MULTI-FIDELITY MACHINE LEARNING FOR PEROVSKITE BAND GAP PREDICTIONS

by

Panayotis T. Manganaris

A Thesis

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Master of Science

School of Materials Engineering

West Lafayette, Indiana

May 2023

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Arun Mannodi-Kanakkithodi, Chair

School of Materials Engineering

Dr. Alejandro Strachan

School of Materials Engineering

Dr. Kendra Erk

School of Materials Engineering

Approved by:

Pending FORM 9 Approval

To my family

ACKNOWLEDGMENTS

I am grateful for the guidance of my advisor, Dr. Arun Mannodi-Kanakkithodi. The Ross Fellowship awarded to me by Purdue University Graduate School helped fund my work. The Rosen Center of Advanced Computing provided the computational resources needed to conduct simulations, store data, and train models.

PREFACE

This is the preface.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF LISTINGS

LIST OF PROTOCOLS

LIST OF SCHEMES

LIST OF SYMBOLS

γ photon

ABBREVIATIONS

HaP	halide perovskite
VASP	Vienna Ab initio Simulation Package
QM/ML	quantum mechanics machine learning
SLME	spectroscopic limited maximum efficiency
PCE	power conversion efficiency
DFT	density functional theory
GGA	generalized gradient approximation
PBE	Perdew-Burke-Ernzerhof Functional
HSE06	Heyd-Scuseria-Ernzerhof Functional
PCA	principal component analysis
t-SNE	t-distributed stochastic neighbor embedding
UMAP	uniform manifold approximation and projection
GPR	Gaussian Process Regression
RFR	Random Forest Regression
SISSO	Sure Independence Screening and Sparsifying Operator
SQS	special quasi-random structures
PAW	projector augmented wave
NIST	National Institute of Standards and Technology
PES	Potential Energy Surface
SHAP	Shapley Additive Explaination
GNN	Graph Neural Networks

NOMENCLATURE

MA Methylamonium (Cationic Methylamine) $CH_3NH_3^+$

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

GLOSSARY

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

deep learning	a paradigm of machine learning differing from classical learning in that the features of the input data are themselves learned by the algorithm
latent space	a multidimensional abstraction of a problem space. the relationship between coordinates in this space and observation in the real world can be formulated to guarantee the viability of solutions in the abstraction
evolutionary algorithms	a class of nature-inspire algorithms often applied to optimization in high dimensional discontinuous functions
ALIGNN	the Atomistic Line Graph Neural Network considers relative positions of atoms in a crystal as well as the relative angles between bonds by creating two related node and edge graphs and convluting them in a staggered manner together
Gaussian Process	Any function which returns samples from an underlying multivariate normal distribution
FAIR	Findable Accessible Interoperable and Reusable Data
multi-task learning	A type of machine learning where an algorithm learns multiple functions simultaneously, while exploiting commonalities and differences between the functions
Spin Orbit Coupling	An additional term intended to account for the increased relevance of quantum angular momentum to electromagnetic response in heavy atoms

ABSTRACT

A wide range of optoelectronic applications demand semiconductors optimized for purpose. My research focused on data-driven identification of ABX_3 Halide perovskite compositions for optimum photovoltaic absorption in solar cells. I identified mixtures of candidate constituents at the A, B or X sites of the perovskite supercell which leveraged how mixed perovskite band gaps "bow" away 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 2.5 eV for visible light spectrum absorption.

I trained machine learning models on previously reported datasets of halide perovskite band gaps based on first principles computations performed at different fidelities. The primary objective of these models was to predict the perovskite band gap using the composition and inherent elemental properties as descriptors, eventually leading to accurate prediction and screening across 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 ~40,000 hypothetical compounds, based on which we screened ~3000 stable compounds with band gap predicted to be between 1 and 2.5 eV at experimental accuracy.

Multi-fidelity Machine Learning for Perovskite Band Gap Predictions

Introduction

Multi-Fidelity Learning

The state of the art in materials modeling favors Graph Neural Network (GNN) architectures. (Chen et al., 2019; Choudhary & DeCost, 2021; Xie & Grossman, 2018) These deep learning models have sufficient flexibility to capture the continuous variability in relative positions of crystals and molecules. They are effective models, but they are difficult to use with physical materials. Accurately characterizing structures at a level of atomic granularity cannot be achieved even with state of the art 3D Electron Tomography techniques. (Ercius et al., 2015) Yet, characterization of chemical composition is a well established practice, for example using X-ray spectroscopy.

Graph convolutional neural networks can power more accurate structure-target predictions at multiple fidelities (Chen et al., 2020) by performing Multi-Task Learning (MTL). For instance, this multiple-fidelity machine learning technique can infer the relationships between more plentiful PBE GGA data and rarer but more accurate HSE06 data based on a shared set of predicting features. This relationship, if sufficiently general, can be used to reliably extrapolate from known points on the PBE co-domain to the unknown HSE co-domain. Of course, while this is implemented successfully in neural networks, the concept holds for any model architecture that can simultaneously regress multidimensional targets which do not need to constitute one rectangular data structure.

Additionally, there are alternative approaches for learning from multiple fidelities of data that can be implemented on the domain side. This circumvents the requirement for flexibility in encoding the co-domain. For instance modeling the multiple outcomes as varying depending on a categorical variable representing the fidelity makes it possible to use a single target regression methods. Our problem of accurately modeling low availability, high fidelity targets is approached in this way.

We will employ the domain-side approach where the largest, lowest fidelity component of our dataset consists of density functional theory (DFT) band gap predictions made at the generalized gradient approximation (GGA) Perdew-Burke-Ernzerhof (PBE) level of theory. On the other end, the smallest and highest fidelity subset of the sample consists of experimental measurements of physical devices collected from the literature.

Multiple Fidelity Dataset

Perovskite Band Gaps

We aim to accurately predict performance-relevant halide perovskite (HaP) band gaps, which strongly predict photovoltaic performance. (Mannodi-Kanakkithodi et al., 2019)

Furthermore, using multi-fidelity modeling we aim to predict the experimentally measured band gaps of compounds that have only been simulated to date. Our fidelity hierarchy climbs from DFT simulations performed using the basic PBE GGA functional, to results obtained from physical experiments aggregated in literature see table 1. (Almora et al., 2020; Kim et al., 2014; Swanson et al., 2017)

While we acknowledge the advantages of GNNs, we aim to express band gap as functions primarily of the perovskite composition. It is known the octahedral arrangement of perovskites is most relevant to their electronic structure, nevertheless, a strong understanding of the influence of chemical composition on performance will continue to be a priority as it is expected to aid in Feature Engineering.

Table 1. Sample counts by density functional represented in dataset

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

A detailed analysis of this combined hybrid organic-inorganic and purely inorganic HaP DFT dataset is covered in a prior article by J. Yang and Mannodi-Kanakkithodi (2022) and

in DFT Details. Naturally, the statistics obtained from each fidelity vary (figure 1). This is the primary challenge we will address with the categorically dimensioned multi-fidelity models discussed in Methods.

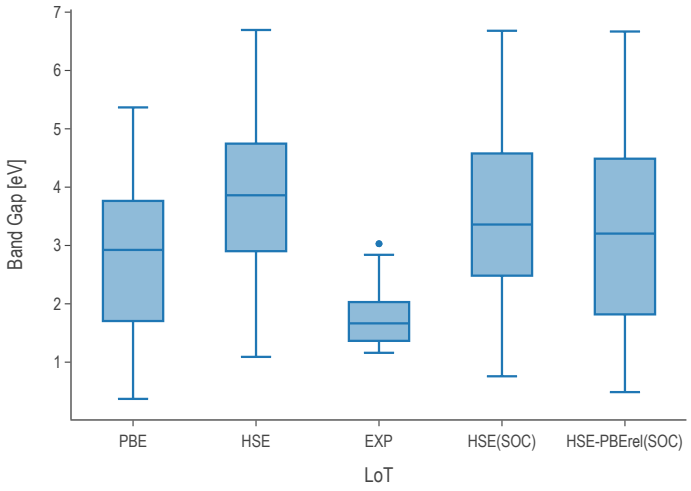

Figure 1. Variability in band gaps at each fidelity

Sampling

The simulations for each level of theory are performed on some number of members to a fixed subset of the total 37785 compositions that can be combinatorially generated in a 2x2x2 perovskite supercell when allowing *at most* single-site alloying with our 14 constituent candidates for 3 sites (table 2).

Table 2. ABX_3 Chemical Domain

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

Within this domain space, we try to maintain a balance in the share of samples that represent each one of the "cardinal mixing" categories. Additionally, within each mix we try

to maintain a reasonable balance of purely inorganic samples versus hybrid organic-inorganic samples. See Methods for details on how these categories were utilized in model development.

This previously reported dataset demonstrates very even coverage of the cardinal mixing domain as shown in figure 2. The clusters in this figure correspond to the alloying scheme of the data points

This sample provides the opportunity to comfortably interpolate the properties of other members of the cardinal mixing domain. Concurrently, we test the ability of our models to extrapolate with respect to alloying scheme as well as level of theory. It also provides an opportunity to investigate the statistical impact of constituent compounds on perovskite property prediction. See Discussion.

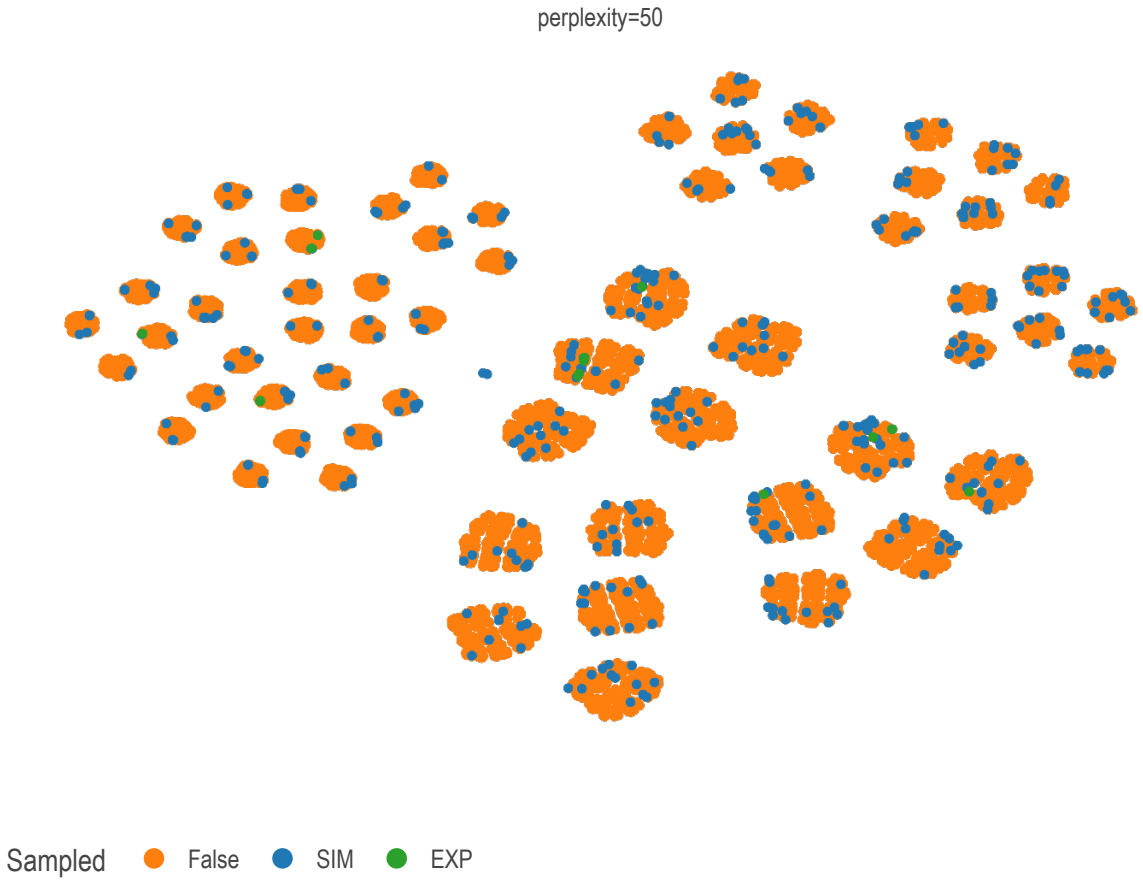

Figure 2. Samples overlaid on cardinal mixing chemical domain projected from fourteen to two dimensions via t-SNE

Model Optimization

The rigorous hyper-parameter Optimization (HPO) of any feature engineering and modeling pipeline is a problem discussed extensively in the literature. HPO approaches can be broadly separated into exhaustive and efficient optimization strategies. (L. Yang & Shami, 2020) We use a two-stage procedure for selecting the best model parameters.

The first stage is an exhaustive grid-search over diversely sampled parameter space. Each combination of parameters instantiates a model which is then fit to each of a set of stratified training subsets generated by a K=3 K-fold split cross-validation strategy. Every fitted model is subsequently tested against the cross-validation test sets and a suite of regression scoring metrics are applied to each member category simultaneously using a custom SciKit-learn score adapter[1]. The grid search is then narrowed to a high performance quadrant of the search space by the model evaluator based on recommendations made by a simple entropy minimization algorithm[1].

The recommended grid quickly eliminates under-performing settings based on the sample probability of a setting appearing in a set of finalists according to the scoring rankings. The selection score is additionally influenced by a weighted sum of the scoring ranks allowing for considerably tuning the selection criterion. For best results, a few different grid spaces should be 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.

Methods

DFT Details

The largest subdivision of 1400 compounds correspond to a series of optoelectronic DFT simulations. The simulated experiments are performed on some subset (table 1) of ~500,

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

Listing 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)
```

exclusively pseudo-cubic, ABX_3 supercells obtained by geometry optimization of modified structure files (Pilania et al., 2016) originally obtained from the Computational Materials Repository[2]. Each cell demonstrates an SQS mixed composition at none or one of each of the A, B, or X sites. See the coverage of this sample in figure 2.

Each relaxed structure is made in two ways. Once with the PBE GGA functional and once with the HSE06 functional. Band gaps are obtained using a static band structure calculation performed at the same and at higher levels of theory.

The chosen functionals each offer strengths and weaknesses. PBE is inexpensive but typically underestimates band gaps. HSE06 is orders of magnitude more expensive and may fail to converge structure relaxations but tends to be more trustworthy for electronic structure properties. HSE06 on PBE relaxation attempts to mitigate the disadvantages of each individually. The use of Spin Orbit Coupling helps to better electronic properties simulation in compounds containing Lead.

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[3] which robustly coverts the composition string of each data point into a $\alpha+\beta+\gamma$ dimensional composition vector. In the case of our total dataset description $\alpha+\beta+\gamma = 14$. (J. Yang & Mannodi-Kanakkithodi, 2022) In a subset of the data, the chemical vector (listing 2) is produced using cmcl (listing 1).

[2]↑https://cmr.fysik.dtu.dk/
[3]↑https://github.com/PanayotisManganaris/cmcl

Listing 2. Data frame of composition vectors generated by cmcl

Formula	FA	Pb	Sn	I	MA	Br
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. (Mannodi-Kanakkithodi & Chan, 2022) However, a spare descriptor is generally bad for extrapolative modeling. (Ghiringhelli et al., 2015)

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

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

- 14 sparse composition vectors extracted from chemical formula using `cmcl`[3]
- 36 dense site-averaged property space computed as a linear combination of composition vectors and measured elemental properties (Mentel, 2014)
- 5 categorical dimensions one-hot-encoding level of theory.
  - this provides the categorical axis for multi-task learning
  - see table 1

Machine Learning Algorithms and Parameter Optimization

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

We are careful to maintain the diversity of mixing types and hybrid-organic/inorganic samples within each fidelity subset. We expect this will help to ensure the models learn relationships between fidelities, not differences in alloy scheme or constituency distributions within each fidelity.

Each model architecture is rigorously optimized with regard to both 1) generality over the domains of Perovskite compositions and site-averaged constituent properties and 2) generality over the domain of alloy classifications.

In order to monitor for possible categorical biases effecting regressions, nine metrics are used to evaluate the performance of each model over all alloy types at every stage of the hyper-parameter optimization. This is done simultaneously, only models that perform uniformly well on all alloy types are selected.

We expect perovskites of a given alloy class and of a given hybrid-organic/inorganic status will perform significantly differently with respect to a particular application compared to perovskites of a another class or status. We attempt to make models that reasonably explain this high entropy mixing diversity by utilizing the cardinal mixing represented in our sample.

We do this by training each model using two test/train splits. First, the optimal model parameters are chosen for their performance under a random split. A minimum of 3-fold cross-validation is performed for every set of model parameters that is considered.

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

Two separate cross validation schemes are employed at each stage of the design process. First, the sample set is shuffled once and split to mitigate the models tendency to fit on sample order, then, stratified K-folds are generated in manner consistent with the types of each sample. The regressor is then trained on the subsets of each class. Its ability to extrapolate is independently metered on each validation fold consisting of members of the other classes.

Second, the ability for a model trained on samples belonging to one class/status to extrapolate to samples of another class/status is tested as well. The samples again are shuffled and split. then the training set is separated using a grouping K-fold split strategy.

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

Feature Engineering

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

SIS[4] is a powerful application of compressed sensing. (Ghiringhelli et al., 2017) The SIS operator is a potent dimensionality reduction technique. It does not perform any mathematical decomposition but instead picks existent dimensions that begin to approximate an orthogonal basis. It outperforms CUR decomposition by functioning effectively in extremely high rank vector spaces. (Hamm & Huang, 2019; Ray et al., 2021) This is accomplished by posing the decomposition as a compressed sensing problem in the correlation metric space.

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

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

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

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

SISSO is run for our dataset on the same partitioning scheme used by the previous models via an SciKit-learn compliant (Buitinck et al., 2013) interface[5] extensively modified from the original Matgenix[6] code. Additionally, the algorithm is informed of features units so that it is restricted to meaningful linear combinations. SIS features complexity is restricted to

[4]↑https://github.com/rouyang2017/SISSO
[5]↑https://github.com/PanayotisManganaris/pysisso
[6]↑https://github.com/Matgenix/pysisso

a maximum of 3 operations primarily to encourage parsimonious descriptions. The available operation set is outlined in table 3.

Table 3. operations for formation of combinatorial super-space

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

Results

Best Models on Raw Domain

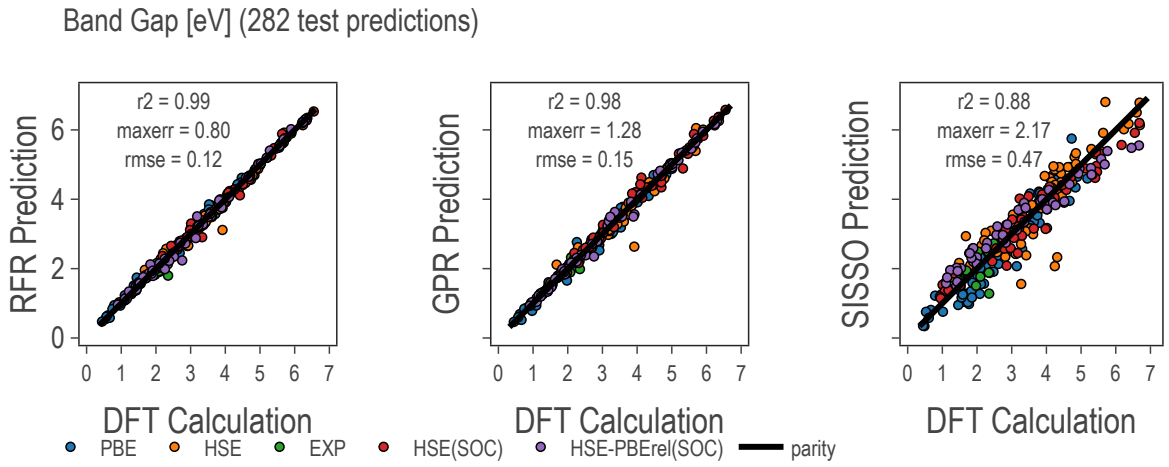

Figure 3. model predictions vs true values at multiple fidelities

The optimized models are high performing (Table 4). The RFR hyper-parameters are listed in the appendix (Table **??**).

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

SISSO Model and SIS Engineered Features

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

The best SISSO model for band gap involving 3 SIS features (each composed of up to 4 basic features) has an unremarkable RMSE of 0.476 eV, barely outperforming an OLS regression on 55 dimensions (see Table 4). It is expressed in equation 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.

$$
\begin{aligned}
bg\ [\text{eV}] = {} & 1.752393064((X; \text{electronegativity} * A; \text{heat of fusion}) - (B; \text{electron affinity} + B; \text{ionization ener} \\
& + - 0.5862929089((B; Sn - \text{HSE}) + (\text{PBE} - X; \text{electronegativity})) \\
& +1.063684923((A; \text{electronegativity} - B; Ca) * (B; \text{heat of vap} - X; \text{electron affinity})) \\
& +4.657097107 \qquad (1)
\end{aligned}
$$

Table 4. RMSE of models on raw domain calculated per LoT subset

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

Computing and combining more than 3 SIS features is not rewarding of the computational expense. Residuals are increasingly uncorrelated with the generated SIS features and model accuracy gains do not outstrip complexity. However, in the process of creating Equa-

tion 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.

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 Featurization of Chemistries.

fitting models to SIS features may leverage the denser and more continuous domain to improve extrapolative predictions. Potentially into the high-entropy domain, or simply Theory. However using the SIS subspaces in this way compromises on SISSO's explicability and necessitates SHAP analysis. Unfortunately, whatever the gains in training time complexity and extrapolative ability, the models underperformed in predicting band gap in the cardinal mixing domain (see Table 4). 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.

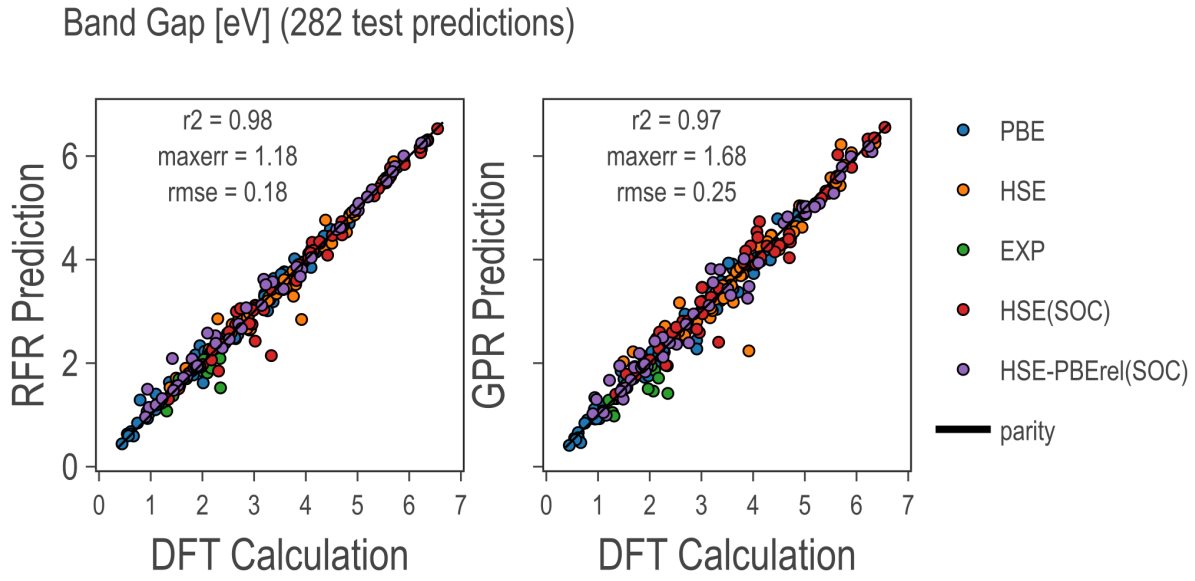

Figure 4. SIS-based model predictions vs true values at multiple fidelities

Discussion

SHAP Analysis of Domain

SHAP scores are computed automatically for every dimension of every sample in the domain by the python SHAP package[7]. The sum of the expectation value of the target conditioned on the model features and the SHAP scores computed for each predictor variable of a sample is the model's prediction for that sample target. (Lundberg & Lee, 2017) For the perovskite band gap the expectation value is 2.836 when conditioned on the raw features and 2.863 when conditioned on the SIS features. The raw features' SHAP values are more centered around zero while engineered features are more often scored decisively positive or negative.

Figures 5 and 6 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 5, the B-site Electronegativity is often a strongly positive contributor to the RFR prediction. However, almost always in this case it is out-contributed by other features, it does not mostly determine the result but it is still valuable. On the contrary, when it is a strongly negative contributor it effectively determines the result. It is interesting to see how models make use of features in light of basic bi-variate correlations. The only features that correlate strongly with band gap are illustrated in figure 7. 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,

[7]↑https://github.com/slundberg/shap

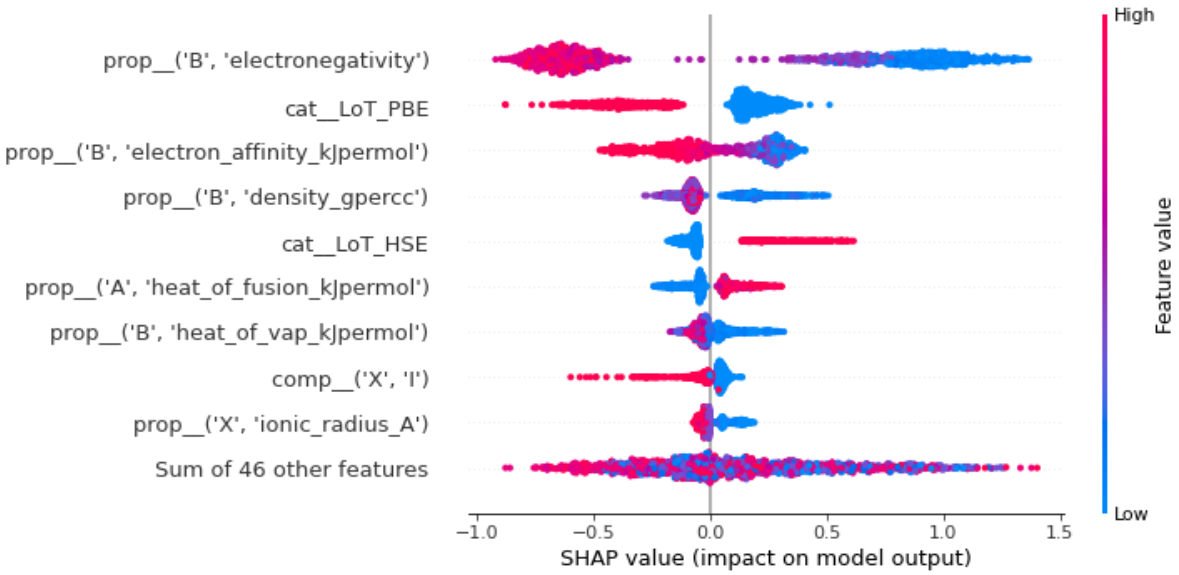

Figure 5. Random Forest Regression Band Gap SHAP Values

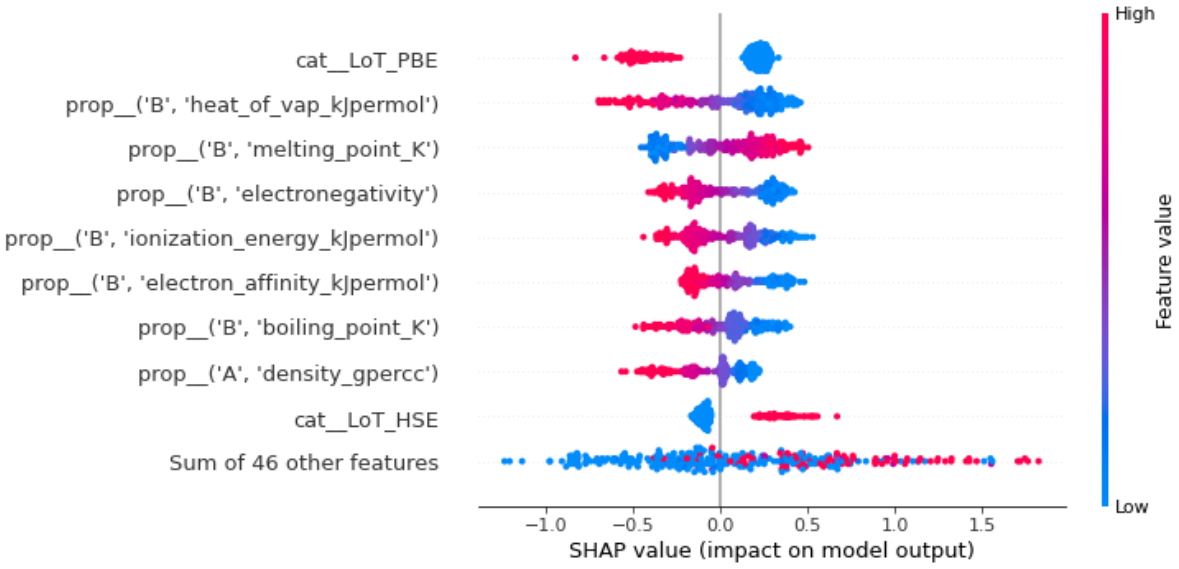

Figure 6. Gaussian Process Regression Band Gap SHAP Values

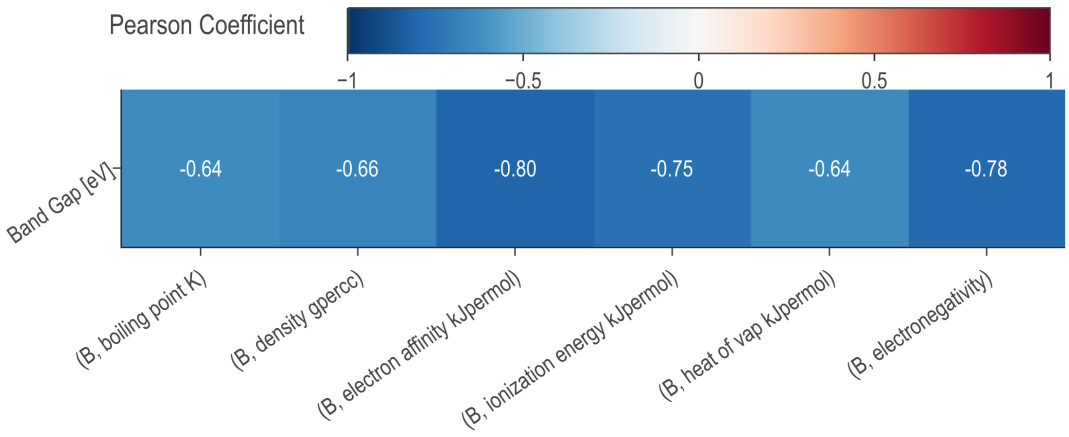

Figure 7. raw features with ($|p| > 0.5$) against band gap

alloy scheme, or presence of organic A-site occupants. This analysis confirms the difficulty of deducing a rule of thumb for the synthesis of perovskites with desirable properties. If anything, figure 8 confirms that the Iodine at the X site tends to slightly increase band gaps.

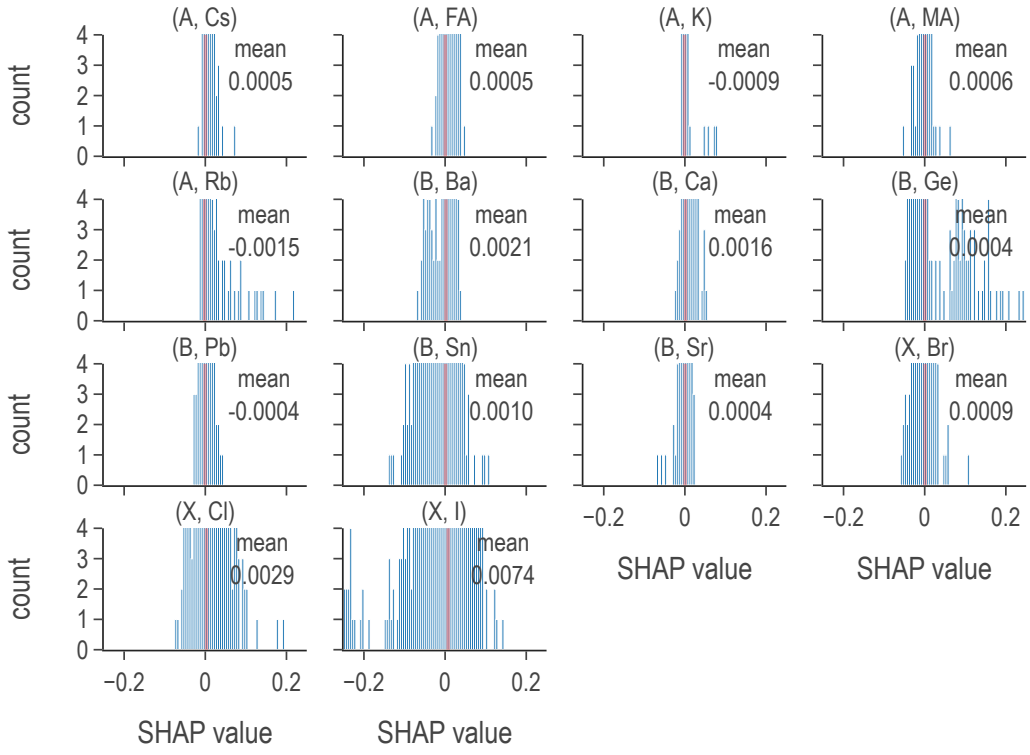

Figure 8. SHAP score distributions reveal effects of individual constituents

Predictions and Screening

Using the superior RFR model, we made predictions of 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) and according to the scores reported in Table 4, the RFR model is capable of predicting band gaps at the experimental fidelity with a 0.154448

rmse. (J. Yang et al., 2023) These predictions were projected on the sample space in Figure 9.

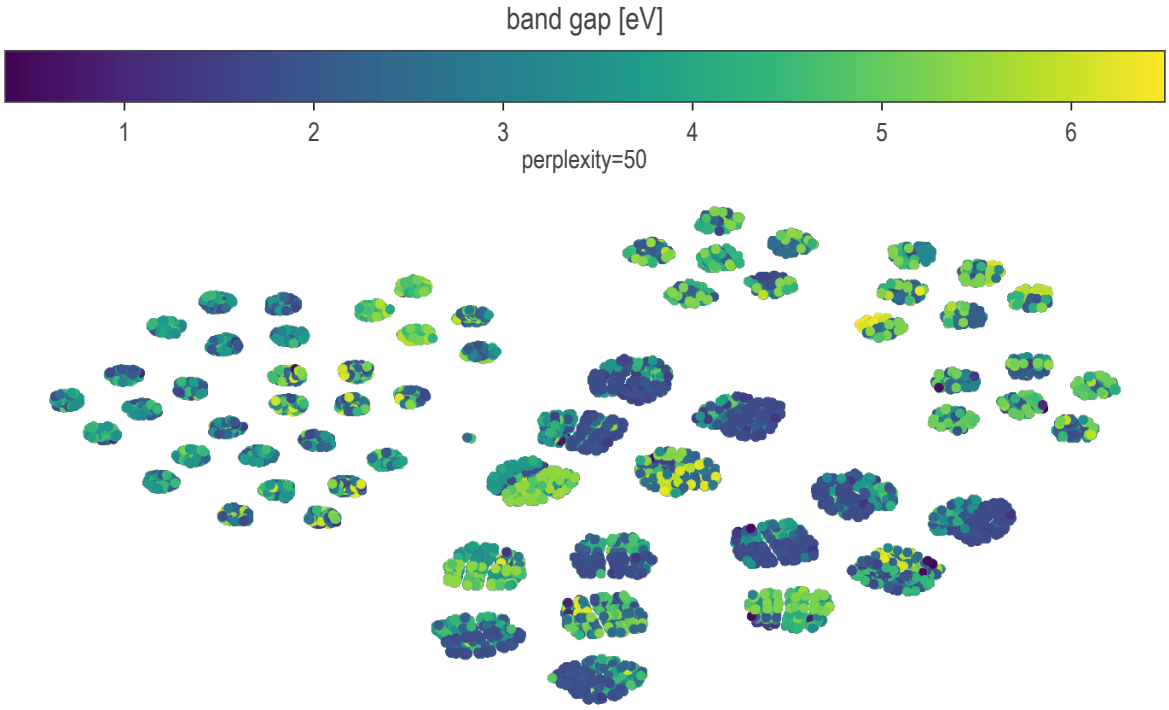

Figure 9. Band gap predictions overlaid on cardinal mixing chemical domain projected from fourteen to two dimensions via t-SNE

We followed a similar high-throughput screening procedure to that laid out in prior works. (Mannodi-Kanakkithodi & Chan, 2021; J. Yang et al., 2023) We selected for band gaps between 1.0 and 2.5 eV as this range is expected to yield the best power conversion efficiency (PCE) in the visible spectrum. (Shockley & Queisser, 1961; Yu & Zunger, 2012) Perovskite compounds were selected for their predicted stability by cutting on each of three tolerance factors. Namely the Goldschmidt's tolerance, the octahedral tolerance, and the tolerance proposed by Bartel Christopher et al. (2019).

These cuts trimmed the data set from ~40000 points to a subset of 3247 viable candidates. These selected candidates were projected onto the domain space in Figure **??**. 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 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 10.

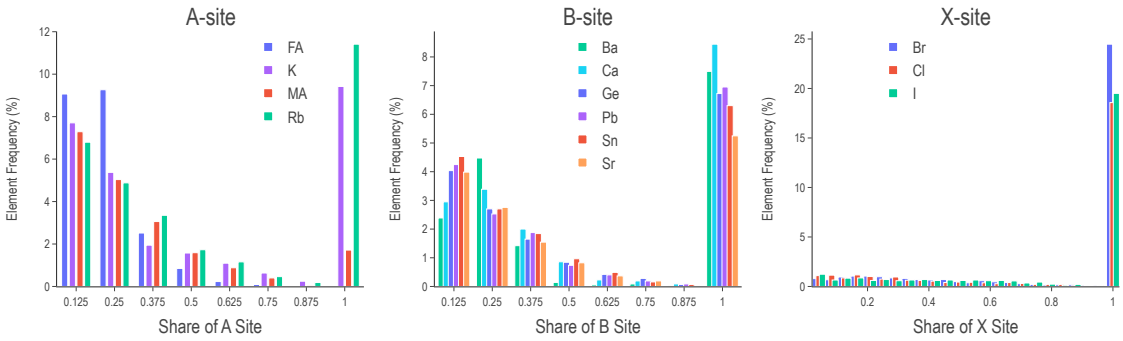

Figure 10. Frequency of mixing fractions of species at the A, B, and X sites across the ~3000 screen compounds

REFERENCES

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

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

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

Chen, C., Ye, W., Zuo, Y., Zheng, C., & Ong, S. P. (2019). Graph networks as a universal machine learning framework for molecules and crystals. *Chemistry of Materials*, *31*(9), 3564–3572. https://doi.org/10.1021/acs.chemmater.9b01294

Chen, C., Zuo, Y., Ye, W., Li, X., & Ong, S. P. (2020). Multi-fidelity graph networks for machine learning the experimental properties of ordered and disordered materials. *CoRR*. http://arxiv.org/abs/2005.04338v1

Choudhary, K., & DeCost, B. (2021). Atomistic line graph neural network for improved materials property predictions. *npj Computational Materials*, *7*(1), 185. https://doi.org/10.1038/s41524-021-00650-1

Ercius, P., Alaidi, O., Rames, M. J., & Ren, G. (2015). Electron tomography: A three-dimensional analytic tool for hard and soft materials research. *Advanced Materials*, *27*(38), 5638–5663. https://doi.org/10.1002/adma.201501015

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Pilania, G., Mannodi-Kanakkithodi, A., Uberuaga, B. P., Ramprasad, R., Gubernatis, J. E., & Lookman, T. (2016). Machine learning bandgaps of double perovskites. *Scientific Reports*, *6*(1). https://doi.org/10.1038/srep19375

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

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

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

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

Xie, T., & Grossman, J. C. (2018). Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical Review Letters*, *120*(14). https://doi.org/10.1103/physrevlett.120.145301

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

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

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

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