Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Digital Discovery*, 2023, **2**, 856

A high-throughput computational dataset of halide perovskite alloys†

Jiaqi Yang, Panayotis Manganaris and Arun Mannodi-Kanakkithodi *

Novel halide perovskites with improved stability and optoelectronic properties can be designed *via* composition engineering at cation and/or anion sites. Data-driven methods, especially involving high-throughput first principles computations and subsequent analysis based on unique materials descriptors, are key to achieving this goal. In this work, we report a density functional theory (DFT) dataset of 495 ABX_3 halide perovskite compounds, with monovalent organic or inorganic cations as A, divalent Group 2 or Group 14 elements as B, and I, Br, or Cl as X, and different amounts of mixing applied at each site using the special quasirandom structures (SQS) approach. We perform GGA-PBE calculations on all 495 pseudo-cubic perovskite structures and between 250 and 300 calculations each using the more expensive HSE06 functional, with and without spin–orbit coupling, both including full geometry optimization and static calculations on PBE optimized structures. Lattice parameters, decomposition energy, band gap, and theoretical photovoltaic efficiency derived from computed optical absorption spectra, are determined from each level of theory, and some comparisons are made with collected experimental values. Trends in the data are unraveled in terms of the effects of mixing at different sites, fractions of specific elemental or molecular species present in the compound, and averaged physical properties of species at different sites. We perform screening across the perovskite dataset based on multiple known definitions of stability factors, deviation from cubicity in the optimization cell, and computed stability and optoelectronic properties, leading to a list of promising compositions as well as design principles for achieving multiple desired properties. Our multi-objective, multi-fidelity, computational halide perovskite alloy dataset, one of the most comprehensive to date, is available open-source, and currently being used to train predictive and optimization models for accelerating the design of novel compositions for superior performance across many optoelectronic applications.

Received 9th February 2023
Accepted 4th May 2023

DOI: 10.1039/d3dd00015j

rsc.li/digitaldiscovery

1 Introduction

Perovskites have historically been materials of immense interest for a variety of industrial applications. With a general formula of ABX_3, a perovskite cubic unit cell contains two cations A and B at the corners and body center, and an anion X at each of the face centers. The symbolic 3D perovskite structure is a network of BX_6 octahedra robustly held together by large A-site cations. This unique structure means that perovskite properties are incredibly tunable, by changing the size and number of A/B/X species, by manipulating relative octahedral arrangements, and by creating non-cubic and metastable phases. Numerous research efforts have been devoted to halide perovskites (HaPs), especially as photovoltaic (PV) absorbers.[1–4] In ABX_3 HaPs, X-site anions are halogens such as I and Br, B-site cations may be divalent elements such as Pb and Sn, and the A-site is occupied by large monovalent cations that are either inorganic (*e.g.* Cs, K, and Rb) elements or organic molecules (*e.g.* Methylammonium (MA) and Formamidinium (FA)). The most commonly studied halide hybrid organic–inorganic perovskites (HOIPs), $MAPbI_3$ and $FAPbI_3$, have demonstrated large power conversion efficiency (PCE) values between 20% and 25% when used as absorbers in single- or multi-junction solar cells.[5,6] This is a five-fold improvement over the efficiencies first reported in 2009 and shows the most attractive feature of HaPs, their unique tunability. A perovskite structure is considered stable if the ionic radii of A, B, and X-site species satisfy the well-known tolerance (t) and octahedral (μ) factors.[7] Even under these restrictions, the chemical space of HaP structures, alloying ratios, ionic ordering, and possible defects is still combinatorial and poses a highly multidimensional optimization problem.

Three of the most common ways of tuning the properties of HaPs are described below:

(1) Composition: the most promising HaP compositions for PV absorption explored to date usually contain a mix of MA, FA, and Cs at the A-site, primarily Pb at the B-site with minor fractions of other divalent cations such as Sn and Ge, and I or Br at

School of Materials Engineering, Purdue University, West Lafayette 47907, IN, USA. E-mail: amannodi@purdue.edu

† Electronic supplementary information (ESI) available: https://github.com/PanayotisManganaris/manuscript-PIP-data-manifest. See DOI: https://doi.org/10.1039/d3dd00015j

the X-site often with little Cl. The discovery of novel HaP compositions with attractive properties is on the rise as researchers expand the search into more complex alloys, novel A-site organic molecules, and substitutes for Pb at the B-site from Group IV, Group II, or transition elements.[8–11] Mixing at the A site improves the general stability to degradation, while B site and X site mixing can tune and optimize band gaps and optical absorption. The allure of A/B/X-site mixing, even creating high entropy perovskite alloys, is in obtaining starkly different properties compared to pure compositions, possibly eliminating the harmful effect of defects, and improving the long-term stability and consequent optoelectronic performance.

(2) Structure and phases: while the canonical perovskite phase is cubic, many HaPs are most stable in tetragonal, orthorhombic, or hexagonal phases.[12] For a given composition and phase, there may exist many local minima on the potential energy surface, typically sampled *via* rigorous application of evolutionary or minima hopping algorithms, atomic perturbations within larger supercells, or by varying degrees of distortion and rotation in the octahedral networks. Stable or metastable structures thus obtained may show better properties than previously studied ground state structures.[13] In addition, HaPs may also manifest as double perovskites or 2D layered perovskites which include large organic spacer ligands, providing another means of tailoring the stability and optoelectronic properties.

(3) Defects: investigation of the electronic structure of crystalline materials is incomplete without consideration of point defects, either native or impurity, which will affect the optoelectronic properties by modifying charge carrier lifetimes, equilibrium conductivity, and resulting trap-limited efficiencies.[14,15] Point defects manifest as vacancies, interstitials, or substitutions, and the same defect may behave very differently in different compositions or structures, highlighting the need to include the presence of defects as another variable towards tuning HaP properties.

The chemical design space of HaPs is very much combinatorial and raises challenges for experimentalists to perform effective screening. First principles-based density functional theory (DFT) simulations have been systematically performed to study the optoelectronic properties of HaPs as a function of structure, composition, and defects. The expense of standard DFT computations is reasonable when searching for new promising candidates in such a boundless space. Recently, DFT simulations have been reliably used for modeling structure, heat of formation or decomposition, band gaps, optical absorption spectra, and defect formation energies of a variety of HaPs.[2,16] High-throughput DFT (HT-DFT) computations provide the most effective way to screen across a large space of hybrid and inorganic ABX_3 halide perovskites. An examination of the HaP-related computational literature reveals that there have been numerous medium ($\sim 10^2$ data points) to large ($\sim 10^3$ or more data points) DFT datasets reported for HaPs, which have been successfully used to screen promising materials with desired stability and formability as well as PV-suitable band gaps, among other properties.[12,17–19]

A clear limitation of HT-DFT-driven screening is the computational expense of applying a suitably advanced level of theory across a large number of materials. This problem is typically addressed by coupling DFT computations with state-of-the-art machine learning (ML) or artificial intelligence (AI) techniques. Within the area of perovskites, there are many examples in the literature where DFT datasets and suitable atomic/structural/compositional descriptors have been used to train a variety of ML-based predictive and classification models, leading to accelerated prediction of lattice constants, formation energies, band gaps, and other important properties.[18,20,21] Such DFT-ML models, once rigorously trained and tested, are deployed for high-throughput screening across massive datasets of unknown perovskites. We recently published a thorough overview of many such efforts applying DFT and/or ML towards halide perovskite discovery.[22]

In this work, we report a large HT-DFT dataset of 495 chemically distinct, pseudo-cubic, halide perovskite alloys. This dataset builds upon the 229 compounds reported in prior work by Mannodi-Kanakkithodi and Chan,[16] adding more types of mixing, better property estimates, and detailed analysis of trends and correlations. The relatively large size of this dataset is intended to provide an initial sampling suitable for a guided search within the HaP alloy space. In this dataset, all perovskite structures are cubic or pseudo-cubic, and the focus is more on investigating the dependence of computed properties on composition, and specifically the type of alloying.

Based on the generated perovskite structures, we perform GGA-PBE calculations and report the computed decomposition energy, band gap, and theoretical PV efficiency. In addition, around 250 to 300 calculations are performed using the HSE06 functional (henceforth referred to as HSE), in three different versions: using full geometry optimization, with and without spin–orbit coupling (SOC), and static calculations on GGA-optimized structures with SOC. The same properties are computed from all three types of HSE computations, which enables the comparison of PBE and multiple HSE estimates with experiments, as well as an understanding of the importance of SOC for certain compositions. Pearson correlation analysis is performed to study the contribution of specific A/B/X species and their known elemental/molecular properties on the DFT computed properties, leading to some useful design rules. We further combine DFT-computed properties with perovskite stability factors such as the octahedral and tolerance factors and determine a deviation from cubicity for all optimized structures, to obtain a list of promising candidates for solar absorption and related optoelectronic applications. We emphasize that this chemically diverse, multi-objective, multi-fidelity dataset of HaP alloys will serve many ML endeavors in the future for prediction and inverse design, be used as the foundation for extended datasets of non-cubic structures and new chemistries, and drive the experimental discovery of novel HaP compositions with targeted properties.

2 Methodology

2.1 Devising a halide perovskite chemical space

The dataset we report is based on the standard cubic ABX_3 perovskite structure. Fourteen common perovskite constituents

are selected to form the chemical space. The five constituents making up the A-site occupants include three inorganic and two organic cations. Six divalent metals represent the possible B-site occupants and three halogen anions make up the possible X-site occupants. The elemental and molecular space used to construct the data set is shown in Fig. 1(a). In total, these component vectors form a constrained 14 dimensional space within which all perovskite compounds consisting of the species shown in Fig. 1(a) must exist.

The pure (non-alloyed) possibilities are exhaustively sampled using $5 \times 6 \times 3 = 90$ compounds. Starting from these pure perovskite structures, we perform systematic mixing at the A, B, and X sites. For simulating perovskite alloys, the special quasi-random structures (SQS) method[23] is applied to build periodic structures that make the first nearest-neighbor shells as similar to the target random alloy as possible. The SQS can be considered the best possible periodic supercell representing a given mixed-composition perovskite. The distribution of different types of mixing across our dataset is shown in Fig. 1(b). For simplicity, only one type of mixing at a time is considered in this study; that is, mixing is not performed at multiple (A/B/X) sites simultaneously. In total, we performed GGA-PBE computations on 90 pure, 126 A-site mixed, 151 B-site mixed, and 127 X-site mixed HaPs.

Each HaP composition is simulated using a $2 \times 2 \times 2$ supercell, which allows A- and B-site mixing to be performed in discrete 1/8th fractions of the total site occupancy, and X-site mixing in 1/24th fractions, though for simplicity, we restrict X-site mixing to fractions of $3 \times (1/24)$. At these mixing levels, these systems may be referred to as perovskite alloys. Fig. 2 shows the distribution of various types of mixing of the 14 total species at the A, B, and X-sites, across the dataset of 495 compounds. Since mixing is only allowed on one out of the three sites at a time, there is a higher prevalence of the 8/8 fraction for each species. We also find a larger occurrence of the smallest fractions of mixing at A and B sites as compared to intermediate fractions; overall, every type of mixing is represented within the dataset a few times. Using the procedure presented in Fig. 1(c), we calculate the stability and optoelectronic properties for the HaP dataset using both the semilocal GGA-PBE functional and the hybrid HSE06 functionals. Ultimately, we generated a dataset of 495 points at the PBE level, and between 244 and 299 points each at the HSE–PBE + SOC (referring to HSE + SOC on PBE relaxed structures), HSE-

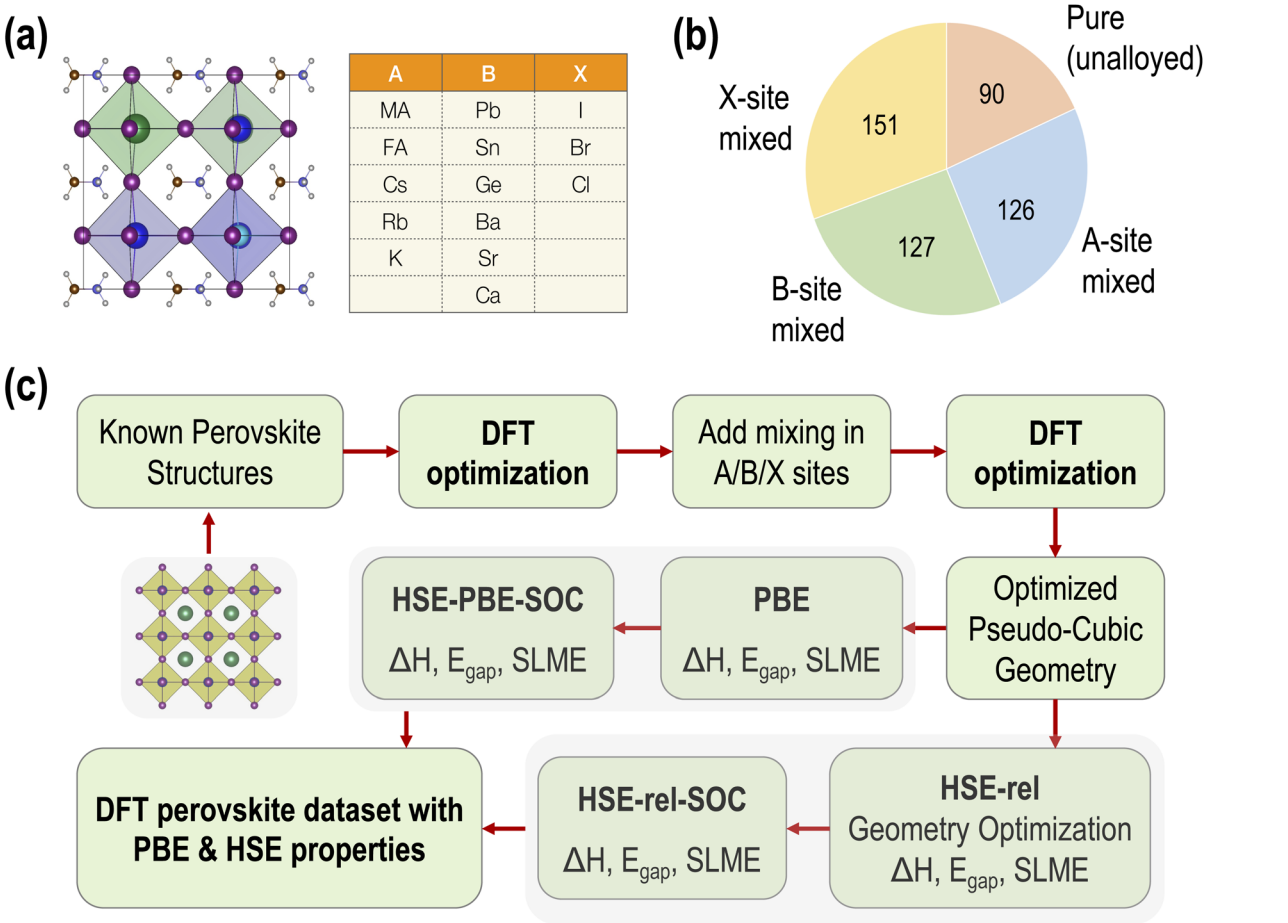

Fig. 1 (a) Chemical space of ABX_3 perovskites studied in this work. (b) Number of samples representing each kind of primary alloy. (c) Steps involved in generating the PBE and HSE datasets of three kinds of properties, namely the decomposition energy (ΔH), band gap (E_{gap}), and spectroscopic limited maximum efficiency (SLME).

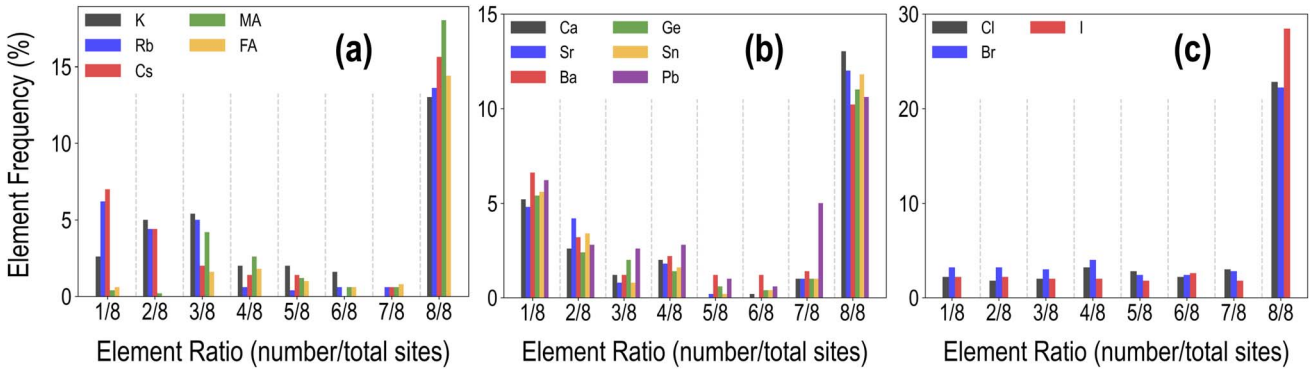

Fig. 2 Distribution of the mixing fractions of various species at the A (a), B (b), and X (c) sites across the PBE dataset of 495 compounds.

relaxed, and HSE-relaxed + SOC levels. The exact constitution of the dataset from different levels of theory is presented in Table 1. The different number of data points from different functionals is a consequence of the number of computations that were completed within the constraints of computing resources and researcher time, but adequate chemical diversity is maintained in each dataset, and as explained later, insights from cheaper functionals can be extended to more expensive theories.

2.2 DFT details

All DFT computations were performed using VASP version 6.2 (ref. 24–26) employing the projector augmented wave (PAW) pseudopotentials.[27,28] The Perdew–Burke–Ernzerhof (PBE) functional within the generalized gradient approximation (GGA)[29] as well as the hybrid HSE06 (ref. 30) ($\alpha = 0.25$ and $\omega = 0.2$) functionals are used for exchange–correlation energy. The energy cutoff for the plane-wave basis is set to 500 eV. For all PBE geometry optimization calculations, the Brillouin zone was sampled using a $6 \times 6 \times 6$ Monkhorst–Pack mesh for unit cells and a $3 \times 3 \times 3$ mesh for supercells. Using the PBE optimized structure as input, the electronic band structure is calculated along high-symmetry k-points[31,32] to obtain accurate band gaps, and the optical absorption spectrum is further calculated using the LOPTICS tag, setting the number of energy bands to 1000 for each structure. For HSE calculations, geometry optimization was performed using only the Gamma point, and the subsequent electronic structure computations used a reduced $2 \times 2 \times 2$ Monkhorst–Pack mesh. The force convergence threshold is set to be -0.05 eV Å^{-1}. Spin–orbit coupling is also applied to two types of HSE computations using the LORBIT tag and the non-collinear magnetic version of VASP 6.2.[33] We obtain optical absorption spectra from different HSE functionals by using the difference between the respective PBE and HSE band gaps, and shifting the PBE-computed spectrum.

Table 1 Number of HaP compounds studied using each of the 4 theories applied in this work

Functional	Number of data points
PBE	495
HSE-rel	299
HSE-rel-SOC	282
HSE–PBE-SOC	244

2.3 DFT computed properties

2.3.1 Decomposition energy. In this work, we estimate the stability of any ABX_3 compound based on the energy of decomposition to all possible AX and BX_2 phases. In addition, we add a mixing entropy term for all alloys, assuming the temperature to be 300 K. The decomposition energy (ΔH) is thus calculated using eqn (1), individually from each level of theory.

$$\Delta H = E_{\text{opt}}(\text{ABX}_3) - \sum_i x_i E_{\text{opt}}(\text{AX}) - \sum_i x_i E_{\text{opt}}(\text{BX}_2) + k_{\text{B}}T\left(\sum_i x_i \ln(x_i)\right) \quad (1)$$

Here, E_{opt}(M) refers to the total DFT energy of any compound M, k_{B} is the Boltzmann constant, T is the temperature (fixed to be 300 K in this work), and x_i is the fraction of any particular species mixed at the A/B/X site. The weighted sums over E_{opt}(AX) and $E_{\text{opt}}(\text{BX}_2)$ signify that an ABX_3 alloy is assumed to decompose to multiple AX and BX_2 phases, based on the number of species mixed at the A, B, or X site. Taking $\text{A(B}_1)_x(\text{B}_2)_{1-x}\text{X}_3$ as an example, the decomposition energy would be calculated using eqn (2). We assume that the "BX_2" decomposition products for B1–B2 mixed perovskite are $(B_1)X_2$ and $(B_2)X_2$.

$$\Delta H\left[\text{A(B}_1)_x(\text{B}_2)_{1-x}\text{X}_3\right] = E_{\text{opt}}(\text{AB}_1\text{B}_2\text{X}_3) - E_{\text{opt}}(\text{AX}) - x \times E_{\text{opt}}(\text{B}_1\text{X}_2) - (1-x) \times E_{\text{opt}}(\text{B}_2\text{X}_2) + k_{\text{B}}T(x\ln(x) + (1-x)\ln(1-x)) \quad (2)$$

The decomposition energy is calculated from 4 different levels of theory, namely PBE (ΔH^{PBE}), HSE-relaxed ($\Delta H^{\text{HSE-rel}}$), HSE-relaxed-SOC ($\Delta H^{\text{HSE-rel-SOC}}$), and HSE–PBE-SOC ($\Delta H^{\text{HSE-PBE-SOC}}$). All decomposition energy values are reported per ABX_3 formula unit. Calculating ΔH for X-site mixed compounds involves some additional work because of the multiple choices

for AX and BX_2 phases; more details are provided in the ESI and in Fig. S1 and S2.†

2.3.2 Band gap. From the PBE band structure calculations and the static HSE calculations using the 2 × 2 × 2 Monkhorst–Pack mesh, four types of electronic band gaps are computed in eV: PBE (E_{gap}^{PBE}), HSE-relaxed ($E_{gap}^{HSE\text{-}rel}$), HSE-relaxed-SOC ($E_{gap}^{HSE\text{-}rel\text{-}SOC}$), and HSE–PBE-SOC ($E_{gap}^{HSE\text{-}PBE\text{-}SOC}$).

2.3.3 Spectroscopic limited maximum efficiency (SLME). Introduced by Yu and Zunger,[34] the SLME is a convenient metric for evaluating a semiconductor's suitability for single junction photovoltaic (PV) absorption. In this work, SLME is calculated considering a 5 μm sample thickness for every perovskite using eqn (3)–(5), combining the original SL3ME.py code from Yu *et al.*[34] with our DFT computed absorption spectra and band gaps.

$$a(E) = 1 - e^{-2\alpha(E)L} \quad (3)$$

Here, $\alpha(E)$ is the DFT computed optical absorption coefficient as a function of incident photon energy and L is the thickness of the absorber.

$$J = e\int_0^\infty a(E)I_{sun}(E)dE - J_0\left(1 - e^{\frac{eV}{kT}}\right) \quad (4)$$

$$\eta = \frac{P_m}{P_{in}} = \frac{\max(J \times V)}{P_{in}} \quad (5)$$

J is the current density, I_{sun} is the light spectrum intensity of sunlight, and P refers to the power used to calculate SLME efficiency. Using the DFT (PBE) computed optical absorption spectrum as well as the magnitude and type (direct or indirect) of the band gap as input, SLME is directly calculated using an open-source package.[35] This calculation is performed using PBE as well as the 3 different HSE functionals based on shifting the band gap, resulting in 4 theoretical estimates of PV efficiency, denoted as $SLME^{PBE}$, $SLME^{HSE\text{-}rel}$, $SLME^{HSE\text{-}rel\text{-}SOC}$, and $SLME^{HSE\text{-}PBE\text{-}SOC}$ (Table 2).

Table 2 RMSE values of band gaps computed from different functionals compared with experimental (Exp) values

Functional	Band gap RMSE *vs.* Exp (eV)
PBE	0.78
HSE-rel	0.93
HSE-rel-SOC	0.74
HSE–PBE-SOC	0.70

3 Results and discussion

3.1 Comparing DFT with experiments

In Fig. 3, we compare the various PBE and HSE calculated lattice constant and band gap values with the corresponding experimental results collected from Tao *et al.*[36] and Almora *et al.*[37] We find that the root mean square error (RMSE) of PBE lattice constants compared to experiments is 0.27 Å, while the corresponding HSE RMSE is 0.31 Å. The percentage error of PBE-relaxed lattice constants compared to experiments is 2.21%, and the corresponding HSE-relaxed percentage error is 3.91%. These results show that hybrid functional-based geometry optimization is unnecessary for obtaining accurate crystal structure information. We note that the accuracy of optimized geometry may be further improved by using the PBEsol functional[38] or by incorporating van der Waals interactions with the PBE functional using DFT-D3 (ref. 39) or a similar approach, especially for hybrid perovskites.

Fig. 3(b) shows that E_{gap}^{PBE} is generally an underestimation compared to experiments, as expected, showing an RMSE of 0.78 eV. The corresponding RMSEs of $E_{gap}^{HSE\text{-}rel}$, $E_{gap}^{HSE\text{-}rel\text{-}SOC}$, and $E_{gap}^{HSE\text{-}PBE\text{-}SOC}$ are respectively 0.93 eV, 0.74 eV, and 0.70 eV. We find that on average, HSE–PBE-SOC is the best approach out of the four for reproducing band gaps, but other functionals may be more accurate for certain types of compositions (such as purely inorganic *vs.* organic–inorganic, Pb-based or Pb-free, *etc.*), as will be discussed further later in this article. HSE-relaxed band gaps are heavily overestimated and brought down by the inclusion of SOC. It should also be noted that phase information was not always available for certain

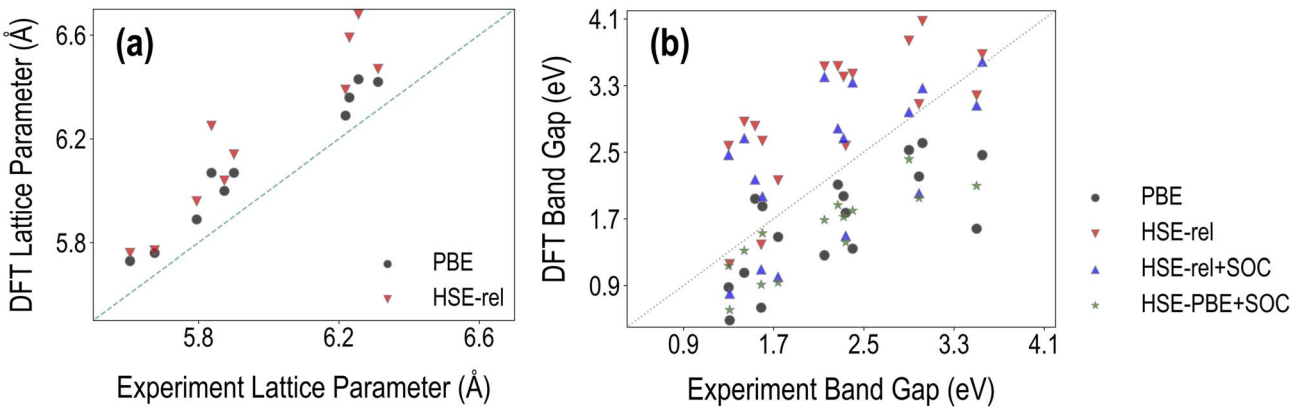

Fig. 3 Comparison between the DFT computed and experimentally measured properties of the selected HaPs: (a) cubic lattice constants and (b) band gaps.

experimental data points collected from the literature, and non-cubic phases may affect the accuracy of the computational results here. We additionally performed a comparison for a much smaller dataset of 9 compounds known to be cubic from experiments; a plot showing these band gaps is presented in Fig. S4,† and the corresponding RMSE values in Table S1† show that $E_{gap}^{HSE\text{-}PBE\text{-}SOC}$ has a respectable RMSE against experiments of 0.4 eV. Finally, it should be noted here that the PBE RMSE is not significantly different from the RMSE of HSE–PBE-SOC, which comes from the accidental accuracy of semi-local functionals without SOC for HOIPs.[16,40]

3.2 Visualizing the PBE dataset

Fig. 4 presents a visualization of the PBE computed properties across the dataset of 495 compounds. The data can be distinguished in terms of purely inorganic *vs.* hybrid organic–inorganic compounds, as well as in terms of the type of mixing. A broad range of values is observed for the three properties, which is a testament to the chemical diversity in our dataset. ΔH^{PBE} varies from ∼−1.5 eV to ∼4 eV, with a majority of the data points in the unstable >0 eV region, while E_{gap}^{PBE} goes from ∼0.5 eV to ∼5.5 eV. $SLME^{PBE}$ goes from a low of 0 (when band gaps are too large for visible range absorption) to a maximum of 0.25 (or 25% efficiency). Fig. 4(a) shows E_{gap}^{PBE} plotted against ΔH^{PBE}, with the shaded region showing the ranges of favorability, chosen here as $\Delta H^{PBE} < 0$ eV and 1 eV < E_{gap}^{PBE} < 2.5 eV.

We find that stable compositions ($\Delta H^{PBE} < 0$ eV) are predominantly occupied by HOIPs, with a fair few pure, B-site mixed, and X-site mixed compounds. A large number of A-site mixed HOIPs as well as a majority of inorganic HaPs occupy the unstable region, indicating that although the presence of organic cations is desirable to prevent perovskite decomposition, mixing at the A-site may not always be beneficial. The band gap shows less clear trends, and as will be explained later, is largely dependent on the type and number of specific B and X-site ions. The region of desirable E_{gap}^{PBE} and ΔH^{PBE} is largely populated by HOIPs with B-site or X-site mixing. Furthermore, Fig. 4(b) shows $SLME^{PBE}$ plotted against E_{gap}^{PBE}, showing the characteristic relationship that has been explored in past studies.[41,42] SLME rises initially as the band gap increases, reaching a peak of ∼25% around E_{gap}^{PBE} = 1.5 eV, and subsequently goes down until it goes to 0 for E_{gap}^{PBE} > ∼3 eV. The largest $SLME^{PBE}$ values are shown by pure hybrid and B-site mixed compounds, both hybrid and inorganic.

3.3 Composition–property correlations

To obtain a qualitative understanding of how different constituents at the A, B, and X sites contribute to the properties of interest, we encode each compound in the dataset using a set of descriptors and calculate the Person coefficient of linear correlation[43] between each descriptor dimension and each property. Since all HaPs in this study are cubic or pseudo-cubic, the essential distinguishing feature from one compound to another is the composition or the chemical formula. Every compound is thus encoded using two types of descriptors: a 14-dimensional composition vector representing fractions of every species (Cs, MA, Pb, Br, *etc.*) in the compound, and a 36-dimensional "elemental properties" vector, representing weighted averages of 12 elemental (or molecular) properties each (such as ionic radii, electron affinity, ionization energy, *etc.*) of the respective species at A, B, and X sites. A complete list of all 50 descriptors is provided in Table SIII.†

Fig. 5(a) shows the linear correlation between composition descriptors and PBE properties, namely lattice constant, decomposition energy, band gap and SLME. In the heatmap, darker red implies large positive correlation, darker blue implies large negative correlation, and white means there is little or no correlation. A few important relationships immediately stand out from this plot: large ions like Ba and I lead to an increase in the lattice constant, while Cl has the reverse effect.

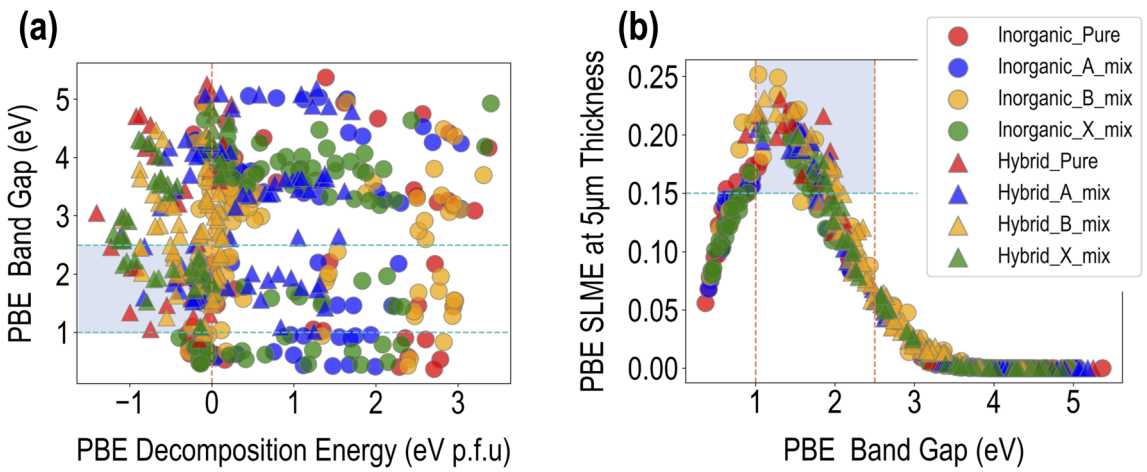

Fig. 4 Visualization of the PBE dataset: (a) band gap against decomposition energy, and (b) SLME at 5 μm thickness sample thickness against the band gap. Different colors represent different types of mixing and different symbols are used to distinguish between purely inorganic HaPs and HOIPs. The shaded regions attempt to capture compounds with negative decomposition energy, a band gap between 1 eV and 2.5 eV, and SLME larger than 15%.

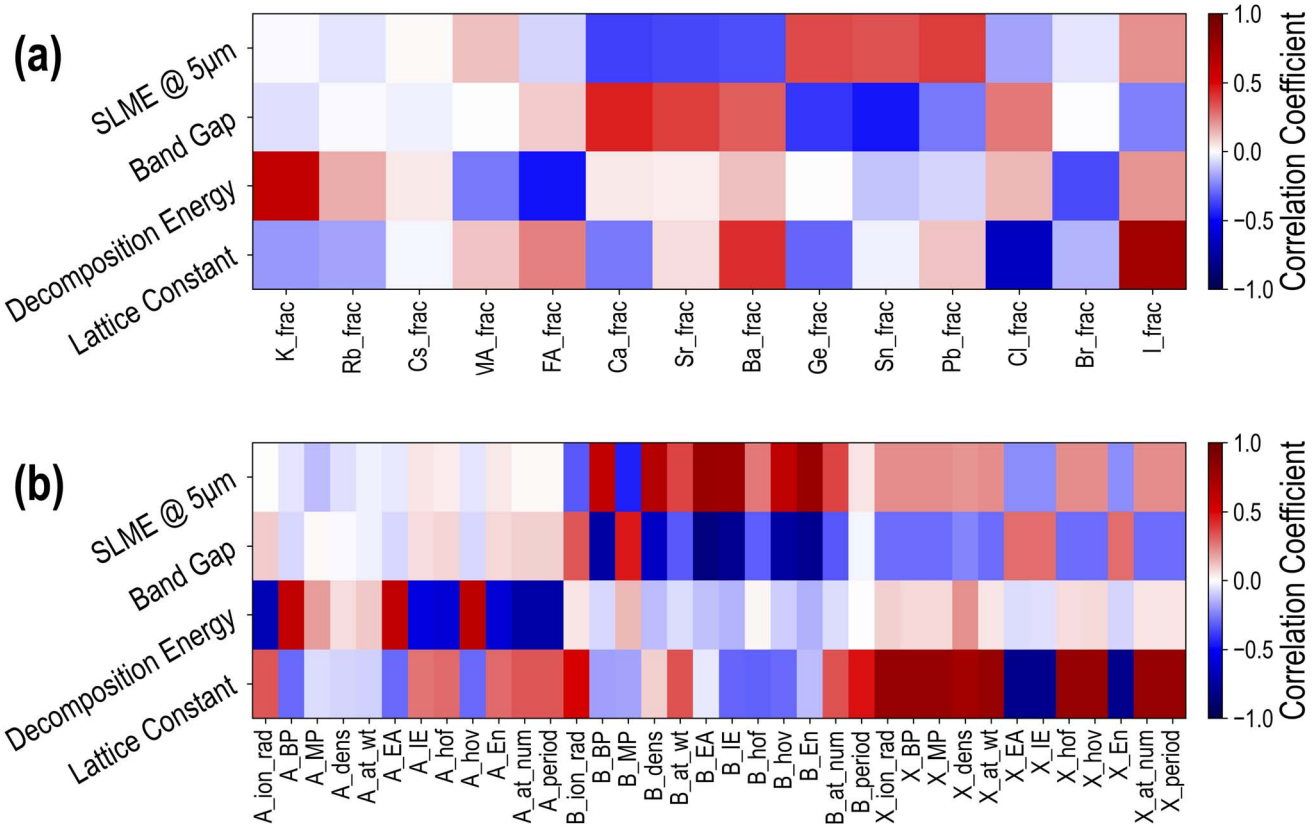

Fig. 5 Pearson coefficients of linear correlation between 4 PBE computed properties and (a) 14 compositional descriptors and (b) 36 elemental property descriptors.

An increase in the fraction of K at the A-site increases ΔH^{PBE} and thus makes the compound more unstable, while increasing the fraction of FA will make it more stable. B-site elements generally have little effect on the stability, but have much larger correlations with E_{gap}^{PBE} and SLME^{PBE}. While Ca, Sr, and Ba increase E_{gap}^{PBE} and decrease SLME^{PBE}, Ge and Sn decrease E_{gap}^{PBE} and increase SLME^{PBE}. Pb shows a large positive correlation with SLME^{PBE}, which is not surprising given that FA/MA/Cs-based Pb iodide or bromide perovskites are most common in optoelectronic applications. Finally, X-site species show more modest correlation with E_{gap}^{PBE} and SLME^{PBE}, with Br showing virtually no correlation with the band gap, which is an effect of Br lying between I and Cl in the band gap spectrum. The lower values of correlation between X-site constituents and the band gap and SLME reveal that mixing or complete substitution at the B-site has a more significant effect on the optoelectronic properties. These correlations provide confirmation for some well-known effects and some simple design principles for HaP compositions with targeted properties.

Next, we calculated the linear correlations between the 36-dimensional elemental property descriptors and the 4 PBE properties, and the results are presented in Fig. 5(b). Once again, it can be seen that the biggest contributors to ΔH^{PBE} are A-site properties: specifically, increasing the ionic radius, ionization energy, or atomic number of A-site species makes the compound more stable, whereas increasing the boiling point, electron affinity, or heat of vaporization makes it less stable. The largest correlations with the lattice constant are from X-site features, with higher electron affinity, ionization energy, or electronegativity of the X-site constituent reducing the lattice constant and all other properties increasing it. When it comes to the band gap and SLME, we once again notice an overwhelming contribution from the B-site species. Increasing the boiling point, electron affinity, ionization energy, or electronegativity of B-site species helps decrease E_{gap}^{PBE} and increase SLME^{PBE}, explaining why Pb/Sn/Ge are clearly more beneficial in PV applications than Ba/Sr/Ca at the B-site. These correlations help expand our design principles based purely on the fractions of different species and provide an opportunity to train predictive models for various properties.[16,22]

3.4 Improving property predictions using HSE06 and spin–orbit coupling

It was shown in Section 3.1 that for a set of selected HaP compositions, while PBE-optimized lattice constants match well with experiments, PBE band gaps are underestimated, and HSE–PBE-SOC band gaps match better with measured values. GGA-PBE computations are generally reliable for the structure and stability (formation or decomposition energy) of both hybrid and purely inorganic HaPs, but advanced levels of theory such as the HSE06 functional or GW approximation, with the inclusion of SOC to account for the relativistic effects of heavy

atoms such as Pb, are of paramount importance when it comes to electronic and optical properties. Here, we perform a series of expensive HSE calculations across the HaP dataset and report trends and major observations, specifically the effect of full geometry optimization with HSE compared to using the PBE-optimized structures, and the effect of incorporating SOC in the calculation. Overall, we generate HSE datasets of the decomposition energy, band gap, and SLME, for HSE-relaxed (299 data points), HSE-relaxed + SOC (282 data points), and HSE–PBE + SOC (244 data points).

A visualization of the types of mixing per A/B/X-site species is presented in Fig. S5,† and different properties are plotted against each other for the three types of HSE datasets in Fig. S9–S11.† We find similar distributions to the PBE data, with notable differences coming from HSE band gaps being generally larger and eliminating a lot of the low SLME data points. Very similar ΔH values are obtained for all compositions from the 4 methods, showing that PBE-based stability metrics should be more than reliable. We note here that SLME from the different HSE functionals is obtained using the PBE-computed optical absorption spectrum shifted along the energy axis by the difference between E_{gap}^{PBE} and the corresponding HSE E_{gap}; this is a method that helps us determine a theoretical efficiency from HSE without performing a full optical absorption calculation using HSE. Fig. 6 presents a comparison between the different types of HSE and PBE band gaps, dividing the data in terms of the nature of A-site species: purely organic, purely inorganic, or mixed organic–inorganic. While it is clear that B-site and X-site species are the major contributors to the band gap, we divide the data like this mainly to observe how important HSE *vs.* PBE geometry optimization is for hybrid *vs.* inorganic HaPs and the magnitudes of difference between PBE and HSE band gaps and between using and not using SOC. Furthermore, we observe from our dataset that HSE-relaxation might be superfluous, as HSE-relaxed lattice parameters, decomposition energies, and band gaps largely correlate with the corresponding PBE-relaxed values, as shown in Fig. S6–S8.†

It can be seen from the 299 data points in Fig. 6(a) that $E_{gap}^{HSE\text{-}rel}$ is, on average, 1 eV or more greater than E_{gap}^{PBE}, with larger differences appearing when A-site contains only organic molecules; we attribute this to the larger degree of geometry optimization from HSE in the presence of organic cations than when only inorganic cations are present, leading to larger differences in the band gap. Fig. 6(b) shows $E_{gap}^{HSE\text{-}rel\text{-}SOC}$ plotted against $E_{gap}^{HSE\text{-}rel}$ for 282 data points. As expected, SOC brings down E_{gap} for many of the compounds, and keeps E_{gap} the same for many other compounds, confirming that SOC is certainly vital for certain compositions but can be ignored in others, as has been discussed in past studies.[16,40,44] Interestingly, we observe that for several purely inorganic HaPs with lower E_{gap} < 3 eV, SOC significantly reduces the gap. Next, we plot in Fig. 6(c) $E_{gap}^{HSE\text{-}PBE\text{-}SOC}$ *vs.* $E_{gap}^{HSE\text{-}rel\text{-}SOC}$ for 244 data points, in an attempt to understand the difference between HSE-relaxed and HSE-on-PBE-relaxed band gaps (with the inclusion of SOC in both). We find virtually identical band gaps from both methods for all pure inorganic HaPs, but large differences when organic cations exist at the A-site, which can once again be explained by the more severe geometry optimization from HSE in the latter. While $SLME^{PBE}$ peaked at around 25%, the corresponding HSE peaks appear around 16% as a consequence of shifting the optical absorption spectrum by the difference between the PBE and HSE band gaps. For compounds with the highest $SLME^{PBE}$ values, E_{gap}^{PBE} is around 1 eV, while the corresponding HSE-rel band gaps are higher and the HSE-rel + SOC and HSE–PBE-SOC are lower. These band gap differences take $SLME^{HSE}$ lower than the PBE peak. This effect holds for all compounds with $SLME^{PBE} > 15\%$, such that $SLME^{HSE}$ following the band gap shift always tends to be lower and peaks at 15 to 16%.

Finally, we examine the relationships between HaP composition and various types of band gaps discussed above, by calculating Pearson coefficients of linear correlation. Fig. 7 shows the correlations for five types of properties, namely E_{gap}^{PBE} (PBE Gap), $E_{gap}^{HSE\text{-}rel}$ (HSE Gap), $E_{gap}^{HSE\text{-}rel\text{-}SOC}$ (SOC Gap), $E_{gap}^{HSE\text{-}rel} - E_{gap}^{PBE}$ (Δ(HSE–PBE)), and $E_{gap}^{HSE\text{-}rel\text{-}SOC} - E_{gap}^{HSE\text{-}rel}$ (Δ(SOC-HSE)). For the first three quantities, we find virtually identical behavior, and it can be concluded that various A/B/X-site species have the same increasing or decreasing influence on any PBE or HSE E_{gap}. Correlations with the (HSE–PBE) E_{gap} difference values show that certain constituents such as FA, Cs, Sn, or I could have marginal influence, but the differences are largely uniform across the

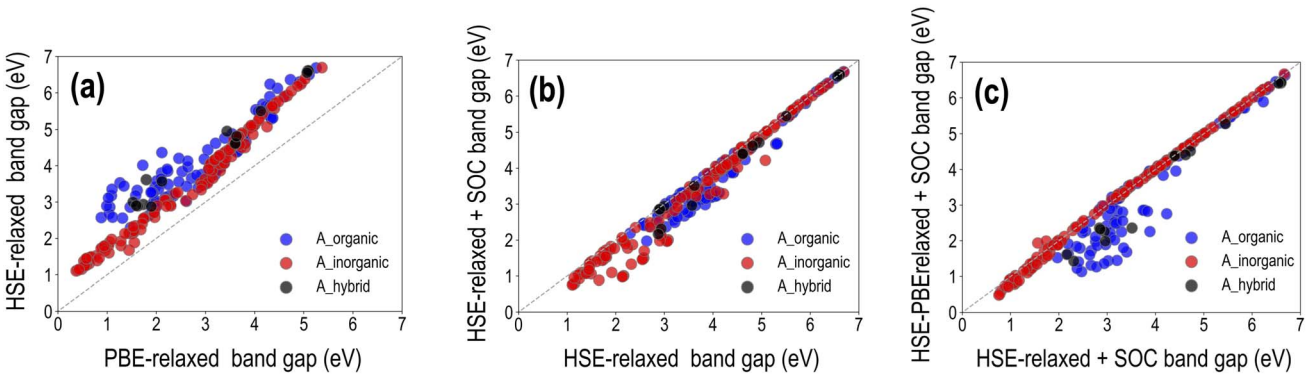

Fig. 6 Visualization of band gaps computed from various HSE06 approaches: (a) HSE-relaxed band gap *vs.* PBE band gap. (b) HSE-relaxed band gap with SOC *vs.* HSE-relaxed band gap without SOC. (c) HSE–PBE with SOC band gap *vs.* HSE-relaxed with SOC band gap.

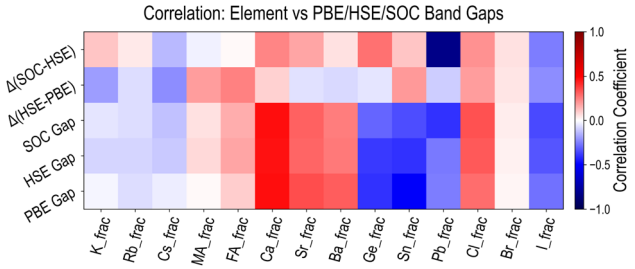

Fig. 7 Pearson coefficients of linear correlation between various types of band gaps and the 14 compositional descriptors. HSE gap refers to the HSE-relaxed band gap, SOC gap refers to the HSE-relaxed with SOC band gap, Δ(HSE−PBE) is the difference between PBE and HSE-relaxed band gaps, and Δ(SOC-HSE) is the difference between the HSE-relaxed band gap with and without SOC.

dataset. The effect of SOC is very evident in the correlation analysis for $E_{gap}^{HSE\text{-}rel\text{-}SOC} - E_{gap}^{HSE\text{-}rel}$. While A-site species have no influence here, Pb has by far the highest negative correlation, Ge has a slightly positive correlation, and I has a slightly negative correlation. We conclude that the inclusion of SOC is of utmost importance for Pb-based HaPs, and E_{gap} values will be significantly lower (and more accurate) when using SOC. Fig. S12–S14† present the complete correlation analysis for all properties computed from the three types of HSE functionals and the 50-dimensional descriptors introduced earlier, showing very similar trends as compared to the PBE dataset.

3.5 Deviation from cubicity

As alluded to earlier, some of the PBE and/or HSE geometry optimization calculations, especially when multiple organic cations are present in the HaP supercell, could lead to significant distortions of the cubic perovskite structure. Beyond the use of a perfectly cubic supercell as the starting geometry, the cubic shape is not enforced in the computations, but a vast majority of the structures in the dataset are cubic or pseudo-cubic. Here, we investigate how far any structure deviates from an acceptable pseudo-cubic shape, and use this information to subsequently screen out severely deformed, visibly non-cubic, or non-perovskite phases, even if the energy may be low. We define a metric known as Deviation from Cubicity (DC), estimated by how different $\bar{b}$ and $\bar{c}$ lattice constant values are compared to lattice constant a, as shown in eqn (6). Similarly, an angular deviation is calculated by measuring how different angles α, β, and γ are from 90°, as shown in eqn (7). DC values greater than 10% for the lattice constant and greater than 5% are considered too non-cubic and excluded during the screening process, which will be explained in a later section.

$$DC_b = \frac{|b-a|}{a} \tag{6}$$

$$DC_\alpha = \frac{\alpha - 90^\circ}{90^\circ} \tag{7}$$

$$DC_{avg} = \frac{DC_b + DC_c + DC_\alpha + DC_\beta + DC_\gamma}{5} \tag{8}$$

Fig. 8(a) shows ΔH^{PBE} plotted against the average deviation from cubicity (DC_{avg}), calculated using eqn (8). Fig. S15 and S16† show individual plots of ΔH^{PBE} against DC corresponding to $\bar{b}$, $\bar{c}$, α, β, and γ. It can be seen that ~90% of the compounds show a DC_{avg} of <2%, reinforcing confidence in the cubic/pseudo-cubic nature of a majority of the dataset. Around 20 compounds show a DC_{avg} of >5%, and all of them are HOIPs with A-site, B-site, or X-site mixing. The non-symmetricity introduced in the supercell when large organic molecules are mixed with other organic or inorganic cations, and when complex mixing is performed at the B or X sites in the presence of large organic cations, leads to elongation, contraction, or twist along one or more directions. A consequence of the high-throughput nature of our computational work is the inability to visually examine the cubicity of every structure: the current analysis helps reveal some unfavorable deviations in certain compounds, which will be used as one of the factors while determining suitable compositions in terms of perovskite formability, stability, and optoelectronic properties. It should be noted that larger DC_{avg} values tend to correspond to negative ΔH^{PBE}, but despite their stability from DFT, such compounds have a non-perovskite like phase and are thus excluded from current screening and saved for future analysis.

3.6 Comparing perovskite formability factors with decomposition energy

The formability of an ABX_3 perovskite is typically predicted using the Goldschmidt tolerance and octahedral factors, which depend on the ionic radii of A, B, and X-site species. In recent years, there have been newer factors devised through analysis of large quantities of experimental and computational perovskite data, often using machine learning techniques;[22] one such factor was suggested by Bartel *et al.*[7] Here, we utilize three factors, namely the traditional tolerance factor (t), the octahedral factor (o), and the Bartel tolerance factor (t_B), defined using eqn (9)–(11) respectively, to quantify the formability of all perovskites in our dataset and compare these values with DFT computed ΔH. For compounds with mixing, the weighted averages of A-site (r_A), B-site (r_B), and X-site (r_X) radii are considered.

Octahedral factor:

$$o = \frac{r_B}{r_X} \tag{9}$$

Tolerance factor:

$$t = \frac{r_A + r_X}{\sqrt{2}(r_B + r_X)} \tag{10}$$

Bartel[7] tolerance factor:

$$t_{Bartel} = \frac{r_X}{r_B} - \left[1 - \frac{\frac{r_A}{r_B}}{\ln\left(\frac{r_A}{r_B}\right)}\right] \tag{11}$$

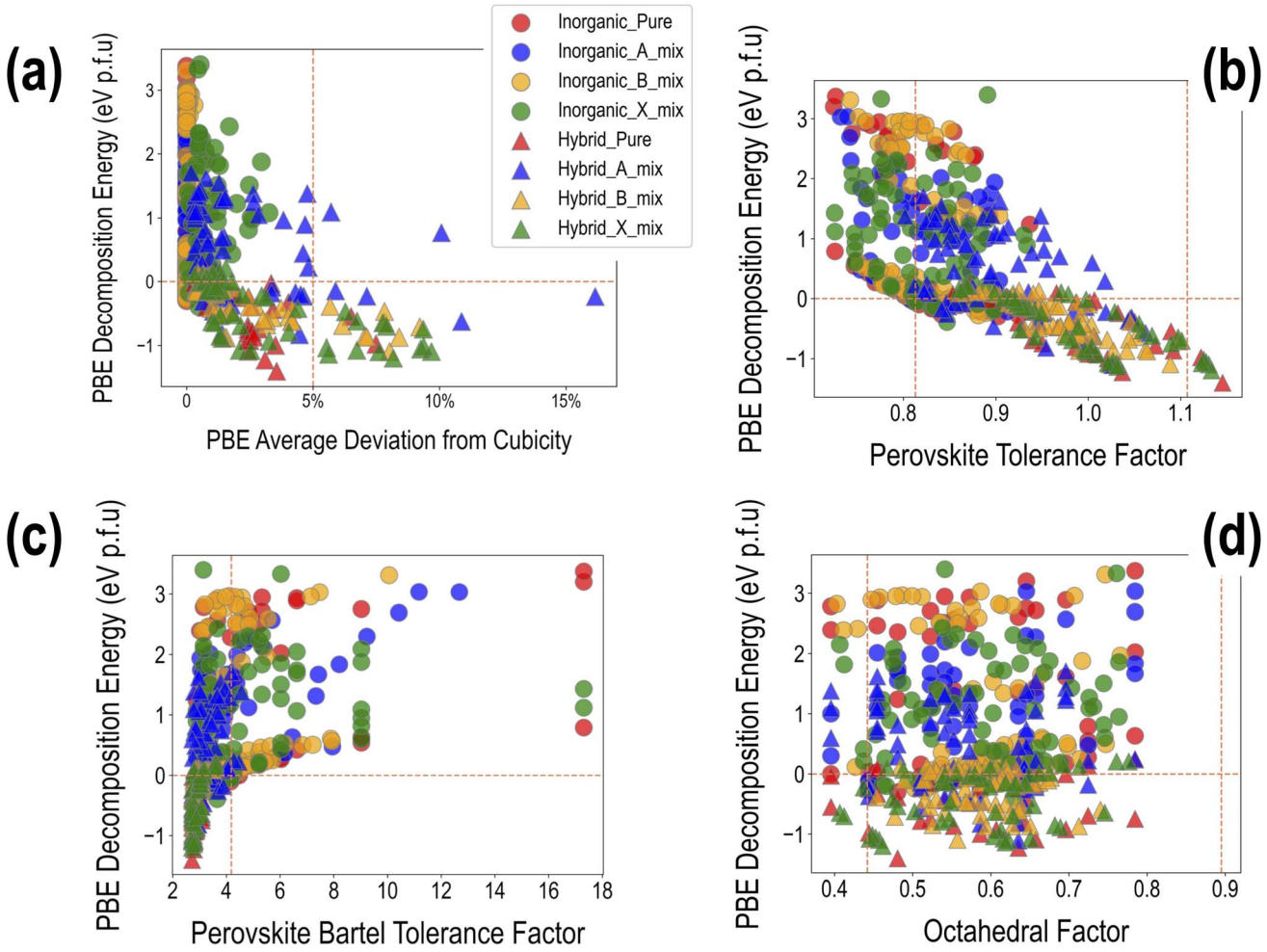

Fig. 8 The PBE computed decomposition energy plotted against (a) average deviation from cubicity, (b) Goldschmidt tolerance factor, (c) Bartel tolerance factor, and (d) octahedral factor. The vertical and horizontal dashed lines aim to distinguish between negative and positive decomposition energies and highlight desirable ranges of other quantities, namely $DC_{avg} < 5\%$ (a), $t \in (0.813–1.107)$ (b), $t_B < 4.18$ (c), and $o \in (0.442–0.895)$ (d).

The suggested ranges for perovskite formability are $o \in (0.442–0.895)$, $t \in (0.813–1.107)$, and $t_B < 4.18$. Fig. 8(b)–(d) respectively show t, t_B, and o values for the entire dataset plotted against ΔH^{PBE}. It can be seen that there is a roughly inverse relationship between t and ΔH^{PBE}, which is to be expected as larger values of t mean more favorable perovskite formability and should thus also correspond to negative decomposition energies. A small number of compounds lie in the $t > 1.1$ range and also show negative ΔH^{PBE}; upon closer inspection, we find that they are HOIPs with significantly distorted structures showing large deviation from cubicity. One example of such a compound is $FAGeBr_{2.25}Cl_{0.75}$, which has $t = 1.13$ and $\Delta H^{PBE} = -1.06$ eV, but a highly distorted PBE optimized structure with $DC_b = 16.4\%$. Fig. 8(c) shows that essentially all HaP compositions with negative ΔH^{PBE} fall within the desirable $t_B < 4.18$ region. Similarly, Fig. 8(d) shows that nearly all compounds with negative ΔH^{PBE} lie in the desirable range of o values, barring a very small number of distorted structures. It should be noted from Fig. 8(b)–(d) that hundreds of compositions that satisfy the formability factor conditions show positive (often very large positive) ΔH^{PBE} values, which means that they will easily decompose to other halide phases. Our observations point to the idea that such factors may be necessary but not sufficient conditions for perovskite formability and stability.

3.7 High-throughput screening of compositions with favorable properties

So far, we have visualized and analyzed an HT-DFT dataset of HaP alloys using PBE and multiple HSE functionals. In performing an initial screening of candidates with promise for single-junction solar absorption, we must consider ΔH (a necessary but not complete description of perovskite stability), E_{gap}, and SLME; in addition, established perovskite formability factors as well as deviation from cubicity should be considered. We note here that the DFT dataset covers as wide a compositional spectrum of HaPs as possible, within the 14-dimensional chemical space. There are, of course, innumerable compositions that could be generated which are intermediate to those currently being studied, such as by mixing in fractions other

than $n/8$ (where n is a positive integer), which may require simulations in larger supercells and potentially lead to even more desirable combinations of properties. We tackle this issue in future studies by building upon our current dataset and applying state-of-the-art ML algorithms for prediction and inverse design. For the moment, we use the criteria/properties described in the previous section to screen for promising materials within the DFT datasets.

We apply the following screening criteria on the PBE dataset:

1. Formability: $o \in 0.442$–0.895, $t \in 0.813$–1.107, and $t_B < 4.18$.

2. (Pseudo) cubicity: $DC_b < 10\%$, $DC_c < 10\%$, $DC_\alpha < 5\%$, $DC_\beta < 5\%$, and $DC_\gamma < 5\%$,

3. Thermodynamic stability: $\Delta H^{PBE} < 0$ eV. Negative ΔH is a necessary but not complete metric for preventing ABX_3 decomposition to phases AX and BX_2; decomposition could occur to other phases, and the effects of kinetics, ion segregation, defects, *etc.* are ignored in this work.

4. Band gap: $E_{gap}^{PBE} \in 1$–2.5 eV. PV-suitable band gaps lie close to 1.5 eV. We use a wide range here to account for the various inadequacies of the PBE band gap description: it will underestimate gaps of inorganic compounds but either accidently be accurate or slightly overestimate the gaps of HOIPs: this effect has been studied in prior studies.[16,40]

5. PV efficiency: $SLME^{PBE} > 0.10$. This criterion goes hand-in-hand with the band gap requirement, as it can be seen from Fig. 4(b) that the highest SLME values correspond roughly to the band gap range described above.

Fig. 9 shows our five-fold screening process, based on which we obtain 32 candidates (out of 495) that fulfill each requirement. Also shown is a pie chart with the distribution of various types of mixing in the screened list of compounds. It can be seen that a majority of the screened compounds, 19 in total, are B-site mixed, and there are only 6 unalloyed compositions. Fig. 10 further shows the relative distributions of various A-site, B-site, and X-site species. We find that MA followed by FA is by far the most common A-site cation, often occupying all of the A-site (8/8 mixing fraction), followed by Cs and Rb. There are no pure K-based compounds; K, as well as Rb and Cs, occur in small fractions in some of the compounds. Pb and Sn appear in an overwhelming majority of the compounds, with Ge, Ca, Sr, and Ba only occurring in smaller fractions of 3/8 or less. This is consistent with the observation that Pb and Sn, and sometimes Ge, are most beneficial for ideal optoelectronic properties, whereas Ca/Sr/Ba should occur in minor fractions to keep the band gap small. Pb has a high preference for 7/8 and 8/8 occupation, hinting at the difficulty in developing Pb-free perovskites with ideal properties. At the X-site, Br and I without any mixing are most common, and in the 4 compounds with X-site mixing, I, Br, and Cl are found in various fractions.

We performed a similar screening procedure using the HSE–PBE-SOC dataset, which was found to compare best with experiments for the band gap. Applying the very same criteria as shown in Fig. 9 leads to a list of 14 stable and formable compounds with desirable band gaps and SLME, out of which 4

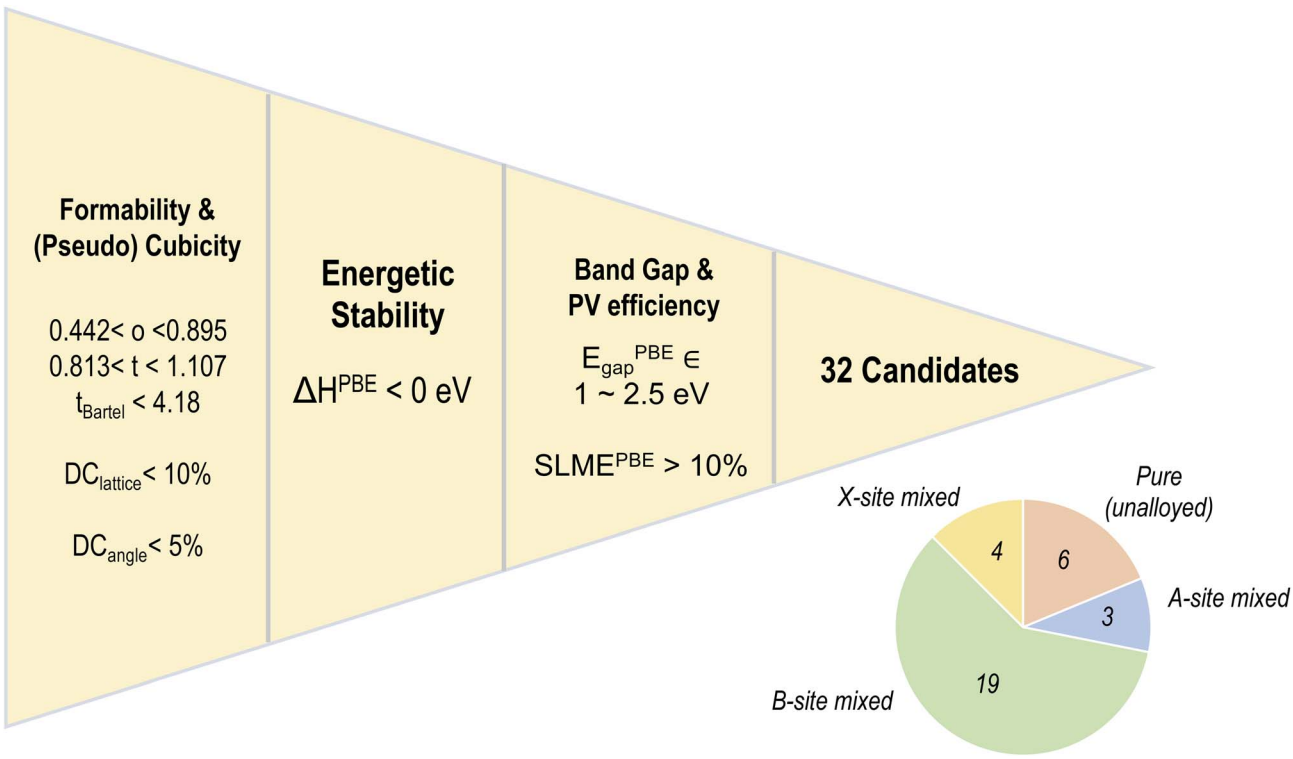

Fig. 9 Screening methodology applied on the PBE dataset, yielding 32 candidates that satisfy all perovskite formability and cubicity conditions, show negative decomposition energy, and PV-appropriate band gaps and SLME. The pie chart shows the distribution of various alloy types in the screened list of compounds.

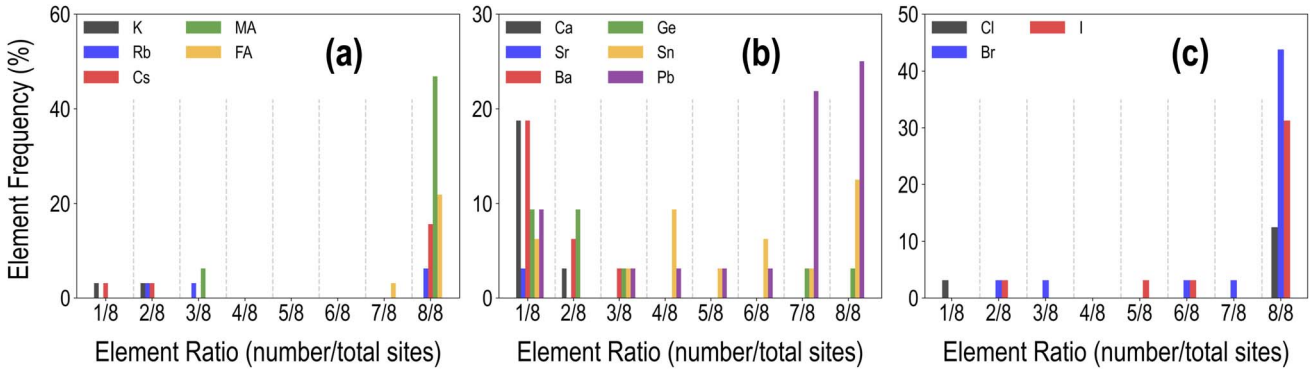

Fig. 10 Distribution of mixing fractions of various species at the A (a), B (b), and X (c) sites across the list of 32 promising compounds selected from the PBE dataset.

are pure unalloyed compounds, 1 is (purely inorganic) A-site mixed, 8 are B-site mixed, and 1 is X-site mixed. Distributions of the types of mixing fractions of various species in this screened list of compounds are shown in Fig. S17.† Although the HSE–PBE-SOC screened list is much smaller than the PBE screening list due to a smaller overall dataset, some similar trends are found in both screening procedures. B-site mixing is most prevalent, as is high fractions of Pb, and sometimes Sn and Ge, at the B-site. Ca/Sr/Ba prefer mixing in small fractions. Most compounds are MA-based and nearly all of them contain Br or I. We note that our conclusions on the populations of different species and types of mixing in the screened set of compounds may change slightly in the future if a much larger dataset is available, such as *via* machine learning-based predictions.

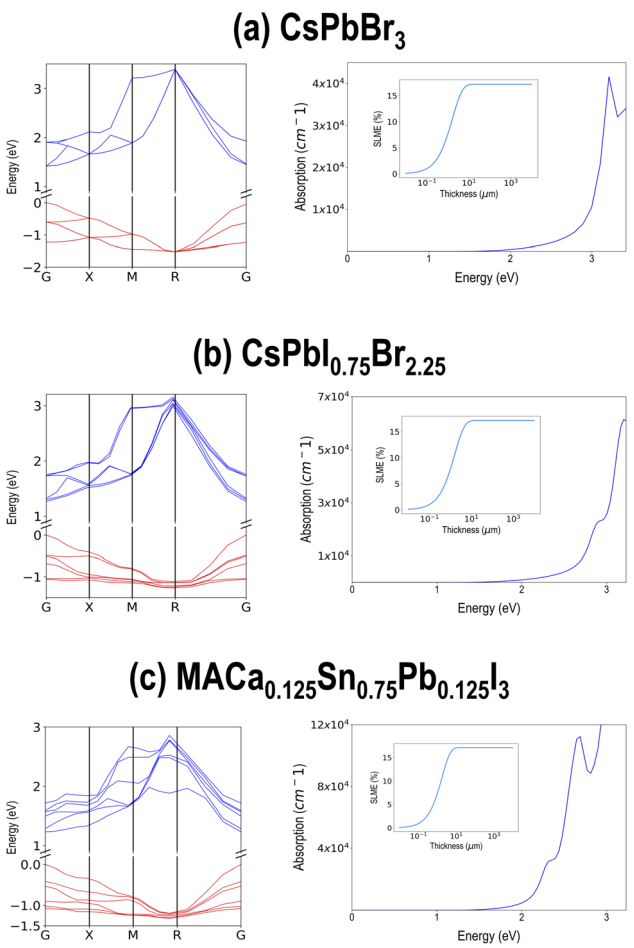

Fig. 11 HSE–PBE-SOC calculated electronic band structures, optical absorption spectra, and SLME *vs.* sample thickness (inset) plots for three promising compounds, namely (a) $CsPbBr_3$, (b) $CsPbI_{0.75}Br_{2.25}$, and (c) $MACa_{0.125}Sn_{0.75}Pb_{0.125}I_3$.

The entire screened lists of compounds from PBE and HSE–PBE-SOC are presented in Tables SII and SIII† respectively, along with their chemical formula and (PBE or HSE–PBE-SOC) computed ΔH, E_{gap}, and SLME at 5 μm sample thickness. Interestingly, all the compounds in the HSE–PBE-SOC list appear in the PBE list as well. Three of the best performing compounds are selected and their HSE–PBE-SOC computed electronic band structures, optical absorption spectra, and SLME *vs.* sample thickness plots are pictured in Fig. 11. These compounds, namely $CsPbBr_3$, $CsPbI_{0.75}Br_{2.25}$, and $MACa_{0.125}Sn_{0.75}Pb_{0.125}I_3$, show direct band gaps around 1.5 eV and SLME > 15% in their cubic or pseudo-cubic phases.

4 Prospects and future work

What we reported in this work is one of the largest DFT datasets to date of pseudo-cubic HaP alloys containing some of the most commonly used cation and anion species. These data enabled us to understand the dependence of stability and optoelectronic properties on perovskite composition, specifically the type of mixing. However, this work is the first step in a very long process that will involve extensions to non-cubic phases, other properties of interest, more improved levels of theory, alternative cation and anion choices, and other perovskite forms such as double perovskites and 2D perovskites, ultimately leading to more universal prediction, screening, and design. It is important to note that while the bulk stability, band gap, and theoretical single-junction PV efficiency provide essential parameters for initial screening of PV-relevant HaPs, extensions need to be made to other crucial properties, including electron and hole transport properties, formation energies and

electronic levels of point defects, and the behavior of relevant perovskite surfaces and interfaces.

In our work, the immediate next extension is towards non-cubic perovskite phases. For instance, $CsPbBr_3$ may prefer the orthorhombic phase, while $MAPbI_3$ and $MA(Pb–Sn)I_3$ may assume the tetragonal phase, and this work considers all such compounds only in a cubic or pseudo-cubic rendition. In previous work,[16] it was shown that for the same composition, unalloyed or with mixing, changing the phase could modify the band gap by 0.5 eV or more in many cases. Cubic perovskite phases are often not the ground state, and are in many cases very unstable—as indicated by the large positive decomposition energies for many compounds in our dataset. Non-cubic phases are either the most stable, or metastable/competing phases for most of the compositions studied in this work. Currently, we have high-throughput computations ongoing for tetragonal, orthorhombic, and hexagonal phases of several mixed HaPs; the perovskite phase itself can be added as an input to the compositional and elemental descriptors to obtain new correlations. In addition, computations are being performed for further tailoring of properties by accessing polymorphs within each phase, *e.g.*, *via* octahedral distortion and rotation,[45] or *via* re-optimization of the same composition in larger supercells with slight distortions.[46] As an example, plots showing the computed decomposition energy for selected perovskites with varying degrees of octahedral distortion as well as in different prototypical phases are presented in Fig. S18 and S19;† it can be seen that some amount of distortions can keep the perovskite stable, the cubic phase is not always the ground state, and sometimes the range of decomposition energies for a given composition can be quite broad.

We further anticipate significant improvements in DFT predictions of various properties. Our attempt to utilize a few different functionals to benchmark properties against experiments was hindered by a number of factors discussed in the manuscript, including the perovskite phase and lack of additional corrections. Our ongoing computations involve testing the influence of the PBEsol[38] and PBE-D3 (ref. 39) functionals, combined with static HSE06 or GW computations with SOC,[47] for better optical and electronic properties. Furthermore, the inclusion of new types of elemental or molecular species – such as transition metals (Cd, Zn, Ni, *etc.*) at the B-site would necessitate the use of specific levels of theory, such as GGA + U.[48] The consideration of other important properties, such as defect formation energies and carrier mobilities, would involve testing and deploying multiple functionals as well.

It should be noted again that there might not be one best functional that works for the entire chemical space when considering organic *vs.* inorganic A-site cations, and Pb/Sn *vs.* other B-site cations. A likely solution is the use of an ensemble of functionals as well as experimental estimates (which might need to be averaged as well, given the range of values generally reported by different experimental researchers for the same materials) for hundreds of HaP compositions, and training of multi-fidelity ML models.[49] Large quantities of low-fidelity data combined with more modest amounts of high-fidelity data can lead to highly accurate predictions of experiment-level property estimates.

In general, ML has a massive role to play here, as has been demonstrated for HaPs in multiple prior studies.[16,22] Concurrent manuscripts are planned to report rigorously optimized predictive models for multiple properties and fidelities, based on the datasets and descriptors discussed in this work. Such models can easily be extended to new choices for A/B/X ions such as transition metals,[16] as well as other phases, by addition of new dimensions to the descriptors. The inclusion of more general crystalline structure representations as inputs for ML, such as using crystal graphs and graph neural networks,[50–52] would be essential for treating same compositions and structures with a variety of distortions or lattice strains. Once composition-based and/or structure-based ML predictive models are rigorously optimized and validated, they could be deployed for the prediction of over thousands of ABX_3 compounds available in databases such as the Materials Project[53] or Open Quantum Materials Database,[54] as well as over millions of hypothetical materials, for prediction, screening, and discovery. Finally, inverse design techniques, such as using the genetic algorithm[55] or generative neural networks,[56] could be applied upon the DFT-ML surrogate models to drive the efficient discovery of new HaP compositions/structures with multiple desired properties. For instance, our ongoing work involves generating populations of novel HaP compositions using GA while optimizing a fitness function that includes metrics for chemical feasibility, negative ΔH, E_{gap} between 1 and 2 eV, and SLME > 15%; this process can yield thousands of promising compounds beyond the scope of the current work and beyond brute-force enumeration. The dataset and analysis presented in this work serve as a springboard for efforts that are currently underway, to ultimately accelerate the prediction and design of novel perovskites for optoelectronics and to extend such approaches to other material classes and applications.

5 Conclusions

In this work, we present a high-throughput DFT dataset of pseudo-cubic ABX_3 halide perovskite alloys, with mixing of multiple ions permitted at the A, B, or X sites, using the GGA-PBE functional and three types of hybrid HSE06 approaches. This dataset contains 495 unique compositions with PBE computed decomposition energies, band gaps, and spectroscopic maximum limited efficiencies (SLME) from the optical absorption spectra, and the same properties for 299 compounds from full HSE relaxation, 282 compounds from HSE relaxation with spin–orbit coupling (SOC), and 244 compounds from static HSE computations on PBE relaxed structures with SOC. Pearson correlation analysis reveals the extent of positive or negative correlation of the amount of any A/B/X species as well as their well-known elemental/molecular properties with the computed stability and optoelectronic properties, reproducing known trends and unraveling interesting new relationships. Screening is performed for materials resistant to decomposition, with photovoltaic-suitable band gaps and high SLME, as well as including other perovskite formability factors such as

Goldschmidt tolerance and octahedral factors and the deviation of the perovskite structure from cubicity, to obtain 32 promising compounds from PBE and 14 from HSE–PBE-SOC. This work forms the basis for predictive machine learning models which will accelerate the design of novel perovskites with attractive properties.

Data availability

This statement explains the availability and details for data and codes used in the manuscript.

(1) All Density Functional Theory (DFT) computations are performed using the Vienna Ab initio Simulation Package (VASP), version 6.2. The software can be found at **https://www.vasp.at/**.

(2) Starting structures for this work, as well as a subset of reported structures and properties, can be found in our previous publication: A. Mannodi-Kanakkithodi and M. K. Y. Chan, *Energy Environ. Sci.*, 2022, **15**, 1930–1949, **https://doi.org/10.1039/D1EE02971A**.

(3) Tabulated data and scripts for extracting all properties of perovskites are available on our Github repo: **https://github.com/yjq829/perovskite_dataset.git**. The Tabulated data is also provided as an .xlsx file in the supporting documents.

(4) The code for calculating Spectroscopic Limited Maximum Efficiency (SLME) analysis can be found at **https://github.com/ldwillia/SL3ME.git**, based on the publication **https://doi.org/10.1103/PhysRevLett.108.068701**.

(5) All raw DFT data, including input and output files, can be found on Materials Data Facility:[57] A. Mannodi-Kanakkithodi, M. K. Chan, J. Yang and P. Manganaris, High-Throughput DFT Dataset of Halide Perovskite Alloys, 2022, **https://petreldata.net/mdf/detail/abx3_perovs_alloys_v1.1**.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

Extensive discussions with and scientific feedback from Prof. David Fenning (UC San Diego), Dr Rishi Kumar (Berkeley lab), and Dr Maria Chan (Argonne National Lab) are acknowledged. This work was performed at Purdue University, under startup account F.10023800.05.002 from the Materials Engineering department. This research used resources of the National Energy Research Scientific Computing Center, the Laboratory Computing Resource Center at the Argonne National Laboratory, and the RCAC clusters at Purdue.

References

1 M. I. H. Ansari, A. Qurashi and M. K. Nazeeruddin, *J. Photochem. Photobiol., C*, 2018, **35**, 1–24.
2 W.-J. Yin, J.-H. Yang, J. Kang, Y. Yan and S.-H. Wei, *J. Mater. Chem. A*, 2015, **3**, 8926–8942.
3 J. S. Manser, J. A. Christians and P. V. Kamat, *Chem. Rev.*, 2016, **116**, 12956–13008.
4 T. M. Brenner, D. A. Egger, L. Kronik, G. Hodes and D. Cahen, *Nat. Rev. Mater.*, 2016, **1**, 15007.
5 P. Cui, D. Wei, J. Ji, H. Huang, E. Jia, S. Dou, T. Wang, W. Wang and M. Li, *Nat. Energy*, 2019, **4**, 150–159.
6 M. Jeong, W. C. In, M. G. Eun, Y. Cho, M. Kim, B. Lee, S. Jeong, Y. Jo, W. C. Hye, J. Lee, J.-H. Bae, K. K. Sang, S. K. Dong and C. Yang, *Science*, 2020, **369**, 1615–1620.
7 J. Bartel Christopher, C. Sutton, R. Goldsmith Bryan, R. Ouyang, B. Musgrave Charles, M. Ghiringhelli Luca and M. Scheffler, *Sci. Adv.*, 2019, **5**, eaav0693.
8 S. Zhu, J. Ye, Y. Zhao and Y. Qiu, *J. Phys. Chem. C*, 2019, **123**, 20476–20487.
9 A. Banerjee, S. Chakraborty and R. Ahuja, *ACS Appl. Energy Mater.*, 2019, **2**, 6990–6997.
10 J. Ding, S. Du, T. Zhou, Y. Yuan, X. Cheng, L. Jing, Q. Yao, J. Zhang, Q. He, H. Cui, X. Zhan and H. Sun, *J. Phys. Chem. C*, 2019, **123**, 14969–14975.
11 C. Greenland, A. Shnier, S. K. Rajendran, J. A. Smith, O. S. Game, D. Wamwangi, G. A. Turnbull, I. D. W. Samuel, D. G. Billing and D. G. Lidzey, *Adv. Energy Mater.*, 2020, **10**, 1901350.
12 M. Kar and T. Körzdörfer, *J. Chem. Phys.*, 2018, **149**, 214701.
13 C. Kim, T. D. Huan, S. Krishnan and R. Ramprasad, *Sci. Data*, 2017, **4**, 170057.
14 D. Dahliah, G. Brunin, J. George, V.-A. Ha, G.-M. Rignanese and G. Hautier, *Energy Environ. Sci.*, 2021, **14**, 5057–5073.
15 S. Kim, J. A. Márquez, T. Unold and A. Walsh, *Energy Environ. Sci.*, 2020, **13**, 1481–1491.
16 A. Mannodi-Kanakkithodi and M. K. Y. Chan, *Energy Environ. Sci.*, 2022, **15**, 1930–1949.
17 I. E. Castelli, J. M. García-Lastra, K. S. Thygesen and K. W. Jacobsen, *APL Mater.*, 2014, **2**, 081514.
18 H. Park, R. Mall, F. H. Alharbi, S. Sanvito, N. Tabet, H. Bensmail and F. El-Mellouhi, *Phys. Chem. Chem. Phys.*, 2019, **21**, 1078–1088.
19 W. Pu, W. Xiao, J.-W. Wang, X.-W. Li and L. Wang, *Mater. Des.*, 2021, **198**, 109387.
20 J. C. Stanley, F. Mayr and A. Gagliardi, *Adv. Theory Simul.*, 2020, **3**, 1900178.
21 B. D. Lee, W. B. Park, J.-W. Lee, M. Kim, M. Pyo and K.-S. Sohn, *Chem. Mater.*, 2021, **33**, 782–798.
22 J. Yang and A. Mannodi-Kanakkithodi, *MRS Bull.*, 2022, **47**, 940–948.
23 Z. Jiang, Y. Nahas, B. Xu, S. Prosandeev, D. Wang and L. Bellaiche, *J. Phys.: Condens. Matter*, 2016, **28**, 475901.
24 G. Kresse and J. Furthmüller, *Phys. Rev. B*, 1996, **54**, 11169–11186.
25 G. Kresse and J. Hafner, *Phys. Rev. B*, 1993, **47**, 558–561.
26 G. Kresse and J. Furthmüller, *Comput. Mater. Sci.*, 1996, **6**, 15–50.
27 G. Kresse and D. Joubert, *Phys. Rev. B*, 1999, **59**, 1758–1775.
28 G. Kresse and J. Hafner, *J. Phys.: Condens. Matter*, 1994, **6**, 8245–8257.
29 J. P. Perdew, K. Burke and M. Ernzerhof, *Phys. Rev. Lett.*, 1996, **77**, 3865–3868.

30 J. Heyd, G. E. Scuseria and M. Ernzerhof, *J. Chem. Phys.*, 2003, **118**, 8207–8215.
31 Y. Hinuma, G. Pizzi, Y. Kumagai, F. Oba and I. Tanaka, *Comput. Mater. Sci.*, 2017, **128**, 140–184.
32 A. M. Ganose, A. J. Jackson and D. O. Scanlon, *J. Open Source Softw.*, 2018, **3**, 717.
33 S. Steiner, S. Khmelevskyi, M. Marsmann and G. Kresse, *Phys. Rev. B*, 2016, **93**, 224425.
34 L. Yu and A. Zunger, *Phys. Rev. Lett.*, 2012, **108**(6), 068701.
35 L. Williams, *Sl3me – a Python3 Implementation of the Spectroscopic Limited Maximum Efficiency (SLME) Analysis of Solar Absorbers*, **https://github.com/ldwillia/SL3ME**.
36 S. Tao, I. Schmidt, G. Brocks, J. Jiang, I. Tranca, K. Meerholz and S. Olthof, *Nat. Commun.*, 2019, **10**, 2560.
37 O. Almora, D. Baran, G. C. Bazan, C. Berger, C. I. Cabrera, K. R. Catchpole, S. Erten-Ela, F. Guo, J. Hauch, A. W. Y. Ho-Baillie, T. J. Jacobsson, R. A. J. Janssen, T. Kirchartz, N. Kopidakis, Y. Li, M. A. Loi, R. R. Lunt, X. Mathew, M. D. McGehee, J. Min, D. B. Mitzi, M. K. Nazeeruddin, J. Nelson, A. F. Nogueira, U. W. Paetzold, N.-G. Park, B. P. Rand, U. Rau, H. J. Snaith, E. Unger, L. Vaillant-Roca, H.-L. Yip and C. J. Brabec, *Adv. Energy Mater.*, 2021, **11**, 2002774.
38 G. I. Csonka, J. P. Perdew, A. Ruzsinszky, P. H. T. Philipsen, S. Lebègue, J. Paier, O. A. Vydrov and J. G. Ángyán, *Phys. Rev. B*, 2009, **79**, 155107.
39 S. Grimme, J. Antony, S. Ehrlich and H. Krieg, *J. Chem. Phys.*, 2010, **132**, 154104.
40 A. Mannodi-Kanakkithodi, J.-S. Park, N. Jeon, D. H. Cao, D. J. Gosztola, A. B. F. Martinson and M. K. Y. Chan, *Chem. Mater.*, 2019, **31**, 3599–3612.
41 M. Bercx, N. Sarmadian, R. Saniz, B. Partoens and D. Lamoen, *Phys. Chem. Chem. Phys.*, 2016, **18**, 20542–20549.
42 K. Choudhary, M. Bercx, J. Jiang, R. Pachter, D. Lamoen and F. Tavazza, *Chem. Mater.*, 2019, **31**, 5900–5908.
43 J. Benesty, J. Chen, Y. Huang and I. Cohen, in *Pearson Correlation Coefficient*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2009, pp. 1–4.
44 T. Das, G. Di Liberto and G. Pacchioni, *J. Phys. Chem. C*, 2022, **126**, 2184–2198.
45 M. L. Holekevi Chandrappa, Z. Zhu, D. P. Fenning and S. P. Ong, *Chem. Mater.*, 2021, **33**, 4672–4678.
46 X.-G. Zhao, G. M. Dalpian, Z. Wang and A. Zunger, *Phys. Rev. B*, 2020, **101**, 155137.
47 J. Wiktor, U. Rothlisberger and A. Pasquarello, *J. Phys. Chem. Lett.*, 2017, **8**, 5507–5512.
48 S. A. Tolba, K. M. Gameel, B. A. Ali, H. A. Almossalami and N. K. Allam, *Density Functional Calculations*, IntechOpen, Rijeka, 2018, ch. 1.
49 W. Y. X. L. Chi Chen, Y. Zuo and S. P. Ong, *Nat. Comput. Sci.*, 2021, **1**, 46–53.
50 T. Xie and J. C. Grossman, *Phys. Rev. Lett.*, 2018, **120**, 145301.
51 C. Chen, W. Ye, Y. Zuo, C. Zheng and S. P. Ong, *Chem. Mater.*, 2019, **31**, 3564–3572.
52 K. Choudhary and B. DeCost, *npj Comput. Mater.*, 2021, **7**, 185.
53 A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder and K. A. Persson, *APL Mater.*, 2013, **1**, 011002.
54 J. E. Saal, K. Scott, *et al.*, *JOM*, 2013, **65**, 1501–1509.
55 P. C. Jennings, S. Lysgaard and T. Bligaard, *Nat. Comput. Sci.*, 2019, **5**, 46.
56 Y. Pathak, K. S. Juneja, G. Varma, M. Ehara and U. D. Priyakumar, *Phys. Chem. Chem. Phys.*, 2020, **22**, 26935–26943.
57 A. Mannodi-Kanakkithodi, M. K. Chan, J. Yang and P. Manganaris, *High-Throughput DFT Dataset of Halide Perovskite Alloys*, 2022, **https://petreldata.net/mdf/detail/abx3_perovs_alloys_v1.1**.