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

August 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:

Dr. Nikhilesh Chawla

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 and advice. 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 for which I am honored and thankful. The Rosen Center of Advanced Computing provided the computational resources needed to conduct simulations, store data, and train models. I deeply appreciate the opportunity to study, learn, and grow at Purdue University.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF LISTINGS

LIST OF SYMBOLS

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

ABBREVIATIONS

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

NOMENCLATURE

FA Formamidinium (Cationic Formamidine) $CH(NH_2)_2^+$

MA Methylammonium (Cationic Methylamine) $CH_3NH_3^+$

GLOSSARY

FAIR	Findable Accessible Interoperable and Reusable Data
K-fold split	Data partition divided into K arbitrary groups for use in cross-validation schemes
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
Materials Project	US Government-led multidisciplinary collaboration founded in 2011 as the Materials Genome Initiative.
Spin Orbit Coupling	An additional term intended to account for the increased relevance of quantum angular momentum to electromagnetic response in heavy atoms
cardinal mixing	Describes perovskite alloys where no more than one of the A, B, or X sites is occupied by multiple possible constituents
classical learning	a paradigm of machine learning that is dependent on expert knowledge to extract quality features from samples in a dataset
cross-validation	Method for gathering statistics on the abilities of a model to fit to the parent partition
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
features	attributes of an observed event or object which might empirically explain the event or object
groupwise K-fold	Data partition divided into K-folds where each fold corresponds to a category label
hyper-parameter	a setting that controls how a learning algorithm works
level of theory	Refers to the rank of a DFT functional in the hierarchy of phenomenological comprehensiveness. A proxy for accuracy.

machine learning	a science concerned with algorithms which improve their performance with exposure to new data
multi-task learning	A type of machine learning where an algorithm learns multiple functions simultaneously, while exploiting commonalities and differences between the functions
partition	Portion of sample data reserved for a purpose in model development
surrogate model	a representation which attempts to capture as much of the relationship between a domain and a target property as possible

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. This enables accurate, high fidelity 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 this, we identified 1251 stable compounds with band gap predicted to be between 1 and 2 eV at experimental accuracy, successfully narrowing the candidates 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. [1–4] 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. [5, 6] 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. [7] 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 acceleration. My research focused on accelerating this search through the development and application of data-driven design methods in the composition space of halide perovskites. The most recent work presented in this dissertation is being submitted for publication as P. Manganaris *et al.*, "Multi-fidelity machine

learning for perovskite band gap predictions," In Preparation, Jun. 2023. Publications from earlier work leading to this dissertation include J. Yang *et al.*, "A high-throughput computational dataset of halide perovskite alloys," *Digital Discovery*, 2023, ISSN: 2635-098X. DOI: 10.1039/d3dd00015j. [Online]. Available: http://dx.doi.org/10.1039/D3DD00015J, and P. Manganaris *et al.*, *MRS computational materials science tutorial*, en, 2022. DOI: 10.21981/D1J2-AR65. [Online]. Available: https://nanohub.org/resources/36041?rev=90.

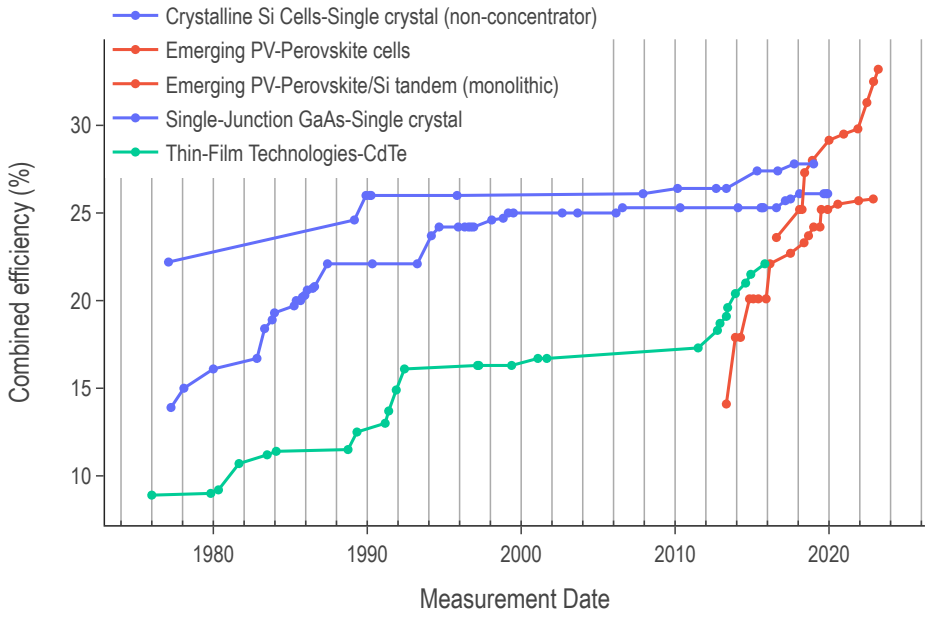

Figure 1.1. Rapid rise in cumulative maximum of HaP PCEs

1.1 Design Goals in Perovskite Absorbers

Perovskite properties may be tuned in various ways. The introduction of dopants and defects [11, 12] and the mutation of their cubic structure [13, 14] 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. [15–18] Mixing at the A-site has been shown to improve formability [19], while B-site and X-site mixing can tune and optimize band gap and therefore the maximum optical wavelength absorbable by the semiconductor. Band gap is the energy required to promote an electron from the valence band to the conduction band. Absorption of visible light from the plentiful green to red portion of the spectrum, by the simple relation $E_{bg} = E_{\gamma} = \frac{hc}{\lambda}$, corresponds to band gaps of 1-2 eV. 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.

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). [20, 21] 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 $2 \times 2 \times 2$ 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

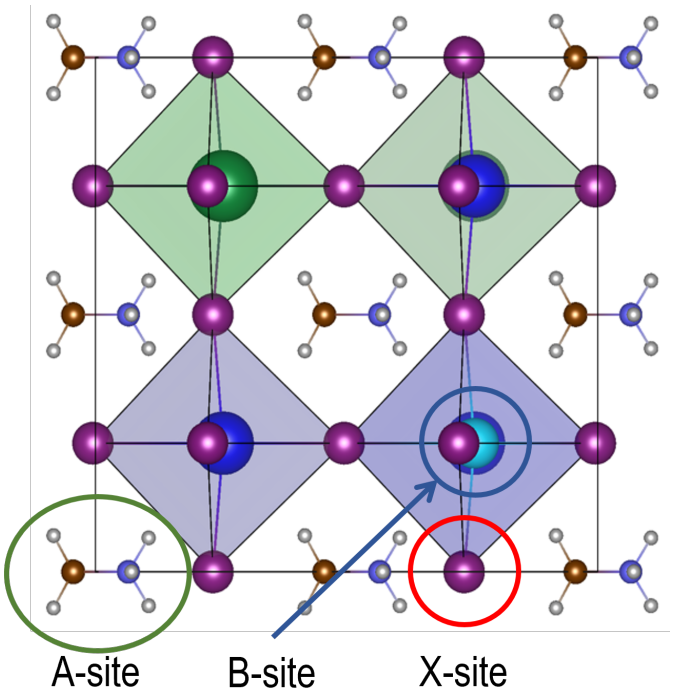

Figure 1.2. $2 \times 2 \times 2$ α-phase supercell with Methylammonium at the A-site

information, band gaps, optical absorption spectra, and defect formation energies of a variety of HaPs with reasonable accuracy. [2, 22] 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. [13, 23–25] 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. [24, 26, 27] Such DFT-ML models, once rigorously trained and tested, are deployed for high-throughput screening across massive sample spaces of unknown perovskites. [28]

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			

1.2 Multi-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. [9, 22] There are more calculations than there are distinct compositions because each composition is simulated multiple times to obtain results of varying fidelity. Also, it is supplemented by an additional ~100 points of data aggregated from reputable sources by Almora *et al.* [29]. In total it is one of the largest

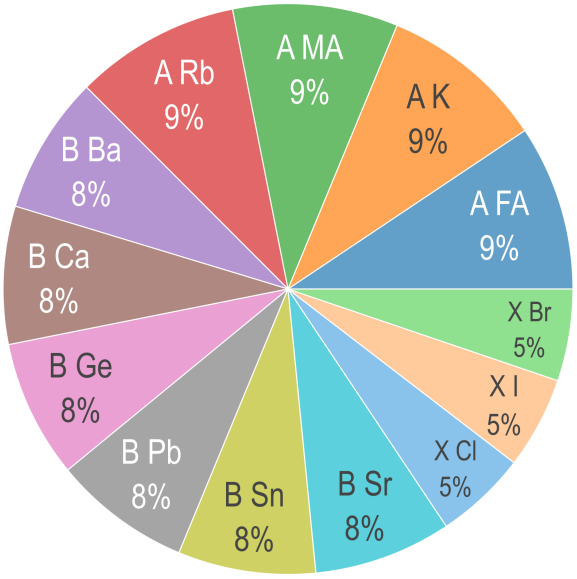

Figure 1.3. The cardinal mixing sample space contains equal fractions of each element

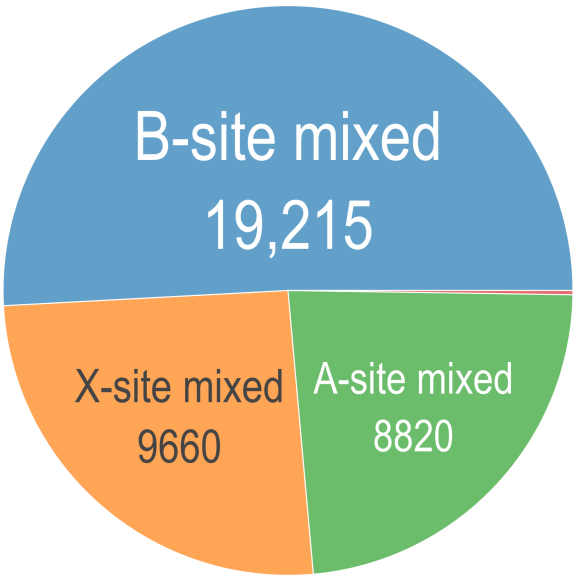

Figure 1.4. The cardinal mixing sample space contains mostly B-site mixed compounds

first principles cubic halide perovskite datasets and represents years of work. It provided an excellent foundation for my 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. [30] Various tolerance factors have been proposed in the literature to describe this constraint. Each tolerance factor is expressed as a function of the ionic radii of the species at each site r_A, r_B, and r_X. The Goldschmidt tolerance is defined as expression 1.1. [2] The octahedral Factor is defined as the simple ratio 1.2. A recent tolerance factor proposed by Bartel Christopher *et al.* [7] is defined as expression 1.3.

$$t = \frac{r_A + r_X}{\sqrt{2} * (r_B + r_X)} \tag{1.1}$$

$$o = \frac{r_B}{r_X} \tag{1.2}$$

$$b = \frac{r_X}{r_B} - \left[1 - \frac{\frac{r_A}{r_B}}{\ln(\frac{r_A}{r_B})} \right] \tag{1.3}$$

The approximate definitions $t \approx 1$, $o \approx 0.67$, and $b \lesssim 4$ quantify perovskite α-phase stability. [2, 7] These criteria are useful for efficiently evaluating if a proposed perovskite can be formed. They supplemented the cuts on band gap we developed for identifying high-performing Perovskite absorbers.

1.4 Thesis Overview

In chapter 2, I review the work we did in "A High-Throughput Computational Dataset of Halide Perovskite Alloys"[9] to create a multi-fidelity dataset of perovskite band gaps. I then detail the methods I used for predicting band gaps and the models I obtained in chapter 3. Results and a discussion of their significance is given in chapter 4. The open-source tools I developed for this work have been made available to the broader materials science community. The essential qualities of these tools have been detailed in Appendix B.

2. DENSITY FUNCTIONAL THEORY SIMULATION

All material properties are fundamentally a function of the motion of electrons and atomic nuclei. These fundamental building blocks of matter can be modeled as point particles in a many bodied system governed by an equation of motion. Due to the quantum nature of these particles, position and momentum are defined in terms of a complex-valued wave function that describes the probability of finding a particle at a point in space. Therefore, the equation of motion is the Schrödinger equation (2.1), the solution of which yields the energy of a single configuration of particles. For electronic and optical properties, the electron configuration is most significant, allowing for some simplification by using the Born-Oppenheimer approximation (BOA).

$$\imath\hbar\frac{\partial}{\partial t}\Psi(R_{3N}, r_{3n}, t) = \mathcal{H}(R_{3N}, r_{3n}, P_{3N}, p_{3n})\Psi(R_{3N}, r_{3n}, t) \overset{BOA}{\Longrightarrow} \hat{H}\psi(r_{3n}) = E\psi(r_{3n}) \quad (2.1)$$

The solution to the Schrödinger equation is unfortunately intractable for even tens of electrons due to the huge expense of integrating the $3n$ dimensional electron wave function $\psi(r_{3n})$. Density functional theory is the leading method for tractably computing the energy of many electron interactions in quantum systems. DFT is founded on the Hohenberg Kohn relation proving an electron configuration fully determines the potential energy due to the nuclear configuration. To illustrate, the electron density at a position is the expectation value of the position of all electrons not at that position. This density emerges when computing the expectation value of the potential energy of a single electron r_k

$$\langle\psi|\hat{V}_{nuc}(r_k)|\psi\rangle = \int \mathrm{d}^3 r_k \hat{V}_{nuc}(r_k) \int \Pi_{\mathrm{i}=1}^{k-1}\, \mathrm{d}^3 r_{\mathrm{i}} \Pi_{\mathrm{i}=k+1}^{n}\, \mathrm{d}^3 r_{\mathrm{i}} |\psi(r_{\mathrm{i}})|^2 = \int \hat{V}_{nuc}(r_k) n(r_k)\, \mathrm{d}^3 r = \hat{V}(n) \quad (2.2)$$

due to the nuclei. This expression is reversible, showing that for a given density there must be a unique $\psi(n(r))$ expressed as a functional of the density. With the establishment of this density functional, any property of interest may be found entirely in terms of the density. The major benefit of this re-framing is that the density functional lives in 3-D

space, whereas the wave function lives in 3n-D space. By circumventing the wave function, the many-electron problem solving for the configurational energy

$$E = \langle\psi|\hat{H}|\psi\rangle = \langle\psi(n)|\hat{T}(n) + \hat{U}(n) + \hat{V}(n)|\psi(n)\rangle \implies E(n) = \hat{T}(n) + \hat{U}(n) + \hat{V}(n) \quad (2.3)$$

can be effectively cast to a sum of single-electron problems with theoretically no compromise on accuracy. Minimizing $E(n(r))$ yields the ground state energy and ground state electron configuration. The electron-electron interaction $\hat{U}(n(r))$ lacks a perfect solution and requires approximation. The system potential energy $\hat{V}(n(r))$, as discussed, is uniquely defined in 3D space by the nuclei in the system. The kinetic energy operator $\hat{T}(n(r))$ in combination with the system energy is used to derive the Kohn-Sham equations (2.4).

$$\left(-\frac{\hbar^2}{2m}\nabla + \hat{V}_s\right)\psi(r_k) = \epsilon_k\psi(r_k) \quad (2.4)$$

The solution for these single particle wave functions yield the electron density $n(r) = \sum_{i=1}^{n} |\psi_i(r)|^2$ necessary to solve (2.3). The single-particle potential $\hat{V}_s$ is the sum of the system energy, the electron-electron Coulomb repulsion, and the so called "exchange correlation." This latter functional is what makes a solution for $\hat{U}$ so elusive. The more physics accounted for in the approximations used for exchange correlation, the better the accuracy but the greater the cost. So these approximations are ranked by "level of theory." My work is necessary because perfect simulations do not exist due to the conceptual difficulty in efficiently modeling the exchange correlation. For the remained of this chapter I present how DFT was used to create the multi-fidelity dataset that enabled my work.

All DFT computations were performed using VASP version 6.2 [31, 32] employing projector-augmented-wave (PAW) pseudo-potentials. [33–35] 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 was simulated in a $2 \times 2 \times 2$ supercell, which allowed 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 restricted X-site mixing to fractions of 3/24. For simulating mixed perovskites, the special quasi-random structures (SQS) method was applied to build periodic structures that made the first nearest-neighbor shells as similar to the target random alloy as possible. [36] 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. [37, 38] The energy cutoff for the plane-wave basis was 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 was 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. [39, 40] 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. [41] 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 $2 \times 2 \times 2$ Monkhorst-Pack mesh. These four measures E_{bg}^{PBE}, E_{bg}^{HSE}, $E_{bg}^{HSE(SOC)}$, and $E_{bg}^{HSE\text{-}PBErel(SOC)}$ populated the multi-fidelity dataset I used for my work. I aimed to accurately predict the band gaps of entirely hypothetical HaP compounds at the experimental fidelity using multi-fidelity models. This is motivated by the fact that the absorption spectrum of a perovskite determines its performance in a photovoltaic. [42] A well defined relationship therefore exists between the efficiency of a perovskite absorber and its true band gap. [43] The actual band gap (measured at the experimental fidelity) thus strongly predicts photovoltaic performance, and analysis of this relation informs the screening criterion. My work aims to accelerate the design of high-performing photovoltaic devices by enabling more accurate rapid identification of candidate perovskites with desirable band gap properties.

2.2 Spectroscopic Limited Maximum Efficiency (SLME)

Introduced by Yu and Zunger [43], the SLME is a convenient metric for evaluating a semiconductor's suitability for single junction photovoltaic (PV) absorption. In this work, SLME was calculated considering a 5ţm sample thickness for every perovskite using equations 2.5, 2.6, and 2.7, combining the original SL3ME.py code from Yu and Zunger [43] with our DFT computed absorption spectra and band gaps.

$$a(E) = 1 - \mathrm{e}^{-2\alpha(E)L} \tag{2.5}$$

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.

$$J = \mathrm{e} \int_0^\infty a(E) I_{sun}(E) dE - J_0(1 - \mathrm{e}^{\frac{eV}{kT}}) \tag{2.6}$$

$$\eta = \frac{P_m}{P_{in}} = \frac{\max(J \times V)}{P_{in}} \tag{2.7}$$

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. [44] This calculation was 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. [43, p.1] Experimental data [29] broadly agrees with PBE simulation, so the range of 1 to 2 eV was justified (see figure 2.1). Also, notice that even in just the sample dataset, there were 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 conducted 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.

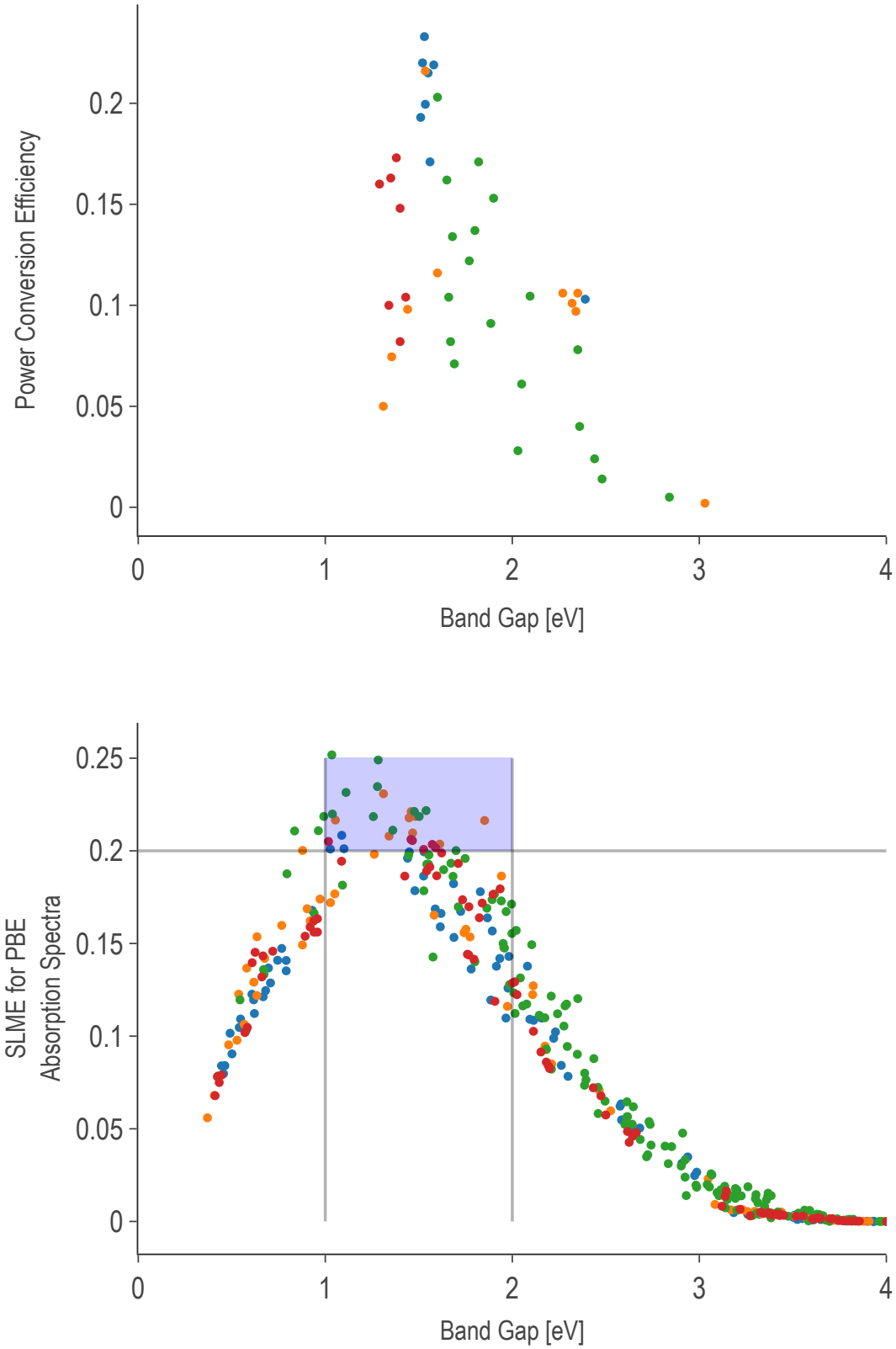

Figure 2.1. PBE SLME of sample compares to experimental PCE and cleanly demarcates competitive range of band gaps

The range of band gaps sampled by each simulation method are similar and are characterized by similar variance. The descriptive statistics of each greatly exceeded those of the experimental subset (see figure 2.2). Nevertheless, the latter undoubtedly represented the smallest error from truth. The types of mixing per level of theory were apportioned as in figure 2.4. This is the primary challenge I addressed with the multi-fidelity models discussed in chapter 3.

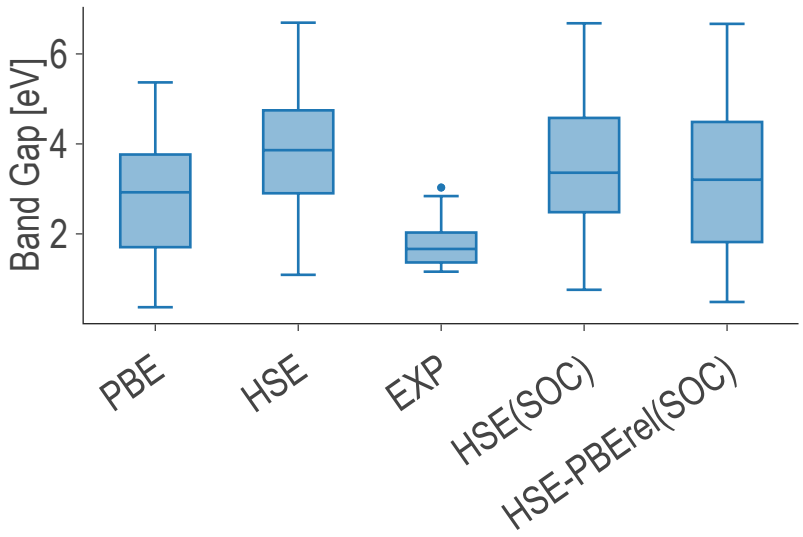

Figure 2.2. Variability in sampled band gaps at each fidelity

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. [22, 42]

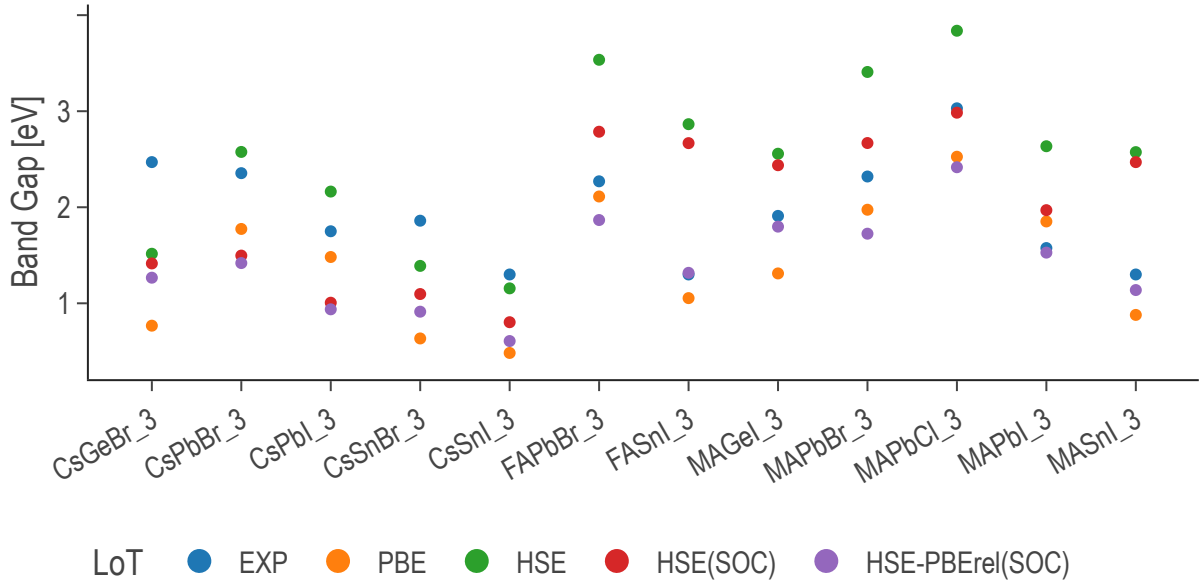

Figure 2.3. Effect of level of theory on band gap measurement

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 were 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 was 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 helped 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.

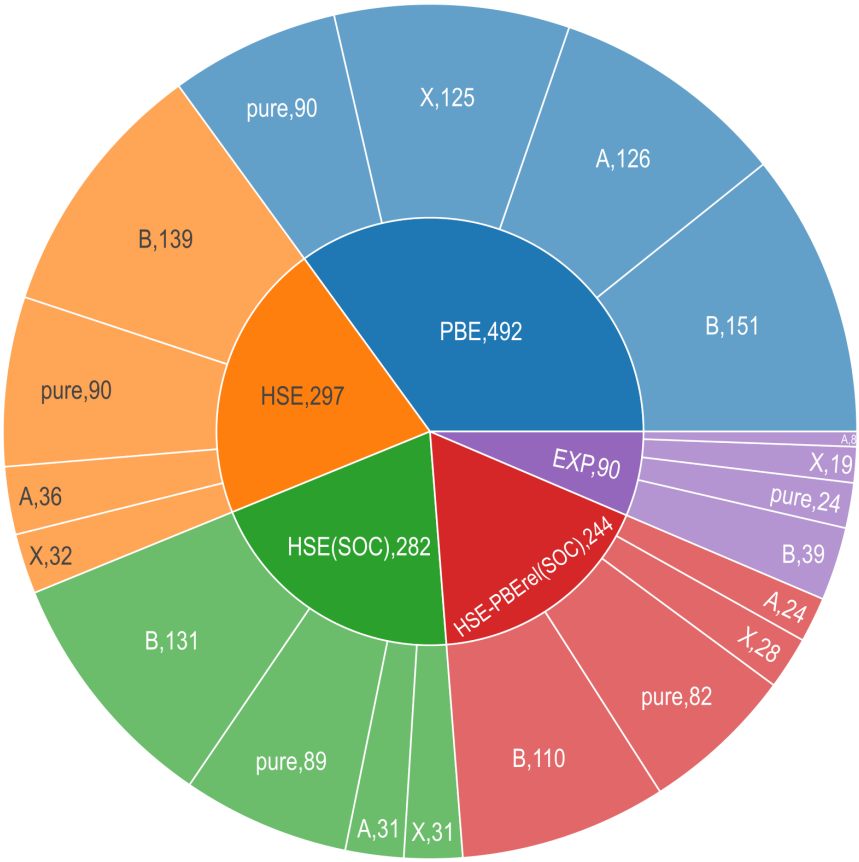

Figure 2.4. Share by count of total data apportioned from each experimental subcategory

Most importantly, this sample gave even coverage of the cardinal mixing domain as shown in figure 2.5. The clusters in this figure were 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.

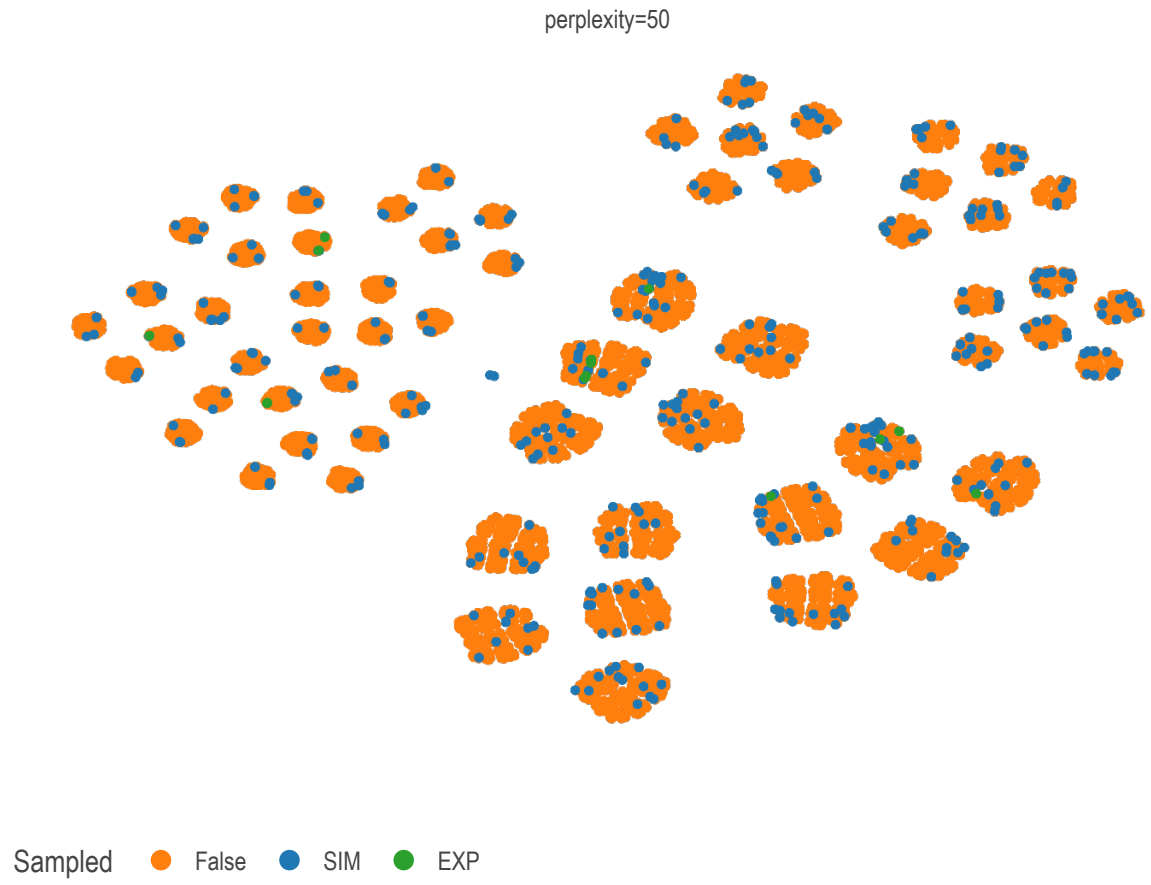

Figure 2.5. Samples overlaid on cardinal mixing chemical domain projected from fourteen to two dimensions via t-SNE

Our multi-fidelity computational halide perovskite alloy dataset generated with the methods described here 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. It serves as the foundation for the modeling work presented in the following chapter.

3. MODELS OF PEROVSKITE BAND GAP

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 ML modeling using unique material descriptors, are key to achieving this goal. I used 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.

The fidelity hierarchy in our data sample climbs from DFT simulations performed using the basic PBE GGA functional, to results obtained from physical experiments aggregated in literature. [29, 45, 46] 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. [22] 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. [47]

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. [48] 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

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)
```

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 cross-validated using a suite of regression scoring metrics applied to each LoT subset 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 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 considerable tuning of 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 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 $\alpha + \beta + \gamma = 14$. [28] In a subset of the data, the chemical vector (listing 3.2) is produced using cmcl (listing 3.1).

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

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 interpolating the properties of irregularly mixed large supercells. [22] However, a spare descriptor is generally bad for extrapolative modeling. [49]

When extrapolation is the aim, continuously distributed, unique, and linearly independent features are much more reliable. [50]

Our attempts to provide a domain with these characteristics results in a raw feature space containing the following. Fourteen sparse composition vectors extracted from chemical formula using `cmcl`[2]. See figure 3.1. Thirty-six dense site-averaged-property vectors computed as a linear combination of composition vectors and measured elemental properties. [47] See figure 3.2. Finally, 5 categorical dimensions one-hot-encoding level of theory. This provides the categorical axis for multi-task learning. See table 2.1.

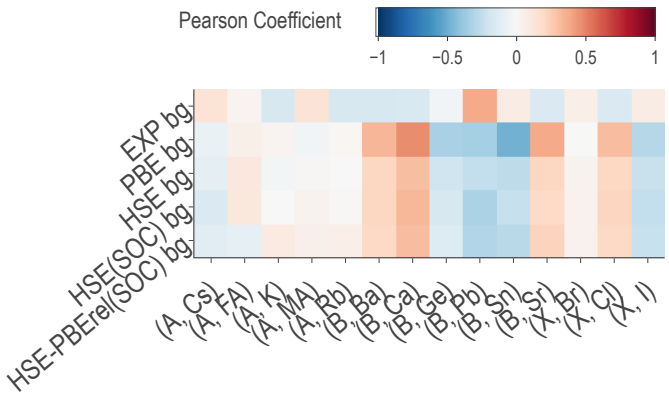

Figure 3.1. How composition vectors correlate with target bandgaps

3.3 Machine Learning Algorithms and Scoring Methodology

I trained Random Forest Regression (RFR) and Gaussian Process Regression (GPR) models to predict the band gap from the union of predictor features previously discussed. I chose to use the implementations of each of these algorithms packaged with the SciKit-Learn v1.2 package for python. [51] The hypothesis sets offered to the task by each of these

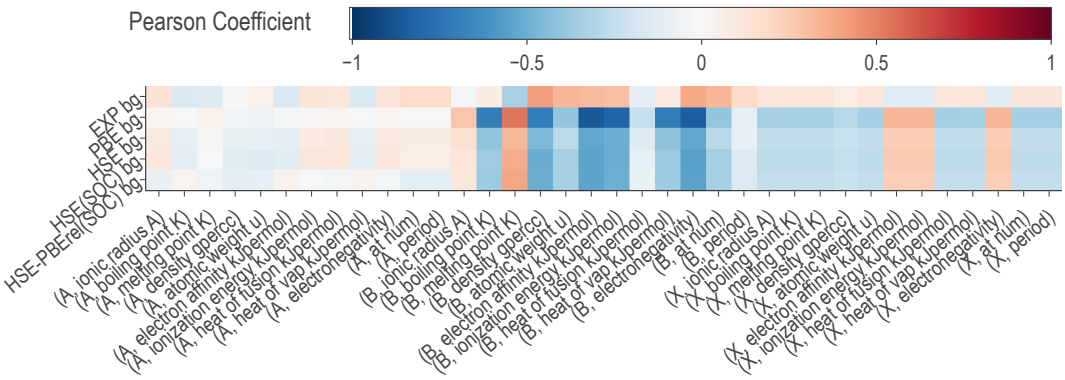

Figure 3.2. How site-averaged-properties vectors correlate with target bandgaps

architectures differ dramatically. A RFR is a flexible nonlinear ensemble model consisting of decision trees, a simple algorithm that captures interactions between descriptors in the form of a comparative algorithm. Each tree is trained on a random subsample of the training data, resulting in different algorithms. The RFR prediction is made by following the algorithm of each tree to the target and averaging the results. The advantage of this approach is that each tree is highly biased to the data allowing even minute outliers to be accounted for. However, by using many of them, the variance in data can also be explained, thus reducing the variance of error. Naturally, the RFR benefits from using as much data as possible with as many estimators as possible.

A GPR model is a principled linear model functioning very differently. Informally, GPRs "remember" the training examples and judge unlabeled data by its similarity to that aggregate memory. This is implemented as a kernel method leveraging some similarity function $k(x, x')$. It works out that This function defines a "universe" of functions with varying characteristics and, simultaneously a density of functions which can be interpreted as a Bayesian prior on function space. Naturally, kernels require engineering to accommodate prior expectations. Additionally, they offer limited options for capturing non-differentiable stepwise functions. Of course, the random forest offers nothing but stepwise functions thus posing as the opposite end of the two extremes. By applying Bayes' rule with a likelihood function defined using the sampled data and solving for the posterior function distribution, a principled set of predictive functions is obtained. Averaging these functions yields the mean prediction (identical to a Kriging model's ridge) and their variability gives principled estimates of the error at each point in the domain. Due to this, the algorithm saturates after sufficient training and additional data ceases to benefit. The primary advantage of this method is simply that it works for any two quantifiably similar x, potentially vectors, text, or graphs. This offers some possibilities for future research with enhanced features. However this comes at the cost of $\mathcal{O}(N^3)$ training time complexity and a break down in efficacy in sparse, high-dimensional spaces.

In order to monitor the performance of the regression during training nine metrics were used simultaneously to evaluate performance with respect to each fidelity and in the overall with attention to overall accuracy and maximum inaccuracy. These scores were used

throughout the hyper-parameter optimization to judge which parameters resulted in the best validation performance. To train models to be more faithful to the highest fidelity, the score for that subset was weighted as more important. So, eventually, only models that performed uniformly well on all alloy types and better on predicting the experimental dataset were selected. From the various approaches tried, the best optimized model was selected to make experimental-quality predictions on all 37785 points in the sample space. This procedure in demonstrated in an online notebook by Manganaris *et al.* [10] hosted on the Purdue nanoHUB.

3.4 Feature Engineering

There has been success in creating analytical expressions for perovskite properties, particularly lattice parameters. [52] In an attempt to find an analytical predictor for band gap we employ the Sure Independence Screening and Sparsifying Operator (SISSO). [53] SISSO is a generalization of "greedy pursuit" algorithms previously used for this purpose, namely orthogonal matching pursuit (OMP) and the Least Absolute Shrinkage and Selection Operator (LASSO) otherwise known as basis pursuit. [54] The SIS[3] operator is a powerful application of compressed sensing and used to find a conceptually orthogonal basis of compound features that best explain the signal in some function. [55] The SO is a potent dimensionality reduction, it does not perform any mathematical decomposition but instead picks existent dimensions that begin to approximate an orthogonal basis set. It outperforms CUR decomposition by functioning effectively in extremely high rank vector spaces. [56, 57] This is accomplished by posing the decomposition as a compressed sensing problem in the correlation metric space. Together, these operators allows the program to effectively find candidates for a linearly independent basis in a vector space of immense size. Unlike legacy techniques it does not suffer when features are correlated. [54, 58] This high performance handling of highly correlated vectors makes it particularly appealing for use with the perovskite features. The features illustrated in figures 3.2 are derived from those in figure 3.1.

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

Those composition vectors, due to the fact that they represent a unit formula are themselves correlated. At the least, they trace a bounded space.

A full SISSO model produces a parsimonious model of the target property which is easy to interpret. Subsequent applications of the SISSO operator to the residuals of the previous model serve as a clever interrogation of error[59] yielding additional terms that, at the cost of simplicity, better explain the target. Notice, however, the large space of features contains more good explanatory features than are used in the final expression. I extensively modified a SciKit-learn compliant [60] interface[4] to the SISSO program originally developed by Matgenix[5] for the purpose of better leveraging these cut explanatory features. The goal of this approach was to overcome the limitations of the raw feature space by finding a basis of varied, unique, and descriptive features which could serve as the domains for more powerful estimators. I planned to use this strategy to train SIS-augmented versions of the RFR and GPR models previously discussed. I additionally hoped that this could help to cut down the total number of descriptors necessary, especially the sparse features. To improve the interpretability of these, the SIS algorithm was restricted to combining raw features in ways that preserved their units, so to preserve overall interpretability. SIS features complexity was restricted to a maximum of 3 operations primarily to encourage parsimonious descriptions. The operations in table 3.1 were used to create compound features.

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

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

3.5 Training and Evaluation Methodology

First, the sample set is shuffled to mitigate the models tendency to fit on sampling order. Model training proceeded only after partitioning the dataset using an 80/20 train-test split. The split was made in a stratified manner, ensuring both partitions contained a proportionate fraction of samples from each fidelity subset. The test set of 282 points was held out for final evaluation. The training used the remaining 1123 data points. In order to specify the learning algorithm best at predicting the experimental fidelity from simulated data, a thorough grid-search of each algorithm's hyper-parameters was performed. First an estimator pipeline was constructed as in figure 3.3. Any sparse vectors were made dense and NaNs were filled with zeros by a `SimpleImputer`. The feature vectors were subject to l1 normalization so that the compound's stoichiometry converted to ratios. The estimator at the end of the pipeline was instantiated with default parameters and later had optimal settings injected.

To find these settings, I opted for a K-folds validation strategy first necessitating an optimal value for K be found. This was done empirically, first by generating learning curves with $K = 10$ for each estimator. The knee on these plots indicates the minimum number of samples needed to train effectively on average. The size of one fold would be the size of the validation set, with the remaining folds used in training. So, I doubled the number of samples at the knee, subtracted it from the size of the total training partition and used the result to determine the size of one fold. This method resulted in setting $K = 3$ for GPR training and $K = 4$ for RFR training.

Finally I performed hyper-parameter optimization by exhaustive grid search using methods discussed in section 3.1. Optimizing a model this way was expected to result in a good ability to interpolate perovskites properties in the well covered sample space detailed in figure 2.5. Ideally, the resulting model need only extrapolate over the one-hot-encoded LoT dimensions. To confirm this, validation requires two approaches.

In order, the first approach was critical to confirm the model was capable of predicting properties of entirely unseen compositions. A modified method of leave-one-out cross-validation was used to understand the distribution of possible errors. The optimally

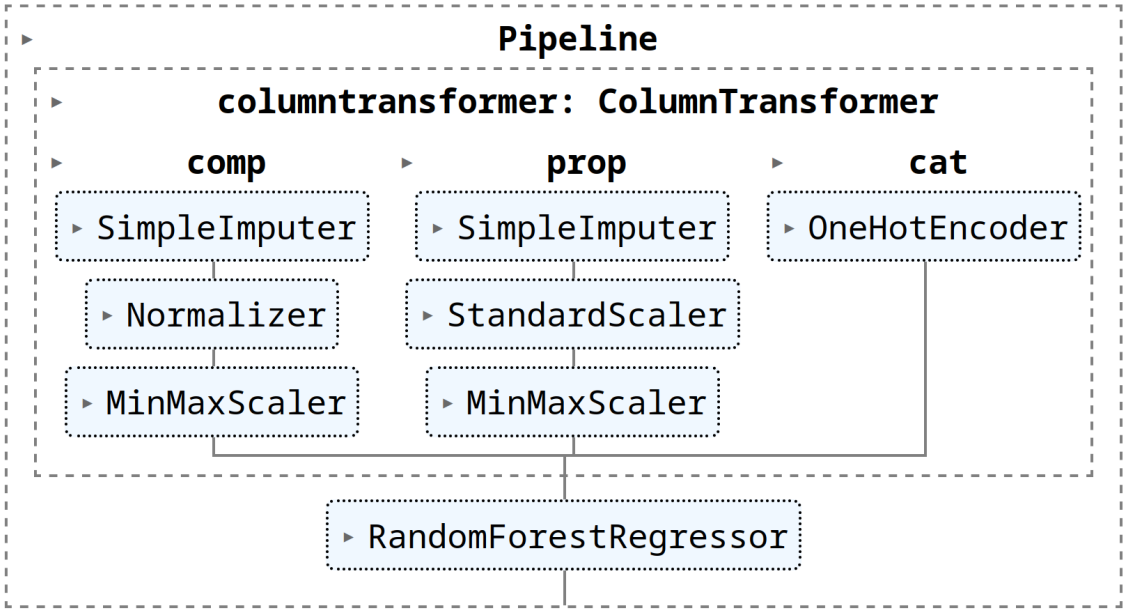

Figure 3.3. SciKit-Learn pipeline terminating in a default random forest estimator

parametrized model was retrained on a dataset derived from the train set where one particular combination of elements is not present at all in any stoichiometry and tasked to make predictions on a validation set of the excluded compounds. To be clear, individual elements may be represented in the training set as long as they don't appear together in a single compound. This strategy better tested the model's ability to comprehend novel compounds which it saw many of in the total sample space. Additionally, this avoided the expense of true leave-one-out validation. The second validation was to test that the model can accurately predict the values for specific unseen elements at different levels of theory. In this case, test compounds need not be entirely unique because it was expected that the extrapolated predictions at higher levels of theory will leverage the better coverage of lower fidelity sample sets. So, finally, the test set is used and the resulting predictions are evaluated for accuracy. Shapley Additive Explaination (SHAP) analysis of the models lends insight to the average physical impacts of 1) site-specific alloying, 2) using organic molecules in the Perovskite superstructure, and 3) the distribution of effects that a level of theory has on the prediction.

3.6 Results

3.6.1 Best Models on Raw Domain

Following HPO these models are finally validated against the test sets originally split off from the sample for both their extrapolative ability and consistency. The random forest model was the best performing compared to the Gaussian process and SISSO regressions. The analysis of its ability to make extrapolative predictions on completely unknown compositions is discussed in the following section 3.7. The RFR boasts an RMSE error on the total dataset of only 0.12 eV. This combined with its RMSE error of 0.15 eV on the experimental fidelity subset promises this model can make quality predictions of the band gap at the experimental fidelity. See table 3.2. This error compares favorably with the best models currently ranked by the Materials Project's MatBench standard. [61] Notably, these models use only the stoichiometry of the compositions while most MatBench models require atomic structures in addition to stoichiometric information. Of course, these models are highly spe-

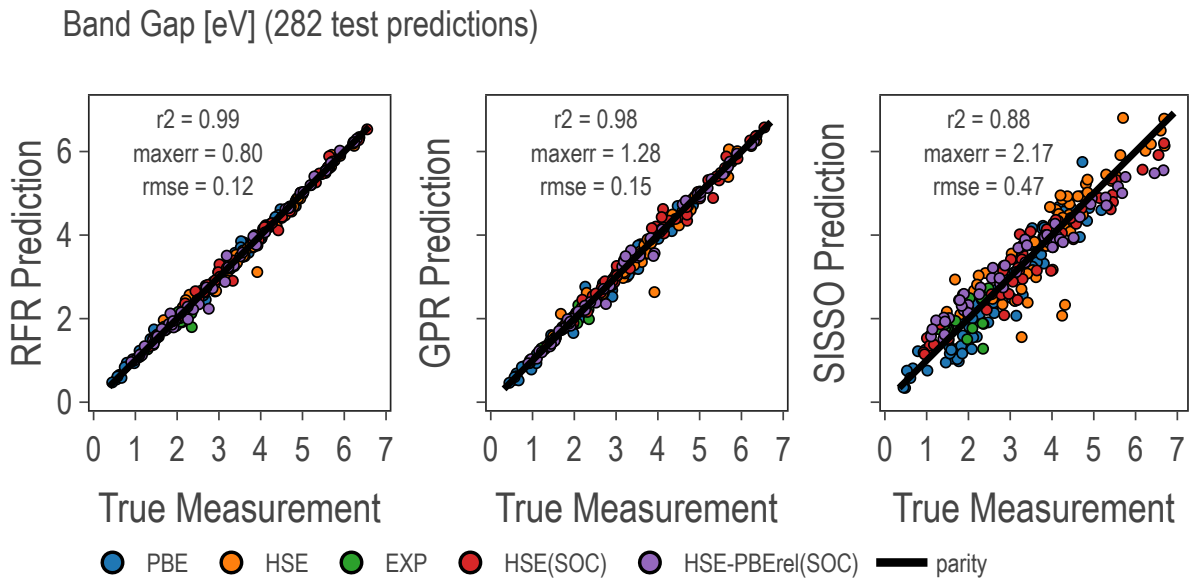

Figure 3.4. model predictions vs true values at multiple fidelities

cialized to this sample space. The RFR hyper-parameters are listed in the appendix (Table A.1).

The GPR model was tried with various kernels, both stationary and non-stationary, where each essentially describe the differentiability of the functions used in the posterior. Ultimately, the best was a non-stationary Matern kernel with $\nu = \frac{3}{2}$. This kernel defines "rough" functions that are only once-differentiable. This makes sense considering the inability of GPR to capture step functions and the stepwise nature of the LoT encoding dimensions. Additionally, both the LoT dimensions and the composition vectors are sparse, which challenges the algorithm Nevertheless, it is a close second to the RFR and offers a propitious start to models utilizing more advanced features.

3.6.2 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. [53, 55]

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.

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

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	0.15	0.12	0.49	0.47	0.25	0.18
EXP	0.12	0.15	0.30	0.33	0.33	0.23
PBE	0.12	0.10	0.47	0.39	0.17	0.13
HSE	0.21	0.15	0.55	0.51	0.30	0.20
HSE(SOC)	0.15	0.10	0.53	0.57	0.27	0.22
HSE-PBE(SOC)	0.13	0.13	0.46	0.47	0.25	0.18

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.

3.6.3 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 features obtained in section 3.2.

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 The-

ory. 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.

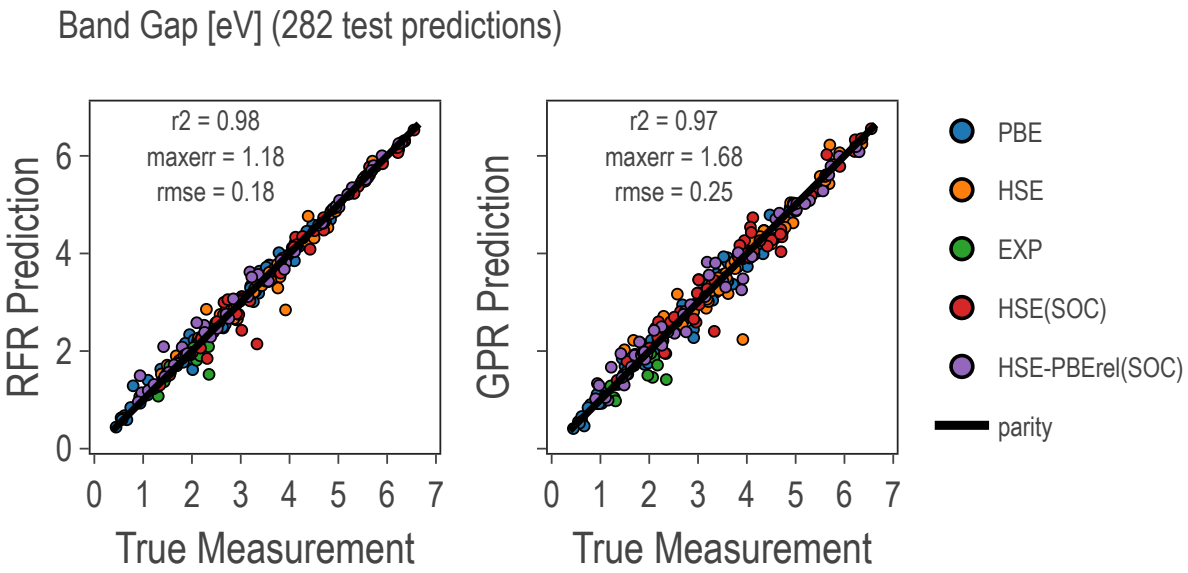

Figure 3.5. SIS-based model predictions vs true values at multiple fidelities

3.7 Discussion

3.7.1 Validation of Expected Error

The errors reported in the prior section are promising but questionable due to the lack of testing of the model's ability to make predictions on completely novel compounds. Following the appropriate validation methodology outlined in section 3.5 addresses this concern. This results in a distribution of errors with a mean EXP RMSE of about 0.2 eV, which is only slightly worse than error obtained in the testing. The mean scores are better than the median, but this is mostly due to a small number of outlier compounds which are not very exotic, so it is unrealistic not to train on at least one example of them at a simulation

fidelity. See figure 3.6. There is definitely a loss in validation performance as compared to the scores on the train set, see table 3.3. Notice, the R^2 score is missing because it can not be computed on validation sets that contain only a single sample. The explained variance (ev) score is poor on the test set, but nearly perfect on the train set. This is the biggest difference in the scoring by far, simply due to the higher variability of errors in the test sets. Certainly, the interpolation demanded of the model will not be perfect on wholly unseen compositions, but it seems that in the majority of instances, the prediction can be justifiably expected to be reasonably accurate.

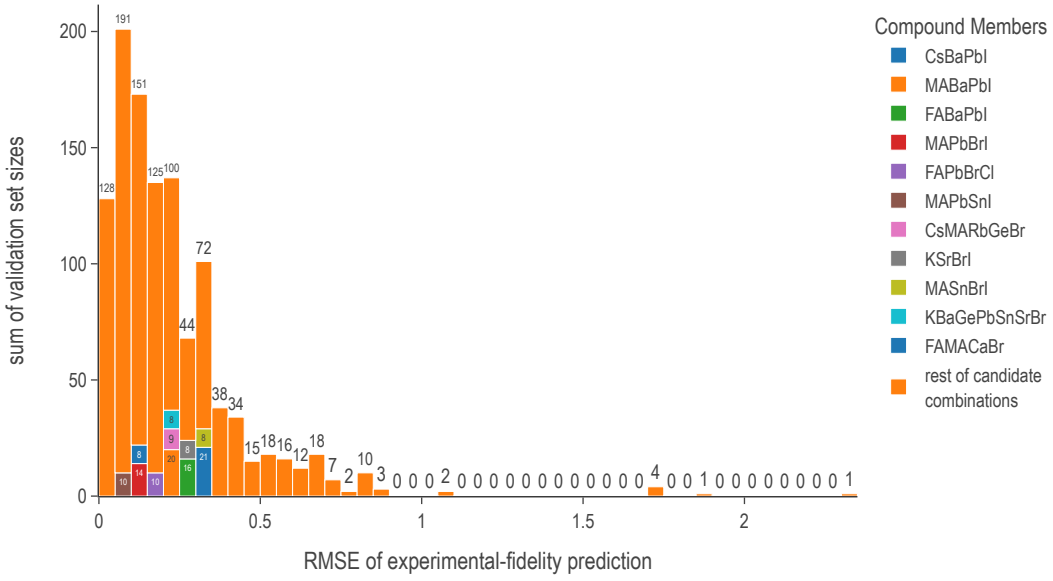

Figure 3.6. Distribution of leave-one-composition-out cross-validation errors weighted by the size of validation sets

Table 3.3. Leave-one-composition-out cross-validation scored by complete suite

partition	test	train
ev	-2.48	0.99
maxerr	0.32	0.96
rmse	0.22	0.10
rmse EXP	0.22	0.06
rmse PBE	0.18	0.11
rmse HSE	0.22	0.11
rmse HSE(SOC)	0.21	0.10
rmse HSE-PBErel(SOC)	0.20	0.10

3.7.2 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. [62] 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 3.7 and 3.8 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 3.7, the B-site Electronegativity is a strongly positive contributor to the RFR prediction. Large B-site electronegativities tend to result in a subtraction from the mean band gap of about 0.5 eV. Small electronegativities tend to result in an addition to the mean band gap averaging 1.5 eV but with much wider variability. 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 summarized in figure 3.9. 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, 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

[6]↑https://github.com/slundberg/shap

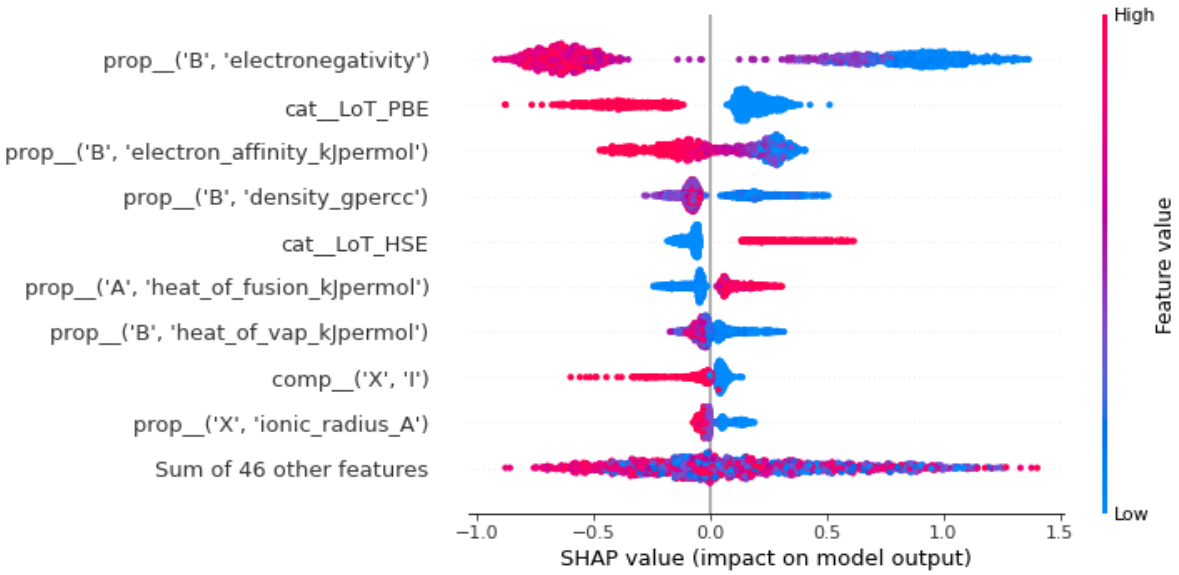

Figure 3.7. Random Forest Regression Band Gap SHAP Values

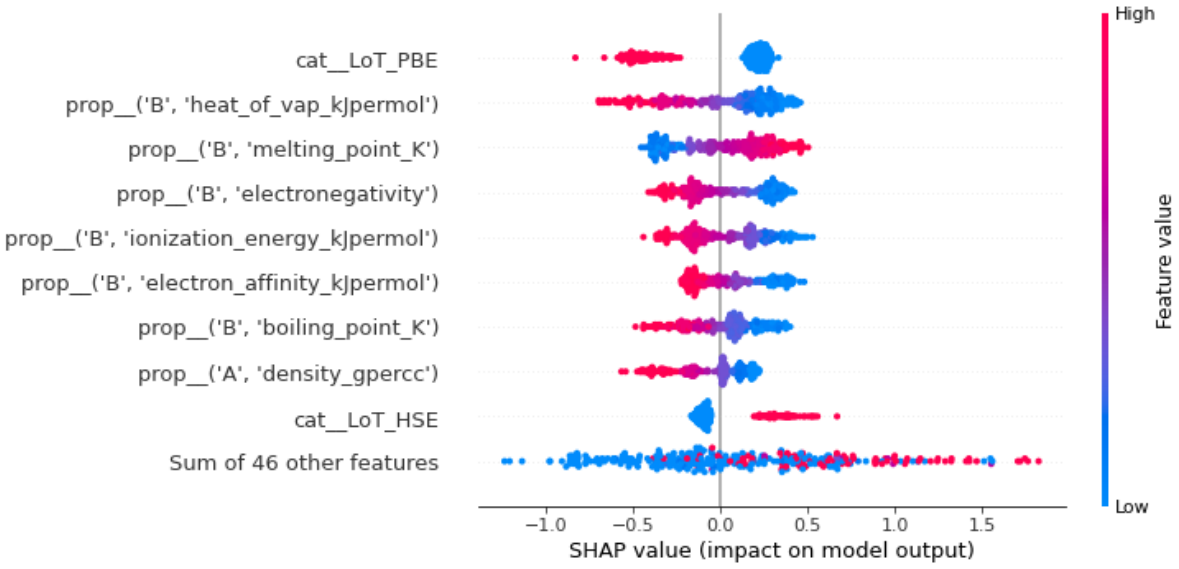

Figure 3.8. Gaussian Process Regression Band Gap SHAP Values

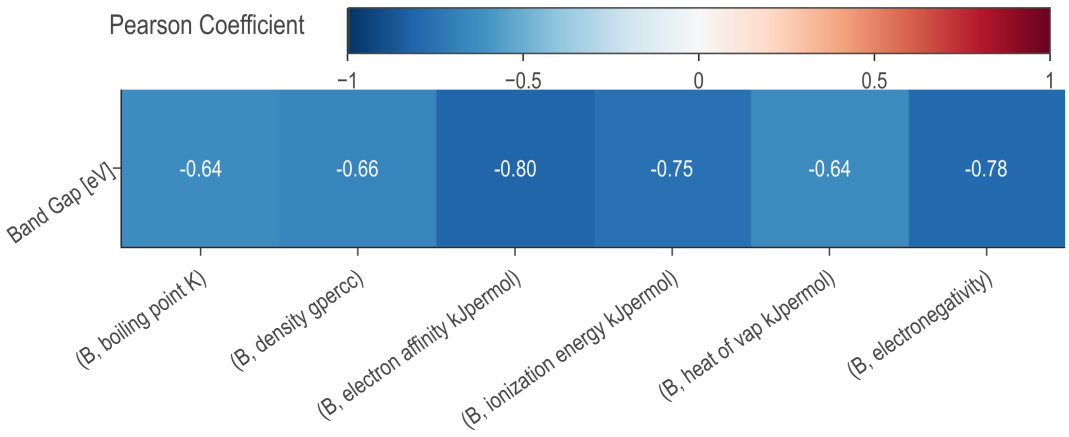

Figure 3.9. raw features with ($|p| > 0.5$) against band gap

anything, figure 3.10 confirms that the Iodine at the X site tends to slightly increase band gaps.

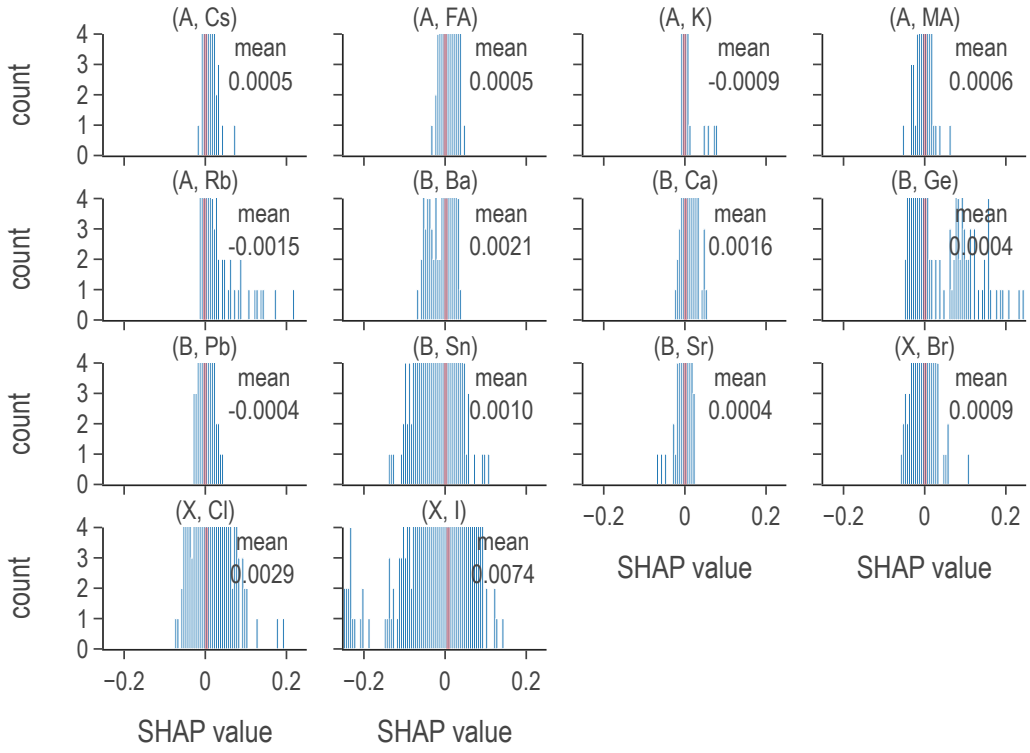

Figure 3.10. SHAP score distributions reveal effects of individual constituents

3.7.3 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 3.11.

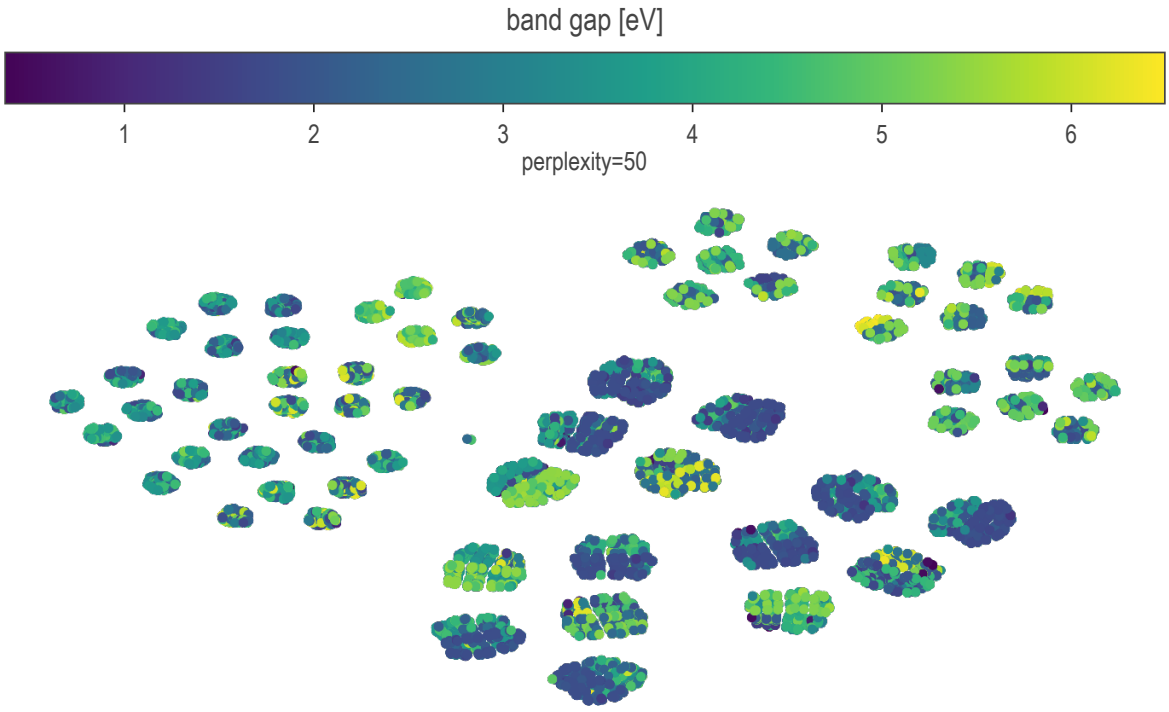

Figure 3.11. Band gap predictions overlaid on cardinal mixing chemical domain projected from fourteen to two dimensions via t-SNE

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. [9, 63] See figure 3.12. 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. [43, 64] 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 3.13 may be juxtaposed with figure 1.3.

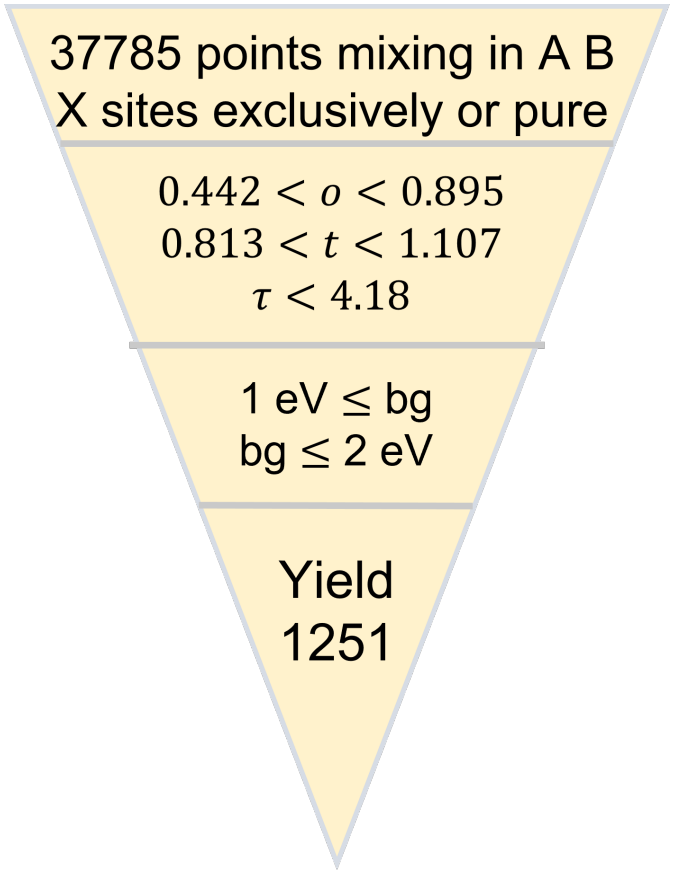

Figure 3.12. Summary of screening operations used to identify candidate compounds

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 A.10. 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 $1/8^{\text{th}}$ of their site at a rate of about 8%, with Potassium and Rb also preferring

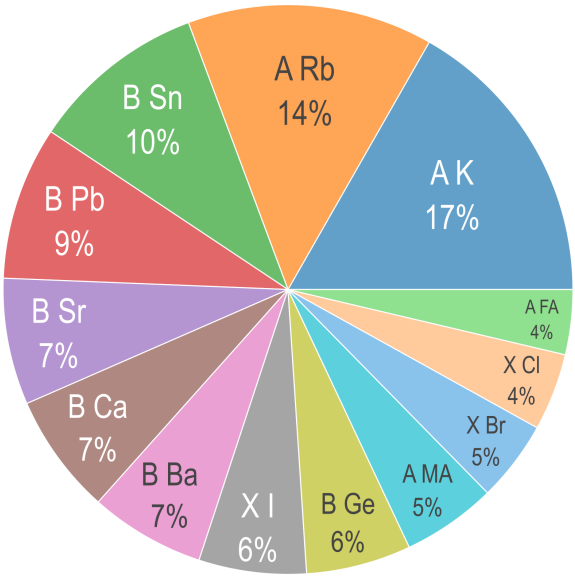

Figure 3.13. The compounds selected from the cardinal mixing sample space contain varying fractions of each element

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 3.14. The strong preference for pure sites simply reflects that this sample space contained compositions mixed at no more than one site simultaneously.

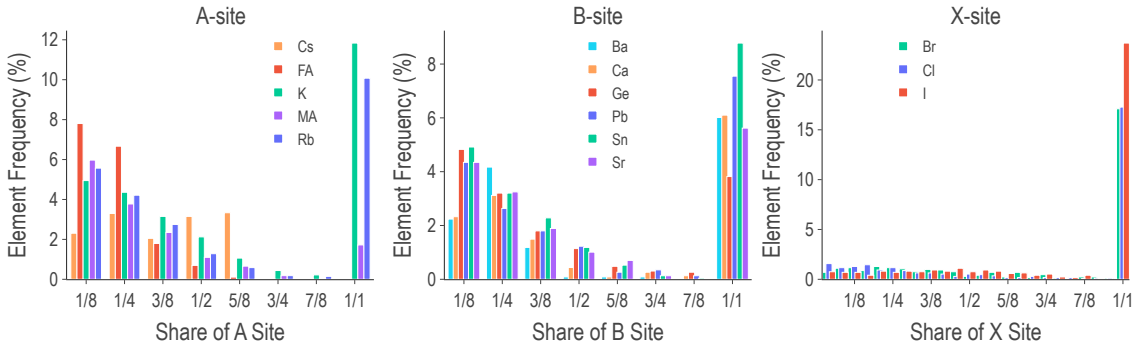

Figure 3.14. Frequency of mixing fractions of species at the A, B, and X sites across the ~1200 screen compounds

4. CONCLUSIONS

A set of promising hypothetical compositions is identified by this work. I identified a set of 1251 promising perovskite compositions with desirable band gaps for Photovoltaic applications using a novel data-driven approach. The selected band gaps are predicted at an experimental fidelity with error likely ranging from 0.15 to 0.3 eV. Of the selected compositions 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. The 30 most stable lead-free compounds identified with band-gap in the 1-2 eV range are listed in table 4.2.

I evaluated a variety of machine learning techniques and implemented simple but effective models for estimating band gap for perovskites from multi-fidelity data.

I found RFR models performed best in this setting, and believe this is because the architecture is an ensemble model so it reduces the variance of error by design. Additionally, the decision trees that make up the forest best identify and capture relevant feature interactions as corroborated by Pearson correlations. Trees also flexibly represent complex nonlinear relationships between feature interactions and band gap. RFR models may be data hungry but are generally better about accommodating outliers. The RFR properties that helped it perform best in this setting are not domain-specific and therefore are likely to apply to other material properties and for other types of compounds.

In summary, I experimented with training multiple models to differentiate and treat appropriately observations with different fidelities and demonstrated good results with the RFR in particular. The Gaussian Process Regression was a very close second. The SISSO model and the RFR and GPR based on the SIS-selected superspace do not perform well, contrary to expectations. The GPRs ability to theoretically deal with more complex descriptors deserves further study to realize its full potential. 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]
19290	FA0.375Rb0.625Sn1.000I1.000	1.98
19309	FA0.375MA0.125Rb0.500Sn1.000Cl1.000	1.99
19310	FA0.375MA0.125Rb0.500Sn1.000Br1.000	1.95
19306	FA0.375MA0.125Rb0.500Sr1.000Cl1.000	1.95
19308	FA0.375MA0.125Rb0.500Sn1.000I1.000	1.70
19307	FA0.375MA0.125Rb0.500Sr1.000Br1.000	1.96
19304	FA0.375Rb0.625Ba1.000Br1.000	1.93
19303	FA0.375Rb0.625Ba1.000Cl1.000	1.89
19305	FA0.375MA0.125Rb0.500Sr1.000I1.000	1.74
19302	FA0.375Rb0.625Ba1.000I1.000	1.68
19076	FA0.250K0.250MA0.375Rb0.125Sn1.000Br1.000	1.93
19301	FA0.375Rb0.625Ca1.000Br1.000	1.69
19300	FA0.375Rb0.625Ca1.000Cl1.000	1.72
19324	FA0.375MA0.250Rb0.375Sr1.000Cl1.000	1.72
19008	FA0.250K0.125MA0.625Ge1.000I1.000	1.93
19056	FA0.250K0.250MA0.250Rb0.250Sn1.000I1.000	1.84
19004	FA0.250K0.125MA0.625Sn1.000Br1.000	1.77
19032	FA0.250K0.250Rb0.500Ba1.000I1.000	1.95
19073	FA0.250K0.250MA0.375Rb0.125Sr1.000Br1.000	1.94
19094	FA0.250K0.250MA0.500Sn1.000Br1.000	1.87
19049	FA0.250K0.250MA0.125Rb0.375Ca1.000Br1.000	1.85
19003	FA0.250K0.125MA0.625Sn1.000Cl1.000	1.75
19074	FA0.250K0.250MA0.375Rb0.125Sn1.000I1.000	1.68
19002	FA0.250K0.125MA0.625Sn1.000I1.000	1.46
19053	FA0.250K0.250MA0.250Rb0.250Sr1.000I1.000	1.88
19075	FA0.250K0.250MA0.375Rb0.125Sn1.000Cl1.000	1.95
19070	FA0.250K0.250MA0.250Rb0.250Ba1.000Br1.000	1.86
19093	FA0.250K0.250MA0.500Sn1.000Cl1.000	1.86
19001	FA0.250K0.125MA0.625Sr1.000Br1.000	1.71
19091	FA0.250K0.250MA0.500Sr1.000Br1.000	1.89

REFERENCES

[1] M. I. H. Ansari, A. Qurashi, and M. K. Nazeeruddin, "Frontiers, opportunities, and challenges in perovskite solar cells: A critical review," *Journal of Photochemistry and Photobiology C: Photochemistry Reviews*, vol. 35, pp. 1–24, 2018, ISSN: 1389-5567. DOI: 10.1016/j.jphotochemrev.2017.11.002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1389556717301144.

[2] W.-J. Yin, J.-H. Yang, J. Kang, Y. Yan, and S.-H. Wei, "Halide perovskite materials for solar cells: A theoretical review," *Journal of Materials Chemistry A*, vol. 3, no. 17, pp. 8926–8942, 2015. DOI: 10.1039/c4ta05033a. [Online]. Available: http://dx.doi.org/10.1039/c4ta05033a.

[3] J. S. Manser, J. A. Christians, and P. V. Kamat, "Intriguing optoelectronic properties of metal halide perovskites," *Chemical Reviews*, vol. 116, no. 21, pp. 12 956–13 008, 2016, ISSN: 0009-2665. DOI: 10.1021/acs.chemrev.6b00136. [Online]. Available: https://doi.org/10.1021/acs.chemrev.6b00136.

[4] T. M. Brenner, D. A. Egger, L. Kronik, G. Hodes, and D. Cahen, "Hybrid organic-inorganic perovskites: Low-cost semiconductors with intriguing charge-transport properties," *Nature Reviews Materials*, vol. 1, no. 1, p. 15 007, 2016, ISSN: 2058-8437. DOI: 10.1038/natrevmats.2015.7. [Online]. Available: https://doi.org/10.1038/natrevmats.2015.7.

[5] P. Cui *et al.*, "Planar p-n homojunction perovskite solar cells with efficiency exceeding 21.3 %," *Nature Energy*, vol. 4, no. 2, pp. 150–159, 2019, ISSN: 2058-7546. DOI: 10.1038/s41560-018-0324-8. [Online]. Available: https://doi.org/10.1038/s41560-018-0324-8.

[6] M. Jeong *et al.*, "Stable perovskite solar cells with efficiency exceeding 24.8 % and 0.3-v voltage loss," *Science*, vol. 369, no. 6511, pp. 1615–1620, 2020. DOI: 10.1126/science.abb7167. [Online]. Available: https://doi.org/10.1126/science.abb7167.

[7] J. Bartel Christopher *et al.*, "New tolerance factor to predict the stability of perovskite oxides and halides," *Science Advances*, vol. 5, no. 2, eaav0693, 2019. DOI: 10.1126/sciadv.aav0693. [Online]. Available: https://doi.org/10.1126/sciadv.aav0693.

[8] P. Manganaris, J. Yang, and A. Mannodi Kanakkithodi, "Multi-fidelity machine learning for perovskite band gap predictions," In Preparation, Jun. 2023.

[9] J. Yang, P. T. Manganaris, and A. K. Mannodi Kanakkithodi, "A high-throughput computational dataset of halide perovskite alloys," *Digital Discovery*, 2023, ISSN: 2635-098X. DOI: 10.1039/d3dd00015j. [Online]. Available: http://dx.doi.org/10.1039/D3DD00015J.

[10] P. Manganaris, S. Desai, and A. Kanakkithodi, *MRS computational materials science tutorial*, en, 2022. DOI: 10.21981/D1J2-AR65. [Online]. Available: https://nanohub.org/resources/36041?rev=90.

[11] S. Kim, J. A. Márquez, T. Unold, and A. Walsh, "Upper limit to the photovoltaic efficiency of imperfect crystals from first principles," *Energy Environ. Sci.*, vol. 13, pp. 1481–1491, 5 2020. DOI: 10.1039/D0EE00291G. [Online]. Available: http://dx.doi.org/10.1039/D0EE00291G.

[12] D. Dahliah, G. Brunin, J. George, V.-A. Ha, G.-M. Rignanese, and G. Hautier, "High-throughput computational search for high carrier lifetime, defect-tolerant solar absorbers," *Energy Environ. Sci.*, vol. 14, pp. 5057–5073, 9 2021. DOI: 10.1039/D1EE00801C. [Online]. Available: http://dx.doi.org/10.1039/D1EE00801C.

[13] M. Kar and T. Körzdörfer, "Computational screening of methylammonium based halide perovskites with bandgaps suitable for perovskite-perovskite tandem solar cells," *The Journal of Chemical Physics*, vol. 149, no. 21, p. 214 701, 2018, ISSN: 0021-9606. DOI: 10.1063/1.5037535. [Online]. Available: https://doi.org/10.1063/1.5037535.

[14] C. Kim, T. D. Huan, S. Krishnan, and R. Ramprasad, "A hybrid organic-inorganic perovskite dataset," *Scientific Data*, vol. 4, no. 1, p. 170 057, 2017, ISSN: 2052-4463. DOI: 10.1038/sdata.2017.57. [Online]. Available: https://doi.org/10.1038/sdata.2017.57.

[15] S. Zhu, J. Ye, Y. Zhao, and Y. Qiu, "Structural, electronic, stability, and optical properties of cspb1-xsnxibr2 perovskites: A first-principles investigation," *The Journal of Physical Chemistry C*, vol. 123, no. 33, pp. 20 476–20 487, 2019, ISSN: 1932-7447. DOI: 10.1021/acs.jpcc.9b04841. [Online]. Available: https://doi.org/10.1021/acs.jpcc.9b04841.

[16] A. Banerjee, S. Chakraborty, and R. Ahuja, "Rashba triggered electronic and optical properties tuning in mixed cation-mixed halide hybrid perovskites," *ACS Applied Energy Materials*, vol. 2, no. 10, pp. 6990–6997, 2019, not free. DOI: 10.1021/acsaem.9b01479. [Online]. Available: https://doi.org/10.1021/acsaem.9b01479.

[17] J. Ding *et al.*, "Cesium decreases defect density and enhances optoelectronic properties of mixed ma1-xcsxpbbr3 single crystal," *The Journal of Physical Chemistry C*, vol. 123, no. 24, pp. 14 969–14 975, 2019, ISSN: 1932-7447. DOI: 10.1021/acs.jpcc.9b03987. [Online]. Available: https://doi.org/10.1021/acs.jpcc.9b03987.

[18] C. Greenland *et al.*, "Correlating phase behavior with photophysical properties in mixed-cation mixed-halide perovskite thin films," *Advanced Energy Materials*, vol. 10, no. 4, p. 1 901 350, 2020, ISSN: 1614-6832. DOI: 10.1002/aenm.201901350. [Online]. Available: https://doi.org/10.1002/aenm.201901350.

[19] H. Zhang, M. K. Nazeeruddin, and W. C. H. Choy, "Perovskite photovoltaics: The significant role of ligands in film formation, passivation, and stability," *Advanced Materials*, vol. 31, no. 8, p. 1 805 702, Jan. 2019, ISSN: 0935-9648. DOI: 10.1002/adma.201805702. [Online]. Available: http://dx.doi.org/10.1002/adma.201805702.

[20] Y. Yan, W.-J. Yin, T. Shi, W. Meng, and C. Feng, "Defect physics of ch3nh3pbx3 (x = i, br, cl) perovskites," *Organic-Inorganic Halide Perovskite Photovoltaics*, pp. 79–105, 2016. DOI: 10.1007/978-3-319-35114-8_4. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-35114-8_4.

[21] L. Dimesso, A. Quintilla, Y.-M. Kim, U. Lemmer, and W. Jaegermann, "Investigation of formamidinium and guanidinium lead tri-iodide powders as precursors for solar cells," *Materials Science and Engineering: B*, vol. 204, pp. 27–33, Feb. 2016, ISSN: 0921-5107. DOI: 10.1016/j.mseb.2015.11.006. [Online]. Available: http://dx.doi.org/10.1016/j.mseb.2015.11.006.

[22] A. Mannodi-Kanakkithodi and M. K. Y. Chan, "Data-driven design of novel halide perovskite alloys," *Energy Environ. Sci.*, vol. 15, pp. 1930–1949, 5 2022. DOI: 10.1039/D1EE02971A. [Online]. Available: http://dx.doi.org/10.1039/D1EE02971A.

[23] I. E. Castelli, J. M. García-Lastra, K. S. Thygesen, and K. W. Jacobsen, "Bandgap calculations and trends of organometal halide perovskites," *APL Materials*, vol. 2, no. 8, p. 081 514, 2014. DOI: 10.1063/1.4893495. [Online]. Available: https://doi.org/10.1063/1.4893495.

[24] H. Park *et al.*, "Exploring new approaches towards the formability of mixed-ion perovskites by dft and machine learning," *Physical Chemistry Chemical Physics*, vol. 21, no. 3, pp. 1078–1088, 2019, ISSN: 1463-9076. DOI: 10.1039/C8CP06528D. [Online]. Available: http://dx.doi.org/10.1039/C8CP06528D.

[25] W. Pu, W. Xiao, J.-W. Wang, X.-W. Li, and L. Wang, "Screening of perovskite materials for solar cell applications by first-principles calculations," *Materials & Design*, vol. 198, p. 109 387, Jan. 2021. DOI: 10.1016/j.matdes.2020.109387.

[26] J. C. Stanley, F. Mayr, and A. Gagliardi, "Machine learning stability and bandgaps of lead-free perovskites for photovoltaics," *Advanced Theory and Simulations*, vol. 3, no. 1, p. 1 900 178, 2020. DOI: 10.1002/adts.201900178. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/adts.201900178. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/adts.201900178.

[27] B. D. Lee, W. B. Park, J.-W. Lee, M. Kim, M. Pyo, and K.-S. Sohn, "Discovery of lead-free hybrid organic/inorganic perovskites using metaheuristic-driven dft calculations," *Chemistry of Materials*, vol. 33, no. 2, pp. 782–798, 2021. DOI: 10.1021/acs.chemmater.0c04499. eprint: https://doi.org/10.1021/acs.chemmater.0c04499. [Online]. Available: https://doi.org/10.1021/acs.chemmater.0c04499.

[28] J. Yang and A. Mannodi-Kanakkithodi, "High-throughput computations and machine learning for halide perovskite discovery," *MRS Bulletin*, vol. 47, no. 9, pp. 940–948, Sep. 2022, ISSN: 1938-1425. DOI: 10.1557/s43577-022-00414-2. [Online]. Available: http://dx.doi.org/10.1557/s43577-022-00414-2.

[29] O. Almora *et al.*, "Device performance of emerging photovoltaic materials (version 1)," *Advanced Energy Materials*, vol. 11, no. 11, p. 2 002 774, 2020. DOI: 10.1002/aenm.202002774. [Online]. Available: http://dx.doi.org/10.1002/aenm.202002774.

[30] G. Kieslich, S. Sun, and A. K. Cheetham, "An extended tolerance factor approach for organic-inorganic perovskites," *Chemical Science*, vol. 6, no. 6, pp. 3430–3433, 2015, ISSN: 2041-6539. DOI: 10.1039/c5sc00961h. [Online]. Available: http://dx.doi.org/10.1039/C5SC00961H.

[31] G. Kresse and J. Furthmüller, "Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set," *Phys. Rev. B*, vol. 54, pp. 11 169–11 186, 16 Oct. 1996. DOI: 10.1103/PhysRevB.54.11169. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.54.11169.

[32] G. Kresse and J. Hafner, "Ab initio molecular dynamics for liquid metals," *Phys. Rev. B*, vol. 47, pp. 558–561, 1 Jan. 1993. DOI: 10.1103/PhysRevB.47.558. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.47.558.

[33] G. Kresse and J. Furthmüller, "Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set," *Computational Materials Science*, vol. 6, no. 1, pp. 15–50, 1996, ISSN: 0927-0256. DOI: 10.1016/0927-0256(96)00008-0. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0927025696000080.

[34] G. Kresse and D. Joubert, "From ultrasoft pseudopotentials to the projector augmented-wave method," *Phys. Rev. B*, vol. 59, pp. 1758–1775, 3 Jan. 1999. DOI: 10.1103/PhysRevB.59.1758. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.59.1758.

[35] G. Kresse and J. Hafner, "Norm-conserving and ultrasoft pseudopotentials for first-row and transition elements," *Journal of Physics: Condensed Matter*, vol. 6, no. 40, pp. 8245–8257, Oct. 1994. DOI: 10.1088/0953-8984/6/40/015. [Online]. Available: https://doi.org/10.1088/0953-8984/6/40/015.

[36] Z. Jiang, Y. Nahas, B. Xu, S. Prosandeev, D. Wang, and L. Bellaiche, "Special quasirandom structures for perovskite solid solutions," *Journal of Physics: Condensed Matter*, vol. 28, no. 47, p. 475 901, 2016. DOI: 10.1088/0953-8984/28/47/475901. [Online]. Available: http://dx.doi.org/10.1088/0953-8984/28/47/475901.

[37] J. P. Perdew, K. Burke, and M. Ernzerhof, "Generalized gradient approximation made simple," *Phys. Rev. Lett.*, vol. 77, pp. 3865–3868, 18 Oct. 1996. DOI: 10.1103/PhysRevLett.77.3865. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.77.3865.

[38] J. Heyd, G. E. Scuseria, and M. Ernzerhof, "Hybrid functionals based on a screened coulomb potential," *The Journal of Chemical Physics*, vol. 118, no. 18, pp. 8207–8215, 2003. DOI: 10.1063/1.1564060. eprint: https://doi.org/10.1063/1.1564060. [Online]. Available: https://doi.org/10.1063/1.1564060.

[39] Y. Hinuma, G. Pizzi, Y. Kumagai, F. Oba, and I. Tanaka, "Band structure diagram paths based on crystallography," *CoRR*, 2016. arXiv: 1602.06402 [cond-mat.mtrl-sci]. [Online]. Available: http://arxiv.org/abs/1602.06402v4.

[40] A. M. Ganose, A. J. Jackson, and D. O. Scanlon, "Sumo: Command-line tools for plotting and analysis of periodic *ab initio* calculations," *Journal of Open Source Software*, vol. 3, no. 28, p. 717, 2018. DOI: 10.21105/joss.00717. [Online]. Available: https://doi.org/10.21105/joss.00717.

[41] S. Steiner, S. Khmelevskyi, M. Marsmann, and G. Kresse, "Calculation of the magnetic anisotropy with projected-augmented-wave methodology and the case study of disordered $Fe_{1-x}Co_x$ alloys," *Phys. Rev. B*, vol. 93, p. 224 425, 22 Jun. 2016. DOI: 10.1103/PhysRevB.93.224425. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.93.224425.

[42] A. Mannodi-Kanakkithodi *et al.*, "Comprehensive computational study of partial lead substitution in methylammonium lead bromide," *Chemistry of Materials*, vol. 31, no. 10, pp. 3599–3612, 2019. DOI: 10.1021/acs.chemmater.8b04017. [Online]. Available: http://dx.doi.org/10.1021/acs.chemmater.8b04017.

[43] L. Yu and A. Zunger, "Identification of potential photovoltaic absorbers based on first-principles spectroscopic screening of materials," *Physical Review Letters*, vol. 108, no. 6, Feb. 2012, ISSN: 1079-7114. DOI: 10.1103/physrevlett.108.068701. [Online]. Available: http://dx.doi.org/10.1103/PhysRevLett.108.068701.

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

[45] J.-P. Kim, J. A. Christians, H. Choi, S. Krishnamurthy, and P. V. Kamat, "Cdses nanowires: Compositionally controlled band gap and exciton dynamics," *The Journal of Physical Chemistry Letters*, vol. 5, no. 7, pp. 1103–1109, 2014, PMID: 26274456. DOI: 10.1021/jz500280g. eprint: https://doi.org/10.1021/jz500280g. [Online]. Available: https://doi.org/10.1021/jz500280g.

[46] D. E. Swanson, J. R. Sites, and W. S. Sampath, "Co-sublimation of cdsexte1-x layers for cdte solar cells," *Solar Energy Materials and Solar Cells*, vol. 159, pp. 389–394, 2017, ISSN: 0927-0248. DOI: 10.1016/j.solmat.2016.09.025. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0927024816303634.

[47] L. Mentel, *mendeleev – a python resource for properties of chemical elements, ions and isotopes*, version 0.9.0, 2014. [Online]. Available: https://github.com/lmmentel/mendeleev.

[48] L. Yang and A. Shami, "On hyperparameter optimization of machine learning algorithms: Theory and practice," *Neurocomputing*, vol. 415, pp. 295–316, 2020, ISSN: 0925-2312. DOI: 10.1016/j.neucom.2020.07.061. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231220311693.

[49] L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, and M. Scheffler, "Big data of materials science: Critical role of the descriptor," *Physical Review Letters*, vol. 114, no. 10, Mar. 2015, ISSN: 1079-7114. DOI: 10.1103/physrevlett.114.105503. [Online]. Available: http://dx.doi.org/10.1103/PhysRevLett.114.105503.

[50] T. C. H. Lux, L. T. Watson, T. H. Chang, Y. Hong, and K. Cameron, "Interpolation of sparse high-dimensional data," *Numerical Algorithms*, vol. 88, no. 1, pp. 281–313, Nov. 2020, ISSN: 1572-9265. DOI: 10.1007/s11075-020-01040-2. [Online]. Available: http://dx.doi.org/10.1007/s11075-020-01040-2.

[51] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[52] L. Jiang *et al.*, "Prediction of lattice constant in cubic perovskites," *Journal of Physics and Chemistry of Solids*, vol. 67, no. 7, pp. 1531–1536, Jul. 2006, ISSN: 0022-3697. DOI: 10.1016/j.jpcs.2006.02.004. [Online]. Available: http://dx.doi.org/10.1016/j.jpcs.2006.02.004.

[53] R. Ouyang, S. Curtarolo, E. Ahmetcik, M. Scheffler, and L. M. Ghiringhelli, "Sisso: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates," *Phys. Rev. Materials*, vol. 2, p. 083 802, 8 Aug. 2018. DOI: 10.1103/PhysRevMaterials.2.083802. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevMaterials.2.083802.

[54] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, Jan. 1996, ISSN: 0035-9246. DOI: 10.1111/j.2517-6161.1996.tb02080.x. [Online]. Available: http://dx.doi.org/10.1111/j.2517-6161.1996.tb02080.x.

[55] L. M. Ghiringhelli *et al.*, "Learning physical descriptors for materials science by compressed sensing," *New Journal of Physics*, vol. 19, no. 2, p. 023 017, Feb. 2017, ISSN: 1367-2630. DOI: 10.1088/1367-2630/aa57bf. [Online]. Available: http://dx.doi.org/10.1088/1367-2630/aa57bf.

[56] P. Ray, S. S. Reddy, and T. Banerjee, "Various dimension reduction techniques for high dimensional data analysis: A review," *Artificial Intelligence Review*, vol. 54, no. 5, pp. 3473–3515, Jan. 2021, ISSN: 1573-7462. DOI: 10.1007/s10462-020-09928-0. [Online]. Available: http://dx.doi.org/10.1007/s10462-020-09928-0.

[57] K. Hamm and L. Huang, "Cur decompositions, approximations, and perturbations," *CoRR*, 2019. arXiv: 1903.09698v2 [math.NA]. [Online]. Available: http://arxiv.org/abs/1903.09698v2.

[58] N. Gauraha, "Introduction to the lasso," *Resonance*, vol. 23, no. 4, pp. 439–464, Apr. 2018, ISSN: 0973-712X. DOI: 10.1007/s12045-018-0635-x. [Online]. Available: http://dx.doi.org/10.1007/s12045-018-0635-x.

[59] D. G. Mayo, *Error and the Growth of Experimental Knowledge*. Apr. 1996, ISBN: 9780226511986. DOI: 10.7208/9780226511993.

[60] L. Buitinck *et al.*, "API design for machine learning software: Experiences from the scikit-learn project," in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 2013, pp. 108–122.

[61] A. Dunn, Q. Wang, A. Ganose, D. Dopp, and A. Jain, "Benchmarking materials property prediction methods: The matbench test set and automatminer reference algorithm," *npj Computational Materials*, vol. 6, no. 1, p. 138, 2020. DOI: 10.1038/s41524-020-00406-3. [Online]. Available: http://dx.doi.org/10.1038/s41524-020-00406-3.

[62] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *CoRR*, 2017. arXiv: 1705.07874 [cs.AI]. [Online]. Available: http://arxiv.org/abs/1705.07874v2.

[63] A. Mannodi-Kanakkithodi and M. K. Chan, "Computational data-driven materials discovery," *Trends in Chemistry*, vol. 3, no. 2, pp. 79–82, 2021. DOI: 10.1016/j.trechm.2020.12.007. [Online]. Available: http://dx.doi.org/10.1016/j.trechm.2020.12.007.

[64] W. Shockley and H. J. Queisser, "Detailed balance limit of efficiency of pn junction solar cells," *Journal of Applied Physics*, vol. 32, no. 3, pp. 510–519, 1961. DOI: 10.1063/1.1736034. eprint: https://doi.org/10.1063/1.1736034. [Online]. Available: https://doi.org/10.1063/1.1736034.

[65] P. Manganaris, J. Yang, and A. Mannodi-Kannakithodi, *Novel halide perovskites described by multiple fidelity models of high-throughput simulations*.

[66] P. Manganaris, *Machine learning modeling of hybrid organic-inorganic perovskites*.

[67] P. Manganaris, S. Desai, A. Mannodi-Kanakkithodi, and G. Kusne, *Machine learning in materials science: From basic concepts to active learning*.

[68] P. Gollapalli, P. Manganaris, and A. Mannodi-Kanakkithodi, "Graph neural network predictions for formation energy of native defects in zinc blende semiconductors," 2023, In Preparation. DOI: 00.0000/00000000.

[69] R. Edlabadkar, J. Yang, H. Rahman, P. Manganaris, E. P. Korimilli, and A. Mannodi-Kanakkithodi, "Driving halide perovskite discovery using graph neural networks," 2023, In Preparation. DOI: 00.0000/00000000.

[70] J. Yang, P. Manganaris, and A. Mannodi-Kanakkithodi, "Discovering novel halide perovskite alloys using multi-fidelity machine learning and genetic algorithm," In Preparation, 2023.

A. ADDITIONAL FIGURES

Feature Distributions

These distributions show how the various descriptors distribute in one dimension. The preponderance of heavy settling in the high end of the range in most features suggests one reason why positive correlations are so popular in the simulated fidelities. Naturally, simulated fidelities have access to more data, which may spread more widely in the high density regions of the distribution. Alternatively the correlation flip seen between the EXP fidelity data and SIM fidelity data might be explained simply by the lack of Experimental measurements in the exotic parts of the domain.

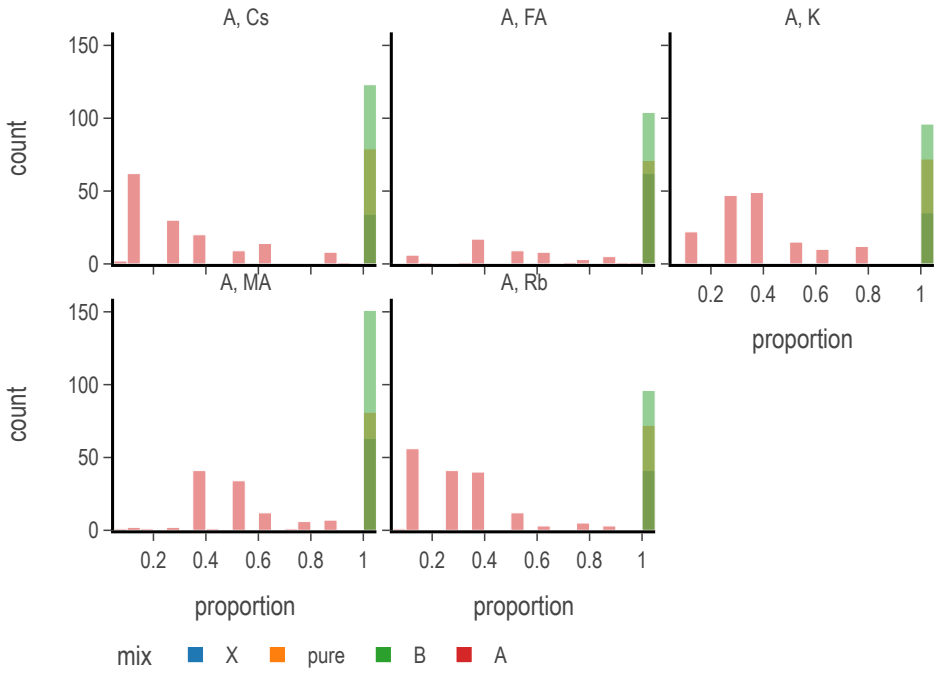

Figure A.1. Normalized Distribution of A-site Constituents

SIS+RFR and SIS+GPR SHAP analysis

The analysis of these models shows how the typical SIS feature is more explanatory. However it also shows that interpretability of such models is limited by the sensibility of the

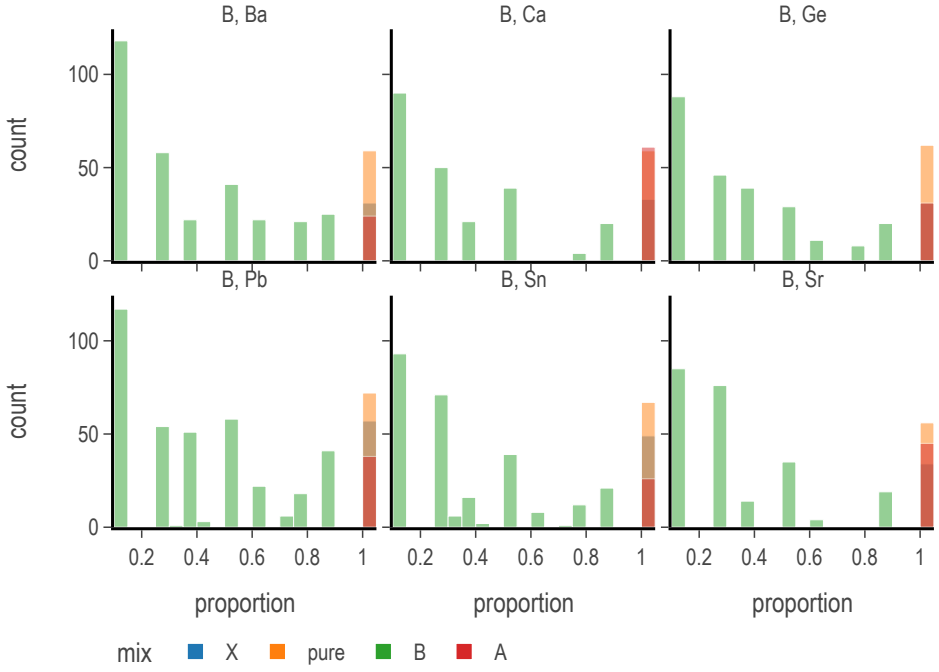

Figure A.2. Normalized Distribution of B-site Constituents

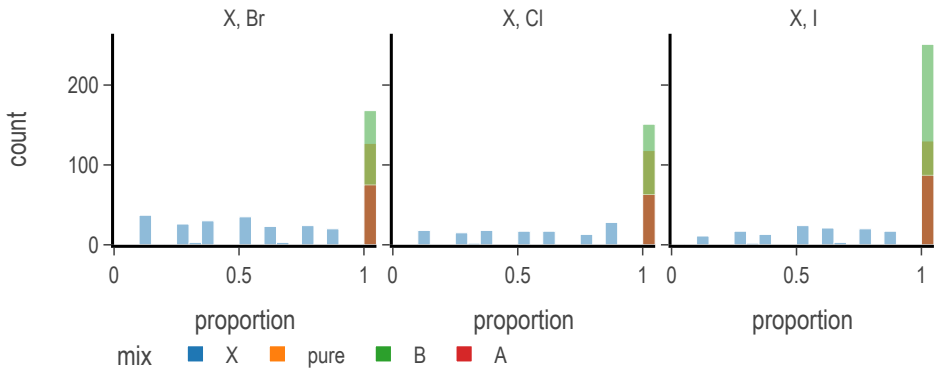

Figure A.3. Normalized Distribution of X-site Constituents

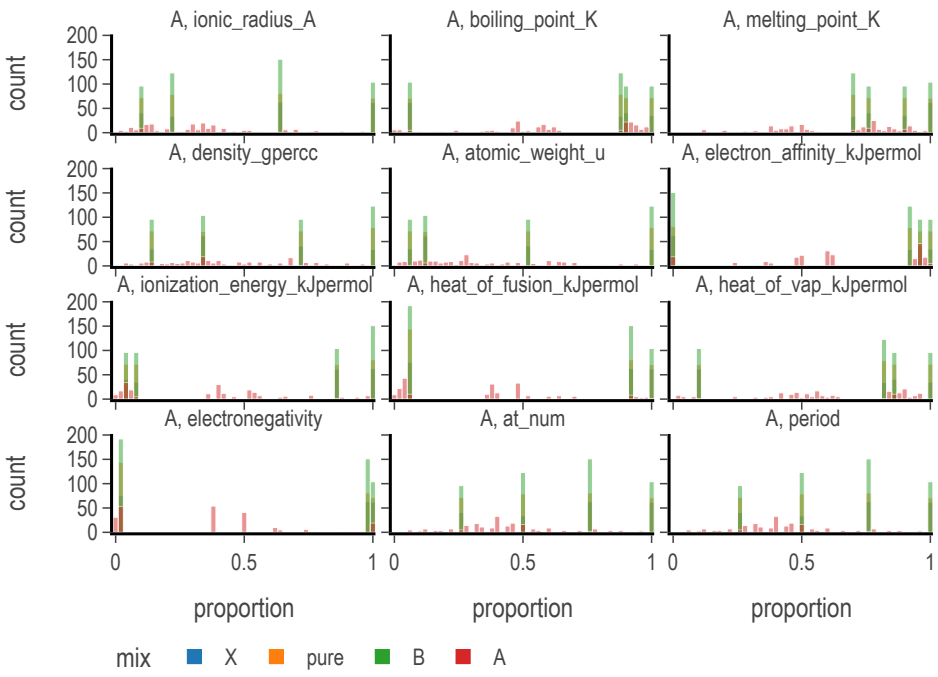

Figure A.4. Distributions of Mean A-Site Properties

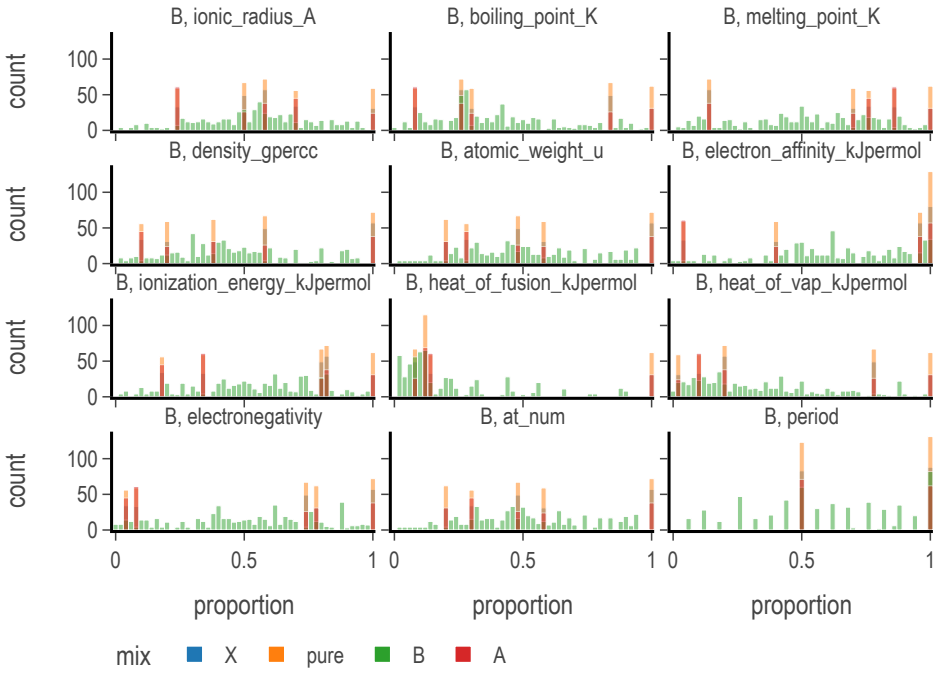

Figure A.5. Distributions of Mean B-Site Properties

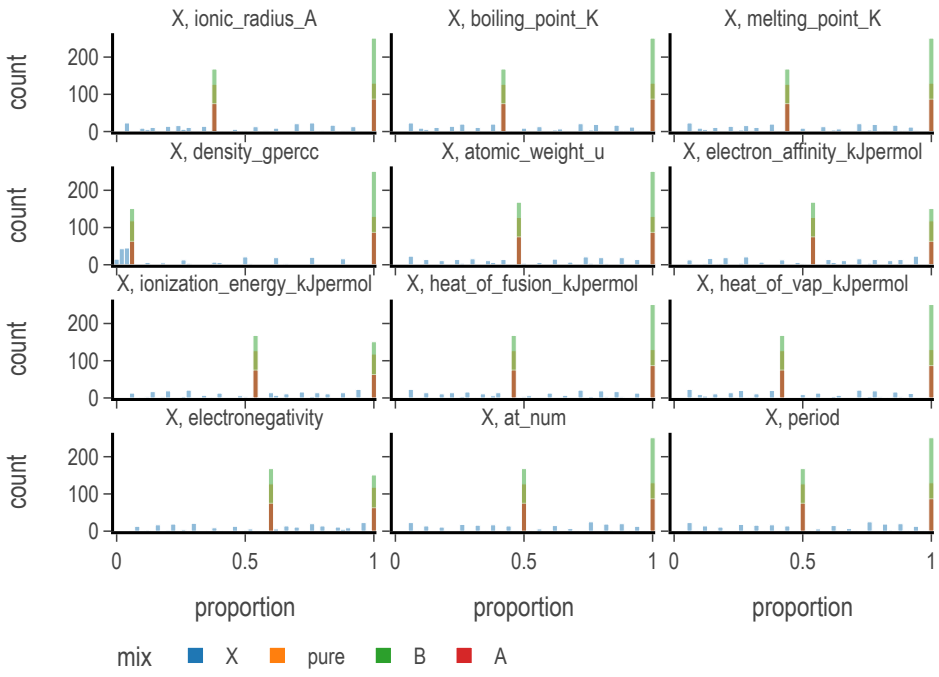

Figure A.6. Distributions of Mean X-Site Properties

combinations used to create the feature. Thankfully, every feature illustrated in figures A.7 and A.8 is constrained to create either unit-less features or features with coherent units.

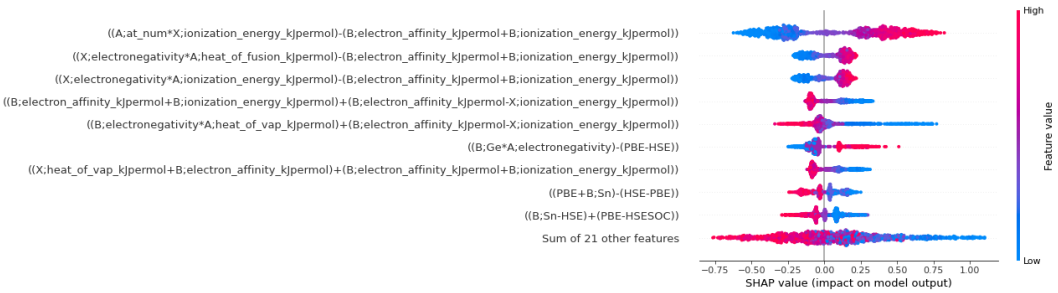

Figure A.7. Random Forest Regression Band Gap on SIS domain SHAP Values

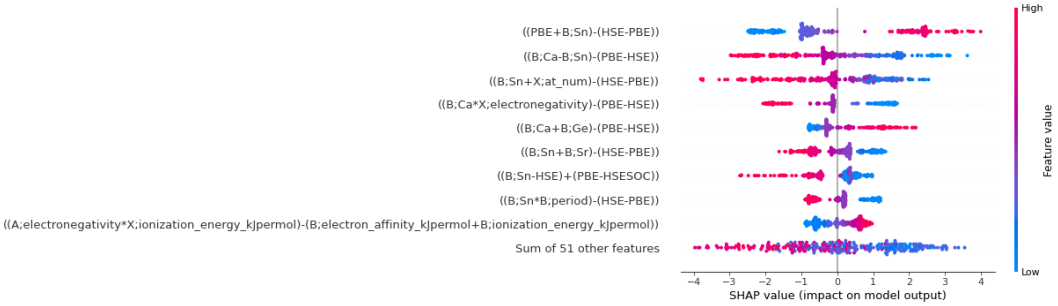

Figure A.8. Gaussian Process Regression Band Gap on SIS domain SHAP Values

Known Clustering in t-SNE Projections

At minimum, this tSNE projection captures the structure in the dataset arising due to the mixing. See figure A.9. Also, it seems some structure may be explained by whether a cluster contains points containing the organic species. Logically, the rest of the sub-clusters separate by the presence of constituent species.

Hyper-parameters of Best Random Forest Estimator

The parameters resulting in a SciKit-Learn RFR estimators best capable of predicting band gaps using my training methodology are given in table A.1.

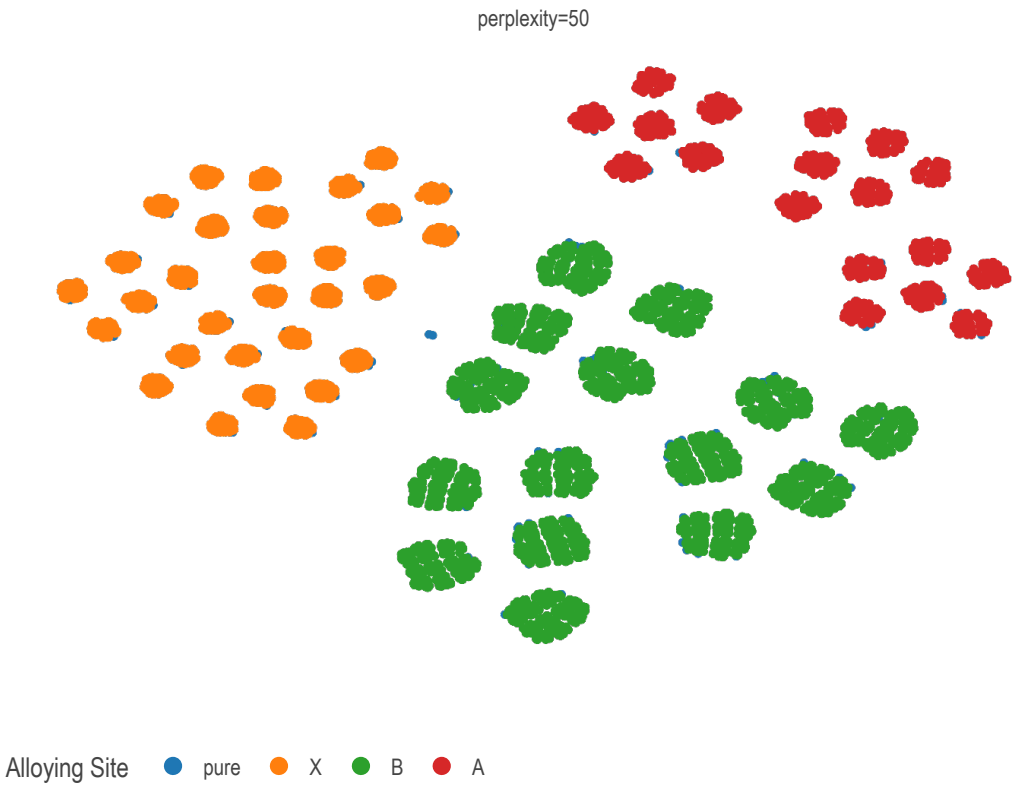

Figure A.9. Projection of sample space via t-SNE overlaid with labels indicating site of mixing

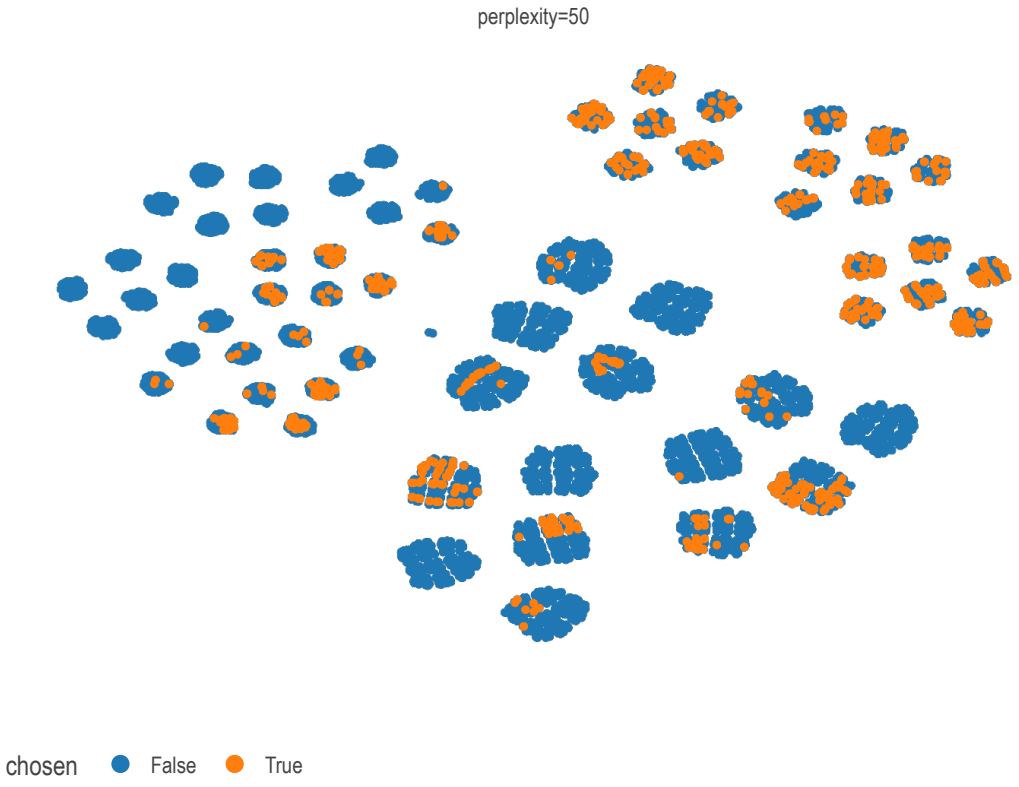

Figure A.10. Projection of sample space via t-SNE overlaid with labels indicating presense of data points in screened subset

Table A.1. Select hyper-parameters from exhaustive search of 31104 models

	Selected Space
normalizer__norm	['l2']
bootstrap	[True]
ccp_alpha	[0.0]
criterion	['poisson']
max_depth	[30]
max_features	[1.0]
max_leaf_nodes	[745]
max_samples	[0.9]
min_impurity_decrease	[0.0]
min_samples_leaf	[1]
min_samples_split	[2]
min_weight_fraction_leaf	[0.0]
n_estimators	[130]
n_jobs	[4]
oob_score	[True]
random_state	[None]
verbose	[0]
warm_start	[False]

B. SOFTWARE AND DATA CONTRIBUTIONS

Data Publication

The prepared sample data I used was published 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.

Software Tools

I have made available the following python libraries 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.

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`

[1] ↑https://ai-materials-and-chemistry.gitbook.io/foundry/
[2] ↑https://github.com/PanayotisManganaris/cmcl
[3] ↑https://github.com/PanayotisManganaris/yogi

Listing B.1. How to load the Mannodi Group halide perovskites data set from the Materials Data Facility repository

```
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]
```

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.* [53]. 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.

A tutorial delivered in the Spring 2022 Materials Research Society conference details the use of these packages for multi-fidelity model training for materials property prediction. This is provided in the form of an iPython notebook P. Manganaris *et al.*, *MRS computational materials science tutorial*, en, 2022. DOI: 10.21981/D1J2-AR65. [Online]. Available: https://nanohub.org/resources/36041?rev=90 hosted on the Purdue University nanoHUB. This provides only a detailed example of the use of `cmcl` and `yogi` in feature extraction and training multi-fidelity random forest models. Examples for the use of the modified pysisso code may be found in the github repository associated with our publication "Multi-Fidelity Machine Learning for Perovskite Band Gap Predictions."

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

C. PRESENTATIONS

- Poster for DS02 symposium at MRS Boston Fall 2022 *Novel Halide Perovskites described by Multiple Fidelity Models of High-Throughput Simulations*[65]
- Talk for Purdue Soft Materials symposium *Machine Learning Modeling of Hybrid Organic-Inorganic Perovskites*[66]
- Developed notebooks [10] for MRS Honolulu Spring 2022 tutorial hosted on nanoHUB *Machine Learning in Materials Science: From Basic Concepts to Active Learning*[67]

D. PUBLICATIONS

[8] P. Manganaris, J. Yang, and A. Mannodi Kanakkithodi, "Multi-fidelity machine learning for perovskite band gap predictions," In Preparation, Jun. 2023.

[9] J. Yang, P. T. Manganaris, and A. K. Mannodi Kanakkithodi, "A high-throughput computational dataset of halide perovskite alloys," *Digital Discovery*, 2023, ISSN: 2635-098X. DOI: 10.1039/d3dd00015j. [Online]. Available: http://dx.doi.org/10.1039/D3DD00015J.

[10] P. Manganaris, S. Desai, and A. Kanakkithodi, *MRS computational materials science tutorial*, en, 2022. DOI: 10.21981/D1J2-AR65. [Online]. Available: https://nanohub.org/resources/36041?rev=90.

[65] P. Manganaris, J. Yang, and A. Mannodi-Kannakithodi, *Novel halide perovskites described by multiple fidelity models of high-throughput simulations.*

[66] P. Manganaris, *Machine learning modeling of hybrid organic-inorganic perovskites.*

[67] P. Manganaris, S. Desai, A. Mannodi-Kanakkithodi, and G. Kusne, *Machine learning in materials science: From basic concepts to active learning.*

[68] P. Gollapalli, P. Manganaris, and A. Mannodi-Kanakkithodi, "Graph neural network predictions for formation energy of native defects in zinc blende semiconductors," 2023, In Preparation. DOI: 00.0000/00000000.

[69] R. Edlabadkar, J. Yang, H. Rahman, P. Manganaris, E. P. Korimilli, and A. Mannodi-Kanakkithodi, "Driving halide perovskite discovery using graph neural networks," 2023, In Preparation. DOI: 00.0000/00000000.

[70] J. Yang, P. Manganaris, and A. Mannodi-Kanakkithodi, "Discovering novel halide perovskite alloys using multi-fidelity machine learning and genetic algorithm," In Preparation, 2023.