A High-Throughput Computational Dataset of Halide Perovskite Alloys[†]

Jiaqi Yang[a], Panayotis Manganaris[a], and Arun Mannodi-Kanakkithodi[a]

Abstract

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 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) based dataset of 495 ABX_3 halide Perovskite compounds, with various atomic and molecular species considered at A, B and X sites, and different amounts of mixing considered at each site generated using the Special Quasirandom Structures (SQS) algorithm for alloys. We perform GGA-PBE calculations on pseudo-cubic Perovskite structures to determine their lattice constants, stability in terms of formation and decomposition energies, electronic band gaps, and properties extracted from optical absorption spectra. To elucidate the importance of the level of theory used, we further perform 299 calculations using the more expensive HSE06 functional and determine lattice constant, stability and band gap, and compare PBE and HSE06 properties with some experimentally measured results. Trends in the datasets are unraveled in terms of the effects of mixing at different sites, the composition in terms of specific atomic or molecular species, and averaged elemental properties of species at different sites. This work presents the most comprehensive DFT perovskite alloy dataset to date and the data, which is open-source, can be exploited to train predictive and optimization models for accelerating the design of completely new compositions that may yield large solar cell efficiencies and improved performance across many optoelectronic applications.

Introduction

The perovskite structure has been widely investigated by material scientists in recent decades for its promising industry applications. A perovskite unit cell contains a generalized three-component ABX3 formula unit wherein A and B are cations with different oxidation states while X is an anion. The symbolic 3D perovskite structure is a network of BX6 octahedra which are held in place by large A-site cations. This unique structure makes the crystal's electronic properties incredibly tunable. Numerous research efforts are devoted to halide perovskite (HaP), specifically, as a promising solar cell absorption material[1–4]. ABX3 halide perovskite X-sites anions consist of 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. The A-site cation can be so large that either inorganic (e.g Cs, K, Rb) elements or organic ligands (e.g Methylammonium (MA) and Formamidinium (FA)). The most commonly studied hybrid organic-inorganic halide perovskite, MAPbI3 and FAPbI3, have relatively recently demonstrated extraordinary power conversion efficiency (PCE) between 20% and 25% when used as absorbers in single or multi junction solar cells[5,6]. This is a five fold improvement over prior art and shows off the most attractive feature of halide perovskites, their unique tunability. Experimentally, it has been shown that a perovskite structure is considered stable if A-site, B-site, and X-site ionic radii satisfy the well-known tolerance (t) and octahedral (μ) relations[7]. Even under these restrictions, the configuration space of HaP structures, alloy ratios, alloy ordering, and defects still poses a highly multidimensional optimization problem. Generally speaking, halide perovskite design can be divided into several aspects.

1. Compounds: The composition of halide perovskites is the most critical aspect of PV performance. The recommended halide perovskites as PV absorber usually contains 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 the X-site often with little Cl. As the researchers expanding the search into complex alloys, more A-site organic ligands and Group IV, Group II, or transition elements, which replace Pb at B site for environmental purpose, are discovered.[8–11] The mixing in A site improves the general stability to degradation, while B site and X site mixing are modifying the properties of band gaps and optical absorption.

2. Structure and phases: Besides the well-known cubic halide perovskites, ABX3 perovskites can also exist in tetragonal, orthorhombic, or hexagonal phases[12]. The structure can further change to corner-shared octahe-

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

[†] Electronic Supplementary Information (ESI) available: `https://github.com/PanayotisManganaris/manuscript--PIP-data-manifest` See DOI: 00.0000/00000000.

dral networks, discovered from evolutionary or minima hopping algorithms[13]. 2D layered halide perovskites are also a hot topic and widely researched phase for halide perovskites.

3. Defects: Vacancies, self-interstitials, or impurities may be spontaneously present within HaPs, or intentionally introduced to change the equilibrium conductivity[14,15]. It's essential to halide perovskites to study the effects of point defects on electronic properties and structural stability.

As the huge number of halide perovskites compounds, it rises challenges for experimentalists to screen over all possibilities. However, first principles simulation, such as density functional theory (DFT), have been systematically performed to model each of the factors mentioned above. More importantly, the expense of DFT modeling is much more acceptable when searching for new promising candidates in such boundaryless space. Recently, DFT simulations are known to be reliably applied for modeling structural information, heat of formation or decomposition, band gaps, optical absorption spectra, and defect formation energies of a variety of HaPs, with mixed accuracy compared to experiments.[2,15] DFT simulation also successfully screening a large set of hybrid and inorganic ABX3 halide perovskites. Castelli *et al.*[16] present a halide perovskite dataset of 240 hybrid and inorganic perovskites, with Cs, MA, and FA as A-site cations, Pb and Sn at the B-site, and some mix of Cl, Br, and I at the X-site, in the cubic, tetragonal, and two types of orthorhombic phases. This work investigates the formation of mixed halide compounds and trends of band gaps, revealing the correct combination of A cation, B cation, perovskite phase, and X-site mixing which yields a PV-suitable band gap and good stability.

The limitation of DFT screening is the expensive cost for suitable accurate level of theory for a huge set of data set. To better explore the perovskite space, coupling DFT calculations with state-of-the-art machine learning (ML) or artificial intelligence (AI) techniques is a new direction for high throughput simulation on halide perovskite screening. Park *et al.*[17] employed a DFT+ML approach to investigate the inherent (dis)similarity of various chalcogenide and halide perovskites. They used GGA-PBE calculations on 120 ABX3 compounds to generate the dataset, and then training random forest regression model with t, μ, and tabulated elemental properties as inputs were able to predict band gaps and octahedral distortion, and reveal the descriptors of most importance to each property. A reliable DFT data set with large numbers of mixed halide perovskites will serve as great training data for ML models and accelerate the exploring of new halide perovskite compounds.

We report a synthetic dataset collected for 495 chemically distinct, pseudo cubic Halide Perovskites. This dataset builds on that of 229 samples which formed the foundation of the prior work by Mannodi-Kanakkithodi and Chan[15]. The density functional theory (DFT) computed properties we collected and the levels of theory used are discussed in Methodology. The relatively large size of this dataset is intended to provide an initial sampling suitable for guiding exploration of the alloy space. In this dataset, structures are constantly pseudo-cubic, instead we focus on studying the effects of sample composition and combinatorial alloying possibilities on the computed properties.

This exploration is focused on analyzing data features derived from the perovskite compositions. Composition vectors are inherently bounded – they have a maximum magnitude – and they are strongly correlate with each other in that they sweep through their bounded domain according to the design of the experiment. There is much difficulty in "unfolding" this hyper-geometric volume in such a way as to reveal patterns in composition that relate to properties targeted for prediction.

In any data analysis, the objective is to extract patterns which directly aid the training of a predictive model. In this interest, a relatively small number of unique, uncorrelated, and "influential"[18] features are desirable. To this end, we report a variety of feature engineering pipelines with an emphasis on dimensionality reduction.

The ubiquitous principal component analysis (PCA) has been employed for materials discovery with success. In one prominent instance, under a very different dataset, it aided with identifying patterns relating superconductor transition temperature with valency in ceramic oxides[19]. here, pca simply identifies 3 out of 5 variables which dominate the variance in the data, thereby aiding the formulation of a physically meaningful description of the phenomenon.

The method of PCA is the Singular Value Decomposition, a unitary transform which generalizes the familiar eigen decomposition. Conceptually, PCA will "rotate" the N data points in M-dimensional space until their widest k-dimensional cross section is found. Naturally, the first k principal components may be chosen by the user, but they are determined in an entirely unsupervised, mathematical manner. So long as $k < M$ they represent a lossy, lower-dimensional projection of the data. Even better, they inherently gather the major variance into a minimal number of orthogonal axes. This makes PCA a robust and powerful feature engineering technique for (1) constructing uncorrelated features[18] and (2) de-noising data.

For this, the study by Rajan *et al.*[19] is noteworthy because it uses the second and third principal components in a 5 dimensional space to formulate its model. This clearly demonstrates that meaningful patterns can emerge even in PCA cross sections that capture less overall variance. This is something we are attentive to in our "round" PCA space with many similarly potent principal components (see figure).

Considerably more complex data projections are also examined. Two relatively novel dimensionality reduction techniques, namely t-distributed stochastic neighbor embedding (t-SNE)[20] and Uniform manifold approximation and projection (UMAP)[21] are employed mainly for the purpose of algorithmically clustering feature spaces. This is because these methods are non-parametric, as such they transform a data space rather than a data point. This makes it more difficult to use them as a feature engineering method, but it is not impossible. An attempt at enabling UMAP to be used in this way exists as an implementation of "parametric UMAP" by McInnes *et al.*[22] which trains a neural network to learn a mapping between the data space and manifold projection.

We do see the older t-SNE method clusters HaP decomposition energies effectively on the embedding space of composition vectors. This quality is examined in multiple ways, but we try to recreate it with parametric UMAP so as to create a viable feature engineering pipeline utilizing the transformation.

Methodology

Building Perovskite Dataset

The dataset we report is based on standard cubic phase ABX_3 Perovskite structures originally obtained from Computational Materials Repository (CMR)[23]. Fourteen common Perovskite constituents are selected to form our Halide Perovskite composition space (Figure 1). Five constituents including MA and FA cations represent the possible A-site occupants. Six metals represent the possible B-site occupants. Three halides represent the possible X-site occupants. In total, these component vectors form a constrained 14 dimensional space (Figure 2) within which all Perovskite compounds consisting of the elements in Figure 1 (a) must exist.

The pure (non-alloyed) possibilities are exhaustively sampled using $5*6*3 = 90$ Perovskites. Based on these pure Perovskite structures, we mix candidates for A, B, and X sites systematically. The alloy space sees combinatorial scaling and must be sparsely sampled (Figure **??**).

To generate site-mixing structures with high representativeness, special quasi-random structures (SQS)[24] are applied. The SQS method is to build a special periodic structure and make the first nearest-neighbor shells as similar to the target random alloy as possible. The SQS can be considered the best possible periodic unit cell representing a given random alloy.

Each computational run is performed using a 2x2x2 supercell, this allows A and B site doping to be modeled in discrete 1/8th fractions of the total site occupancy, and it allows X site doping to be modeled in 1/24th fractions. At these mixing levels, it is appropriate to call all these Perovskites alloys.

Following this procedure, 126 A-site mixing samples, 151 B-site mixing samples and 127 X-site mixing samples are generated. All resulting structures are optimized using a DFT variable-cell relaxation under Perdew-Burke-Ernzerhof (PBE). These same initial structures also underwent a full Heyd-Scuseria-Ernzerhof (HSE) relaxation to help ensure the validity of the PBE relaxations, however this was only computationally tractable for 299 samples.

Calculation Details

DFT calculations are performed with VASP version 6.2. The projector augmented wave potentials were used as pseudopotentials. The generalized gradient approximation of PBE and the hybrid HSE06 (α=0.25 and ω=0.2) functionals are used as exchange-correlation energy. The energy cutoff for the plane-wave basis is set to 500 eV. For PBE the Brillouin zone was sampled by 6x6x6 reciprocal mesh using the Monkhorst-Pack k-point mesh. For HSE the Brillouin zone was sampled by 2x2x2 reciprocal mesh using the Monkhorst-Pack k-point mesh. The structural force convergence threshold is set to be 0.005 eV/Å.

Discussion of DFT Computed Properties

Decomposition Energy

The decomposition energy indicates the stability of a compound. To calculate the decomposition energy for an ABX3 Perovskite, we assume it will decompose to two phases, AX and BX2. Using DFT calculations, we can get the optimized energy of a Perovskite and that of its constituent phases. The decomposition energy is calculated using equation (1). This calculation is performed separately for each level of theory.

$$E_{decomp} = E_{opt}(ABX_3) - E_{opt}(AX) - E_{opt}(BX2) \qquad (1)$$

SLME calculations

The SLME metric developed by Yu and Zunger[25] is used as the primary criterion screening perovskites for their photovoltaic merits. The SLME value is computed considering a 5μm absorption layer for every Perovskite according to equation (**??**).

./figs/outline.png

Fig. 1 (a) Chemical space of ABX_3 perovskites. (b) Number of samples representing each kind of primary alloy. (c) Detailed outline.

./domains/variability_of_composition_vectors.png

Fig. 2 Plots showing number of Perovskites representing a constituent at a certain atomic fraction of a complete Perovskite.

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

Where $\alpha(E)$ is the calculated absorption coefficient, L is the thickness of the absorber.

This calculation is performed using Logan William's SL3ME[26]. Based on absorption spectra obtained from from VASP. This calculation is performed separately for the PBE band gaps and the HSE band gaps, resulting in two synthetic efficiencies for each record.

Results and discussion

Visualization of DFT Data

Lattice constant PBE vs HSE

Fig 3 presents the lattice parameter comparison of PBE calculation and HSE calculations. Full relaxations at both levels of theory mostly agree on the lattice parameters. At least, any deviation appears to be very linearly explained. This suggests the accuracy of PBE relaxation is enough to optimize most Perovskite samples.

Note this also served as a reasonable validation for our results. A few samples did have significantly differing lattice parameters. This prompted checking the optimized structures. We found those Perovskite structures were substantially deformed and no longer had obvious octahedral structures. Thus, we exclude these outliers from any analysis concerned with the dominant pseudo cubic structures which are the focus of this report.

./expval/pbe_v_hse_LC.png

Fig. 3 Comparing lattice constants obtain by full PBE and HSE relax calculations

Comparing synthetic to physical data

./expval/pbe_v_exp_LC.png

Fig. 4 Comparison of PBE computed pseudo-cubic lattice constants with crystallographic measures for lattice constants

./expval/hse_v_exp_LC.png

Fig. 5 Comparison of HSE computed pseudo-cubic lattice constants with crystallographic measures for lattice constants

The physical data used in the Lattice Constant Comparisons is collected from the works of Briones *et al.*[27], Jiang *et al.*[28], Chen *et al.*[29].

The synthetic lattice constants do mostly agree with experiment. The PBE lattice constants are better than the HSE measures 1.

It should be noted that the physical $FASnI_3$ measure reported by Chen *et al.*[29] is clearly an orthorhombic phase. The validating lattice constant here is obtained by averaging lattice parameters. However, per the Deviation from Cubicity metric, this phase is still approximately cubic, with angles near enough to 90 degrees that we consider the data point valid. However, this does explain the disagreement in the parity plots.

Table 1 Root Mean Squared Error

HSE v EXP	PBE v EXP

Decomposition energy vs band gap PBE and HSE

Fig 2(b) is showing the PBE band gap compared to the PBE decomposition energy. It presents the diversity of our Perovskite dataset. The data sets cover a large range of band gap and decomposition energy. For example, we have Perovskite with low decomposition energy (good stability) and suitable band gap value (between 1 eV to 2.5 eV for PBE calculations). And we can also find samples with low stability and large band gap. The distribution of band gap and decomposition energy shows a great diversity of all Perovskite samples and indicates that our data set can statistically represent a sufficient Perovskite space. Fig 2(c) shows the plot of PBE band gaps and HSE decomposition energy. Since we applied HSE calculations for part of the samples, it shows some grouping on low decomposition energy. Compared to PBE plot, more data should be added in

./figs/screening_properties.png

Fig. 6 DFT Results: PBE and HSE properties; lattice constants, decomposition energies, band gaps, and converter efficiencies.

high decomposition energy region and suitable band gap region.

Spectroscopic Limited Maximum Efficiency (SLME) vs PBE Band gap

Fig 2(d) presents the SLME values related to the PBE band gap. SLME is a proven proxy for photovoltaic performance[25]. SLME measures the absorption efficiency of light for the Perovskite. As Fig 2(d) showing, a peak around 1.5 eV is obvious. The peak indicates that these samples with 1.5 eV PBE band gap will also have best absorption efficiency as photovoltaic materials. As the band gap increases, the SLME value decreases and eventually goes to zero due to the high band gap values.

Pearson Correlation Results

No target is adequately explained by a single composition or composition derived axis. But there are helpful relations that aid understanding the significance of key composition derived descriptors.

A Pearson correlation map is produced to check for strong relations. Those that exist, when plotted in detail show some trending, but always with extensive variability. Evidently, an accurate model will have to be formed on a multidimensional domain.

PCA

However, in a homogeneous domain of the sort outlined by the ABX3 perovskite, pca is most effective only for identifying the major topology of the data distribution. While this tends to involve uninformative biases in a regression, it is still useful for revealing global trends in the explored space.

however, these variables are unsurprising and may not as strongly explain the properties targeted for statistical modeling

PCA is a method of projecting high dimensional data onto a plane defined by the two linear combinations of axes that explain as much of the variance as possible.

At this point it is readily apparent that this dataset is highly topological. The data exists on a mostly bounded domain in high dimensions, so there is some geometry the features constitute.

Our models will prefer to use this this geometric structure in their explanation for why Perovskite properties vary, this can be useful for accuracy, it can also be a bias-inducing hindrance.

t-SNE visualization of stability clustering

Screening process of all samples in database

The synthetic Perovskite data is collected for sampling variety and without regard for each compound's viability. In

./expval/PBE_v_comp_pearson2.png

Fig. 7 Pearson linear correlation coefficients between 14 composition variables, and 4 PBE computed properties

./expval/PBE_v_site_prop_pearson.png

Fig. 8 Pearson linear correlation coefficients between 36 elemental descriptors and 4 PBE computed properties

./multivar/comp_ratio_projection_annot.png

Fig. 9 PCA

./multivar/tsne_comp_DecoE_clusters.png

Fig. 10 t-SNE

order to extract potentially high-performing candidates for synthesis, we screen all samples according to certain constraints, thereby obtaining some Perovskites worth examining with higher Level of Theory DFT calculations and future physical experiments by brute force. This dataset can only confidently claim to sample cubic phase perovskites. Therefore, a promising candidate should have low deformation. Additionally, a lower decomposition energy combined with a maximal SLME value would be ideal for physical testing. Our screening procedure consists of cutting based on a custom Deviation from Cubicity metric, octahedral factor, Goldschmidt tolerance factor, Bartel tolerance factor, decomposition energy, and band gap – acting as less stringent proxy for the SLME spectrum. The following sections will discuss the details of each constraint.

Deviation from Cubicity

For all the compositions we tested in our data set, some of them will have large strain and deformation because of the combination of elements. For these largely deformed samples, they are no longer remain a cubic perovskite structure. In this section, we want to analyst how the structure is apart from the cubic perovskite structure and rule out these largely deformed samples. Firstly, we need to define deviation of cubicity. For each perovskite sample, we measure the difference between b, c lattice parameter against a lattice, showing in Equation X. If the lattice deviation of cubicity is larger than 10%, we will consider this perovskite is no longer remain a cubic perovskite and exclude it. Similarly, we also need to consider 3 angles, , and , to make sure the angle also remain 90 degrees. We calculated the difference of , and versus 90 degrees, showing in Equation X, and take this as angular deviation of cubicity. We also consider the samples that have more than 5% of angular deviation of cubicity are not cubic perovskites and exclude them. Fig 7 (d) shows the screening results for angle deviation of cubicity. SI Fig X shows the screening results for b, c lattice and , angle. Most of the samples with non-cubic perovskites structures are organic-inorganic hybrid perovskites. A large part of the excluded samples are A-mixing hybrid perovskites. It indicates that organic ligands in A site sometimes increase the lattice along some direction and make the perovskite deformed.

Octahedral factor, tolerance factor, and Bartel tolerance factor

The stability of Perovskite can be predicted by using the atom radius of all components. There are 3 types of factors are usually considered, Octahedral factor, tolerance factor, and Bartel tolerance factor. The formula of these three factors are shown in equation XXX. In our screening process, we set the criteria for Octahedral factor as 0.442 – 0.895. The criteria for tolerance factor is set to be 0.813 – 1.107. The criteria for Bartel tolerance factor is set to be less than 4.18. Fig 7 (a) shows the Octahedral factor versus decomposition energy plot. Fig 7 (b) shows the tolerance factor versus decomposition energy plot. The tolerance factor shows a trend that as the tolerance factor increases, the decomposition energy decreases. Fig 7 (c) shows the Bartel tolerance factor versus decomposition energy plot. Within the criteria of Bartel tolerance factor, the decomposition energy is rapidly decreased.

Screening Results

The table XXX shows all 41 promising candidates we screened from our dataset by using the cutoff criteria mentioned above. FigXXX shows the element distribution of all decent samples. It indicates that MA and FA are the most common A site choices for a Perovskite with great stability and suitable band gap and photovoltaic properties. For inorganic A sites, Cs also contributes a large portion. In B site element selection, Pb and Sn are the most selected than other B site elements. For X site halide elements, Br is appears the most, and I is the second. It tells us that Br and I are preferred in a stable Perovskite structure. In figure XXX, the mixing distribution indicates that X-site mixing is more favorable to generate a stable Perovskite with required properties. Tuning in A sites and B sites can also lead to a promising results but more risky. The FigXXX and SI FIGXXX shows the element fraction for each of the screened samples. It indicates if certain element prefer to be mixing or non-mixing. For example, Cl is preferred to be pure in X site than mixing with other elements. Also there is some elements that only appears in pure case, for instance Sr at B site. We can also learn from the result that elements, like K, Cs, Cl and Ba, are more preferred to have small fraction if they are mixing with other elements. Popular choice, like Pb and Sn, are very flexible in element fraction. These elements can have a variety of possible fraction in promising Perovskite samples. The screening results is showing the trends of how the combinations of elements are selected and what mixing features are preferred in decent Perovskite structures. It will also provide great help on design new Perovskite compounds and tell us what kinds of compounds we should put in to test next.

Perspective and Future Work

The high-throughput DFT Perovskite dataset is containing huge amount of information of perovskites compounds, structure and properties. It leads us to perform screening, clustering and other process so that lots of trends are revealed. However, there are still many improvements we

./figs/screening_factor.png

Fig. 11 Screening results of Decomposition Energy for (a) octahedral factor (b) Tolerance Factor (c) Bartel Tolerance Factor (d) Deviation of Cubicity.

./figs/screening_pie.png

Fig. 12 49 Screened perovskites analysis, (a) pie chart of element distribution, (b) pie chart of site mixing

Fig. 13 Plots of element fraction for A, B, X sites of all screened perovskites.

are eager to accomplish in the future.

Firstly, the dataset are generally constructed based on cubic perovskites structures. But not all perovskites are remaining cubic as their most stable phases. In the future, we are going to expand our dataset that will include high-throughput DFT calculations for multiple phases of perovskites, for example tetragonal, orthorhombic, and hexagonal phases. This will tells us which phase we should choose for each compound and provide more information for analysis.

Also the level of DFT functional should also be included in the future work. Important corrections, like Van der Waals forces, spin orbital coupling, should also be considered for different compound. We are going to expand the dataset with information of multiple level of DFT functional and corrections, which will eventually lead us to decide the best suitable functional for each compounds and get more accurate prediction.

Structural information is another critical part for tuning the properties of Perovskite. Octahedral distortion and rotation is the main part that causes the change in Perovskite structures. We have ongoing work to reveal the relationship between octahedral information and properties. These part of information will also be include in the future dataset.

The design of this dataset is uniquely suited to the exploration of alloying effects on Perovskite properties. The combinatorial space of possible alloys has been sparsely but systematically sampled along four primary alloy schemes. This sample space affords the opportunity for a QM/ML surrogate model to form the basis of an active learning strategy which can begin selecting potentially high performing multi-site alloy candidates based on the current sample set.

For the calculation of future data points obtained by a surrogate optimizer, we will follow a strategy of performing full structural optimization at a PBE level of theory unless circumstances demand otherwise. This is justified by Figure 3 and the section Comparing synthetic to physical data

Also, we will explore methods for combining the insights of the PBE and HSE datasets in training surrogates, e.g. in section SLME calculations the two SLME spectra could be used together to converge on a physically accurate PCE value.

We anticipate our principal challenge will be in extracting useful predictor variables from the composition information. The basic feature sets examined here are highly correlated, but nonetheless show promise both as a basic screening criterion, and as good classifier features under t-SNE transformation.

Modeling pipelines capable of predicting Perovskite decomposition energy will likely be very achievable using a transduction and invertible equivalent of the t-SNE algorithm, potentially SONG.

For this reason, kernel learning methods appear to be particularly promising for high speed optimization of the space.

Conclusions

In this work, we present a DFT calculated perovskite data set with 495 pseudo-cubic perovskite samples with A-site, B-site and X-site mixing. The mixing arrangements are built using SQS methods. All perovskite structures are optimized by GGA-PBE calculations and furthermore, 299 samples are also relaxed by HSE06 functional. Validating the efficacy of both relaxation methods against experimentally determined structures shows that, conveniently, PBE-GGA is more effective for determining the pseudo-cubic lattice constant. Both PBE and HSE functionals are not very effective at predicting band gaps. In addition, we analyze trends in stability and various electronic properties and optical properties by identifying linear correlations and utilizing clustering methods. These analyses show promise for using compositional properties alone to predict perovskite stability by predicting decomposition energy. Finally, screening on derived factors and predicted targets is also used to filter the 495 perovskite data set down to some promising perovskite candidates with high stability and suitable electronic and optical properties. This data set will be used in the future to train statistical models and inform inverse design pipelines.

Conflicts of Interest

There are no conflicts to declare.

Acknowledgements

Extensive discussions with and scientific feedback from UC San Diego researchers David Fenning and Rishi Kumar and Argonne National Lab scientist Maria Chan 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 Argonne National Laboratory, and the RCAC clusters at Purdue.

References

1 M. I. H. Ansari, A. Qurashi and M. K. Nazeeruddin, *Journal of Photochemistry and Photobiology C: Photochemistry Reviews*, 2018, **35**, 1–24.

2 W.-J. Yin, J.-H. Yang, J. Kang, Y. Yan and S.-H. Wei, *Journal of Materials Chemistry A*, 2015, **3**, 8926–8942.

3 J. S. Manser, J. A. Christians and P. V. Kamat, *Chemical Reviews*, 2016, **116**, 12956–13008.

4 T. M. Brenner, D. A. Egger, L. Kronik, G. Hodes and D. Cahen, *Nature Reviews Materials*, 2016, **1**, 15007.

5 P. Cui, D. Wei, J. Ji, H. Huang, E. Jia, S. Dou, T. Wang, W. Wang and M. Li, *Nature Energy*, 2019, **4**, 150–159.

6 M. Jeong, W. Choi In, M. Go Eun, Y. Cho, M. Kim, B. Lee, S. Jeong, Y. Jo, W. Choi Hye, J. Lee, J.-H. Bae, K. Kwak Sang, S. Kim 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, *Science Advances*, 2019, **5**, eaav0693.

8 S. Zhu, J. Ye, Y. Zhao and Y. Qiu, *The Journal of Physical Chemistry C*, 2019, **123**, 20476–20487.

9 A. Banerjee, S. Chakraborty and R. Ahuja, *ACS Applied Energy Materials*, 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, *The Journal of Physical Chemistry 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, *Advanced Energy Materials*, 2020, **10**, 1901350.

12 M. Kar and T. Körzdörfer, *The Journal of Chemical Physics*, 2018, **149**, 214701.

13 C. Kim, T. D. Huan, S. Krishnan and R. Ramprasad, *Scientific Data*, 2017, **4**, 170057.

14 A. Mannodi-Kanakkithodi, J.-S. Park, N. Jeon, D. H. Cao, D. J. Gosztola, A. B. F. Martinson and M. K. Y. Chan, *Chemistry of Materials*, 2019, **31**, 3599–3612.

15 A. Mannodi-Kanakkithodi and M. K. Y. Chan, *Energy Environ. Sci.*, 2022, **15**, 1930–1949.

16 I. E. Castelli, J. M. García-Lastra, K. S. Thygesen and K. W. Jacobsen, *APL Materials*, 2014, **2**, 081514.

17 H. Park, R. Mall, F. H. Alharbi, S. Sanvito, N. Tabet, H. Bensmail and F. El-Mellouhi, *Physical Chemistry Chemical Physics*, 2019, **21**, 1078–1088.

18 M. W. Mahoney and P. Drineas, *Proceedings of the National Academy of Sciences*, 2009, **106**, 697–702.

19 K. Rajan, C. Suh and P. F. Mendez, *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 2009, **1**, 361–371.

20 L. van der Maaten and G. E. Hinton, *Journal of Machine Learning Research*, 2008, **9**, 2579–2605.

21 L. McInnes, J. Healy and J. Melville, *ArXiv e-prints*, 2018.

22 L. McInnes, J. Healy, N. Saul and L. Grossberger, *Python UMAP: Uniform Manifold Approximation and Projection*, 2018.

23 G. Pilania, A. Mannodi-Kanakkithodi, B. P. Uberuaga, R. Ramprasad, J. E. Gubernatis and T. Lookman, *Scien-*

tific Reports, 2016, **6**, year.

24 Z. Jiang, Y. Nahas, B. Xu, S. Prosandeev, D. Wang and L. Bellaiche, *Journal of Physics: Condensed Matter*, 2016, **28**, 475901.

25 L. Yu and A. Zunger, *Physical Review Letters*, 2012, **108**, year.

26 L. Williams, *Sl3me – a Python3 Implementation of the Spectroscopic Limited Maximum Efficiency (SLME) Analysis of Solar Absorbers*, `https://github.com/ldwillia/SL3ME`.

27 J. Briones, M. C. Guinto and C. M. Pelicano, *Materials Letters*, 2021, **298**, 130040.

28 L. Jiang, J. Guo, H. Liu, M. Zhu, X. Zhou, P. Wu and C. Li, *Journal of Physics and Chemistry of Solids*, 2006, **67**, 1531–1536.

29 Q. Chen, N. D. Marco, Y. M. Yang, T.-B. Song, C.-C. Chen, H. Zhao, Z. Hong, H. Zhou and Y. Yang, *Nano Today*, 2015, **10**, 355–396.

Supplemental Material

Glossary

PCA principal component analysis.

QM/ML quantum mechanics machine learning.

SLME spectroscopic limited maximum efficiency.

t-SNE t-distributed stochastic neighbor embedding.

VASP Vienna Ab initio Simulation Package.

Acronyms

DFT density functional theory.

FA Formamidinium.

GGA generalized gradient approximation.

HSE Heyd-Scuseria-Ernzerhof.

MA Methylammonium.

PAW projector augmented wave.

PBE Perdew-Burke-Ernzerhof.

PCE power conversion efficiency.

SQS special quasi-random structures.

./expval/HSE_v_comp_pearson2.png

Fig. 14 Pearson linear correlation coefficients between 14 composition variables, and 4 HSE computed properties

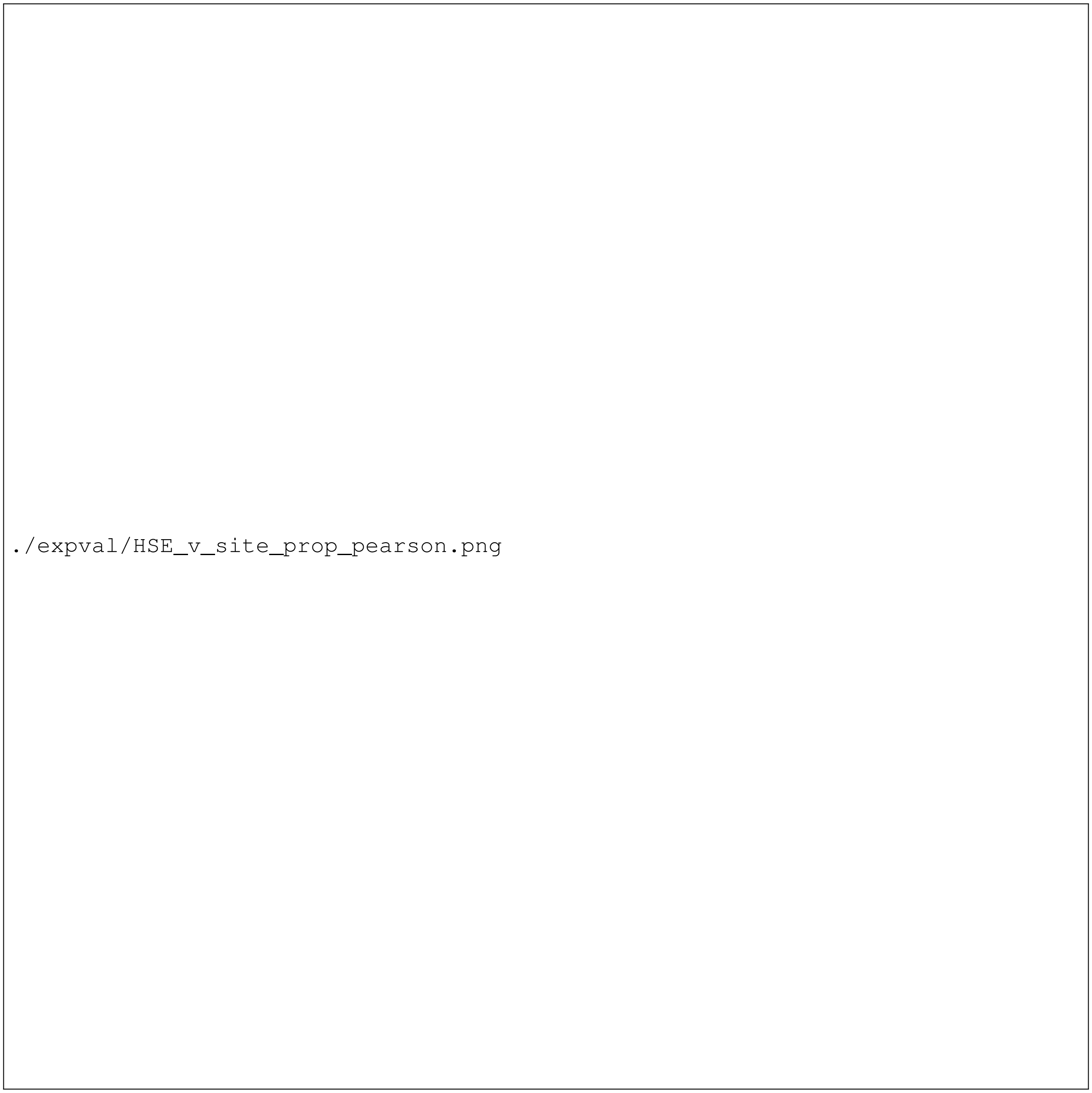

Fig. 15 Pearson linear correlation coefficients between 36 composition variables, and 4 HSE computed properties

./expval/Element_AFrac.jpg

Fig. 16 Element Fraction for A site elements among all screened perovskites

./expval/Element_BFrac.jpg

Fig. 17 Element Fraction for B site elements among all screened perovskites

./expval/Element_XFrac.jpg

Fig. 18 Element Fraction for X site elements among all screened perovskites