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: 13 September 2022)

[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

DONE Multi-Fidelity Modeling

- State "DONE" from "TODO" *[2022-09-13 Tue 12:51]*

The problem with modeling low availability, high fidelity targets is approached in multiple ways in the literature.

- Semi-supervised learning approaches

  need details

- sequential approaches

  sequential approaches involve training a model architecture on larger quantities of low fidelity data, gaining insight into rough patterns which are then conducive to improving the quality of predictions

We will employ this sequential 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.

The dataset utilized for the development of these models is accumulated from multiple experiments conducted at various fidelities. Therefore, naturally, the statistics obtained from each fidelity are unequal. This is the primary challenge we will address with a multi-fidelity model development approach.

Beyond this, we are careful to maintain the diversity of categories member to 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.

TODO Dataset Overview

DONE Sample Representation Summary

- State "DONE" from *[2022-09-13 Tue 12:51]*

For each fidelity, each experiment is performed on some number of members to a fixed subset of the total 37785 compositions that can be made in a 2x2x2 perovskite supercell when allowing at most single-site alloying with our 14 constituents (table I).

TABLE I: ABX_3 Chemical Domain

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

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 1. See section METHODS for details on how these categories were utilized in model development.

LoT	count
EXP	44
HSC	299
HSE	297
PBE	490

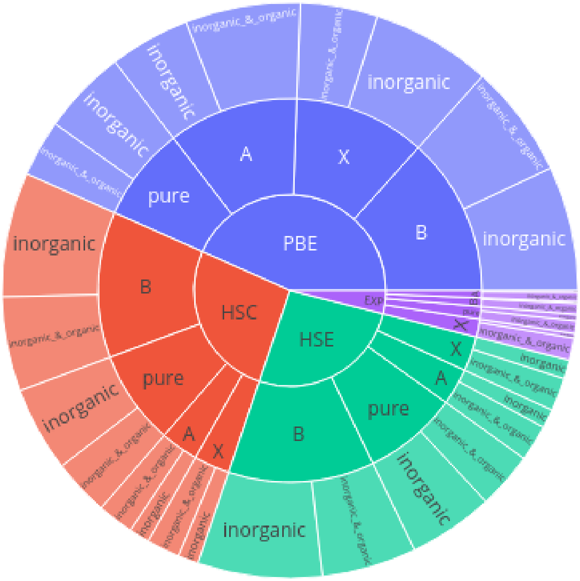

FIG. 1: Share by count of total data apportioned from each experimental subcategory

mix	count
A	202
B	437
X	209
pure	282

The design of this dataset provides an opportunity to assess the ability of our models to extrapolate with respect to alloying scheme. It also provides an opportunity to investigate the statistical impact of constituent compounds on perovskite property prediction. See section RESULTS AND DISCUSSION.

TODO overview of constituent correlations with bandgap

TODO Developing Extrapolative Models

need to introduce the significance of predicting band gaps more effectively

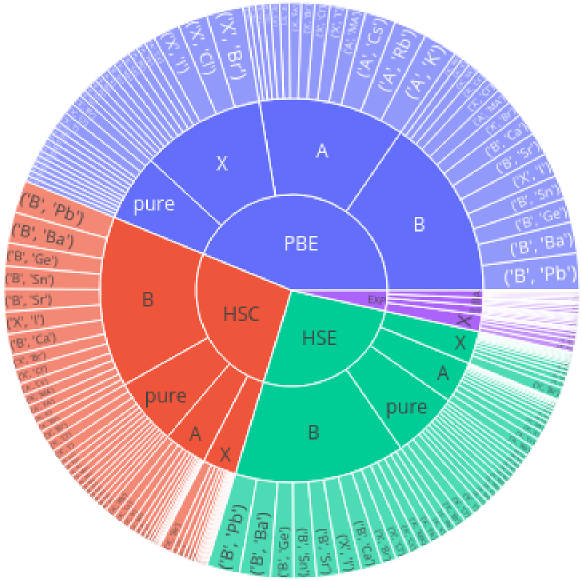

FIG. 2: Tally of constituents present in total dataset as aggregate over all alloy groups

reported experimental band gap values[7,8]

TODO 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[9]. 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.

The scoring metrics we choose vary by model architecture. See the HPO Summary Tables.

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.

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.

RESULTS AND DISCUSSION

Domain analysis

most explanatory basic features

engineered features

naturally, by utilizing GPR in conjunction with these new feature combinations, we can create models with the efficiency of SISSO while obtaining uncertainty estimates.

Best Models

- band gap parity plots
- summary scores
    - best scores
    - best extrapolative ability

Screening

A set of high throughput screening criteria has been previously developed[10]. We apply the band gap criterion to the predictions made the by these enhanced models.

TODO Compare

screening results to the previous batch produced by older models.

Summary

GPR performs best and

METHODS

Several machine learning architectures are rigorously optimized with regard to both generality over the domains of Perovskite compositions and site-averaged atomic properties and generality over the domain of alloy classifications.

In order to control for the classification biases potentially acting on the parameter space of regression models, nine metrics are used to evaluate the performance of each model over all alloy types at every stage of the hyper-parameter optimization simultaneously. Only models that perform uniformly well on all alloy classes are selected.

Validation curves are computed for hyper-parameters to which a given model is particularly sensitive.

TODO Objectives

Our objectives are two fold. First, we aim to accurately predict performance-relevant Perovskite band gaps.

We will follow a multi-fidelity approach, where the bulk of affordable low level-of-theory data will inform and improve the extrapolative ability of models trained on higher fidelity measurements.

Our fidelity hierarchy climbs from simulations performed using the basic PBE functional, to results obtained from physical experiments aggregated in literature[3].

We aim to express these variables as functions of the perovskite composition. Schemes for incorporating structural information will be developed in future work. Nevertheless, a strong understanding of the influence of chemical composition on performance will continue to be a priority as it is expected to aid both in Feature Engineering and in screening the combinatorial chemical space for viable high-entropy compounds.

Second, we hope to better understand the average physical impacts of 1) site-specific alloying and 2) using organic molecules in the Perovskite superstructure.

These goals are not entirely separate from the first goal of expressing various properties as functions of composition, but they can be more simply approached as problems of addressing dependencies in the data statistics. Our model development will test the hypothesis that formula that fall within one of these classifications will share some distributed qualities with others that fit their classification.

TODO Considerations

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 diversity by utilizing the low entropy 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-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 classification of each sample. However, this fold is not used in a classification problem, the regressor is trained on the subsets of each class, and it's 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.

Per architecture, a model is instantiated using the – in aggregate – 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.**

TODO **DFT Details**

The current work is focused on cubic phase compounds only.

TODO VASP

TODO Levels of Theory

The largest table of 490 records contains computed electronic properties. Each is obtained for a unique composition using static Density Functional Theory (DFT) simulations (relaxation, electronic) conducted at a PBE level of theory. Additionally, 299 of the compositions tested in the first 490 records are fully simulated at an HSE level of theory. Another **SUBSET** of these are also examined, following PBE relaxation, using HSE with Spin Orbit Coupling (SOC).

Each simulated structure is made in two ways. Once with the GGA-PBE functional and once with the HSE06 functional. For structure, band gaps are obtained using a static band structure calculation performed at the same and at higher levels-of-theory. Specifically, ~300 of the same compounds underwent HSE06 bandstructure computations, both with and without spin-orbit coupling (SOC), for a total of approximately 900 experiments with enhanced accuracy.

SOC is performed for better better electronic properties.

TODO Featurization of Chemistries

The largest subdivision of ~1400 compounds correspond to a series of optoelectronic properties simulations performed using density functional theory (DFT). The simulated experiments are performed on ~500 pseudo-cubic ABX_3 supercells obtained by geometry optimization. Each cell demonstrates mixed compositions at none or one of each of the A, B, or X sites. See Figure.

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 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[11] $\alpha+\beta+\gamma=14$.

it is easy to make composition based multi-fidelity predictors. Involving structure is a challenge for experimentally collected data because experimental measurements have to be accurately matched to structure graphs, which generally must be either exhaustively validated or generated using specialized equipment.

TODO Feature Engineering

The Basic Feature Space

- 14 dimensional composition vectors extracted from chemical formula
- 36 dimensional site-averaged property space computed from composition space
- categorical dimension one-hot-encoding level of theory
  - PBE on PBE
  - HSE on HSE
  - HSE+SOC on HSE
  - HSE+soc on PBE
- models of band gap trained on union of features
  - Linear
  - RFR
  - GPR

The Engineered Feature Space

- SISSO

the tools

- converting formula strings to vectors
- converting vectors to structures

TODO Model Training Procedure

TODO partitioning

A stratified shuffle split is used to make partitions. This split preserves the proportion of each level of theory in the test and train partitions, which helps with the final model evaluation.

- all decisions about model optimization will be made using only the dedicated training partition
- The test partition will be reserved until a final model pipeline is parametrized and fit
- the predictions made on the test partition will either confirm or deny the model's ability to work outside of the training domain

TODO cross validation strategy

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

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

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

- RFR results

- GPR results

TODO Hyper-Parameter Optimization

- performance with diverse samples
- performance in extrapolation
- grid search with diverse samples
- validation curves

FUTURE WORK

- broaden model domain to include alternative phases
- improve accuracy of surrogate model – RFR/GPR/NN?
  - incorporate much more experimental data[5,12]
  - implement delta learning strategy to utilize more indicators besides SLME/PCE

Results suggest that either alternative semi-supervised learning strategies or much more experimental data is needed to improve the quality of experimental-fidelity predictions.

Our first objective is to utilize much more experimental data in the construction of regressions.

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

- utilize an active learning approach leveraging GPR models to extensively grow the cubic perovskites dataset

The GA fitness function should account for uncertainty in the surrogate model. GA recommendations can be tuned to explore uncertainty. Testing these recommendations with DFT forms a loop that improves the ML models incrementally with exposure to new data, thereby efficiently discovering the best possible cubic perovskite compounds.

identify novel compounds

already, screening has been used to successfully identify some curious options

what more can I add to this?

utilize convolutional graph neural networks

enabling structure->target models

- MEGnet scalable graph networks for polymorph-sensitive property prediction[13]
- ALIGNN[14]
- CGCNN

Extending models to perform effectively on non-cubic Perovskite phases is best enabled by the sophisticated graph neural networks being developed for exactly this purpose.

Already, multi-fidelity "MFGnet" models are being developed under the graph learning paradigm[15].

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.

Model development and feature extraction is performed using python and SciKit Learn v1.2. A library of model evaluation tools is being maintained in the "yogi" repository at `http://github.com/PanayotisManganaris/yogi`

ACKNOWLEDGMENTS

We acknowledge funding from the US Department of Energy SunShot program under contract #DOE DEEE005956. Use of the Center for Nanoscale Materials, an Office of Science user facility, was supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy

Sciences, under Contract No. DE-AC02-06CH11357. We gratefully acknowledge the computing resources provided on Bebop, a high-performance computing cluster operated by the Laboratory Computing Resource Center at Argonne National Laboratory. This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. MYT would like to acknowledge support from the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Science Undergraduate Laboratory Internship (SULI) program. MJD was was supported by the U. S. Department of Energy , Office of Basic Energy Sciences, Division of Chemical Sciences, Geosciences, and Biosciences, under Contract No. DE-AC02-06CH11357.

Author Contributions

A.M.K. conceived the idea. A.M.K. and J.Y. performed the DFT computations. P.T.M. Trained ML models. All authors contributed to the discussion and writing of the manuscript.

Data Availability

DFT data and ML models are available from the corresponding author upon reasonable request. All documents are tracked in the `https://github.com/PanayotisManganaris/manusciprt--multifidelity-dft-ml.git` online repository

Additional Information

The authors declare no competing financial or non-financial interests.

Correspondence and requests for materials should be addressed to A.M.K. (email:amannodi@purdue.edu).

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] 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.

[8] D. E. Swanson, J. R. Sites, and W. S. Sampath, Solar Energy Materials and Solar Cells **159**, 389 (2017).

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

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

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

[12] 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.

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

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

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

APPENDIX

Learning Curves

band gap scores vs training set size

Notice that the error metrics are negated so that, consistently with the R^2 and ev scores, the greater the number, the better the model performs.

HPO Summary Tables

RFR

GPR