MULTI-FIDELITY MACHINE LEARNING FOR PEROVSKITE BAND GAP PREDICTIONS

by

Panayotis T. Manganaris

A Thesis

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Master of Science

School of Materials Engineering

West Lafayette, Indiana

May 2023

THE PURDUE UNIVERSITY GRADUATE SCHOOL STATEMENT OF COMMITTEE APPROVAL

Dr. Arun Mannodi-Kanakkithodi, Chair

School of Materials Engineering

Dr. Alejandro Strachan

School of Materials Engineering

Dr. Kendra Erk

School of Materials Engineering

Approved by:

Pending FORM 9 Approval

To my family, especially my brother Tassos.

ACKNOWLEDGMENTS

I am grateful to Professor Arun Mannodi-Kanakkithodi for his mentorship, support, and knowledge, Professor Erk for her guidance and counsel, and Professor Strachan for his instruction. I also thank my colleague Jiaqi Yang for his collaboration, and Habibur Rahman for his friendship and curiosity. My work was funded in part by my advisor's startup grant F.10023800.05.002 and the Ross Fellowship awarded to me by Purdue University. The Rosen Center of Advanced Computing provided the computational resources needed to conduct simulations, store data, and train models.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF LISTINGS

LIST OF SYMBOLS

γ	photon
J	Current Density
I_{sun}	Light Spectrum Intensity of Sunlight
P	power

ABBREVIATIONS

HaP	halide perovskite
VASP	Vienna Ab initio Simulation Package
SLME	spectroscopic limited maximum efficiency
PCE	power conversion efficiency
DFT	density functional theory
GGA	generalized gradient approximation
PBE	Perdew-Burke-Ernzerhof Functional
HSE06	Heyd-Scuseria-Ernzerhof Functional
PCA	principal component analysis
t-SNE	t-distributed stochastic neighbor embedding
UMAP	uniform manifold approximation and projection
GPR	Gaussian Process Regression
RFR	Random Forest Regression
SISSO	Sure Independence Screening and Sparsifying Operator
SQS	special quasi-random structures
PAW	projector augmented wave
SHAP	Shapley Additive Explaination

NOMENCLATURE

MA	Methylammonium (Cationic Methylamine) $CH_3NH_3{}^+$
FA	Formamidinium (Cationic Formamidine) $CH(NH_2)_2{}^+$

GLOSSARY

Law of Mixing	The rule stating properties of materials of mixed compositions may be predicted by linear interpolation of the properties of similar materials with pure compositions
cardinal mixing	Describes perovskite alloys where no more than one of the A, B, or X sites is occupied by multiple possible constituents
partition	Portion of sample data reserved for a purpose in model development
cross-validation	Method for gathering statistics on the abilities of a model to fit to the parent partition
K-fold split	Data partition divided into K arbitrary groups for use in cross-validation schemes
groupwise K-fold	Data partition divided into K-folds where each fold corresponds to a category label
level of theory	Refers to the rank of a DFT functional in the hierarchy of phenomenological comprehensiveness. A proxy for accuracy.
Materials Project	US Government-led multidisciplinary collaboration founded in 2011 as the Materials Genome Initiative.
machine learning	a science concerned with algorithms which improve their performance with exposure to new data
features	attributes of an observed event or object which might empirically explain the event or object
hyper-parameter	a setting that controls how a learning algorithm works
classical learning	a paradigm of machine learning that is dependent on expert knowledge to extract quality features from samples in a dataset
surrogate model	a representation which attempts to capture as much of the relationship between a domain and a target property as possible
deep learning	a paradigm of machine learning differing from classical learning in that the features of the input data are themselves learned by the algorithm

FAIR	Findable Accessible Interoperable and Reusable Data
multi-task learning	A type of machine learning where an algorithm learns multiple functions simultaneously, while exploiting commonalities and differences between the functions
Spin Orbit Coupling	An additional term intended to account for the increased relevance of quantum angular momentum to electromagnetic response in heavy atoms

ABSTRACT

A wide range of optoelectronic applications demand semiconductors optimized for purpose. My research focused on data-driven identification of ABX_3 Halide perovskite compositions for optimum photovoltaic absorption in solar cells. I trained machine learning models on previously reported datasets of halide perovskite band gaps based on first principles computations performed at different fidelities. Using these, I identified mixtures of candidate constituents at the A, B or X sites of the perovskite supercell which leveraged how mixed perovskite band gaps deviate from the linear interpolations predicted by Vegard's law of mixing to obtain a selection of stable perovskites with band gaps in the ideal range of 1 to 2 eV for visible light spectrum absorption. These models predict the perovskite band gap using the composition and inherent elemental properties as descriptors. Eventually enabling accurate prediction and screening of the much larger chemical space from which the data samples were drawn.

I utilized a recently published density functional theory (DFT) dataset of more than 1300 perovskite band gaps from four different levels of theory, added to an experimental perovskite band gap dataset of ~100 points, to train random forest regression (RFR), Gaussian process regression (GPR), and Sure Independence Screening and Sparsifying Operator (SISSO) regression models, with data fidelity added as one-hot encoded features. I found that RFR yields the best model with a band gap root mean square error of 0.12 eV on the total dataset and 0.15 eV on the experimental points. SISSO provided compound features and functions for direct prediction of band gap, but errors were larger than from RFR and GPR. Additional insights gained from Pearson correlation and Shapley additive explanation (SHAP) analysis of learned descriptors suggest the RFR models performed best because of (a) their focus on identifying and capturing relevant feature interactions and (b) their flexibility to represent nonlinear relationships between such interactions and the band gap. The best model was deployed for predicting experimental band gap of 37785 hypothetical compounds, based on which we identified 1251 stable compounds with band gap predicted to be between 1 and 2 eV at experimental accuracy. Successfully narrowing to about 3% of the screened compositions.

1. INTRODUCTION

Perovskites have historically been materials of great interest for a variety of optoelectronic applications with special interest in the past ten years (see figure 1.1) in their potential as photovoltaic absorbers. As absorbers for solar cells, they offer opportunities to reduce cost and environmental impact as well as increase performance. (Ansari et al., 2018; Brenner et al., 2016; Manser et al., 2016; Yin et al., 2015) A cubic phase perovskite unit cell with general formula ABX_3 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 highly 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. Halide perovskites (HaP), as opposed to the oxide perovskites that have been well researched over the past century are so characterized because their X-site anions are halogens. Their B-site cations may be divalent elements, and the A-site is occupied by large monovalent cations that are either inorganic elements or organic molecules.

The most commonly studied hybrid organic-inorganic HaPs, $MAPbI_3$ 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. (Cui et al., 2019; Jeong et al., 2020) This is a five-fold improvement over the efficiencies of the same compositions first reported in 2009 and demonstrates the most attractive feature of HaPs, their unique tunability. A theoretical cubic perovskite structure is considered stable if the ionic radii of A, B, and X-site species satisfy the well-known tolerance (t) and octahedral (o) factors. (Bartel Christopher et al., 2019) Even accounting for stability constraints, the chemical space of perovskites experiences combinatorial scaling with the number of candidate elements which could be incorporated at each site. This poses a multidimensional optimization problem for which determining the optimal atomic fractions for a particular performance target requires data-driven acceleration. My research focused on the development and application of these design methods in the composition space of halide perovskites. The majority of work presented in this dissertation has been previously published or will be submitted for publication (See CONTRIBUTIONS).

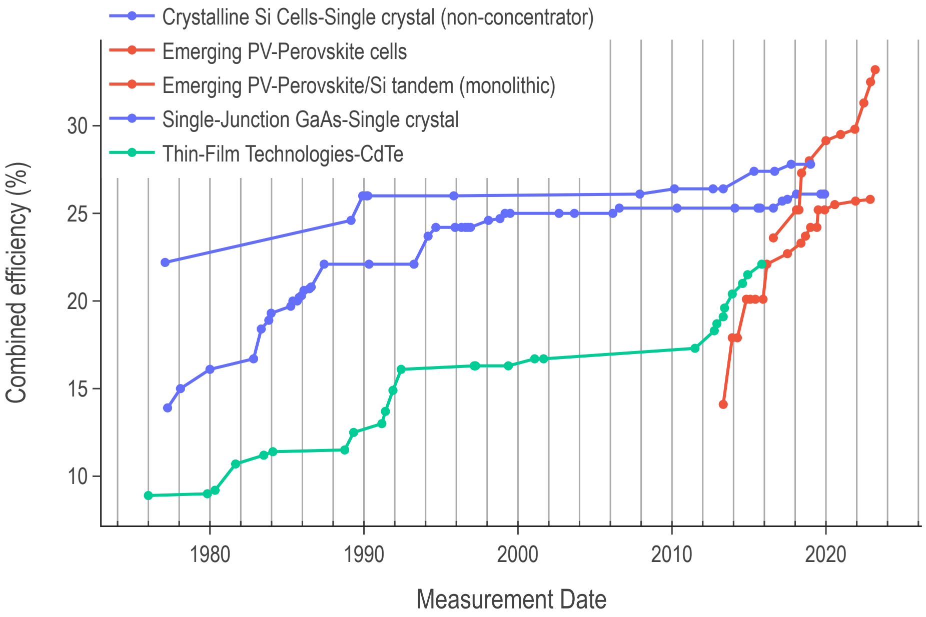

Figure 1.1. Rapid rise in cummulative maximum of HaP PCEs

1.1 Design Goals and Challenges in Perovskite PV Absorbers

Perovskite properties may be tuned in various ways. The introduction of dopants and defects (Dahliah et al., 2021; S. Kim et al., 2020) and the mutation of their cubic structure (Kar & Körzdörfer, 2018; C. Kim et al., 2017) are each promising areas of design. However, the work presented here focuses specifically on the identification of a reasonable number of candidate compositions for future laboratory trials. 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 the X-site often with little Cl. 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. (Banerjee et al., 2019; Ding et al., 2019; Greenland et al., 2020; Zhu et al., 2019) Mixing at A site has been shown to improve formability (Zhang et al., 2019), 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 the creation of high entropy perovskite alloys, is in the promise to obtain dramatically different properties than those of pure compositions. Perovskite properties have demonstrated highly nonlinear responses to changes in composition. It is hoped that exploiting this could lead to possibly eliminating toxic lead, reducing degradation under light exposure, and even improving resistance to adverse environmental conditions while also targeting specific optoelectronic performance markers.

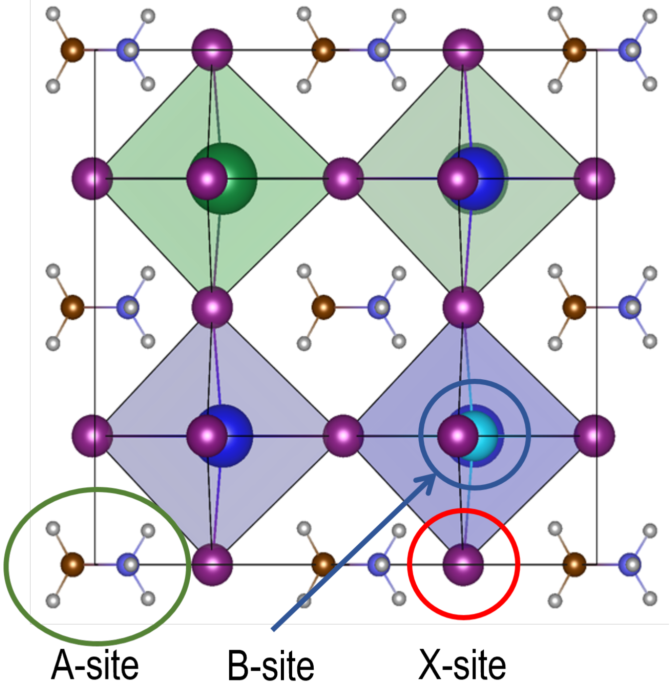

Figure 1.2. $2 \times 2 \times 2$ α-phase supercell with Methylammonium at the A-site

The chemical design space of HaPs is intractably large to effectively screen by physical laboratory methods. The halide perovskite chemical space covered by this dataset was based on fourteen species commonly appearing in study of these materials. The five constituents making up the A-site occupants include three inorganic and two organic cations, namely

$CH_3NH_3^+$ Methylammonium (MA) and $CH(NH_2)_2^+$ Formamidinium (FA). (Dimesso et al., 2016; Yan et al., 2016) Six divalent metals represent the possible B-site occupants and three halogen anions make up the possible X-site occupants. See table 1.1. The total number of distinct compositions possible in a 2 × 2 × 2 supercell is over 207 million. Of these compositions, 37695 contain mixing at only one site and ninety are pure having no mixing at any site. I refer to the combination of these subsets as the "cardinal mixing set" and it equally represents each of the constituent species of interest. See figures 1.3 and 1.4.

First principles density functional theory (DFT) simulations have been systematically performed to study the optoelectronic properties of HaPs as a function of structure, composition, and defects. Recently, DFT simulations have been reliably used to predict structural information, band gaps, optical absorption spectra, and defect formation energies of a variety of HaPs with reasonable accuracy. (Mannodi-Kanakkithodi & Chan, 2022; Yin et al., 2015) An examination of the HaP-related computational literature reveals that there have been numerous medium (~10^2 data points) to large (~10^3 or more data points) DFT datasets reported for HaPs. (Castelli et al., 2014; Kar & Körzdörfer, 2018; Park et al., 2019; Pu et al., 2021) These have been successfully screened to identify promising materials with desired stability and formability as well as PV-suitable band gaps, among other properties.

A clear limitation of High-Throughput (HT) DFT driven screening is the computational expense of applying a suitably advanced level of theory across a very large number of materials. This problem is typically addressed by coupling DFT computations with machine learning (ML) 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 ML-based predictive and classification models, leading to accelerated prediction of lattice constants, formation energies, band gaps, and other important properties. (Lee et al., 2021; Park et al., 2019; Stanley et al., 2020) Such DFT-ML models, once rigorously trained and tested, are deployed for high-throughput screening across massive sample spaces of unknown perovskites. (J. Yang & Mannodi-Kanakkithodi, 2022)

Table 1.1. ABX_3 candidate species per site

A-site	MA	FA	Cs	Rb	K	
B-site	Pb	Sn	Ge	Ba	Sr	Ca
X-site	I	Br	Cl			

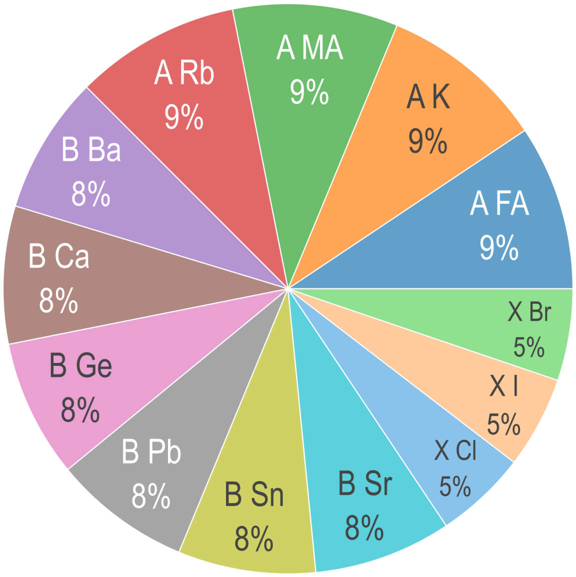

Figure 1.3. The cardinal mixing sample space contains equal fractions of each element

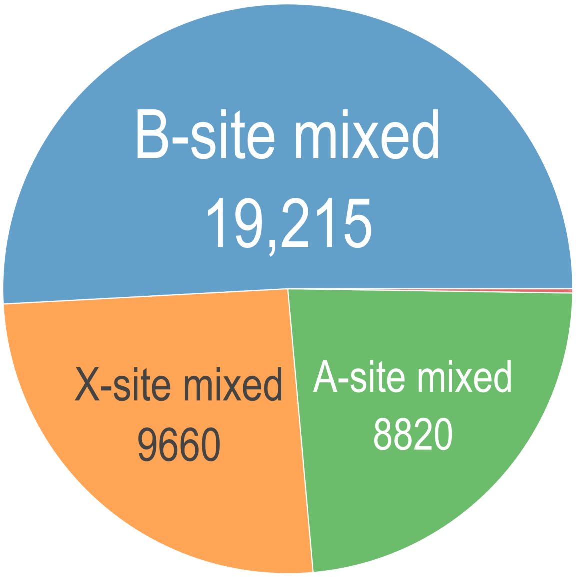

Figure 1.4. The cardinal mixing sample space contains mostly B-site mixed compounds

1.2 Multiple Fidelity Dataset

For my work I used a large DFT dataset collected over the past three years by my advisor and fellow student Jiaqi Yang. This dataset consists of approximately 1300 calculations based on approximately 500 chemically distinct, pseudo-cubic, halide perovskite alloys reported in the Mannodi research group's prior work. (Mannodi-Kanakkithodi & Chan, 2022; J. Yang et al., 2023) There are more calculations than there are distinct compositions because each composition is simulated multiple times. Also, it is supplemented by an additional ~100 points of data aggregated from reputable sources by Almora et al. (2020). In total it contends as one of the largest first principles halide perovskite datasets and represents years of work. The relatively large size of this dataset samples the space of all possible single-site mixed compositions with good coverage. This enables the training of interpolative models in the HaP composition space promising lower and less frequent error than if lesser coverage were used. In this dataset, all perovskite structures are cubic or pseudo-cubic, which aids my focus on investigating effects of composition and alloying on photovoltaic performance.

1.3 Perovskite Formability

In order to simplify the search for viable candidates, some standard empirical rating of perovskite formability are employed. It has been observed that the A-site member must be much larger than its counterpart at the B-site for a Perovskite to be stable. B-site elements are usually large (e.g. Pb, Sn) in Halide Perovskites, which incidentally motivates the use of organic cations at A. (Kieslich et al., 2015) Various tolerance factors have been designed to describe this constraint. The following definitions approximate α phase stability (Bartel Christopher et al., 2019; Yin et al., 2015)

Goldschmidt tolerance

$$1 \approx t = \frac{R_A + R_X}{\sqrt{2} * (R_B + R_X)}$$

Octahedral Factor

$$0.67 \approx o = \frac{r_B}{r_X}$$

Bartel Tolerance

$$4 \gtrsim b = \frac{r_X}{r_B} - \left[1 - \frac{\frac{r_A}{r_B}}{ln(\frac{r_A}{r_B})}\right]$$

2. DENSITY FUNCTIONAL THEORY SIMULATION

All DFT computations were performed using vasp version 6.2 employing projector-augmented-wave (PAW) pseudo-potentials. (Kresse & Furthmüller, 1996a, 1996b; Kresse & Hafner, 1993, 1994; Kresse & Joubert, 1999) Multiple levels of theory (LoT) were used in most computations. Each simulation was conducted on the same set of compositions allowing *at most* single-site mixing of our 14 constituent candidates for 3 sites (table 1.1). Each HaP composition is simulated in a $2 \times 2 \times 2$ supercell, which allows A and B-site mixing to be performed in discrete $1/8^{\text{th}}$ fractions of the total site occupancy, and X-site mixing in $1/24^{\text{th}}$ fractions, though for simplicity, we restrict X-site mixing to fractions of 3x/24. For simulating mixed perovskites, the special quasi-random structures (SQS) method was applied to build periodic structures that make the first nearest-neighbor shells as similar to the target random alloy as possible. (Z. Jiang et al., 2016) The final tally of successfully converged calculations is listed in table 2.1. The Perdew-Burke-Ernzerhof (PBE) functional within the generalized gradient approximation (GGA) as well as the hybrid Heyd-Scuseria-Ernzerhof functional with parameters ($\alpha = 0.25$) and ($\omega = 0.2$) (HSE06) are used for exchange-correlation energy. (Heyd et al., 2003; Perdew et al., 1996) 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 $3 \times 3 \times 3$ for supercells. Using the PBE optimized structure as input, the electronic band structure is calculated along high-symmetry k-points 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. (Ganose et al., 2018; Hinuma et al., 2016) For HSE calculations, geometry optimization was performed using only the Gamma point, and subsequent computations used a reduced $2 \times 2 \times 2$ Monkhorst-Pack mesh. The force convergence threshold is set to be -0.05 eV/Å. Spin-orbit coupling (SOC) is also applied to two flavors of HSE computations using the LORBIT tag and the non-collinear magnetic version of VASP 6.2. (Steiner et al., 2016) Optical absorption spectra from different HSE functionals were obtained by using the difference between the respective PBE and HSE band gap, and shifting the PBE-computed spectrum.

Table 2.1. Sample counts by density functional represented in dataset

	LoT
PBE	492
HSE	297
HSE(SOC)	282
HSE-PBErel(SOC)	244
EXP	90
	1405

2.1 DFT Computed Band Gaps

Four types of electronic band gaps were computed by my advisor and group members using a $2\times2\times2$ Monkhorst-Pack mesh. These four measures $\mathrm{E_{gap}}^{\mathrm{PBE}}$, $\mathrm{E_{gap}}^{\mathrm{HSE}}$, $\mathrm{E_{gap}}^{\mathrm{HSE(SOC)}}$, and $\mathrm{E_{gap}}^{\mathrm{HSE\text{-}PBErel(SOC)}}$ populate the multiple fidelity dataset I model. I aim to accurately predict performance-relevant halide perovskite (HaP) band gaps, which strongly predict photovoltaic performance. (Mannodi-Kanakkithodi et al., 2019) Furthermore, using multi-fidelity modeling, I aim to predict the experimentally measured band gaps of entirely hypothetical compounds, either only ever simulated or not tested at all. This is motivated by a relationship known to exist between the absorption spectra/PCE of the simulated compounds and their band gaps.

2.2 Spectroscopic Limited Maximum Efficiency (SLME)

Introduced by Yu and Zunger (2012), 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 5ţm sample thickness for every perovskite using equations 2.2, 2.2, and 2.2, combining the original SL3ME.py code from Yu and Zunger (2012) with our DFT computed absorption spectra and band gaps.

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

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.

$$J = \mathrm{e}\int_0^\infty a(E)I_{sun}(E)dE - J_0(1 - \mathrm{e}^{\frac{\mathrm{e}V}{kT}})$$

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

To calculate SLME efficiency the current density J, the light spectrum intensity of sunlight I_{sun}, and the power P are all that is needed. Using the DFT computed optical absorption spectrum as well as the magnitude and type (direct or indirect) of band gap as input, SLME is directly calculated using an open-source package. (Williams, 2022) This calculation is performed using all four functionals and compliments the PCE measurements at the experimental fidelity. SLME accounts for more energetic processes than the Shockley-Queisser criterion ($bg \approx 1.3$) allowing for a range of performant band gaps to be identified according to level of theory(Yu & Zunger, 2012, p.1) Experimental data(Almora et al., 2020) broadly agrees with PBE simulation, so the range of 1 to 2 eV. see figure 2.1. Also, notice that even in just the sample dataset, there are candidates with potential to overtake the state of the art absorbers reported by NREL in figure 1.1. This is propitious for the screening I conduct on the 40000 point sample space.

2.3 Improving Property Predictions using HSE06 and Spin-Orbit Coupling

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 reliably compute relaxed structures of both hybrid and purely inorganic HaPs. However, advanced levels of theory such as the HSE06 functional with and without the inclusion of spin orbit coupling (SOC) to account for the relativistic effects of heavy atoms such as Pb, are of paramount importance when it comes to simulating electronic and optical properties.

The data set I used contains a series of ~300 expensive HSE calculations across the 500 sampled compositions. These are intended to yield insight into the effects of full geometry optimization at hybrid levels of theory to those of PBE-optimized structures. Also, the effect of incorporating SOC in the calculation was examined. In review, the sample of 500 band gaps available for training predictors was supplemented by 299 calculations conducted entirely at the HSE level of theory. Furthermore, an additional 282 calculations were performed with HSE in addition to SOC, and 244 calculations were performed by running HSE(SOC) electronic structure calculations on PBE-relaxed structures.

The range of band gaps sampled by each simulation method are similar and are characterized by similar variance. the descriptive statistics of each greatly exceeds those of the

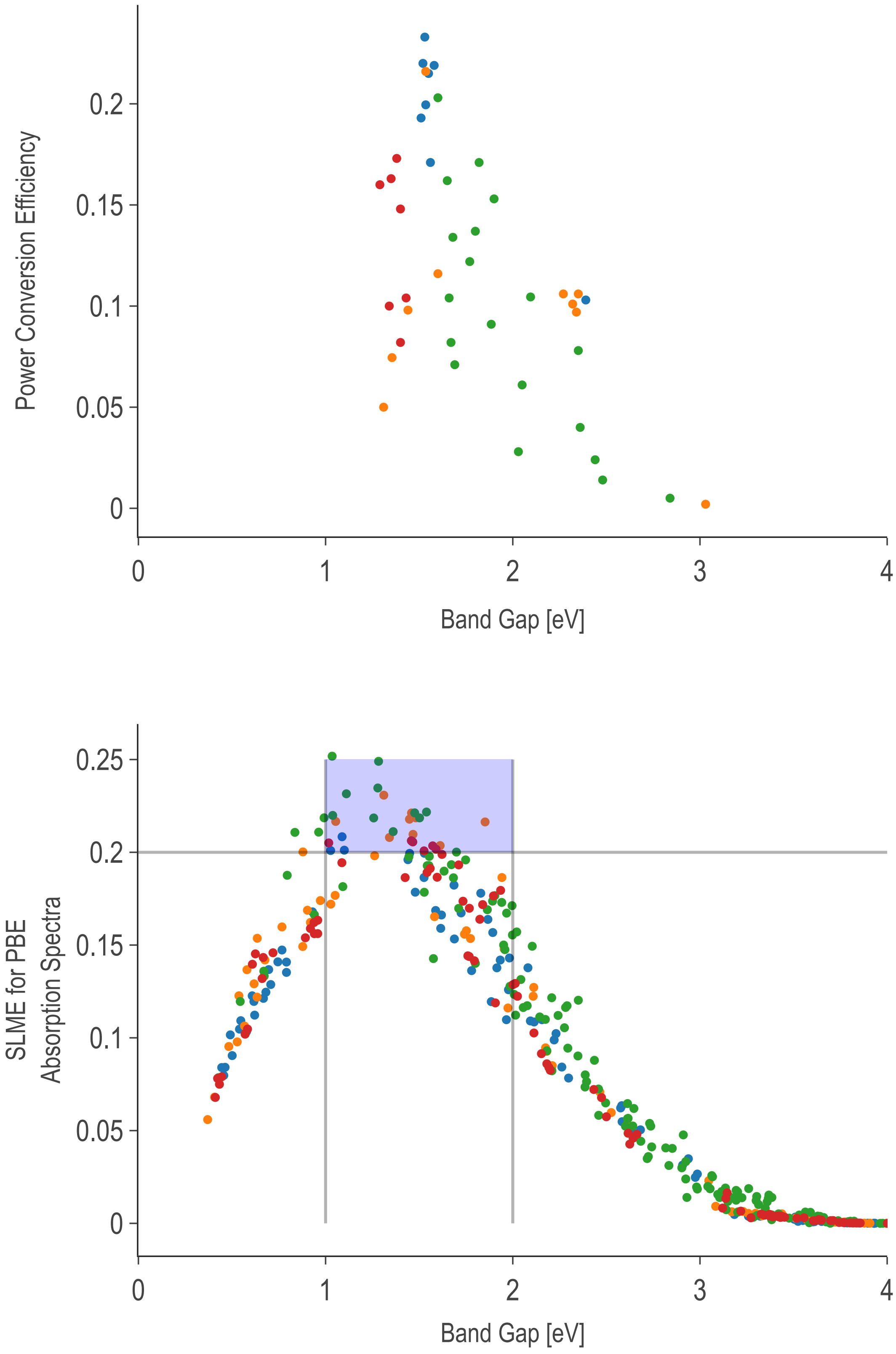

Figure 2.1. PBE SLME of sample compares to experimental PCE and cleanly demarcates competitive range of band gaps

experimental subset (see figure 2.2). Nevertheless, the latter undoubtedly represented the smallest error from truth. The types of mixing per level of theory are apportioned as in figure This is the primary challenge I address with the multi-fidelity models discussed in chapter 3.

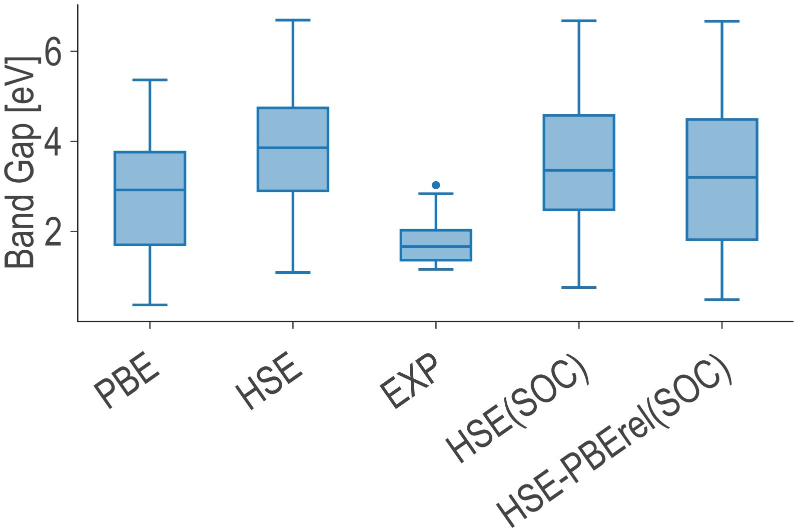

Figure 2.2. Variability in sampled band gaps at each fidelity

It is important to have a notion of which simulation is most accurate to the experimental measurements. Figure 2.3 compares the band gaps obtained for a small subset of elements at all five levels of theory. Theoretically, each functional may be more accurate for certain types of compositions. For instance, organic-inorganic perovskites might benefit from greater account of Van der Waals forces and Pb-based compounds benefit from the use of spin orbit coupling as opposed to Pb-free compounds. Note, phase information was not always available for certain experimental data points collected from the literature, and the inclusion of non-cubic phases in the tables may affect the evaluation of the functionals' accuracy. Also, experimental data is tightly concentrated on the narrow range of performant band gaps likely due to selection bias.

The analysis is summarized in table 2.2. HSE band gaps are heavily overestimated, but may be brought down by the addition of the SOC term. Overall, HSE-PBErel(SOC) is the best approach for simulating band gaps with respect to computational cost and time. PBE root mean square error (RMSE) is not significantly different from the HSE-PBErel(SOC)

RMSE. This is due to the accidental accuracy of semi-local functionals without SOC for hybrid organic-inorganic perovskites. (Mannodi-Kanakkithodi & Chan, 2022; Mannodi-Kanakkithodi et al., 2019)

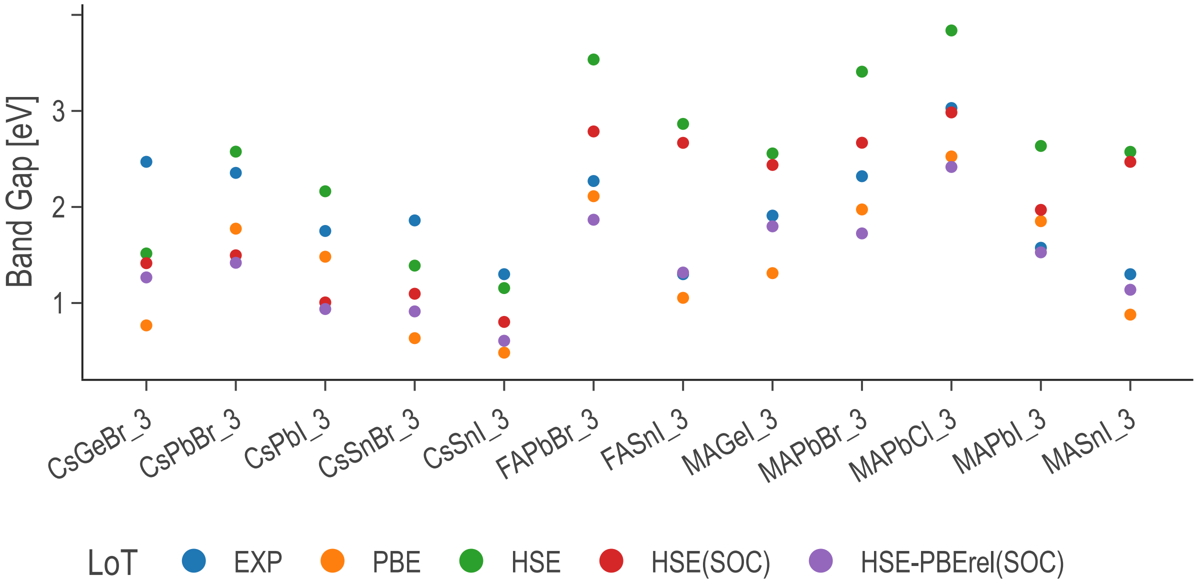

Figure 2.3. Effect of level of theory on band gap measurement

Table 2.2. RMSE values of band gaps computed from different functionals compared with experimental (Exp) values

	RMSE vs EXP
PBE	0.55
HSE	0.87
HSE(SOC)	0.61
HSE-PBErel(SOC)	0.44

2.4 Sampling the Halide Perovskite Chemical Space

Pure (non-alloyed) possibilities are exhaustively sampled using $5*6*3 = 90$ compounds. Starting from these pure perovskite structures systematic mixing was performed at the A, B, and X sites. Figure 2.4 shows the shares of different types of mixing in our sample. Again, for simplicity, only cardinal mixing is considered in this study: that is, mixing is not performed at multiple A/B/X-sites simultaneously. The sample contains a reasonable balance of points representing each one of the cardinal mixing categories. This will help to ensure the ML algorithms learn relationships between fidelities, not differences in mix site or constituency distributions within each fidelity. Additionally, within each mix both purely inorganic samples and hybrid organic-inorganic samples were represented equally.

See the coverage of this sample in figure 2.5.

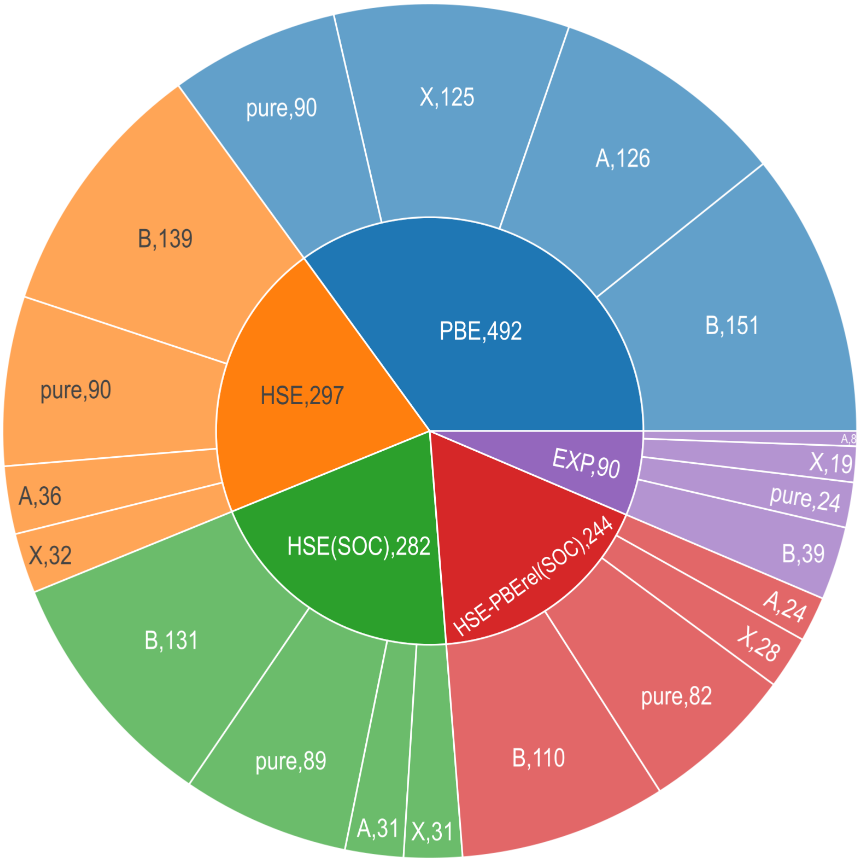

Figure 2.4. Share by count of total data apportioned from each experimental subcategory

Most importantly, this sample gives very even coverage of the cardinal mixing domain as shown in figure 2.5. The clusters in this figure are determined using the t-distributed stochastic neighbor embedding (t-SNE) method. This is a nonparametric dimensionality reduction intended for visualizing statistically relevant clusters in a high dimensional dataset in only two or three dimensions. In this case, the clusters correspond to the mix site of the member data points.

This sample provides the opportunity to comfortably interpolate the properties of other members of the cardinal mixing domain.

See Discussion.

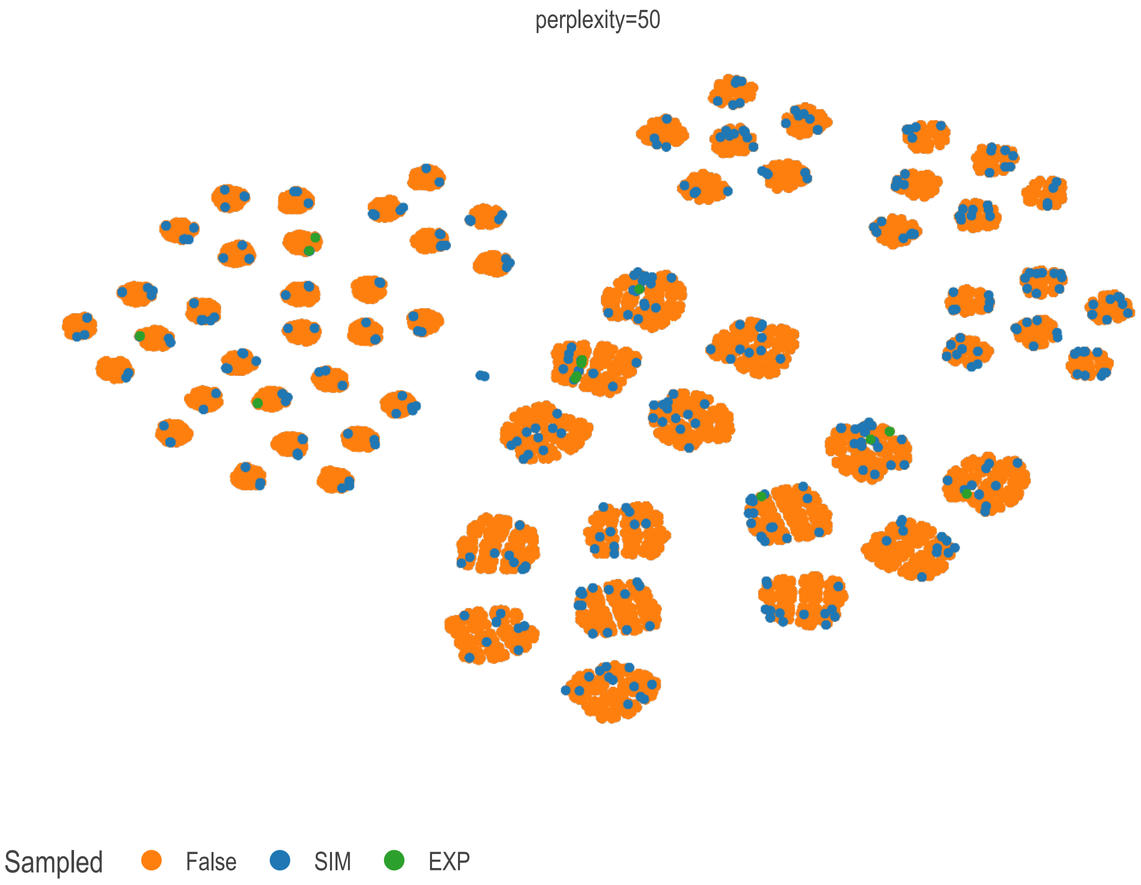

Figure 2.5. Samples overlaid on cardinal mixing chemical domain projected from fourteen to two dimensions via t-SNE

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. I accessed a dataset consisting of – among other characteristic properties – simulated band gaps of a representative sample of halide perovskites (HaP). The effects of mixing at different sites is described by the explicit fraction of a site occupied by a specific atomic or molecular species. Also, a set of abstract features obtained as the weighted averages of these species' bulk physical properties is used to bolster the feature space.

Our multi-objective, multi-fidelity, computational halide perovskite alloy dataset is one of the most comprehensive to date. It is publicly available in the hopes further physical and engineering insights can be extracted by the broader research community.

3. MULTI-FIDELITY MACHINE LEARNING FOR PEROVSKITE BAND GAP PREDICTIONS

The fidelity hierarchy in the sample climbs from DFT simulations performed using the basic PBE GGA functional, to results obtained from physical experiments aggregated in literature. (Almora et al., 2020; J.-P. Kim et al., 2014; Swanson et al., 2017) Low fidelity data makes up the majority of the sample and serves as the foundation for interpolation. However, it does not accuracy reproduce the experimental measurements. My work leverages the data covered in chapters 1 and 2 to predict the band gap of arbitrary perovskite compositions at experimental actuary with little anticipated error.

To do this, a set of interpretable descriptors of each perovskite are used. This takes the form of a 14-dimensional vector containing the atomic fractions of each of the 14 constituent species within the specified perovskite formula. This vector is a sufficient descriptor of a perovskite and has served decent predictions. (Mannodi-Kanakkithodi & Chan, 2022) To improve regression I examine an addition 36 additional predictors derived from linear combinations of compositions and elemental properties obtained from the trusted Mendeleev databases. (Mentel, 2014)

3.1 Model Optimization

The rigorous hyper-parameter Optimization (HPO) of any feature engineering and modeling pipeline is a problem discussed extensively in the literature. HPO approaches can be broadly separated into exhaustive and efficient optimization strategies. (L. Yang & Shami, 2020) We use a two-stage procedure for selecting the best model parameters. The first stage is an exhaustive grid-search over diversely sampled parameter space. Each combination of parameters instantiates a model which is then fit to each of a set of stratified training subsets generated by a K=3 K-fold split cross-validation strategy. Every fitted model is subsequently tested against the cross-validation test sets and a suite of regression scoring metrics are applied to each member category simultaneously using a custom SciKit-learn score adapter[1]. The grid search is then narrowed to a high performance quadrant of the search space by

[1] ↑https://github.com/PanayotisManganaris/yogi

Listing 3.1. An example of the cmcl "ft" feature accessor

```
import cmcl
Y = load_codomain_subset()
df = Y.Formula.to_frame().ft.comp()
df.index = Y.Formula
print(df)
```

the model evaluator based on recommendations made by a simple entropy minimization algorithm[1]. The recommended grid quickly eliminates under-performing settings based on the sample probability of a setting appearing in a set of finalists according to the scoring rankings. The selection score is additionally influenced by a weighted sum of the scoring ranks allowing for considerably tuning the selection criterion. For best results, a few different grid spaces were explored to corroborate eliminations.

After the recommendation is made, the granularity of the grid is increased in the remaining ambiguous parameters and the process is repeated. In general, no more than 2 or 3 exhaustive searches are needed over a given set of grids. Past this point, continuously variable hyper parameters can be individually optimized by plotting validation curves.

3.2 Methods

3.2.1 Featurization of Chemistries

For α total A-site constituents represented in the whole database, β total B-site constituents, and γ total X-site constituents, we provide a Python tool[2] which robustly coverts the composition string of each data point into a $\alpha+\beta+\gamma$ dimensional composition vector. In the case of our total dataset description $\alpha+\beta+\gamma = 14$. (J. Yang & Mannodi-Kanakkithodi, 2022) In a subset of the data, the chemical vector (listing 3.2) is produced using cmcl (listing 3.1).

This is naturally a sparse, relatively high dimensional descriptor. With any growth in the composition space it becomes sparser. This descriptor has been shown to be effective for

[2] ↑https://github.com/PanayotisManganaris/cmcl

Listing 3.2. Data frame of composition vectors generated by cmcl

```
                      FA   Pb   Sn    I   MA   Br
Formula
FAPb_0.7Sn_0.3I_3    1.0  0.7  0.3  3.0  NaN  NaN
MAPb(I0.9Br0.1)3     NaN  1.0  NaN  2.7  1.0  0.3
```

interpolating the properties of irregularly mixed large supercells. (Mannodi-Kanakkithodi & Chan, 2022) However, a spare descriptor is generally bad for extrapolative modeling. (Ghiringhelli et al., 2015)

When extrapolation is the aim, continuously distributed, unique, and linearly independent features are much more reliable. (Lux et al., 2020)

Our attempts to provide a domain with these characteristics results in the following raw feature space.

- 14 sparse composition vectors extracted from chemical formula using `cmcl`[2]
- 36 dense site-averaged property space computed as a linear combination of composition vectors and measured elemental properties (Mentel, 2014)
- 5 categorical dimensions one-hot-encoding level of theory.
    - this provides the categorical axis for multi-task learning
    - see table 2.1

3.2.2 Machine Learning Algorithms and Parameter Optimization

We train Random Forest Regression (RFR) and Gaussian Process Regression (GPR) models of band gap on the union of predictor features previously discussed. The RFR is a flexible nonlinear model, the GPR a principled linear model. Shapley Additive Explaination (SHAP) analysis of the models lends insight to the average physical impacts of 1) site-specific alloying and 2) using organic molecules in the Perovskite superstructure. Model development and feature extraction is performed using Python and SciKit-Learn v1.2. (Pedregosa et al., 2011)

In order to monitor for possible categorical biases effecting regressions, nine metrics are used to evaluate the performance of each model over all alloy types at every stage of the hyper-parameter optimization. This is done simultaneously. in order to train models to be more faithful to the highest fidelity, the score for that subset is weighted as more important. So, only models that perform uniformly well on all alloy types and better on predicting the experimental dataset are selected. Naturally, perovskites mixing at a given site and of a given hybrid-organic/inorganic character were expected to perform significantly differently compared to perovskites containing different mixing/constituents. Classical learning methods require guidance to reasonably explain this mixing diversity and avoid confusion. I provided this guidance by training each model using two test/train splits. First, the optimal model parameters are chosen for their performance under a random split. A minimum of 3-fold cross-validation is performed for every set of model parameters that is considered.

Finally, the optimized model's ability to extrapolate is tested by training/testing on splits determined with a groupwise K-fold splitting strategy.

Two separate cross validation schemes are employed at each stage of the design process. First, the sample set is shuffled once and split to mitigate the models tendency to fit on sample order, then, stratified K-folds are generated in manner consistent with the types of each sample. The regressor is then trained on the subsets of each class. Its ability to extrapolate is independently metered on each validation fold consisting of members of the other classes. Second, the ability for a model trained on samples belonging to one class/status to extrapolate to samples of another class/status is tested as well. The samples again are shuffled and split. then the training set is separated using a grouping K-fold split strategy.

A final best model is instantiated using the overall best performing parameters. These models are finally validated against the test sets originally split off from the sample in both their extrapolative ability and consistency This procedure in demonstrated in an online notebook by Manganaris et al. (2022) hosted on the Purdue nanoHUB.

3.2.3 Feature Engineering

There has been success in creating analytical expressions for perovskite properties, particularly lattice parameters. (L. Jiang et al., 2006) In an attempt to find an analytical predictor for band gap we employ the Sure Independence Screening and Sparsifying Operator (SISSO). (Ouyang et al., 2018)

SIS[3] is a powerful application of compressed sensing. (Ghiringhelli et al., 2017) The SIS operator is a potent dimensionality reduction technique. It does not perform any mathematical decomposition but instead picks existent dimensions that begin to approximate an orthogonal basis. It outperforms CUR decomposition by functioning effectively in extremely high rank vector spaces. (Hamm & Huang, 2019; Ray et al., 2021) This is accomplished by posing the decomposition as a compressed sensing problem in the correlation metric space.

It allows the program to effectively find candidates for a linearly independent basis in a vector space of immense size. unlike legacy techniques, e.g. LASSO, it does not suffer when features are correlated. (Gauraha, 2018; Tibshirani, 1996)

This allows for it to be used in performing a brute force search of a super-space generated by combinatorial operations on the raw predictor variables.

The Sparsifying Operator finds members of the resulting basis set which correlate with the target co-domain. it does this by creating a sparsified linear model, similar again to a LASSO. This process produces an analytic model of the target property, which is easy to interpret and can even be constrained for consistent combination of dimension units.

Subsequent applications of the SIS operator to the residuals of this model are a clever interrogation of error yielding more orthogonal basis sets that can be incorporated into the model. (Mayo, 1996)

SISSO is run for our dataset on the same partitioning scheme used by the previous models via an SciKit-learn compliant (Buitinck et al., 2013) interface[4] extensively modified from the original Matgenix[5] code. Additionally, the algorithm is informed of features units so that it is restricted to meaningful linear combinations. SIS features complexity is restricted to a maximum of 3 operations primarily to encourage parsimonious descriptions. The available operation set is outlined in table 3.1.

Table 3.1. operations for formation of combinatorial super-space

Binary	Unary
addition	reciprocation
subtraction	power 2
multiplication	power 3
division	natural logarithm
	exponentiation
	root 2

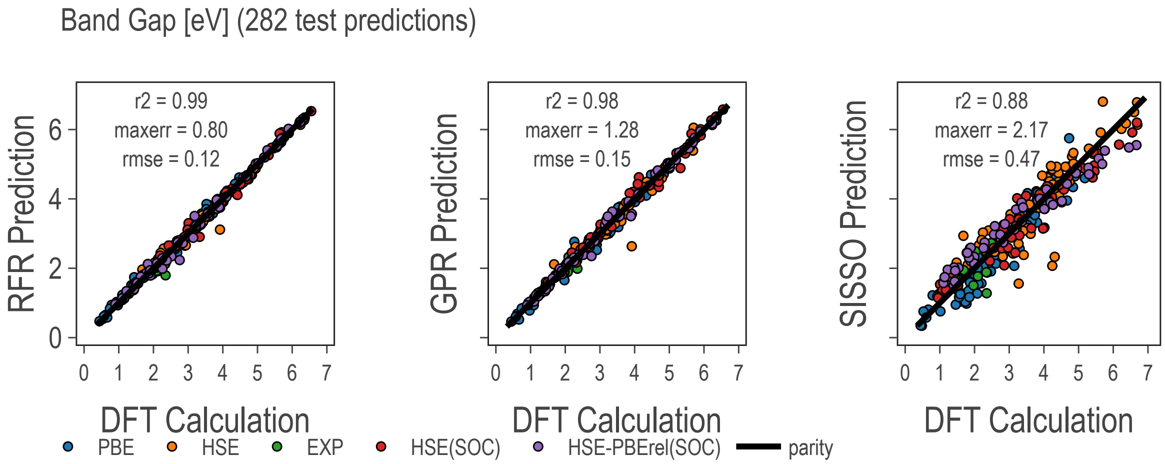

Figure 3.1. model predictions vs true values at multiple fidelities

3.3 Results

3.3.1 Best Models on Raw Domain

The optimized models are high performing (Table 3.2). The RFR hyper-parameters are listed in the appendix (Table **??**).

The GPR model is tried with multiple kernels. Ultimately, the best is a non stationary Matern kernel with $\nu = \frac{3}{2}$.

3.3.2 SISSO Model and SIS Engineered Features

The Sure Independence Screening and Sparsifying Operator (SISSO) is a specific combination of multiple data mining techniques chained together resulting in a symbolically expressed regression model. (Ghiringhelli et al., 2017; Ouyang et al., 2018)

The best SISSO model for band gap involving 3 SIS features (each composed of up to 4 basic features) has an unremarkable RMSE of 0.476 eV, barely outperforming an OLS regression on 55 dimensions (see Table 3.2). It is expressed in equation 3.1. Notably, while the units of the expression do not match the units of band gap as measured (target units are unknown to the algorithm), they are still energy units. This is by design, as the combination of features was restricted so to only allow compatible units to be combined. A separate training session without this restriction was attempted, but the resulting model's performance was worse.

$$\begin{aligned} bg\,\mathrm{eV} = {} & 1.752393064((X;\mathrm{electronegativity} * A;\mathrm{heat\ of\ fusion}) - \\ & \qquad (B;\mathrm{electron\ affinity} + B;\mathrm{ionization\ energy})) \\ & + -0.5862929089((B; Sn - \mathrm{HSE}) + (\mathrm{PBE} - X;\mathrm{electronegativity})) \\ & \quad +1.063684923((A;\mathrm{electronegativity} - B; Ca) * (B;\mathrm{heat\ of\ vap} - X;\mathrm{electron\ affinity})) \\ & \quad +4.657097107 \end{aligned} \tag{3.1}$$

[3] ↑https://github.com/rouyang2017/SISSO

[4] ↑https://github.com/PanayotisManganaris/pysisso

[5] ↑https://github.com/Matgenix/pysisso

Table 3.2. RMSE of models on raw domain calculated per LoT subset

rmse scores	GPR	RFR	Linear OLS	SISSO	SIS + GPR	SIS + RFR
total	0.156214	0.124738	0.499558	0.474754	0.251881	0.187431
EXP	0.120475	0.154448	0.307186	0.330080	0.338949	0.235397
PBE	0.128211	0.101872	0.472430	0.395827	0.171640	0.134529
HSE	0.214920	0.152479	0.558077	0.519706	0.305443	0.208390
HSE(SOC)	0.156785	0.108867	0.535087	0.572644	0.272158	0.221007
HSE-PBE(SOC)	0.130696	0.133027	0.466364	0.470758	0.252624	0.189510

Computing and combining more than 3 SIS features is not rewarding of the computational expense. Residuals are increasingly uncorrelated with the generated SIS features and model accuracy gains do not outstrip complexity. However, in the process of creating Equation 3.1, 150 SIS predictor variables were determined and recorded. 50 primary predictors, 50 first residual predictors, and 50 second residual predictors. These can serve as a high quality, introspective domain for the other architectures to fit on.

3.3.3 Best Models on Engineered Domain

We set the aim of decreasing $\mathcal{O}(n^3)$ computational expense of GPR by ≈10 times. So, we aim to take 30 highly correlated features (slightly more than one half the number used by prior models) from these SIS subspaces. We expected this to solve the problems inherent to the raw 3.2.1.

fitting models to SIS features may leverage the denser and more continuous domain to improve extrapolative predictions. Potentially into the high-entropy domain, or simply Theory. However using the SIS subspaces in this way compromises on SISSO's explicability and necessitates SHAP analysis. Unfortunately, whatever the gains in training time complexity and extrapolative ability, the models underperformed in predicting band gap in the cardinal mixing domain (see Table 3.2). This was unexpected considering the raw features are by their nature highly correlated and presumed redundant. Nevertheless, the RFR model on the higher dimensional, sparser raw features is superior.

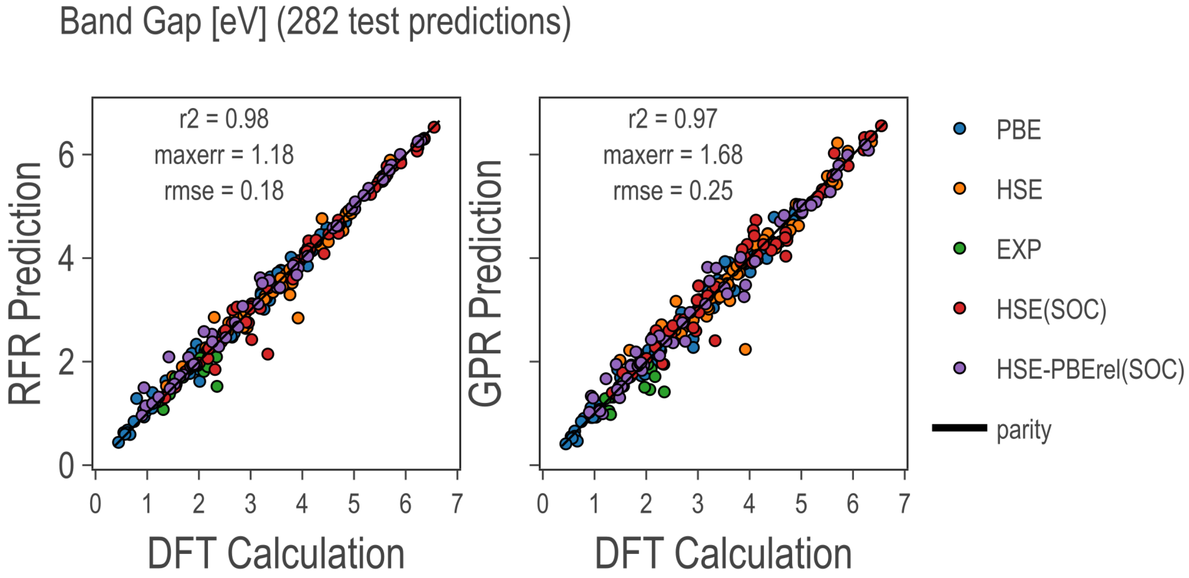

Figure 3.2. SIS-based model predictions vs true values at multiple fidelities

3.4 Discussion

3.4.1 SHAP Analysis of Domain

SHAP scores are computed automatically for every dimension of every sample in the domain by the python SHAP package[6]. The sum of the expectation value of the target conditioned on the model features and the SHAP scores computed for each predictor variable of a sample is the model's prediction for that sample target. (Lundberg & Lee, 2017) For the perovskite band gap the expectation value is 2.836 when conditioned on the raw features and 2.863 when conditioned on the SIS features. The raw features' SHAP values are more centered around zero while engineered features are more often scored decisively positive or negative.

Figures 3.3 and 7 show the top score distributions. In each figure, features are ranked by overall value on the y-axis. The x-axis shows the SHAP score for each point. The points are shaped in a violin plot to show the distribution of effects the presence of the given feature can have. Finally, on the color-axis, feature value specifies whether a particular score is a large or small absolute contributor of the sum to the prediction.

For instance, in figure 3.3, the B-site Electronegativity is often a strongly positive contributor to the RFR prediction. However, almost always in this case it is out-contributed by other features, it does not mostly determine the result but it is still valuable. On the contrary, when it is a strongly negative contributor it effectively determines the result. It is interesting to see how models make use of features in light of basic bi-variate correlations. The only features that correlate strongly with band gap are illustrated in figure 3.5. Notably, the Random Forest Regression (RFR) primarily uses the highly correlated features, while the Gaussian Process Regression (GPR) primarily uses features with lower Pearson correlations.

SHAP scores in principle quantify the contributions of site members and site member properties to the perovskite band gap. On a sample-by-sample basis it is possible to say how much of the bandgap is contributed by the presense of a given quantity of, for example, Germanium. However a clustering analysis reveals no universal patterns. SHAP scores given the raw domain are near zero on average regardless of partitions made by level of theory,

[6] ↑https://github.com/slundberg/shap

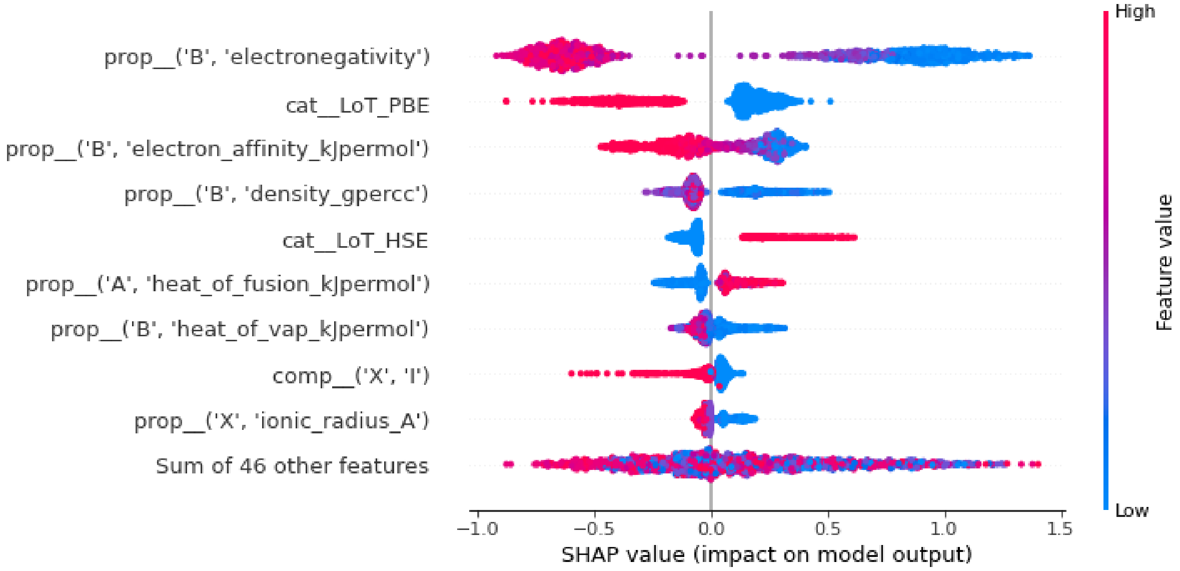

Figure 3.3. Random Forest Regression Band Gap SHAP Values

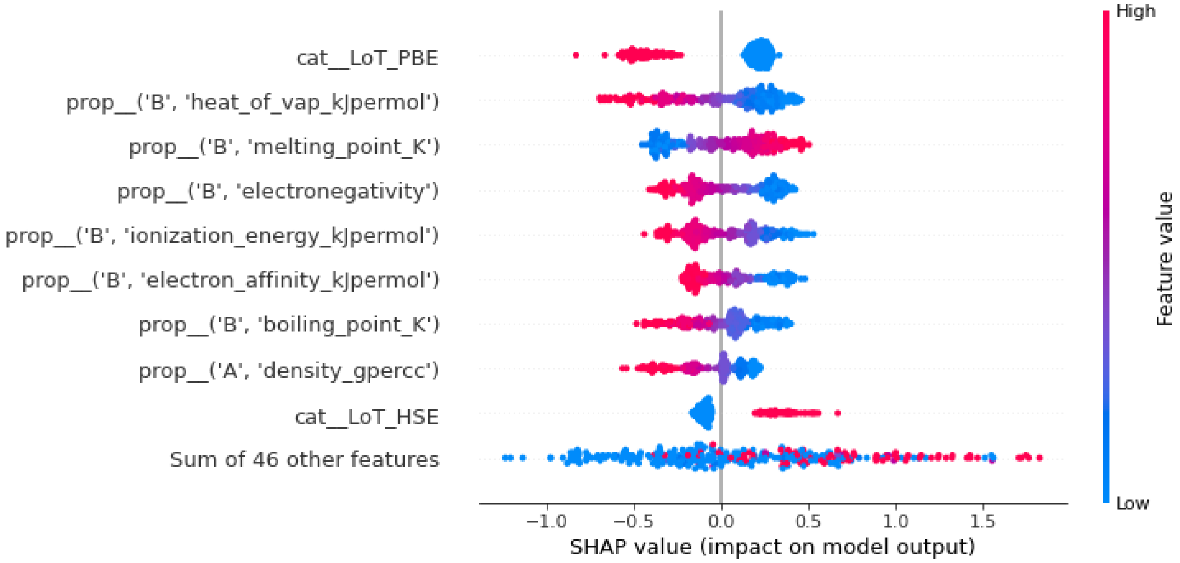

Figure 3.4. Gaussian Process Regression Band Gap SHAP Values

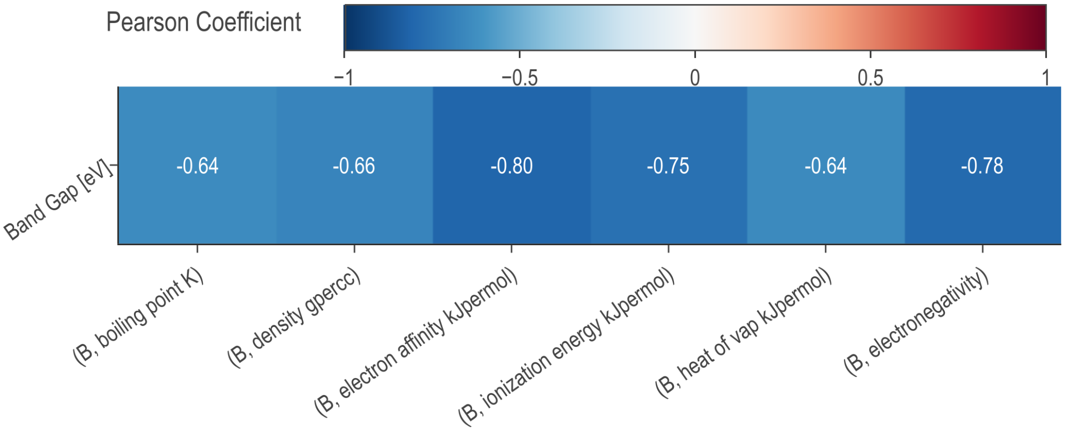

Figure 3.5. raw features with ($|p| > 0.5$) against band gap

alloy scheme, or presence of organic A-site occupants. This analysis confirms the difficulty of deducing a rule of thumb for the synthesis of perovskites with desirable properties. If anything, figure 3.6 confirms that the Iodine at the X site tends to slightly increase band gaps.

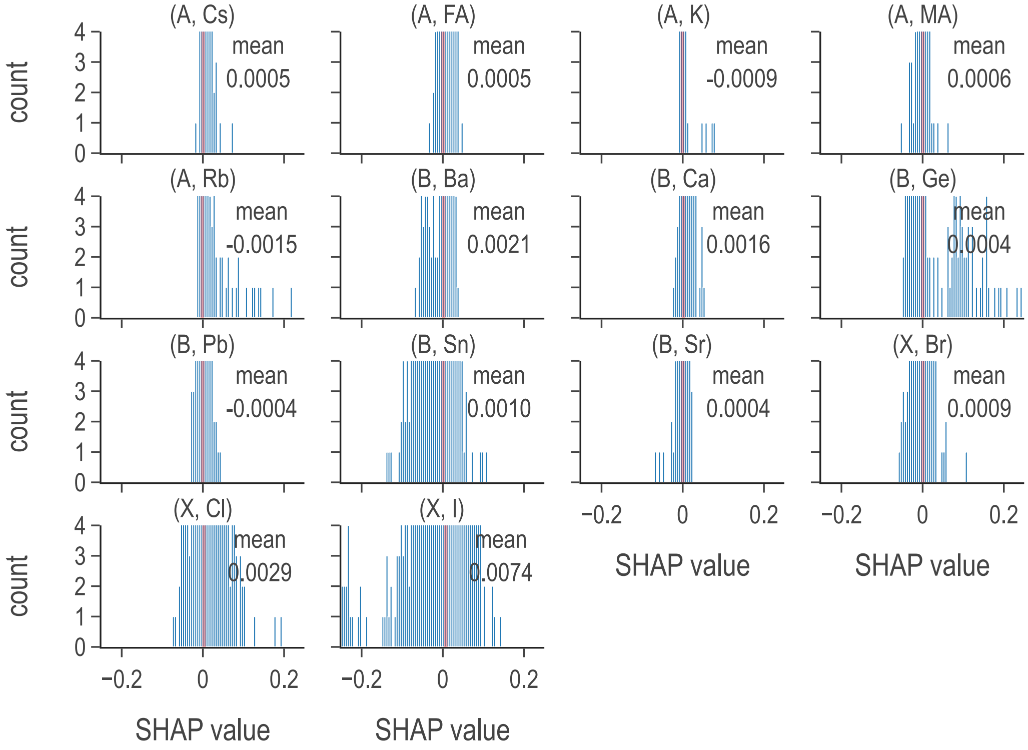

Figure 3.6. SHAP score distributions reveal effects of individual constituents

3.4.2 Predictions and Screening

Using the superior RFR model, I predict the band gap for all 37785 possible compositions demonstrating cardinal mixing within the bounds of a 2x2x2 perovskite super cell. That is eight A-sites shared by up to 5 constituents, 8 B-sites shared by up to six constituents, and 24 X-sites shared by up to 3 constituents. Given the good coverage achieved by our sample dataset (figure 2.5) and according to the scores reported in Table 3.2, the RFR model is capable of predicting band gaps at the experimental fidelity with a 0.15 RMSE. These predictions were projected on the sample space in Figure 3.7.

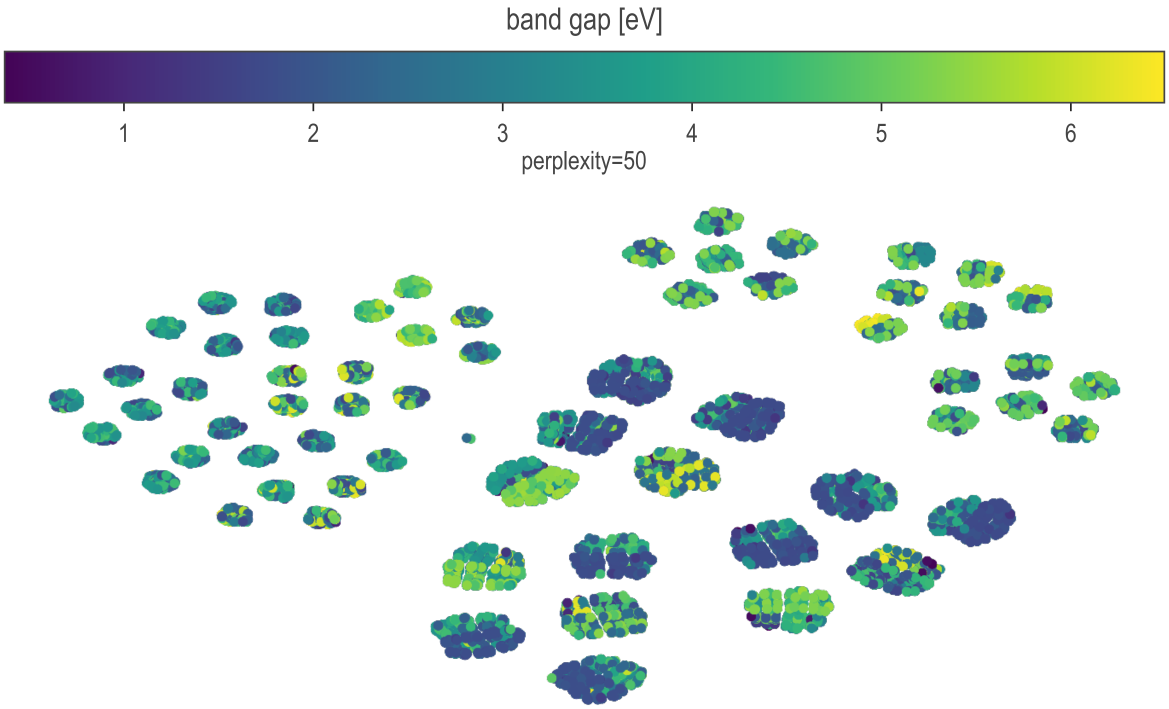

Figure 3.7. Band gap predictions overlaid on cardinal mixing chemical domain projected from fourteen to two dimensions via t-SNE

I followed a similar high-throughput screening procedure to that laid out in prior works, except this covered a large space of purely hypothetical compounds. (Mannodi-Kanakkithodi & Chan, 2021; J. Yang et al., 2023) See figure 3.8. Band gaps between 1 and 2 eV were selected as this range is expected to yield the best power conversion efficiency (PCE) in the visible spectrum. (Shockley & Queisser, 1961; Yu & Zunger, 2012) Perovskite compounds were selected for their predicted stability by cutting on each of three tolerance factors previously established in chapter 1. The constituent ratios of the chosen compositions in figure 3.9 may be juxtaposed with figure 1.3.

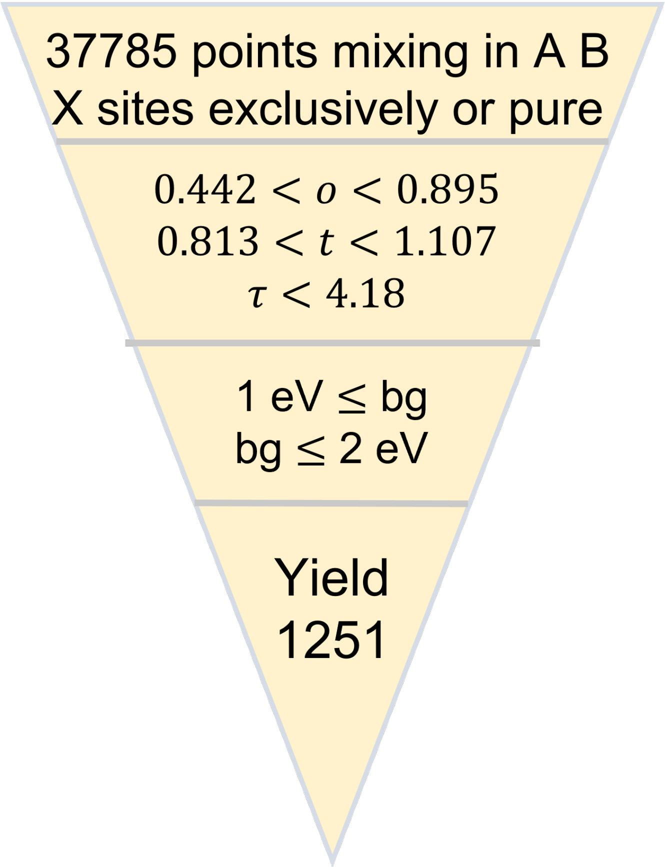

Figure 3.8. Summary of screening operations used to identify candidate compounds

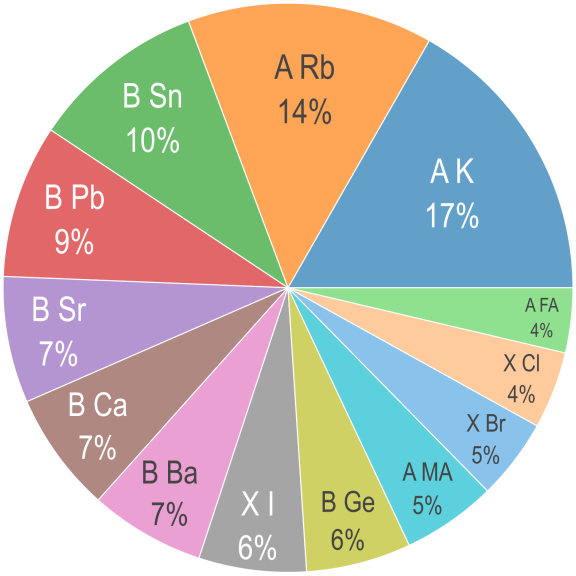

Figure 3.9. The compounds selected from the cardinal mixing sample space contain varying fractions of each element

These cuts trimmed the 37785 points by 97% to a subset of only 1251 viable candidates. These selected candidates were projected onto the t-SNE embedding space in Figure 11. A Frequency analysis revealed the constituent elements of the chosen subset most often occupied either small or large shares of their site. Most A-site constituents preferred occupying 1/8th of their site at a rate of about 8%, with Potassium and Rb also preferring full occupancy 10-12% of the time. B-site constituents favored pure configurations at a rate of 5-8% but also showed some preference for doping configurations. X-site constituents, however showed very strong preference for fully occupying their site 25% of the time. See figure 3.10. The strong preference for pure sites simply reflects that this sample space contained compositions mixed at no more than one site simultaneously.

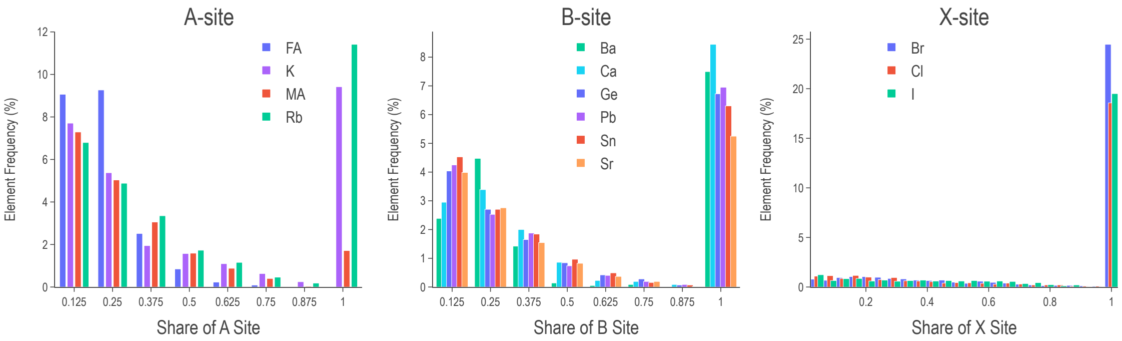

Figure 3.10. Frequency of mixing fractions of species at the A, B, and X sites across the ~3000 screen compounds

4. CONCLUSIONS

A set of promising hypothetical compositions is identified by this work. Of the 1251 compositions selected 640 are purely inorganic, and 611 are hybrid-organic/inorganic. Table 4.1 shows how the set subdivides by mixing. Only 40 of the original expertly designed sample pass the screening, the rest are untested to my knowledge. This shows that there is still much opportunity for discovery in this area and much to be learned about this chemistry. Notably, 834 contain no lead. A random sample of 30 such compounds is listed in table 4.2. Also, no compounds containing Potassium pass the screening.

This was achieved with a very simple regression model. My assessment suggests it is operating as any random forest would. It is simply learning a complex algorithm which can be followed to closely approximate band gaps of a variety of distinct descriptors conditioned on the desired prediction fidelity. Its superior performance is due to its improved flexibility in accounting for outliers. However, the Gaussian Process Regression is a very close second. The GPR's ability to theoretically deal with more complex descriptors makes for an appealing next step. Further improving on the feature engineering and examining what can be learned about the relationship between structure and band gap might begin here.

Table 4.1. Number of selected data points with given mixing site

	count
A	605
B	388
X	256
pure	2

Table 4.2. Thirty hypothetical lead-free formulae and their predicted band gaps

	Formula	band gap [eV]
1472	Rb1.000Ge1.000Cl0.583I0.417	1.610555
13457	K1.000Ge0.125Sn0.500Sr0.375Br1.000	1.588236
14729	K1.000Ca0.250Ge0.625Sn0.125Cl1.000	1.906317
19075	FA0.250K0.250MA0.375Rb0.125Sn1.000Cl1.000	1.951812
20006	FA0.625MA0.375Ba1.000Br1.000	1.751004
17947	K1.000Ba1.000Br0.250Cl0.083I0.667	1.927655
13792	K1.000Ge0.875Sr0.125Cl1.000	1.361000
15452	K1.000Ba0.125Sn0.625Sr0.250I1.000	1.864185
14714	K1.000Ca0.250Ge0.500Sn0.250Cl1.000	1.992692
5862	MA0.500Rb0.500Ba1.000I1.000	1.809483
19050	FA0.250K0.250MA0.125Rb0.375Ba1.000I1.000	1.788057
19109	FA0.250K0.375Rb0.375Sr1.000Br1.000	1.714627
18606	FA0.125K0.500Rb0.375Sn1.000I1.000	1.962433
19148	FA0.250K0.375MA0.250Rb0.125Sn1.000Br1.000	1.815876
2891	Rb1.000Ca1.000Br0.208Cl0.250I0.542	1.705446
26843	Cs0.125FA0.125MA0.750Ba1.000Br1.000	1.950356
31393	Cs0.625Rb0.375Ca1.000Cl1.000	1.878681
29328	Cs0.250FA0.250K0.375Rb0.125Sr1.000I1.000	1.955184
19898	FA0.500K0.375Rb0.125Ba1.000Br1.000	1.727611
2297	Rb1.000Ca0.250Ge0.250Sn0.250Sr0.250Cl1.000	1.803151
3092	Rb1.000Ca1.000Br0.958I0.042	1.691322
28989	Cs0.250FA0.125K0.250MA0.375Sn1.000I1.000	1.744348
13004	K1.000Sn1.000Br0.875Cl0.042I0.083	1.709227
19180	FA0.250K0.500Rb0.250Sr1.000Cl1.000	1.971389
533	Rb1.000Sn1.000Br0.333Cl0.625I0.042	1.788413
3056	Rb1.000Ca1.000Br0.667Cl0.250I0.083	1.706197
19076	FA0.250K0.250MA0.375Rb0.125Sn1.000Br1.000	1.938144
18566	FA0.125K0.375MA0.250Rb0.250Ba1.000Br1.000	1.805977
5396	Rb1.000Ba0.625Ca0.250Sn0.125I1.000	1.737319
26921	Cs0.125FA0.125K0.125MA0.500Rb0.125Sn1.000Br1.000	1.893728

5. PUBLICATIONS

Edlabadkar, R., Yang, J., Rahman, H., Manganaris, P., Korimilli, E. P., & Mannodi-Kanakkithodi, A. (2023). Driving halide perovskite discovery using graph neural networks [In Preparation]. https://doi.org/00.0000/00000000

Gollapalli, P., Manganaris, P., & Mannodi-Kanakkithodi, A. (2023). Graph neural network predictions for formation energy of native defects in zinc blende semiconductors [In Preparation]. https://doi.org/00.0000/00000000

Manganaris, P., Desai, S., & Kanakkithodi, A. (2022). Mrs computational materials science tutorial. https://doi.org/10.21981/D1J2-AR65

Manganaris, P., Yang, J., & Mannodi Kanakkithodi, A. (2023). Multi-fidelity machine learning pervoskite composition vs band gap [In Preparation]. https://doi.org/00.0000/00000000

Yang, J., Manganaris, P. T., & Mannodi Kanakkithodi, A. K. (2023). A high-throughput computational dataset of halide perovskite alloys. *Digital Discovery*. https://doi.org/10.1039/d3dd00015j

6. PRESENTATIONS

- Poster for DS02 symposium MRS fall 2022
- Talk for Purdue Soft Materials symposium
- Developed MRS spring 2022 tutorial hosted on nanoHUB(Manganaris et al., 2022)

7. SOFTWARE AND DATA CONTRIBUTIONS

7.1 Data Publication

I published the prepared sample data I used to the Materials Data Facility. It is available to download with a simple API call following installation and activation[1] of the relevant packages and applications.

7.2 Software Tools

Additionally, a few python libraries began development with this work as the authors' attempt to contribute to the larger effort by the materials science community to ease aggregation, sharing, analysis and reporting of large computational datasets.

The `cmcl` library[2] is under early development under tag v0.1.5. At this stage, cmcl strives to provide an inquisitive interface to perovskite composition feature computers in the style of the pandas API. Listing 3.1 demonstrates its use in extracting composition vectors from the formula strings identifying each compound in a dataset.

[1] ↑https://ai-materials-and-chemistry.gitbook.io/foundry/

[2] ↑https://github.com/PanayotisManganaris/cmcl

Listing 7.1. How to load the Mannodi Group halide perovskites data set from the Materials Data Facility repository

```
# THIS IS PENDING
f = Foundry(
    no_local_server=True,
    no_browser=True,
    globus=True,
    index="mdf"
)
f.load("foundry_mrg_band_gaps_v1.0", globus=globus)
res = f.load_data()
X_mp, y_mp = res['train'][0], res['train'][1]
```

A library of model evaluation tools to assist with exhaustive grid search is being maintained in the `yogi` repository.[3] A grid-search assistant under the `yogi.model_selection.butler` module was used to optimize the hyper-parameters of models reported here. See documentation for the various grid-narrowing strategies available.

Matgenix[4] company initially created a SciKit-learn compliant interface to the SISSO algorithm maintained by Ouyang et al. (2018). This code was forked and modified[5] to enable the creation of the SIS domain engineered regressions primarily by creating a custom `Function Transformer` that may be seeded with the subspace information obtained by first training a SISSO regression using the conventional interface. This was done with the expectation that the resulting model would be superior for the purposes of this work, but it is generally applicable to any other applications demanding this sort of dimensionality reduction.

7.3 Tutorials

I developed and published one tutorial on the Purdue nanoHUB for delivery in the Spring 2022 Materials Research Society conference. Manganaris, P., Desai, S., & Kanakkithodi, A. (2022). MRS Computational Materials Science Tutorial. This tutorial covered a variety of machine learning methods. It provides a detailed example of the use of `cmcl` and `yogi` in feature extraction and training multi-fidelity random forest models by score weighting respectively.

[3] ↑https://github.com/PanayotisManganaris/yogi
[4] ↑https://github.com/Matgenix/pysisso
[5] ↑https://github.com/PanayotisManganaris/pysisso

REFERENCES

Almora, O., Baran, D., Bazan, G. C., Berger, C., Cabrera, C. I., Catchpole, K. R., ErtenEla, S., Guo, F., Hauch, J., HoBaillie, A. W. Y., Jacobsson, T. J., Janssen, R. A. J., Kirchartz, T., Kopidakis, N., Li, Y., Loi, M. A., Lunt, R. R., Mathew, X., McGehee, M. D., … Brabec, C. J. (2020). Device performance of emerging photovoltaic materials (version 1). *Advanced Energy Materials*, *11*(11), 2002774. https://doi.org/10.1002/aenm.202002774

Ansari, M. I. H., Qurashi, A., & Nazeeruddin, M. K. (2018). Frontiers, opportunities, and challenges in perovskite solar cells: A critical review. *Journal of Photochemistry and Photobiology C: Photochemistry Reviews*, *35*, 1–24. https://doi.org/10.1016/j.jphotochemrev.2017.11.002

Banerjee, A., Chakraborty, S., & Ahuja, R. (2019). Rashba triggered electronic and optical properties tuning in mixed cation-mixed halide hybrid perovskites [not free]. *ACS Applied Energy Materials*, *2*(10), 6990–6997. https://doi.org/10.1021/acsaem.9b01479

Bartel Christopher, J., Sutton, C., Goldsmith Bryan, R., Ouyang, R., Musgrave Charles, B., Ghiringhelli Luca, M., & Scheffler, M. (2019). New tolerance factor to predict the stability of perovskite oxides and halides. *Science Advances*, *5*(2), eaav0693. https://doi.org/10.1126/sciadv.aav0693

Brenner, T. M., Egger, D. A., Kronik, L., Hodes, G., & Cahen, D. (2016). Hybrid organic-inorganic perovskites: Low-cost semiconductors with intriguing charge-transport properties. *Nature Reviews Materials*, *1*(1), 15007. https://doi.org/10.1038/natrevmats.2015.7

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., & Varoquaux, G. (2013). API design for machine learning software: Experiences from the scikit-learn project. *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 108–122.

Castelli, I. E., García-Lastra, J. M., Thygesen, K. S., & Jacobsen, K. W. (2014). Bandgap calculations and trends of organometal halide perovskites. *APL Materials*, *2*(8), 081514. https://doi.org/10.1063/1.4893495

Cui, P., Wei, D., Ji, J., Huang, H., Jia, E., Dou, S., Wang, T., Wang, W., & Li, M. (2019). Planar p-n homojunction perovskite solar cells with efficiency exceeding 21.3 %. *Nature Energy*, *4*(2), 150–159. https://doi.org/10.1038/s41560-018-0324-8

Dahliah, D., Brunin, G., George, J., Ha, V.-A., Rignanese, G.-M., & Hautier, G. (2021). High-throughput computational search for high carrier lifetime, defect-tolerant solar absorbers. *Energy Environ. Sci.*, *14*, 5057–5073. https://doi.org/10.1039/D1EE00801C

Dimesso, L., Quintilla, A., Kim, Y.-M., Lemmer, U., & Jaegermann, W. (2016). Investigation of formamidinium and guanidinium lead tri-iodide powders as precursors for solar cells. *Materials Science and Engineering: B*, *204*, 27–33. https://doi.org/10.1016/j.mseb.2015.11.006

Ding, J., Du, S., Zhou, T., Yuan, Y., Cheng, X., Jing, L., Yao, Q., Zhang, J., He, Q., Cui, H., Zhan, X., & Sun, H. (2019). Cesium decreases defect density and enhances optoelectronic properties of mixed ma1-xcsxpbbr3 single crystal. *The Journal of Physical Chemistry C*, *123*(24), 14969–14975. https://doi.org/10.1021/acs.jpcc.9b03987

Edlabadkar, R., Yang, J., Rahman, H., Manganaris, P., Korimilli, E. P., & Mannodi-Kanakkithodi, A. (2023). Driving halide perovskite discovery using graph neural networks [In Preparation]. https://doi.org/00.0000/00000000

Ganose, A. M., Jackson, A. J., & Scanlon, D. O. (2018). Sumo: Command-line tools for plotting and analysis of periodic *ab initio* calculations. *Journal of Open Source Software*, *3*(28), 717. https://doi.org/10.21105/joss.00717

Gauraha, N. (2018). Introduction to the lasso. *Resonance*, *23*(4), 439–464. https://doi.org/10.1007/s12045-018-0635-x

Ghiringhelli, L. M., Vybiral, J., Ahmetcik, E., Ouyang, R., Levchenko, S. V., Draxl, C., & Scheffler, M. (2017). Learning physical descriptors for materials science by compressed sensing. *New Journal of Physics*, *19*(2), 023017. https://doi.org/10.1088/1367-2630/aa57bf

Ghiringhelli, L. M., Vybiral, J., Levchenko, S. V., Draxl, C., & Scheffler, M. (2015). Big data of materials science: Critical role of the descriptor. *Physical Review Letters*, *114*(10). https://doi.org/10.1103/physrevlett.114.105503

Gollapalli, P., Manganaris, P., & Mannodi-Kanakkithodi, A. (2023). Graph neural network predictions for formation energy of native defects in zinc blende semiconductors [In Preparation]. https://doi.org/00.0000/00000000

Greenland, C., Shnier, A., Rajendran, S. K., Smith, J. A., Game, O. S., Wamwangi, D., Turnbull, G. A., Samuel, I. D. W., Billing, D. G., & Lidzey, D. G. (2020). Correlating phase behavior with photophysical properties in mixed-cation mixed-halide perovskite thin films. *Advanced Energy Materials*, *10*(4), 1901350. https://doi.org/10.1002/aenm.201901350

Hamm, K., & Huang, L. (2019). Cur decompositions, approximations, and perturbations. *CoRR*. http://arxiv.org/abs/1903.09698v2

Heyd, J., Scuseria, G. E., & Ernzerhof, M. (2003). Hybrid functionals based on a screened coulomb potential. *The Journal of Chemical Physics*, *118*(18), 8207–8215. https://doi.org/10.1063/1.1564060

Hinuma, Y., Pizzi, G., Kumagai, Y., Oba, F., & Tanaka, I. (2016). Band structure diagram paths based on crystallography. *CoRR*. http://arxiv.org/abs/1602.06402v4

Jeong, M., Choi In, W., Go Eun, M., Cho, Y., Kim, M., Lee, B., Jeong, S., Jo, Y., Choi Hye, W., Lee, J., Bae, J.-H., Kwak Sang, K., Kim Dong, S., & Yang, C. (2020). Stable perovskite solar cells with efficiency exceeding 24.8 % and 0.3-v voltage loss. *Science*, *369*(6511), 1615–1620. https://doi.org/10.1126/science.abb7167

Jiang, L., Guo, J., Liu, H., Zhu, M., Zhou, X., Wu, P., & Li, C. (2006). Prediction of lattice constant in cubic perovskites. *Journal of Physics and Chemistry of Solids*, *67*(7), 1531–1536. https://doi.org/10.1016/j.jpcs.2006.02.004

Jiang, Z., Nahas, Y., Xu, B., Prosandeev, S., Wang, D., & Bellaiche, L. (2016). Special quasirandom structures for perovskite solid solutions. *Journal of Physics: Condensed Matter*, *28*(47), 475901. https://doi.org/10.1088/0953-8984/28/47/475901

Kar, M., & Körzdörfer, T. (2018). Computational screening of methylammonium based halide perovskites with bandgaps suitable for perovskite-perovskite tandem solar cells. *The Journal of Chemical Physics*, *149*(21), 214701. https://doi.org/10.1063/1.5037535

Kieslich, G., Sun, S., & Cheetham, A. K. (2015). An extended tolerance factor approach for organic-inorganic perovskites. *Chemical Science*, *6*(6), 3430–3433. https://doi.org/10.1039/c5sc00961h

Kim, C., Huan, T. D., Krishnan, S., & Ramprasad, R. (2017). A hybrid organic-inorganic perovskite dataset. *Scientific Data*, *4*(1), 170057. https://doi.org/10.1038/sdata.2017.57

Kim, J.-P., Christians, J. A., Choi, H., Krishnamurthy, S., & Kamat, P. V. (2014). Cdses nanowires: Compositionally controlled band gap and exciton dynamics [PMID: 26274456]. *The Journal of Physical Chemistry Letters*, *5*(7), 1103–1109. https://doi.org/10.1021/jz500280g

Kim, S., Márquez, J. A., Unold, T., & Walsh, A. (2020). Upper limit to the photovoltaic efficiency of imperfect crystals from first principles. *Energy Environ. Sci.*, *13*, 1481–1491. https://doi.org/10.1039/D0EE00291G

Kresse, G., & Furthmüller, J. (1996a). Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Computational Materials Science*, *6*(1), 15–50. https://doi.org/10.1016/0927-0256(96)00008-0

Kresse, G., & Furthmüller, J. (1996b). Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B*, *54*, 11169–11186. https://doi.org/10.1103/PhysRevB.54.11169

Kresse, G., & Hafner, J. (1993). Ab initio molecular dynamics for liquid metals. *Phys. Rev. B*, *47*, 558–561. https://doi.org/10.1103/PhysRevB.47.558

Kresse, G., & Hafner, J. (1994). Norm-conserving and ultrasoft pseudopotentials for first-row and transition elements. *Journal of Physics: Condensed Matter*, *6*(40), 8245–8257. https://doi.org/10.1088/0953-8984/6/40/015

Kresse, G., & Joubert, D. (1999). From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B*, *59*, 1758–1775. https://doi.org/10.1103/PhysRevB.59.1758

Lee, B. D., Park, W. B., Lee, J.-W., Kim, M., Pyo, M., & Sohn, K.-S. (2021). Discovery of lead-free hybrid organic/inorganic perovskites using metaheuristic-driven dft calculations. *Chemistry of Materials*, *33*(2), 782–798. https://doi.org/10.1021/acs.chemmater.0c04499

Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *CoRR*. http://arxiv.org/abs/1705.07874v2

Lux, T. C. H., Watson, L. T., Chang, T. H., Hong, Y., & Cameron, K. (2020). Interpolation of sparse high-dimensional data. *Numerical Algorithms*, *88*(1), 281–313. https://doi.org/10.1007/s11075-020-01040-2

Manganaris, P., Desai, S., & Kanakkithodi, A. (2022). Mrs computational materials science tutorial. https://doi.org/10.21981/D1J2-AR65

Manganaris, P., Yang, J., & Mannodi Kanakkithodi, A. (2023). Multi-fidelity machine learning pervoskite composition vs band gap [In Preparation]. https://doi.org/00.0000/00000000

Mannodi-Kanakkithodi, A., & Chan, M. K. Y. (2022). Data-driven design of novel halide perovskite alloys. *Energy Environ. Sci.*, *15*, 1930–1949. https://doi.org/10.1039/D1EE02971A

Mannodi-Kanakkithodi, A., & Chan, M. K. (2021). Computational data-driven materials discovery. *Trends in Chemistry*, *3*(2), 79–82. https://doi.org/10.1016/j.trechm.2020.12.007

Mannodi-Kanakkithodi, A., Park, J.-S., Jeon, N., Cao, D. H., Gosztola, D. J., Martinson, A. B. F., & Chan, M. K. Y. (2019). Comprehensive computational study of partial lead substitution in methylammonium lead bromide. *Chemistry of Materials*, *31*(10), 3599–3612. https://doi.org/10.1021/acs.chemmater.8b04017

Manser, J. S., Christians, J. A., & Kamat, P. V. (2016). Intriguing optoelectronic properties of metal halide perovskites. *Chemical Reviews*, *116*(21), 12956–13008. https://doi.org/10.1021/acs.chemrev.6b00136

Mayo, D. G. (1996, April). *Error and the growth of experimental knowledge*. https://doi.org/10.7208/9780226511993

Mentel, L. (2014). mendeleev – a python resource for properties of chemical elements, ions and isotopes. https://github.com/lmmentel/mendeleev

Ouyang, R., Curtarolo, S., Ahmetcik, E., Scheffler, M., & Ghiringhelli, L. M. (2018). Sisso: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates. *Phys. Rev. Materials*, *2*, 083802. https://doi.org/10.1103/PhysRevMaterials.2.083802

Park, H., Mall, R., Alharbi, F. H., Sanvito, S., Tabet, N., Bensmail, H., & El-Mellouhi, F. (2019). Exploring new approaches towards the formability of mixed-ion perovskites by dft and machine learning. *Physical Chemistry Chemical Physics*, *21*(3), 1078–1088. https://doi.org/10.1039/C8CP06528D

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Perdew, J. P., Burke, K., & Ernzerhof, M. (1996). Generalized gradient approximation made simple. *Phys. Rev. Lett.*, *77*, 3865–3868. https://doi.org/10.1103/PhysRevLett.77.3865

Pu, W., Xiao, W., Wang, J.-W., Li, X.-W., & Wang, L. (2021). Screening of perovskite materials for solar cell applications by first-principles calculations. *Materials & Design*, *198*, 109387. https://doi.org/10.1016/j.matdes.2020.109387

Ray, P., Reddy, S. S., & Banerjee, T. (2021). Various dimension reduction techniques for high dimensional data analysis: A review. *Artificial Intelligence Review*, *54*(5), 3473–3515. https://doi.org/10.1007/s10462-020-09928-0

Shockley, W., & Queisser, H. J. (1961). Detailed balance limit of efficiency of pn junction solar cells. *Journal of Applied Physics*, *32*(3), 510–519. https://doi.org/10.1063/1.1736034

Stanley, J. C., Mayr, F., & Gagliardi, A. (2020). Machine learning stability and bandgaps of lead-free perovskites for photovoltaics. *Advanced Theory and Simulations*, *3*(1), 1900178. https://doi.org/10.1002/adts.201900178

Steiner, S., Khmelevskyi, S., Marsmann, M., & Kresse, G. (2016). Calculation of the magnetic anisotropy with projected-augmented-wave methodology and the case study of disordered $\mathrm{Fe}_{1-x}\mathrm{Co}_x$ alloys. *Phys. Rev. B*, *93*, 224425. https://doi.org/10.1103/PhysRevB.93.224425

Swanson, D. E., Sites, J. R., & Sampath, W. S. (2017). Co-sublimation of cdsexte1-x layers for cdte solar cells. *Solar Energy Materials and Solar Cells*, *159*, 389–394. https://doi.org/10.1016/j.solmat.2016.09.025

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Williams, L. (2022). *Sl3me – a python3 implementation of the spectroscopic limited maximum efficiency (slme) analysis of solar absorbers* (Version 1.0.0). https://github.com/ldwillia/SL3ME

Yan, Y., Yin, W.-J., Shi, T., Meng, W., & Feng, C. (2016). Defect physics of ch3nh3pbx3 (x = i, br, cl) perovskites. *Organic-Inorganic Halide Perovskite Photovoltaics*, 79–105. https://doi.org/10.1007/978-3-319-35114-8_4

Yang, J., Manganaris, P. T., & Mannodi Kanakkithodi, A. K. (2023). A high-throughput computational dataset of halide perovskite alloys. *Digital Discovery*. https://doi.org/10.1039/d3dd00015j

Yang, J., & Mannodi-Kanakkithodi, A. (2022). High-throughput computations and machine learning for halide perovskite discovery. *MRS Bulletin*, *47*(9), 940–948. https://doi.org/10.1557/s43577-022-00414-2

Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, *415*, 295–316. https://doi.org/10.1016/j.neucom.2020.07.061

Yin, W.-J., Yang, J.-H., Kang, J., Yan, Y., & Wei, S.-H. (2015). Halide perovskite materials for solar cells: A theoretical review. *Journal of Materials Chemistry A*, *3*(17), 8926–8942. https://doi.org/10.1039/c4ta05033a

Yu, L., & Zunger, A. (2012). Identification of potential photovoltaic absorbers based on first-principles spectroscopic screening of materials. *Physical Review Letters*, *108*(6). https://doi.org/10.1103/physrevlett.108.068701

Zhang, H., Nazeeruddin, M. K., & Choy, W. C. H. (2019). Perovskite photovoltaics: The significant role of ligands in film formation, passivation, and stability. *Advanced Materials*, *31*(8), 1805702. https://doi.org/10.1002/adma.201805702

Zhu, S., Ye, J., Zhao, Y., & Qiu, Y. (2019). Structural, electronic, stability, and optical properties of cspb1-xsnxibr2 perovskites: A first-principles investigation. *The Journal of Physical Chemistry C*, *123*(33), 20476–20487. https://doi.org/10.1021/acs.jpcc.9b04841

ADDITIONAL FIGURES

Learning Curves

Cross-validation within the training set is the only way of checking the generality of models during the grid search. Identifying the validation split size is necessary to obtain an understanding of how much data is needed to train a model that can generalize.

Learning curves are computed for each scorer. Notice that the error metrics are negated for consistency with the R^2 and ev scores; the greater the number, the better the model performs.

More data offers better chances. However, the smaller the split, the longer and more expensive the loop training becomes, e.g. 10-fold splits makes for 10 sample scores at each partition size. Meaning, 90% of the training set is used for actual training and the remaining 10% is used for validation and this is repeated 10 times.

Shuffling is performed prior to generating each fold. The shuffle is seeded with a deterministic random state to ensure scores are comparable across partition size

Feature Distributions

GPR SHAP analysis

SIS+RFR SHAP analysis

SIS+GPR SHAP analysis

Known Clustering in t-SNE Projections

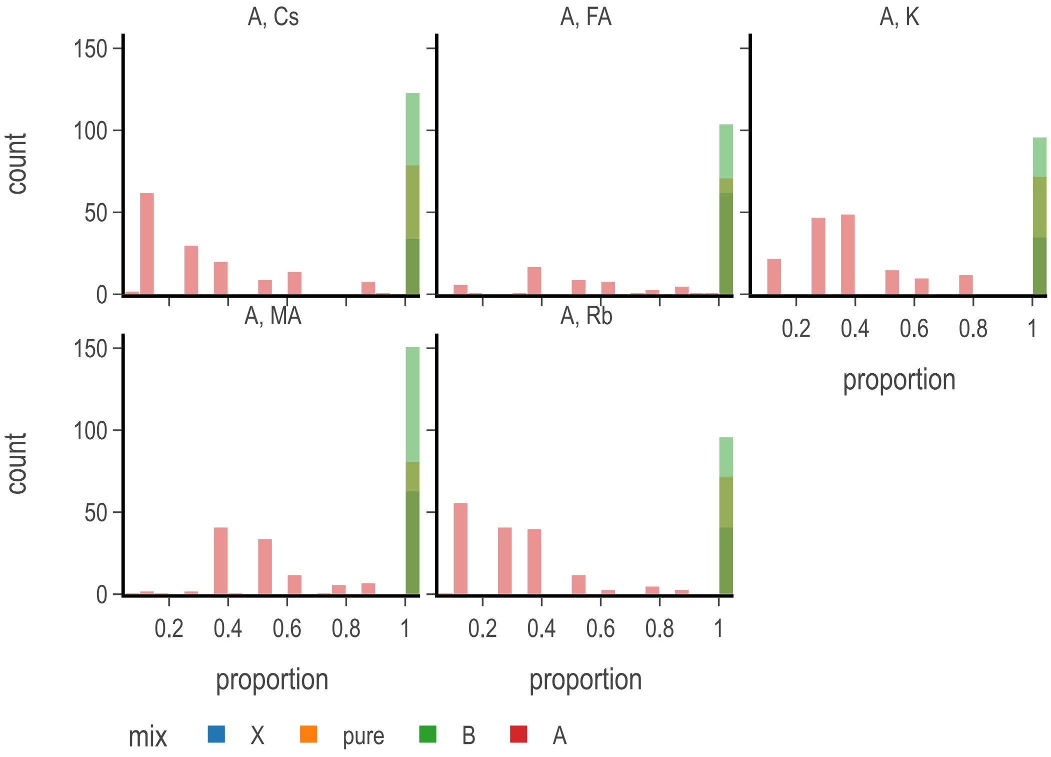

Figure 1. Normalized Distribution of A-site Constituents

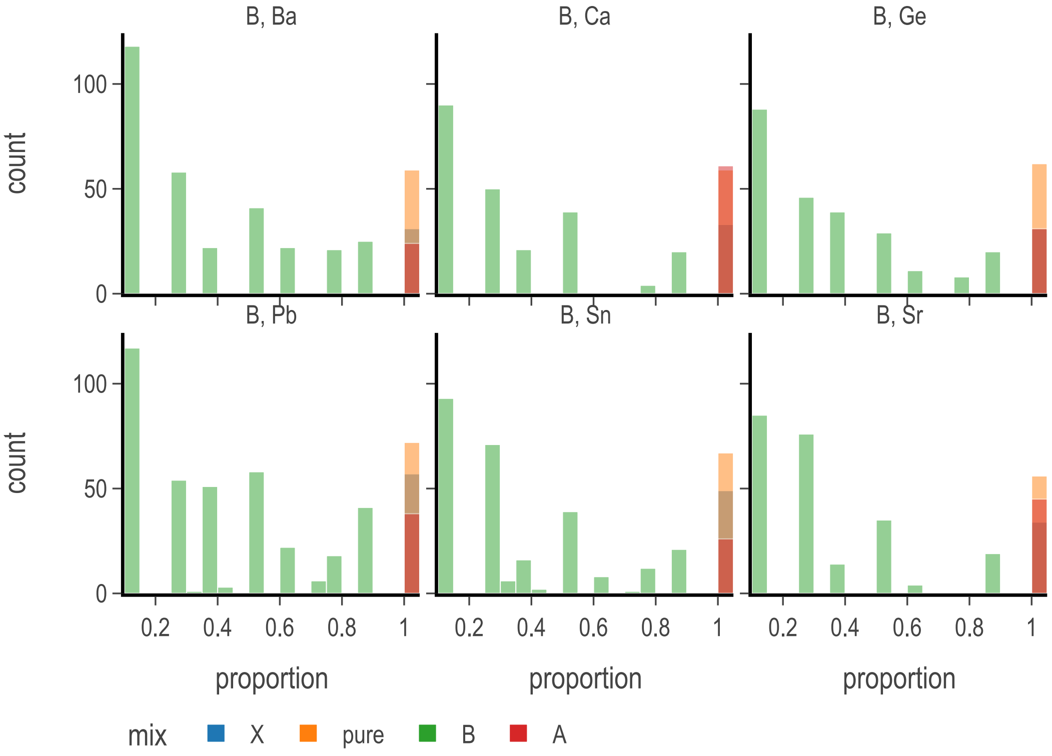

Figure 2. Normalized Distribution of B-site Constituents

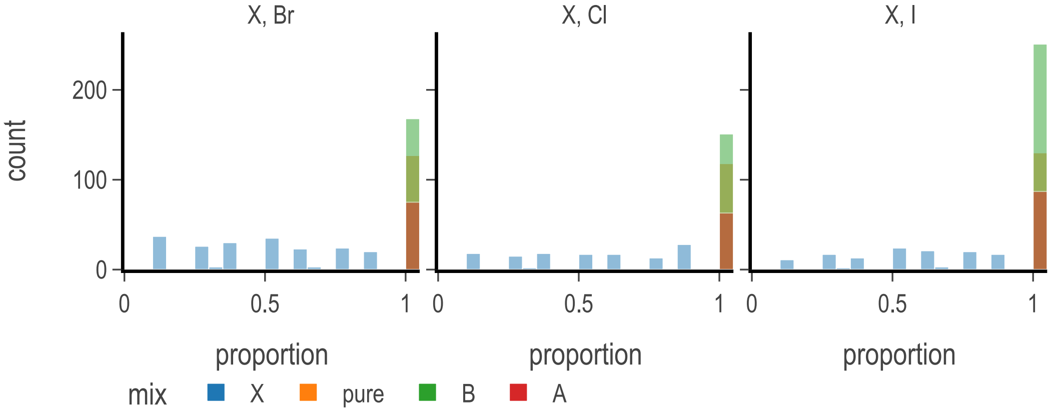

Figure 3. Normalized Distribution of X-site Constituents

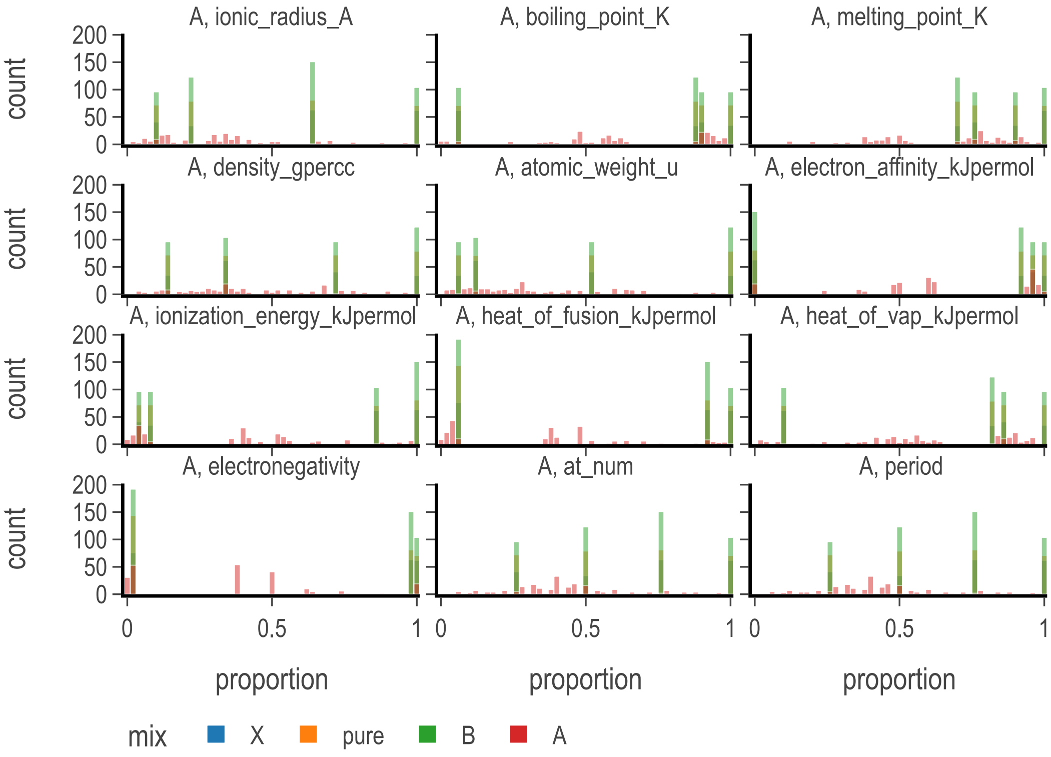

Figure 4. Distributions of Mean A-Site Properties

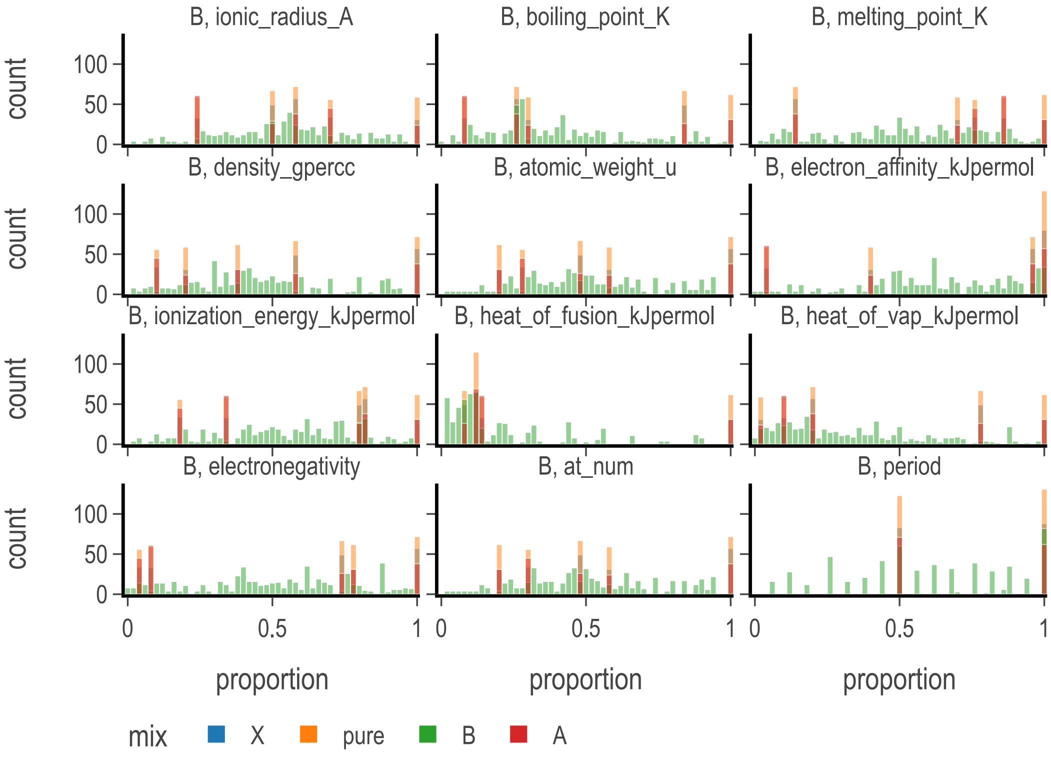

Figure 5. Distributions of Mean B-Site Properties

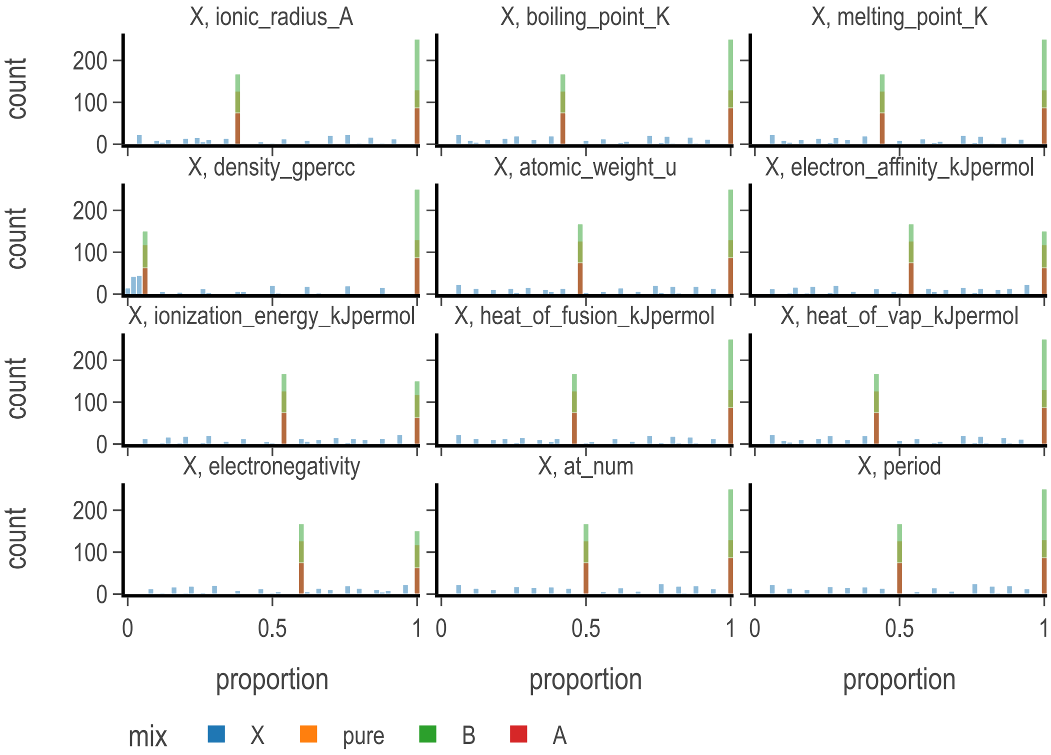

Figure 6. Distributions of Mean X-Site Properties

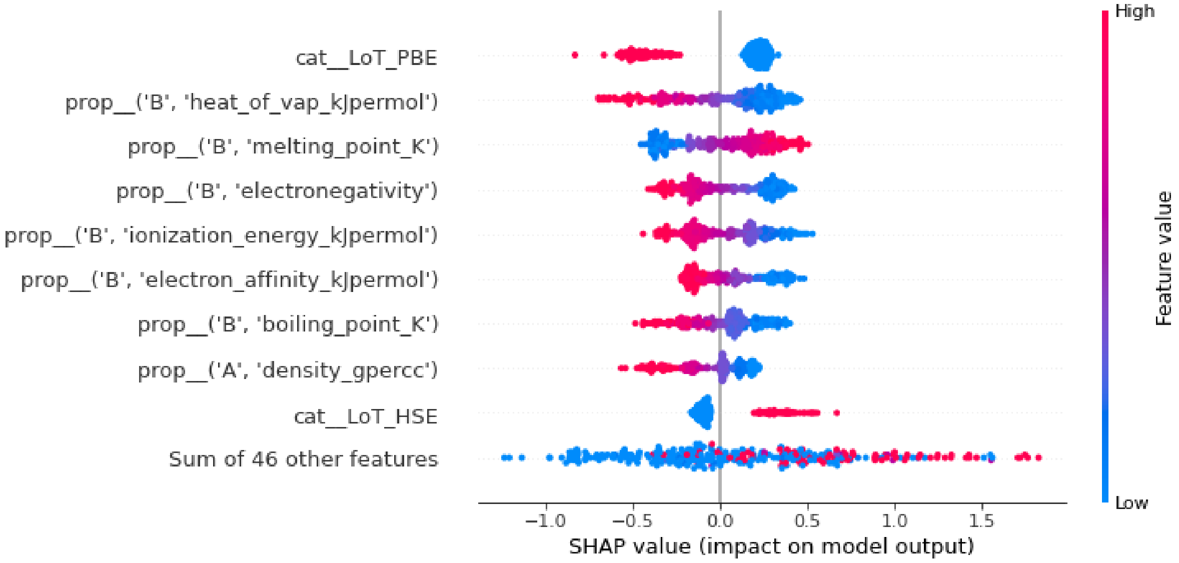

Figure 7. Gaussian Process Regression Band Gap SHAP Values

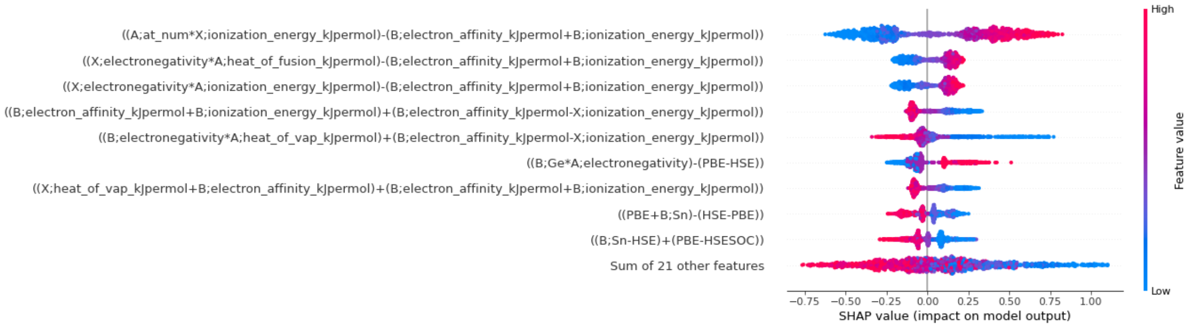

Figure 8. Random Forest Regression Band Gap on SIS domain SHAP Values

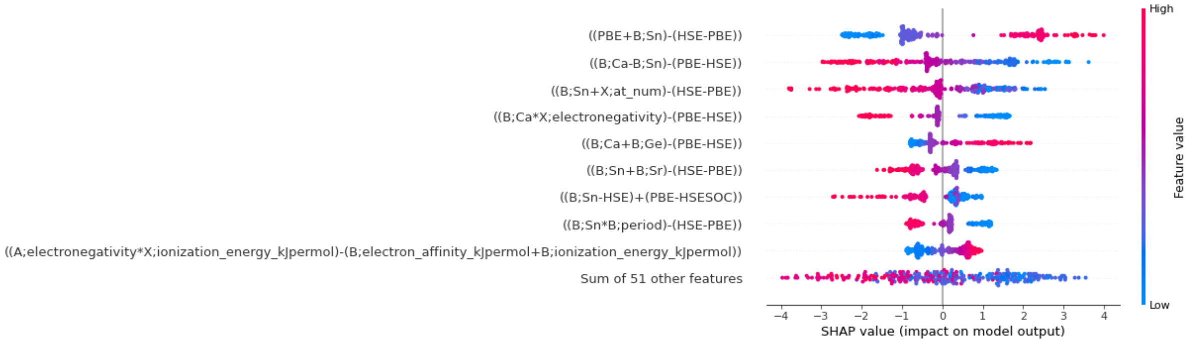

Figure 9. Gaussian Process Regression Band Gap on SIS domain SHAP Values

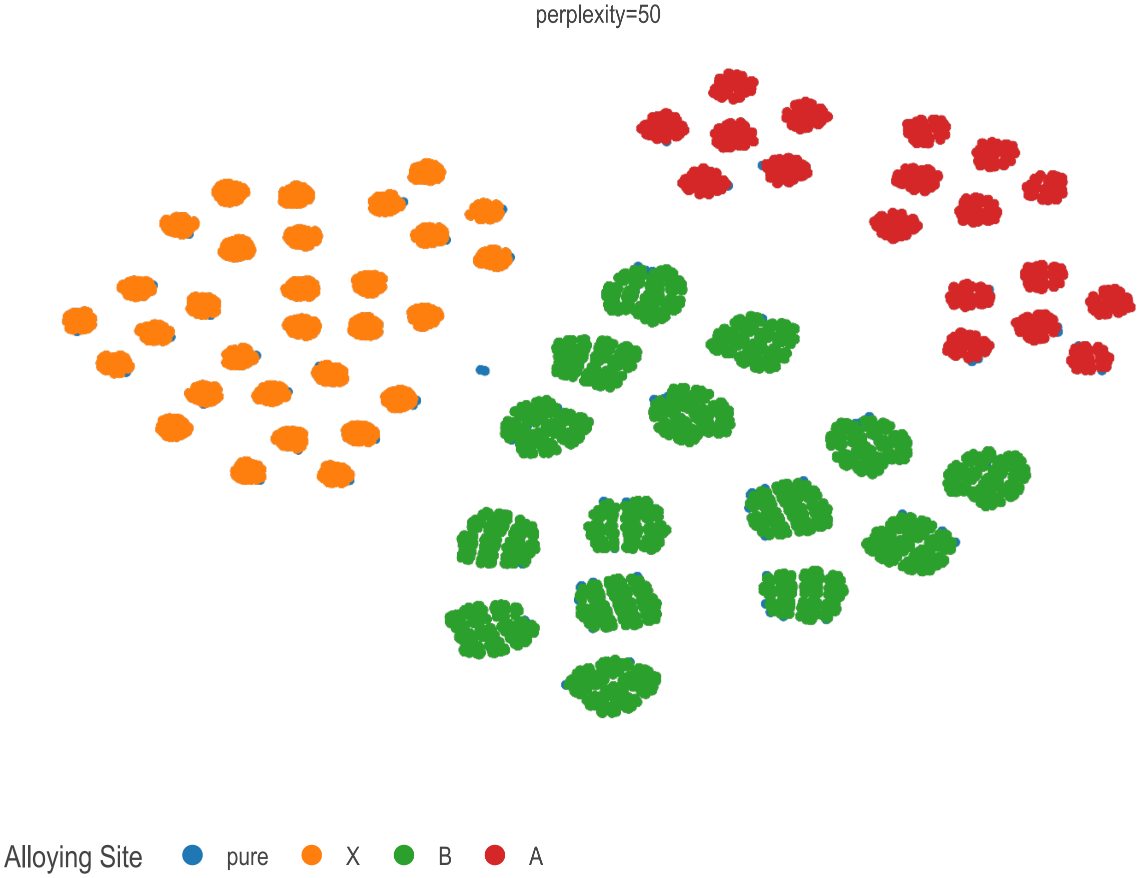

Figure 10. Projection of sample space via t-SNE overlaid with labels indicating site of mixing

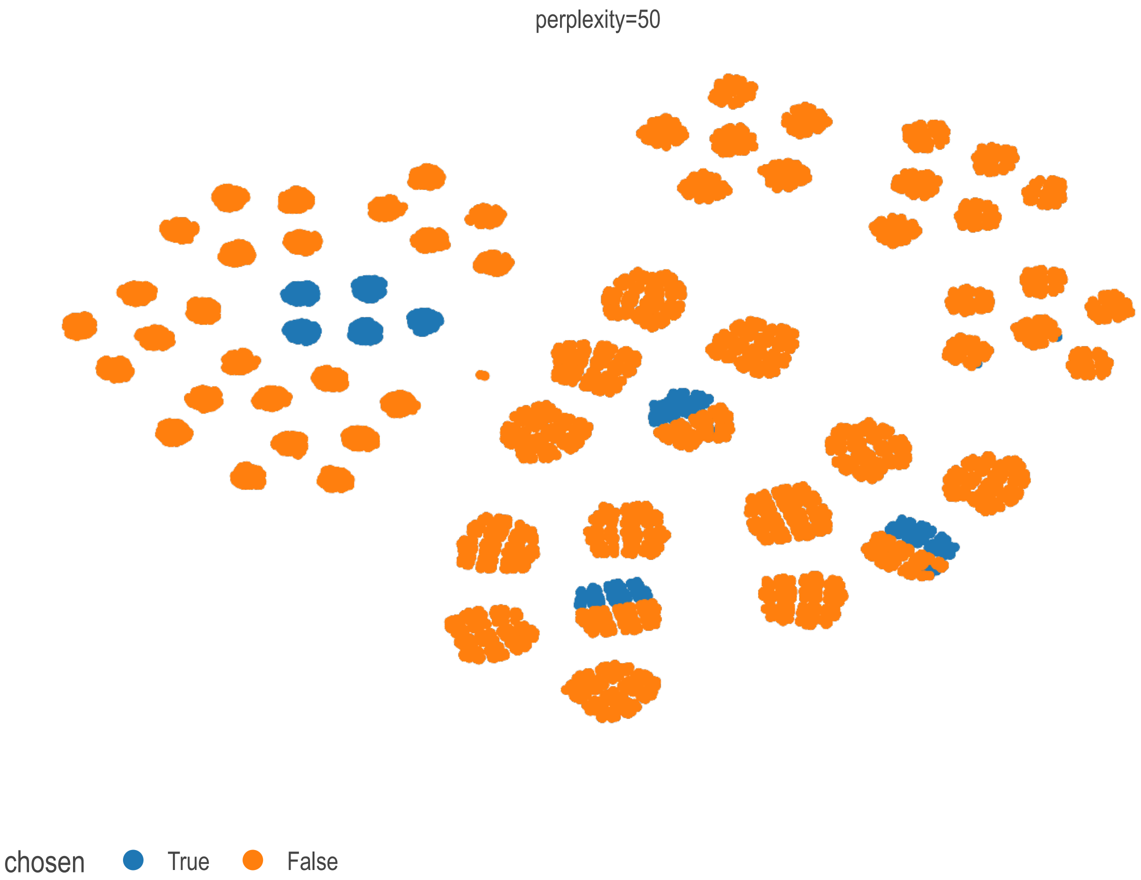

Figure 11. Projection of sample space via t-SNE overlaid with labels indicating presense of data points in screened subset