Multi-fidelity Machine Learning Pervoskite Composition vs Band Gap

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

(Mannodi Research Group)

School of Materials Engineering,

Purdue University, West Lafayette, Indiana 47907, USA

(Dated: 5 January 2023)

[a)]Electronic mail: amannodi@purdue.edu

We report on the details of creating models of halide perovskite properties based on composition and derived descriptors. The primary objective of these models is to eventually recommend perovskite alloy compositions corresponding to targeted properties. Here targets are chosen to yield high photovoltaic (PV) performance. So, we focus on models of the electronic band gap. We leverage the Purdue University nanoHUB, an NSF-funded, Purdue-hosted computational repository, to host literate reproducible notebooks documenting our model development workflow[1]. We thus enable the scientific community to utilize our approach for modeling performance targets for a wider range of promising compounds.

We explore a variety of machine learning (ML) models for the prediction of Perovskite bandgap. A rigorously optimized Random Forest Regressor (RFR), a Gaussian Process (GP) Regressor, and a Sure Independent Screening and Sparsifying Operator[2] (SISSO) regressor.

Approximately 1500 physical and synthetic records spanning various experimental fidelities and alloy schemes are used in model development. All experiments are conducted for one of ~500 perovskite compositions. ~1400 experiments are performed computationally using Density Functional Theory (DFT), ~100 are physical measurements obtained from published literature[3–5].

1. 500 PBE relaxations -> PBE Density of States (DoS) calculation
2. 300 HSE06 relaxations -> HSE06 DoS
3. 300 HSE06 relaxations -> HSE06 + Spin-Orbit Coupling (SOC) DoS
4. 300 PBE relaxations -> HSE06 + SOC DoS
5. 100 experimental band gap measurements

Our models are based primarily on composition information. We implement generic feature extraction by parsing a string encoding the ABX_3 perovskite formula corresponding to each record. The resulting 14-dimensional composition vector is easily obtained for experimental and synthetic data alike. This is a sufficient predictor variable, nonetheless we continue. Secondarily, we also examine 36 additional predictor variables computed as linear combinations of these compositions and certain elemental properties obtained from the trusted Mendeleev databases[6]. Finally, additional fidelity features are one-hot-encoded with the aim of improving model accuracy. In future work, we anticipate adding descriptors based on phase and structural information.

We finally compare the band gap models based on this basic 55 dimensional descriptor and models based on an engineered domain we produced to improve model efficiency, performance, and interpretability.

INTRODUCTION

Multi-Fidelity Learning

The state of the art in materials modeling favors Graph Neural Networks (GNN)[7–9] architectures. These deep learning models have sufficient flexibility to capture the continuous variability in relative positions of crystals and molecules. They are extremely 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[10]. Yet, characterization of chemical composition is a well established practice i.e using X-ray spectroscopy.

Graph convolutional neural networks can power more accurate structure-target predictions at multiple fidelities[11] 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 HSE06 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 multi-task learning approaches that can be implemented on the domain side. This circumvents the requirement for flexibility in encoding the co-domain, making it possible to use a single target regression methods to learn rules for multiple outcomes that vary depending on a categorical variable representing the fidelity. In general, the problem of accurately modeling low availability, high fidelity targets is approached using MTL to learn regressions on datasets compiled either along X or Y from measurements taken at multiple level of theory.

Semi-supervised learning[12,13] is a competing set of methods achieving similar outcomes.

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 Functional (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.

Dataset Overview

Perovskite Band Gaps

we aim to accurately predict performance-relevant Halide perovskite (HaP) band gaps which are strongly predictive of photovoltaic performance[14].

Furthermore, using MTL 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[3,15,16] see table I.

While we acknowledge the advantages of GNN, we aim to express band gap as functions primarily of the perovskite composition. It is known this effort is suboptimal especially as 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 inverse design and in Feature Engineering.

TABLE I: Density Functionals vs Sample Counts

	LoT
PBErel	517
HSErel(SOC)	299
HSErel	297
HSE-PBE(SOC)	244
EXP	44
Total	1401

A detailed analysis of this combined hybrid organic-inorganic and purely inorganic HaP DFT dataset is covered in a prior article by Yang, Manganaris, and Mannodi Kanakkithodi[17] and in DFT Details.

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

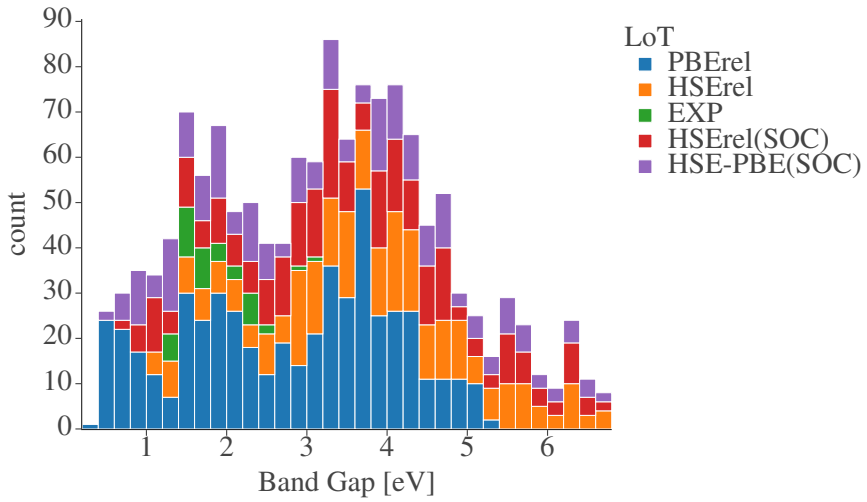


FIG. 1: Variability in band gap per level of theory

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 II).

TABLE II: 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 sample 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 figure 2. See Methods for details on how these categories were utilized in model development.

The design of this dataset provides an opportunity to assess 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 Results and Discussion.

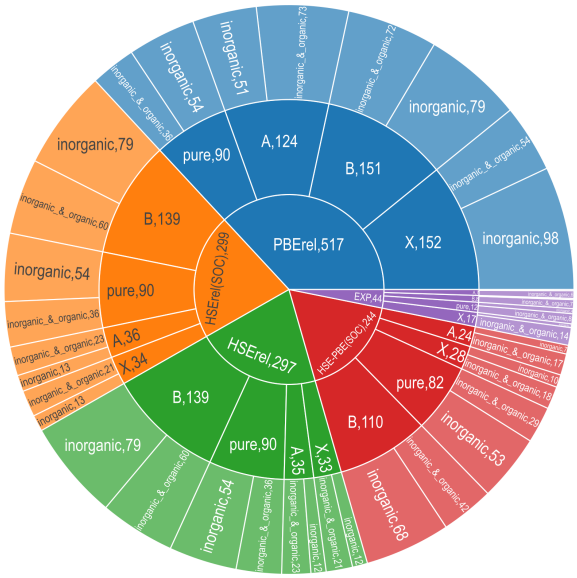


FIG. 2: Share by count of total data apportioned from each experimental subcategory

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[18]. 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].

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

In general, 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 using 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 I) of ~500, exclusively pseudo-cubic, ABX_3 supercells obtained by geometry optimization of modified structure files[19] 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.

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 level 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.

[2] `https://cmr.fysik.dtu.dk/`

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 dataset description[17] $\alpha+\beta+\gamma=14$.

```
subset = [883,886]
df = Y.Formula[subset].to_frame().ft.comp()
df.index = Y.Formula[subset]
print(df)
```

Listing 1: An example of the cmcl "ft" feature accessor

	MA	FA	Pb	I	Br
Formula					
MA0.7FA0.3PbI3	0.7	0.3	1	3.00	NaN
MAPb(I0.41Br0.59)3	1.0	NaN	1	1.23	1.77

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[20], However, a spare descriptor is generally bad for extrapolative modeling[21]. When extrapolation is the aim, continuously distributed, unique, and linearly independent features are much more reliable[22]. Our attempts to provide a domain with these characteristics results in the following raw feature space.

- 14 sparse composition vectors extracted from chemical formula
  - generated using cmcl[3]
  - see Predictor Variables
- 36 dense site-averaged property space
  - computed as a linear combination of composition vectors and measured elemental properties[6]

[3] https://github.com/PanayotisManganaris/cmcl

  - see Predictor Variables

- 5 categorical dimensions one-hot-encoding level of theory.

  - this provides the domain-side categorical axis for multi-task learning

  - see table I

Machine Learning Algorithms and Parameter Optimization

We train RFR and 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. We hope Shapley Additive Explaination (SHAP) analysis of the models will lend 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[23] v1.2.

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 (See Learning Curves). Finally,

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

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

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

A final best model is instantiated using the overall best performing parameters. These models are finally validated against the test sets originally split off from the sample in both their extrapolative ability and consistency across groups.

This procedure in demonstrated in an online notebook by Manganaris, Desai, and Kanakkithodi [1] hosted on the Purdue NanoHUB.

Feature Engineering

There has been success in creating analytical expressions for perovskite properties, particularly lattice parameters[4]. In an attempt to find an analytical predictor for band gap we employ the Sure Independence Screening and Sparsifying Operator (SISSO)[2].

SIS[4] is a powerful application of compressed sensing[24]. 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[25,26] decomposition by functioning effectively in extremely high rank vector spaces. This is accomplished by posing the decomposition as a compressed sensing problem in the correlation metric space. We provide a Sci-Kit Learn

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[27,28]. 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.

[4] `https://github.com/rouyang2017/SISSO`

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[29] yielding more orthogonal basis sets that can be incorporated into the model

SISSO is run for our dataset on the same partitioning scheme used by the previous models via an scikit-learn compliant[30] 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 a maximum of 3 operations primarily to encourage parsimonious descriptions. The available operation set is outlined in table III.

TABLE III: operations for formation of combinatorial super-space

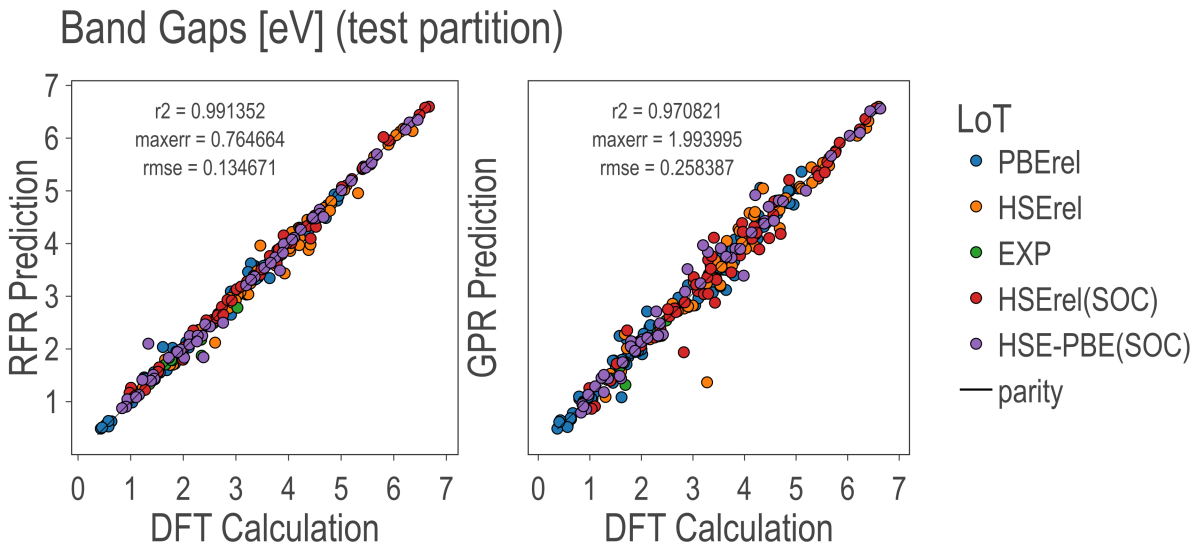
Binary (dimensional)	Unary (dimensionless)
addition	reciprocation
subtraction	power 2
multiplication	power 3
division	natural logarithm
	exponentiation
	root 2

[5] `https://github.com/PanayotisManganaris/pysisso`

[6] `https://github.com/Matgenix/pysisso`

RESULTS AND DISCUSSION

Best Models on Raw Domain



The exhaustively optimized models are obviously high performing (Table IV). The RFR hyperparameter are listed in HPO Summary Tables (Table V). The GPR model is tried with multiple kernels. Ultimately, the best is a non stationary Matern kernel with $\nu = \frac{3}{2}$.

SHAP Analysis of Domain

SHAP scores are computed automatically for every dimension of every sample in the domain. The sum of SHAP scores computed for each predictor variable of a sample is the model's prediction for that sample[31].

Figures 3 and 4 show the aggregated score results. Within 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 z-axis, feature value gives a sense of how often the feature is a relatively minor contributor to a prediction.

For instance, in figure 3, the B-site Electronegativity is fairly 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, the model uses small positive contributions from other features to claw back up to a positive valued band gap.

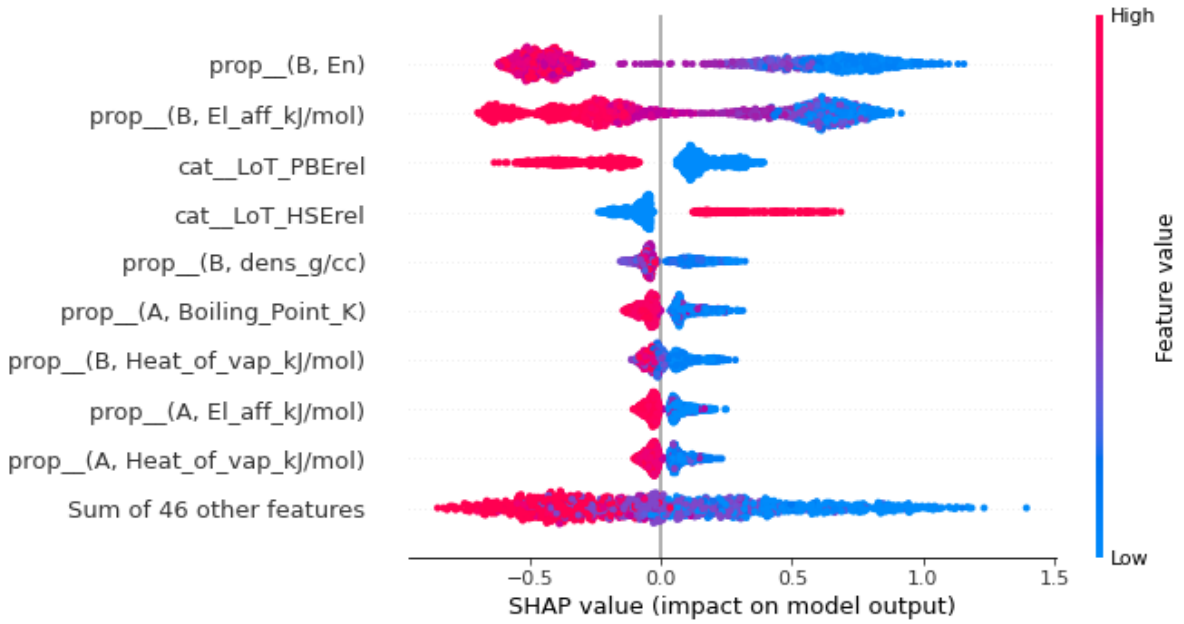


FIG. 3: Random Forest Regression Band Gap SHAP Values

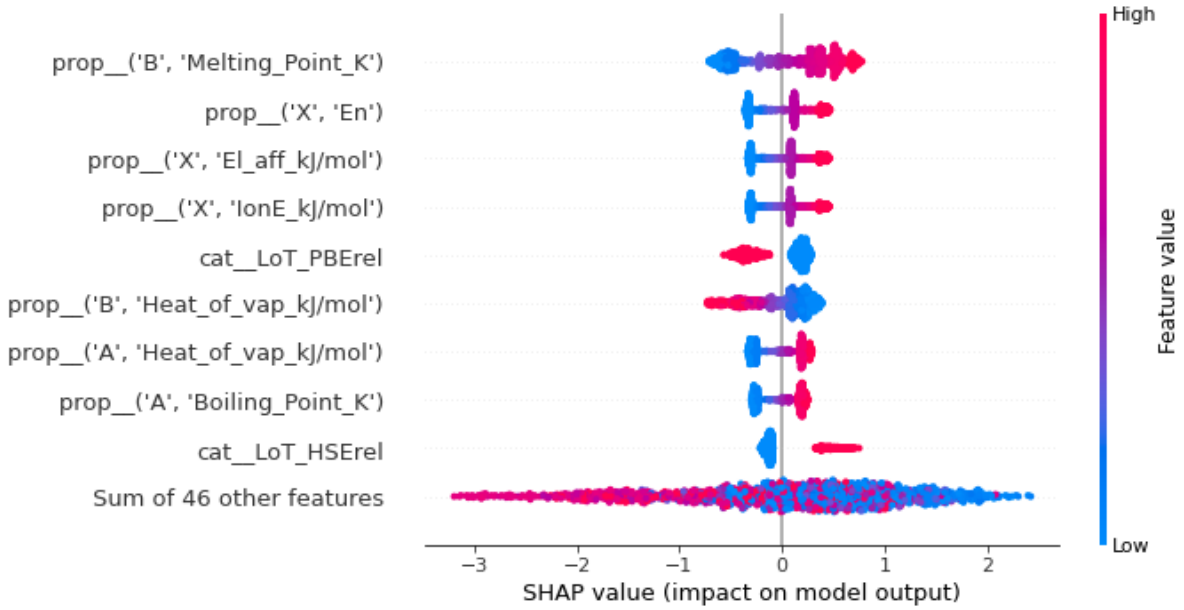


FIG. 4: Gaussian Process Regression Band Gap SHAP Values

These SHAP scores can be used to gain insights into the contributions of site members and site member properties to the perovskite band gap.

Comparing Feature Usefulness to Raw Correlations

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 5.

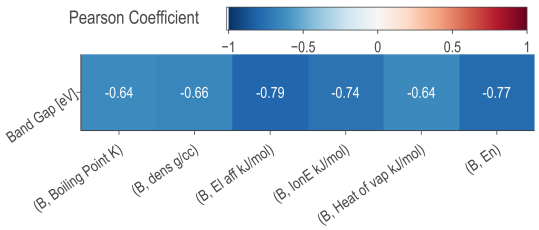


FIG. 5: raw features with ($|p| > 0.5$) against band gap

SISSO Model and Incidental Engineered Features

The Sure Independence Screening and Sparsifying Operator (SISSO)[2,24] is a specific combination of multiple data mining techniques chained together resulting in a symbolically expressed regression model.

The best model for band gap returned by 5 trials of SISSO involves 3 SIS features and has an unremarkable RMSE of 0.4625 eV, barely outperforming an OLS regression on 55 dimensions (see Table IV). 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.

$$\begin{aligned} bg\ [\text{eV}] = 1.752075117&((X;\text{Electronegativity} * A;\text{Heat of Fusion}) \\ &- (B;\text{Ionization Energy} + B;\text{Electron affinity})) \\ -0.5759612116&(B;\text{Sn} + X;\text{Z} - \text{HSErel} - \text{PBErel}) \\ +1.074246385&((A;\text{Electronegativity} - B;\text{Ca}) \\ &\times (B;\text{Heat of Vaporization} - X;\text{Ionization Energy})) \\ +5.254074603&\ [\text{kJ/mol}] \end{aligned} \tag{1}$$

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

TABLE IV: RMSE of models on raw domain calculated per LoT subset

Score Categories	GPR	RFR	Linear OLS	SISSO
rmse	0.258387	0.134671	0.499558	0.474754
rmse EXP	0.157147	0.188727	0.307186	0.330080
rmse PBE	0.204187	0.102778	0.472430	0.395827
rmse HSE	0.337971	0.170726	0.558077	0.519706
rmse HSE(SOC)	0.275642	0.102896	0.535087	0.572644
rmse HSE-PBE(SOC)	0.245003	0.162380	0.466364	0.470758

gains do not outstrip complexity. However, in the process of creating Equation 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.

We set the aim of decreasing $\mathcal{O}(n^3)$ computational expense of GPR by $\approx$10 times. So, we aim to take 20 highly correlated features (slightly less than one half the number used by prior models) from these SIS subspaces. We expect this to solve the problems inherent to the raw Featurization of Chemistries

Computing GPR in conjunction with SIS features, we can create extrapolative models on a much more continuous domain. Using GPR on the SIS subspaces also leverages SISSO's explicability while obtaining uncertainty estimates, which is helpful for our inverse design ambitions.

Best Models on Engineered Domain

FUTURE WORK

Data Science and Materials Engineering

This has been an exercise in explaining a key performance determining property of Halide perovskite (HaP) using only composition information. Of course, the point of this basic approach is to enable experimental data to be more easily integrated into predictive models than is possible with graph-based Convolutional Neural Networks.

Naturally, this invites effort to incorporate much mode experimental data. Some sources by

Briones, Guinto, and Pelicano[5], Jacobsson *et al.*[32] are prime for consideration.

Notably, in this work, we assume that all compounds involved in this model manifest in the same pseudo-cubic phase for simplicity. Unfortunately, this is not nearly accurate to reality. Our experience with experimental collaborators indicates effective application of these models to practical synthesis requires that realistic phase information is somehow incorporated. Ideally this information can be encoded in an interoperable way for both computational and physical experiments.

Finally, these models have iterated on prior composition-only models of band-gap for inverse design[20]. They achieve nearly 50% improvement in accuracy by leveraging more data and, in the case of models on the engineered domain, higher quality predictor variables. Using the best performing model and an accompanying model designed to predict stability in an improved, data-driven perovskites design framework follows naturally. In particular, the GPR should drive an active learning approach.

An ongoing goal will be to grow a database of experimentally examined perovskite photovoltaic prototypes with well defined structures.

Software Tools

A couple of libraries are in development for easing the aggregation, accessing, sharing, and analysis of this data. The current database is packaged in "cmcl" at `http://github.com/PanayotisManganaris/cmcl` under the tag v0.1.5. In this early stage of development, cmcl strives to provide an "inquisitive" interface to perovskite composition feature computers in the style of the pandas API. At its current stage, it has been useful for extracting composition vectors from the formula strings identifying each compound.

A library of model evaluation tools to assist with exhaustive grid search is being maintained in the "yogi" repository at `http://github.com/PanayotisManganaris/yogi`

REFERENCES

[1]P. Manganaris, S. Desai, and A. Kanakkithodi, en"Mrs computational materials science tutorial," (2022).

[2]R. Ouyang, S. Curtarolo, E. Ahmetcik, M. Scheffler, and L. M. Ghiringhelli, Phys. Rev. Materials **2**, 083802 (2018).

[3]O. Almora, D. Baran, G. C. Bazan, C. Berger, C. I. Cabrera, K. R. Catchpole, S. Erten-Ela, F. Guo, J. Hauch, A. W. Y. Ho-Baillie, T. J. Jacobsson, R. A. J. Janssen, T. Kirchartz, N. Kopidakis, Y. Li, M. A. Loi, R. R. Lunt, X. Mathew, M. D. McGehee, J. Min, D. B. Mitzi, M. K. Nazeeruddin, J. Nelson, A. F. Nogueira, U. W. Paetzold, N. Park, B. P. Rand, U. Rau, H. J. Snaith, E. Unger, L. Vaillant-Roca, H. Yip, and C. J. Brabec, Advanced Energy Materials **11**, 2002774 (2020).

[4]L. Jiang, J. Guo, H. Liu, M. Zhu, X. Zhou, P. Wu, and C. Li, Journal of Physics and Chemistry of Solids **67**, 1531 (2006).

[5]J. Briones, M. C. Guinto, and C. M. Pelicano, Materials Letters **298**, 130040 (2021).

[6]L. Mentel, "mendeleev – a python resource for properties of chemical elements, ions and isotopes," .

[7]C. Chen, W. Ye, Y. Zuo, C. Zheng, and S. P. Ong, Chemistry of Materials **31**, 3564 (2019).

[8]K. Choudhary and B. DeCost, npj Computational Materials **7**, 185 (2021).

[9]T. Xie and J. C. Grossman, Physical Review Letters **120** (2018), 10.1103/physrevlett.120.145301.

[10]P. Ercius, O. Alaidi, M. J. Rames, and G. Ren, Advanced Materials **27**, 5638 (2015).

[11]C. Chen, Y. Zuo, W. Ye, X. Li, and S. P. Ong, CoRR (2020), arXiv:2005.04338v1 [cond-mat.mtrl-sci].

[12]O. Chapelle, B. Schölkopf, and A. Zien, *Semi-Supervised Learning* (The MIT Press, Cambridge, Mass, 2006).

[13]D.-H. Lee, ICML 2013 Workshop : Challenges in Representation Learning (WREPL) (2013).

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

[15]J.-P. Kim, J. A. Christians, H. Choi, S. Krishnamurthy, and P. V. Kamat, The Journal of Physical Chemistry Letters **5**, 1103 (2014), pMID: 26274456, https://doi.org/10.1021/jz500280g.

[16]D. E. Swanson, J. R. Sites, and W. S. Sampath, Solar Energy Materials and Solar Cells **159**, 389

(2017).

[17]J. Yang, P. Manganaris, and A. Mannodi Kanakkithodi, "A high-throughput computation dataset of halide perovskite alloys," (2022), in Preparation.

[18]L. Yang and A. Shami, Neurocomputing **415**, 295 (2020).

[19]G. Pilania, A. Mannodi-Kanakkithodi, B. P. Uberuaga, R. Ramprasad, J. E. Gubernatis, and T. Lookman, Scientific Reports **6** (2016), 10.1038/srep19375.

[20]A. Mannodi-Kanakkithodi and M. K. Y. Chan, Energy Environ. Sci. **15**, 1930 (2022).

[21]L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, and M. Scheffler, Physical Review Letters **114** (2015), 10.1103/physrevlett.114.105503.

[22]T. C. H. Lux, L. T. Watson, T. H. Chang, Y. Hong, and K. Cameron, Numerical Algorithms **88**, 281 (2020).

[23]F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, Journal of Machine Learning Research **12**, 2825 (2011).

[24]L. M. Ghiringhelli, J. Vybiral, E. Ahmetcik, R. Ouyang, S. V. Levchenko, C. Draxl, and M. Scheffler, New Journal of Physics **19**, 023017 (2017).

[25]P. Ray, S. S. Reddy, and T. Banerjee, Artificial Intelligence Review **54**, 3473 (2021).

[26]K. Hamm and L. Huang, CoRR (2019), arXiv:1903.09698v2 [math.NA].

[27]R. Tibshirani, Journal of the Royal Statistical Society: Series B (Methodological) **58**, 267 (1996).

[28]N. Gauraha, Resonance **23**, 439 (2018).

[29]D. G. Mayo, *Error and the Growth of Experimental Knowledge* (1996).

[30]L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux, in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning* (2013) pp. 108–122.

[31]S. Lundberg and S.-I. Lee, CoRR (2017), arXiv:1705.07874 [cs.AI].

[32]T. J. Jacobsson, A. Hultqvist, A. García-Fernández, A. Anand, A. Al-Ashouri, A. Hagfeldt, A. Crovetto, A. Abate, A. G. Ricciardulli, A. Vijayan, and et al., Nature Energy (2021), 10.1038/s41560-021-00941-3.

GLOSSARY

cardinal mixing: Describes perovskite alloys where no more than one of the A, B, or X sites is occupied by multiple possible constituents. 5, 9

cross-validation: Method for gathering statistics on the abilities of a model to fit to the parent partition. 6

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. 3

DFT: density functional theory. 4, 7

FA: Formamidinium. 5

features: attributes of an observed event or object which might empirically explain the event or object. 3, 9–12, 14, 15

GGA: generalized gradient approximation. 3, 4, 7

GNN: Graph Neural Networks. 3, 4

GPR: Gaussian Process Regression. 9, 12, 15, 16

groupwise K-fold: Data partition divided into K-folds where each fold corresponds to a category label. 10

HaP: halide perovskite. 4, 15

HSE06: Heyd-Scuseria-Ernzerhof Functional. 3, 7

hyper-parameter: a setting that controls how a learning algorithm works. 6, 12

K-fold split: Data partition divided into K arbitrary groups for use in cross-validation schemes. 6

level of theory: Refers to the rank of a [[ACRshort:dft][DFT]] functional in the hierarchy of phenomenological comprehensiveness. A proxy for accuracy.. 3–7, 9

MA: Methylammonium. 5

multi-task learning: A type of machine learning where an algorithm learns multiple functions simultaneously, while exploiting commonalities and differences between the functions. 3, 9

PBE: Perdew-Burke-Ernzerhof Functional. 3, 4, 7

RFR: Random Forest Regression. 9, 12

SHAP: Shapley Additive Explaination. 9, 12, 13

SISSO: Sure Independence Screening and Sparsifying Operator. 10, 11, 14

Spin Orbit Coupling: An additional term intended to account for the increased relevance of quantum angular momentum to electromagnetic response in heavy atoms. 7

SQS: special quasi-random structures. 7

APPENDIX

Predictor Variables

Site-averaged elemental properties are derived from the composition vectors. Twelve properties are computed per ABX_3 constituent. These are much denser distributions.

1. Ionic Radius
2. Boiling Temperature
3. Melting Temperature
4. Density
5. Atomic Weight
6. Electron Affinity
7. Ionization Energy
8. Heat of Fusion
9. Heat of Vaporization

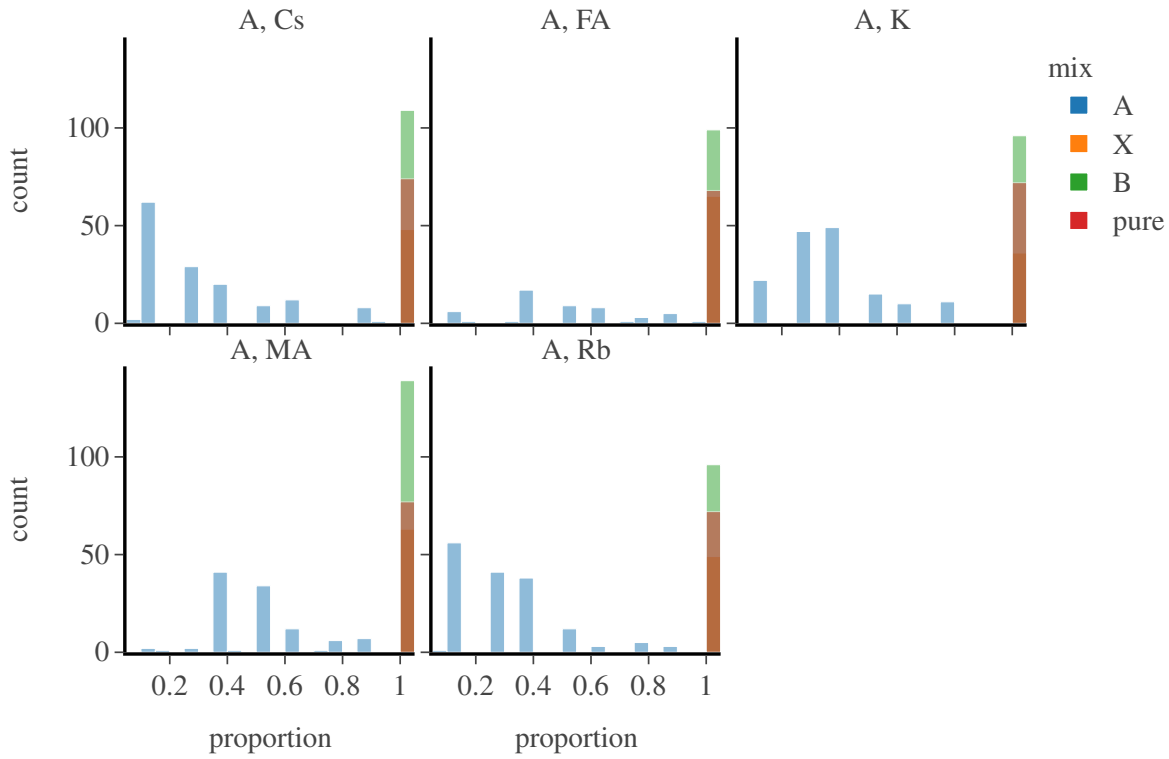


FIG. 6: Normalized Distribution of A-site Constituents

10. Electronegativity

11. Atomic Number

12. Period

Learning Curves

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

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

More data offers better chances. However, the smaller the split, the longer and more expensive the loop training becomes, e.g. 10-fold splits makes for 10 sample scores at each partition size.

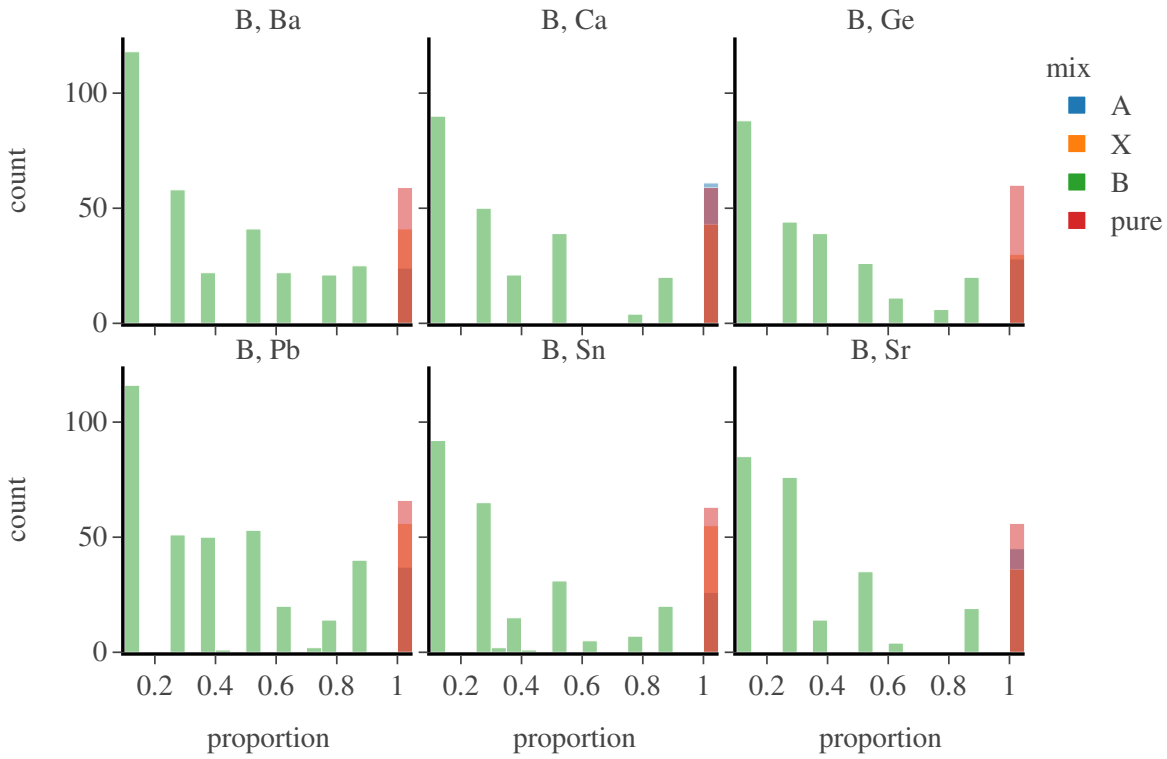


FIG. 7: Normalized Distribution of B-site Constituents

Meaning, 90% of the training set is used for actual training and the remaining 10% is used for validation and this is repeated 10 times.

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

HPO Summary Tables

Random Forest Pipeline Control Parameters

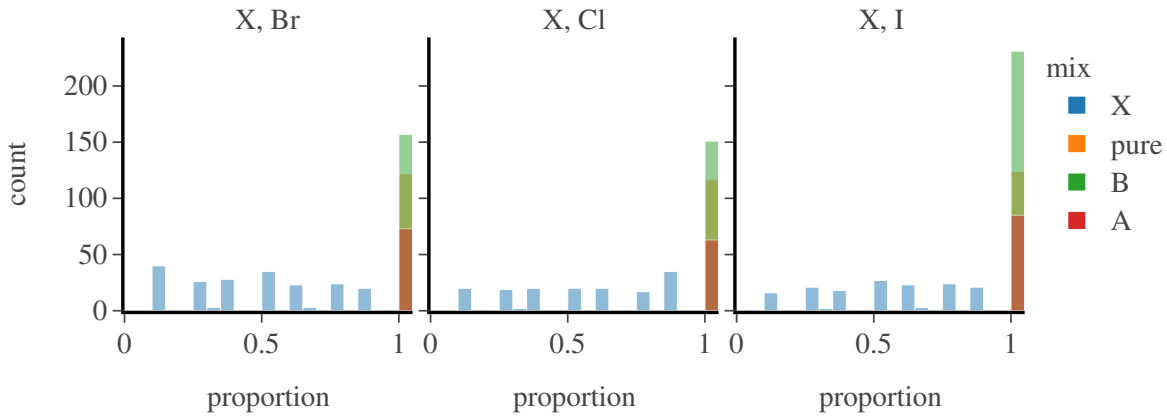


FIG. 8: Normalized Distribution of X-site Constituents

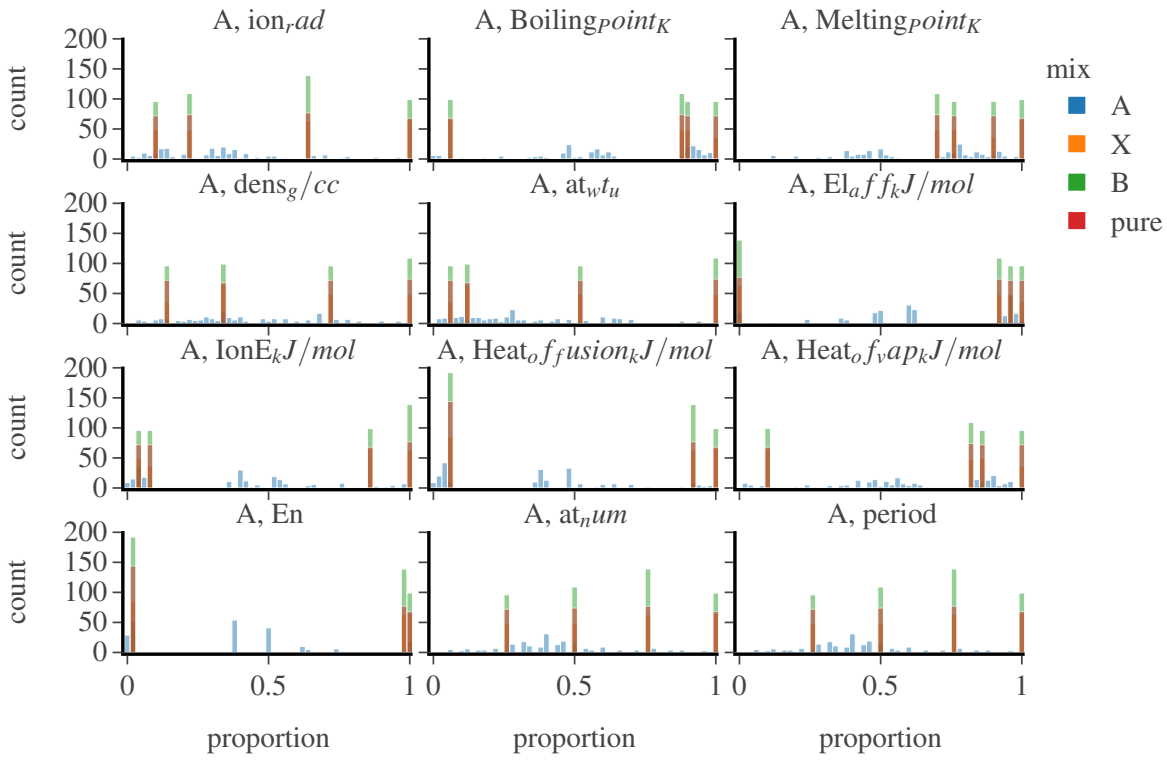


FIG. 9: Distributions of Mean A-Site Properties

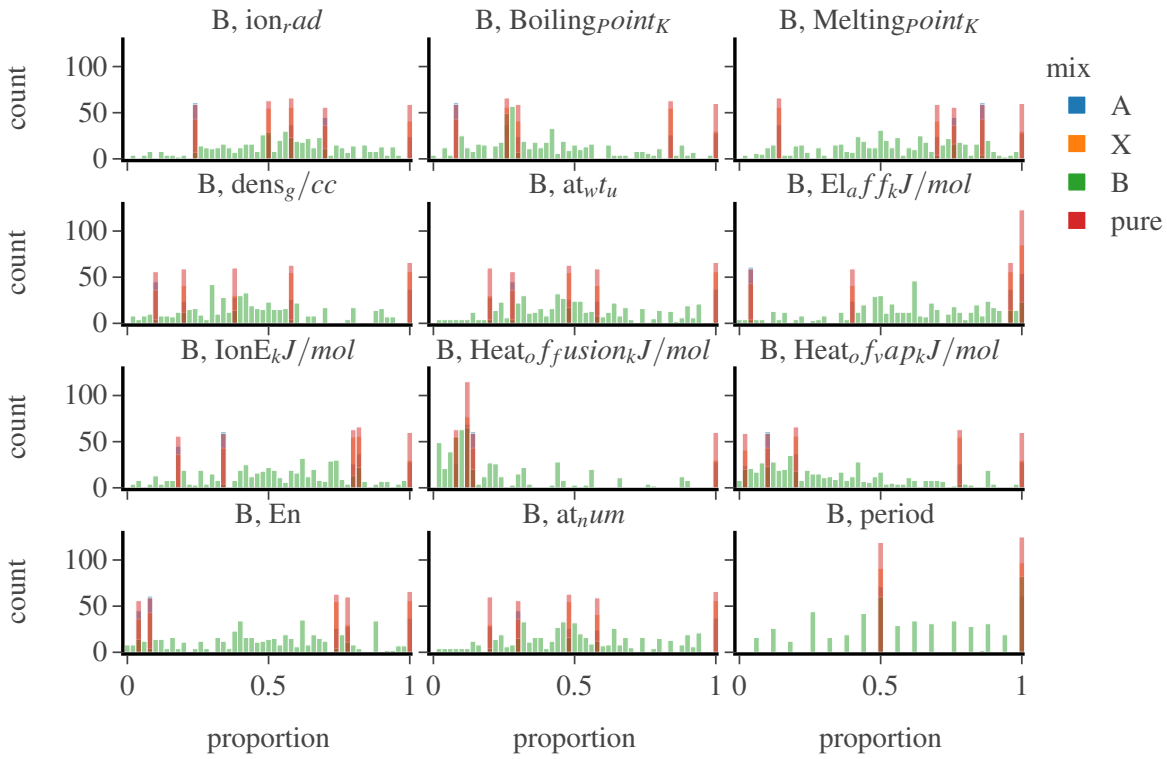


FIG. 10: Distributions of Mean B-Site Properties

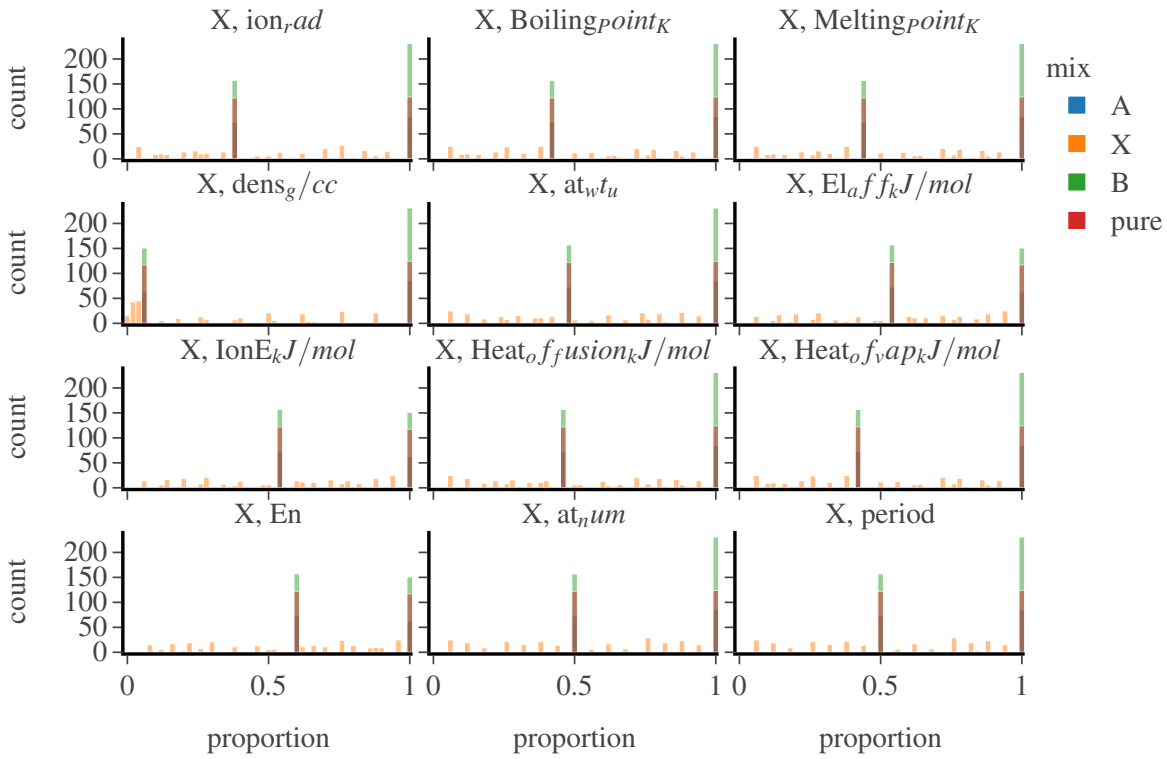


FIG. 11: Distributions of Mean X-Site Properties

TABLE V: Select hyper-parameters from exhaustive search of 10368 models

	Search Space	Selected Space
normalizer_norm	[l1, l2, max]	[l2]
bootstrap	[True]	[True]
ccp_{alpha}	[0.0, 0.002]	[0.0]
criterion	[$squared_{error}$, $absolute_{error}$, poisson]	[$absolute_{error}$]
max_{depth}	[25, 20]	[20]
$max_{features}$	[auto, 3, 5]	[1.0]
$max_{leafnodes}$	[700, 800]	[750]
$max_{samples}$	[0.9, 0.6, 0.3]	[0.9]
$min_{impuritydecrease}$	[0.0, 0.3]	[0.0]
$min_{samplesleaf}$	[1]	[1]
$min_{samplessplit}$	[2, 5]	[2]
$min_{weightfractionleaf}$	[0.0]	[0.0]
$n_{estimators}$	[20, 50, 100]	[150]
n_{jobs}	[4]	[4]
oob_{score}	[True]	[True]
$random_{state}$	[None]	[None]
verbose	[0]	[0]
$warm_{start}$	[False]	[False]