effective mass of electrons and holes, respectively. The minimum and maximum effective masses obtained from calculations in different directions in reciprocal space are presented. Spectroscopic limited maximum efficiency (SLME) is calculated assuming 1 μm thick photon absorber.

CGCNN-predicted and PBESol-calculated thermodynamic stability term ($\Delta H_{decomp} - T\Delta S_{mix}$) and bandgaps ($E_{gap}$) are presented. Band edges of all selected compounds lie on the gamma point

Table 2 | Comparison between PBE0-calculated and experimental bandgap (E_{gap}^{PBE0} and E_{gap}^{exp}, respectively) of compounds in Table 1

Formula	E_{gap}^{PBE0}(eV)	Formula	E_{gap}^{exp}(eV)
$CsGe_{0.5625}Sn_{0.4375}Br_3$	1.39	$CsGe_{0.5}Sn_{0.5}Br_3$	2.0[42]
$CsGe_{0.3125}Sn_{0.6875}I_3$	1.34	$CsGe_{0.5}Sn_{0.5}I_3$	1.50[43]
$CsPb_{0.3125}Sn_{0.6875}Br_3$	1.73	$CsPb_{0.3}Sn_{0.7}Br_3$	1.83[44]
$CsGe_{0.375}Pb_{0.625}I_3$	1.77	$CsGe_{0.1}Pb_{0.9}I_3$	1.698[45]

Thermodynamic stability and bandgap of B-site mixed metal halide perovskites

To estimate the thermodynamic stabilities of the compounds, we used the previously defined decomposition enthalpy, $\Delta H_{\mathrm{decomp}}$[19–22,24]. Decomposition enthalpy is defined as the difference between the energy of $\mathrm{ABX_3}$ and the sum of the energies of the most stable AX and $\mathrm{BX_2}$ phases. Thus, $\Delta H_{\mathrm{decomp}}$ of the B-site mixed MHPs (i.e., $\mathrm{AB^*X_3}$, where $\mathrm{B^*}$ represents mixed B-site species such as $\mathrm{Cd_{0.25}Ge_{0.75}}$ and $\mathrm{Hg_{0.25}Pb_{0.25}Sn_{0.5}}$) can be defined as follows:

$$\Delta H_{\mathrm{decomp}}\left(\mathrm{AB^*X_3}\right) = E\left(\mathrm{AB^*X_3}\right) - E(\mathrm{AX}) - \sum_i x_i E\left(\mathrm{B}_i\mathrm{X_2}\right), \tag{1}$$

where E is the DFT-calculated total energy of the relaxed geometries and x_i is the fraction of the element B_i. From Eq. (1), we can observe that a more negative $\Delta H_{\mathrm{decomp}}$ indicates that the system is more stable when compared to the other binary metal halide phases (i.e., AX and $\mathrm{BX_2}$). The energetically most stable structures for the AX and $\mathrm{BX_2}$ phases were obtained from Materials Project (MP), an open-access database that offers material properties[48]. Conversely, the structure of $\mathrm{GeCl_2}$ was retrieved from the Open Quantum Materials Database (OQMD)[49] because the structure of $\mathrm{GeCl_2}$ is not available in the MP. We present a list of MP or OQMD IDs of AX and $\mathrm{BX_2}$ phases and their DFT (PBESol)-calculated total energies in Supplementary Table 2.

The mixing entropy, ΔS_{mix}, is defined as the configurational entropy based on the ideal solid solution model (Eq. (2)):

$$\Delta S_{\mathrm{mix}} = -k_B \sum_{i=1}^{N} x_i \, \mathrm{ln x_i}, \tag{2}$$

where k_B is the Boltzmann constant, N is the number of components at the B-site, and x_i is the atomic fraction of component i at the B-site.

In addition to the thermodynamic stability, the band type and bandgap (E_{gap}) were calculated to estimate the possibility of application in photovoltaics at the PBEsol level. Because MHPs with indirect bandgaps are not usually suitable for photovoltaics, we classified the band types into two categories: indirect and non-indirect (i.e., direct, metallic, and semi-metallic). Materials determined to have a metallic or semi-metallic bandgap using PBEsol can be identified as having a direct bandgap using PBE0; thus, we did not exclude them from the PBE0 calculations. This approach assumes that the band types calculated using PBEsol and PBE0 are consistent with each other, although the bandgap calculated using PBEsol ($E_{\mathrm{gap}}^{\mathrm{PBEsol}}$) is underestimated when compared to that calculated using PBE0 ($E_{\mathrm{gap}}^{\mathrm{PBE0}}$).

Tolerance factor of B-site mixed metal halide perovskites

According to Bartel et al.[35], the probability of a compound being a perovskite increases as τ becomes smaller than 4.18, and vice versa. τ is calculated using Eq. (3):

$$\tau = \frac{r_X}{r_B} - n_A\left(n_A - \frac{\frac{r_A}{r_B}}{\ln\left(\frac{r_A}{r_B}\right)}\right), \tag{3}$$

where n_A is the oxidation state of A and r_i is the ionic radius of ion i. Here, r_B is the composition-weighted average of the ionic radius of each B element (i.e., $r_{\mathrm{B}} = \sum_{i=1}^{N} x_i r_{B_i}$). We selected the compositions showing $\Delta H_{\mathrm{decomp}} - T\Delta S_{\mathrm{mix}} < 0$ and $\tau < 4.18$ as stable compounds.

Density functional theory calculations

All DFT calculations were performed using the Vienna Ab initio Simulation Package (VASP) 6.1.2.[50], and projector augmented wave pseudopotentials[51]. The pseudopotentials of all elements were chosen according to the official recommendations of the VASP. For training data generation and final DFT validation of the selected compounds, the projection operators of the nonlocal part of the pseudopotential were calculated in reciprocal and real spaces, respectively. The energy cut-off for a plane-wave representation was 500 eV. All the cell parameters and atomic coordinates are relaxed using a convergence criterion of 10^{-5} eV on the energies of the self-consistent electronic step and 0.01 eV/Å on the atomic force. The Brillouin zone was sampled using a Γ-centered automatic k-points generation scheme that was implemented in VASP, with a length of 50 Å for determining the subdivisions. Cell relaxation was performed first; then, the energy was obtained from the consequent fixed-cell relaxation.

The band structure was calculated for the training data and for the selected compounds through screening. For training data generation, the band structure was calculated using PBEsol at high-symmetry k-points in the line mode. For the final DFT validation of the selected compounds, the band structures were calculated using PBE0 with Γ-centered $2 \times 2 \times 2$ Monkhorst–Pack k-points. According to PBEsol, the band edges of all the selected compounds were predicted to lie on the high-symmetry k-points contained in the Γ-centered $2 \times 2 \times 2$ Monkhorst–Pack k-point. Therefore, the Γ-centered $2 \times 2 \times 2$ Monkhorst–Pack k-points were used in the PBE0 calculation to calculate the bandgap and reduce the computational time. Both band structure calculations used the PBEsol-optimized structure as the input and SOC corrections[32]. We compared the bandgaps calculated using PBE0, HSE06[32], and B3LYP[33] functionals for 19 representative perovskite materials with known experimental bandgaps, as shown in Supplementary Table 4 and Supplementary Fig. 5. After comparison, we found that PBE0 best aligns with the experimental bandgaps. The VASPKIT code[52] was used to determine the PBEsol band type and bandgap from the data calculated using VASP.

The effective masses of electrons and holes were calculated based on the PBEsol-calculated band structures using the sumo[53] software, which is a set of command-line tools for plotting and analysis of periodic ab initio calculations. The optical absorption spectra were calculated using the LOPTICS tag[54], setting the number of grid points at which the density of states (NEDOS tag) was set to 2000 for each structure. We obtained optical absorption spectra on Γ-centered $4 \times 4 \times 4$ Monkhorst–Pack k-points with the k-point down-sampling scheme using the PBE0 functional with SOC corrections. Sumo[53] was also used to post-process the optical absorption spectra.

Data availability

The datasets generated and/or analyzed during the current study are available at https://github.com/KRICT-DATA/Perov_CGCNN. Please consult the corresponding author with further questions or requests.

Code availability

The underlying code for this study is available at https://github.com/KRICT-DATA/Perov_CGCNN. Please consult the corresponding author with further questions or requests.

Received: 31 August 2023; Accepted: 11 April 2024;
Published online: 08 May 2024

References

1. Dey, A. et al. State of the art and prospects for halide perovskite nanocrystals. *ACS Nano* **15**, 10775–10981 (2021).
2. Basumatary, P. & Agarwal, P. A short review on progress in perovskite solar cells. *Mater. Res. Bull.* **149**, 111700 (2022).
3. Dong, H. et al. Metal Halide Perovskite for next-generation optoelectronics: progresses and prospects. *eLight* **3**, 3 (2023).
4. Hao, J. & Xiao, X. Recent development of optoelectronic application based on metal halide perovskite nanocrystals. *Front. Chem.* **9**, 822106 (2022).
5. Liao, C.-H., Mahmud, M. A. & Ho-Baillie, A. W. Y. Recent progress in layered metal halide perovskites for solar cells, photodetectors, and field-effect transistors. *Nanoscale* **15**, 4219–4235 (2023).
6. Rao, M. K., Sangeetha, D. N., Selvakumar, M., Sudhakar, Y. N. & Mahesha, M. G. Review on persistent challenges of perovskite solar cells' stability. *Sol. Energy* **218**, 469–491 (2021).

7. Ono, L. K., Juarez-Perez, E. J. & Qi, Y. Progress on perovskite materials and solar cells with mixed cations and halide anions. *ACS Appl. Mater. Interfaces* **9**, 30197–30246 (2017).
8. Xu, F., Zhang, T., Li, G. & Zhao, Y. Mixed cation hybrid lead halide perovskites with enhanced performance and stability. *J. Mater. Chem. A* **5**, 11450–11461 (2017).
9. Wang, Y. et al. High-entropy perovskites for energy conversion and storage: design, synthesis, and potential applications. *Small Methods* **2201138**, 1–27 (2023).
10. George, E. P., Raabe, D. & Ritchie, R. O. High-entropy alloys. *Nat. Rev. Mater.* **4**, 515–534 (2019).
11. Yi, C. et al. Entropic stabilization of mixed A-cation ABX $_3$ metal halide perovskites for high performance perovskite solar cells. *Energy Environ. Sci.* **9**, 656–662 (2016).
12. Li, Z. et al. Stabilizing perovskite structures by tuning tolerance factor: formation of formamidinium and cesium lead iodide solid-state alloys. *Chem. Mater.* **28**, 284–292 (2016).
13. Kumar, S. et al. Mixing entropy-induced layering polydispersity enabling efficient and stable perovskite nanocrystal light-emitting diodes. *ACS Energy Lett.* **4**, 118–125 (2019).
14. Solari, S. F. et al. Stabilization of lead-reduced metal halide perovskite nanocrystals by high-entropy alloying. *J. Am. Chem. Soc.* **144**, 5864–5870 (2022).
15. Sun, S. et al. Accelerated development of perovskite-inspired materials via high-throughput synthesis and machine-learning diagnosis. *Joule* **3**, 1437–1451 (2019).
16. Chen, S. et al. Exploring the Stability of Novel Wide Bandgap Perovskites by a Robot Based High Throughput Approach. *Adv. Energy Mater.* **8**, 1701543 (2018).
17. Jesper Jacobsson, T. et al. Exploration of the compositional space for mixed lead halogen perovskites for high efficiency solar cells. *Energy Environ. Sci.* **9**, 1706–1724 (2016).
18. Yang, J., Manganaris, P. T. & Mannodi Kanakkithodi, A. K. A high-throughput computational dataset of halide perovskite alloys. *Digit. Discov.* **2**, 856–870 (2023).
19. Wang, X. et al. Entropy-driven stabilization of multielement halide double-perovskite alloys. *J. Phys. Chem. Lett.* **13**, 5017–5024 (2022).
20. Pu, W., Xiao, W., Wang, J., Li, X. & Wang, L. Screening of perovskite materials for solar cell applications by first-principles calculations. *Mater. Des.* **198**, 109387 (2021).
21. Liu, D., Li, Q., Jing, H. & Wu, K. First-principles modeling of lead-free perovskites for photovoltaic applications. *J. Phys. Chem. C.* **123**, 3795–3800 (2019).
22. Yamamoto, K., Iikubo, S., Yamasaki, J., Ogomi, Y. & Hayase, S. Structural stability of iodide perovskite: a combined cluster expansion method and first-principles study. *J. Phys. Chem. C.* **121**, 27797–27804 (2017).
23. Choubisa, H. et al. Crystal site feature embedding enables exploration of large chemical spaces. *Matter* **3**, 433–448 (2020).
24. Mannodi-Kanakkithodi, A. & Chan, M. K. Y. Data-driven design of novel halide perovskite alloys. *Energy Environ. Sci.* **15**, 1930–1949 (2022).
25. Zunger, A., Wei, S.-H., Ferreira, L. G. & Bernard, J. E. Special quasirandom structures. *Phys. Rev. Lett.* **65**, 353–356 (1990).
26. van de Walle, A. et al. Efficient stochastic generation of special quasirandom structures. *Calphad* **42**, 13–18 (2013).
27. Yang, Y.-C., Liu, C., Lin, C.-Y. & Xia, Z. Core effect of local atomic configuration and design principles in AlxCoCrFeNi high-entropy alloys. *Scr. Mater.* **178**, 181–186 (2020).
28. Wu, D. et al. Noble-metal high-entropy-alloy nanoparticles: atomic-level insight into the electronic structure. *J. Am. Chem. Soc.* **144**, 3365–3369 (2022).
29. Xie, T. & Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* **120**, 145301 (2018).
30. Perdew, J. P. et al. Restoring the density-gradient expansion for exchange in solids and surfaces. *Phys. Rev. Lett.* **100**, 136406 (2008).
31. Adamo, C. & Barone, V. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *J. Chem. Phys.* **110**, 6158–6170 (1999).
32. Steiner, S., Khmelevskyi, S., Marsmann, M. & Kresse, G. Calculation of the magnetic anisotropy with projected-augmented-wave methodology and the case study of disordered $Fe_{1-x}Co_x$ alloys. *Phys. Rev. B* **93**, 224425 (2016).
33. Krukau, A. V., Vydrov, O. A., Izmaylov, A. F. & Scuseria, G. E. Influence of the exchange screening parameter on the performance of screened hybrid functionals. *J. Chem. Phys.* **125**, 224106 (2006).
34. Stephens, P. J., Devlin, F. J., Chabalowski, C. F. & Frisch, M. J. Ab initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields. *J. Phys. Chem.* **98**, 11623–11627 (1994).
35. Bartel, C. J. et al. New tolerance factor to predict the stability of perovskite oxides and halides. *Sci. Adv.* **5**, eaav0693 (2019).
36. Goldschmidt, V. M. Die Gesetze der Krystallochemie. *Naturwissenschaften* **14**, 477–485 (1926).
37. Filip, M. R. & Giustino, F. The geometric blueprint of perovskites. *Proc. Natl Acad. Sci.* **115**, 5397–5402 (2018).
38. Shockley, W. & Queisser, H. J. Detailed balance limit of efficiency of p-n junction solar cells. *J. Appl. Phys.* **32**, 510–519 (1961).
39. Yu, L. & Zunger, A. Identification of potential photovoltaic absorbers based on first-principles spectroscopic screening of materials. *Phys. Rev. Lett.* **108**, 068701 (2012).
40. Thiele, G., Rotter, H. W. & Schmidt, K. D. Kristallstrukturen und Phasentransformationen von Caesiumtrihalogenogermanaten(II) CsGeX3 (X = Cl, Br, I). *Z. Anorg. Allg. Chem.* **545**, 148–156 (1987).
41. Li, Y. et al. All-inorganic perovskite solar cells featuring mixed group IVA cations. *Nanoscale* **15**, 7249–7260 (2023).
42. Kama, A., Tirosh, S., Itzhak, A., Ejgenberg, M. & Cahen, D. New Pb-free stable Sn–Ge solid solution halide perovskites fabricated by spray deposition. *ACS Appl. Energy Mater.* **5**, 3638–3646 (2022).
43. Chen, M. et al. Highly stable and efficient all-inorganic lead-free perovskite solar cells with native-oxide passivation. *Nat. Commun.* **10**, 16 (2019).
44. Schwartz, H. A. et al. Band-gap tuning in all-inorganic $CsPb_xSn_{1-x}$ Br_3 perovskites. *ACS Appl. Mater. Interfaces* **13**, 4203–4210 (2021).
45. Meng, F. et al. Ge incorporation to stabilize efficient inorganic $CsPbI_3$ perovskite solar cells. *Adv. Energy Mater.* **12**, 2103690 (2022).
46. Krishnamoorthy, T. et al. Lead-free germanium iodide perovskite materials for photovoltaic applications. *J. Mater. Chem. A* **3**, 23829–23832 (2015).
47. Hao, F., Stoumpos, C. C., Cao, D. H., Chang, R. P. H. & Kanatzidis, M. G. Lead-free solid-state organic-inorganic halide perovskite solar cells. *Nat. Photonics* **8**, 489–494 (2014).
48. Jain, A. et al. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
49. 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**, 1501–1509 (2013).
50. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169–11186 (1996).
51. Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B* **50**, 17953–17979 (1994).

52. Wang, V., Xu, N., Liu, J.-C., Tang, G. & Geng, W.-T. VASPKIT: A user-friendly interface facilitating high-throughput computing and analysis using VASP code. *Comput. Phys. Commun.* **267**, 108033 (2021).
53. Ganose, A. M., Jackson, A. J. & Scanlon, D. O. sumo: Command-line tools for plotting and analysis of periodic ab initio calculations. *J. Open Source Softw.* **3**, 717 (2018).
54. Gajdoš, M., Hummer, K., Kresse, G., Furthmüller, J. & Bechstedt, F. Linear optical properties in the projector-augmented wave methodology. *Phys. Rev. B.* **73**, 045112 (2006).

Acknowledgements

This work is supported by the National Research Foundation (NRF) grant funded by the Korean government (MSIT) (RS-2023-00283597) and the National Supercomputing Center with supercomputing resources and technical support (KSC-2019-CRE-0128).

Author contributions

J-S.K. wrote the codes to automate the generation of training data, trained the CGCNN models, implemented chemical exploration, conducted the DFT calculations, surveyed literature, and wrote the paper. J.N. wrote the codes to conduct chemical space exploration using CGCNN, analyzed data, and edited the paper. J.I. provided the main idea of this study, oversaw all work, and edited the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41524-024-01270-1.

Correspondence and requests for materials should be addressed to Jin-Soo Kim, Juhwan Noh or Jino Im.

Reprints and permissions information is available at http://www.nature.com/reprints

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