August 2021

Demystifying Structure-Property Correlations for Perovskite Oxides Using Machine Learning to Discover Lead-Free Alternatives Solar Cells

Afrid Mazhar Shirsekar
Clemson University, afridalpha7@gmail.com

DEMYSTIFYING STRUCTURE-PROPERTY CORRELATIONS FOR PEROVSKITE OXIDES USING MACHINE LEARNING TO DISCOVER LEAD-FREE ALTERNATIVES SOLAR CELLS

A Thesis
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Master of Science
Chemical Engineering

by
Afrid Mazhar Shirsekar
August 2021

Accepted by:
Dr. Rachel Getman, Committee Chair
Dr. Leah Cassabianca
Dr. Mark Roberts

ABSTRACT

Discovering optimal materials for a given application has become extremely difficult due to the vast scope of structure possibilities and requires substantial computational expense to evaluate even one of all possible structures. An increase in the structural complexity can offset combinatorial explosion and could take months, if not years, even with the advanced computational architecture, to screen every candidate. Here in this study, we propose a computational approach based on statistical learning combined with DFT based computations that can effectively screen over the entire range of possible candidates and predict the material's electronic properties with high fidelity. Specifically, we use advanced machine learning algorithms such as gradient boosted decision trees and electronic structure bandgap calculation data to screen perovskite structures as alternatives for lead-free solar cells. Perovskites are compounds with chemical formula ABX_3 where A and B are cations and X is an anion that bonds to both cations. The perovskite class offers compositional flexibility which allows us to tune the structure to obtain better solar absorption efficiency. Using machine learning, we could establish a structure-property correlation, by mapping the attributes of the structures to their bandgaps, which enabled us to screen over all compounds within the perovskite class thus drastically accelerating our search of the optimal lead-free alternatives for solar cells. With purpose of making the dataset more robust, this study explored the complexity of the composition effect by evaluating the substitution of one or more elements in different proportions and arrangements over the possible sites available in the base perovskite structure on the material's bandgap.

DEDICATION

This work is dedicated to my parents, family, professors, mentors, and friends whose constant guidance, love, inspiration, care, and support were the key ingredients in the study. This work is also dedicated in the loving memory of Paul’s Father.

ACKNOWLEDGMENTS

I would like to extend my sincere gratitude to my mentors Dr Rachel Getman, Dr Paul Morales, Dr Cassabianca and Dr Roberts who supported me throughout my master's journey. Their technical and professional guidance helped me grow better at both professional and personal levels.

Special thanks to the one-of-a-kind Getman research group and Dr Getman for providing the right environment for key information transfer and safer platform for brainstorming.

I can't thank enough Palmetto Cluster for ensuring that my numerous jobs were run efficiently. Also, huge love for the CP2k community and enormous admiration for their numerical optimization methods for bringing down the eta from months to weeks.

I also acknowledge the amazing work done by the material's project community for their scripts on recursion and permutations.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

List of Figures (Continued)

CHAPTER ONE

INTRODUCTION

Screening optimal materials for a given application is laborious and time-consuming due to the vast scope of materials possible because of the material's compositional and configurational degrees of freedom[12]. Computing even a subclass of materials will be an insurmountable task. In recent years computing speed has reached astronomical heights and combining this with the data-driven, sophisticated statistical and machine learning algorithms, one can make significant inroads in interpreting the hidden correlations between material property and the structure, thereby accelerating our search for novel materials with optimal properties per application[9]. Such informatics-based statistical learning model development has rapidly advanced research in variety of material science problems in many directions[2,22] and has led to material property predictions such as potentials[2], transition states[33], dielectric constants[40], and bandgaps[32].

The machine learning (ML) approach helps us to model a material chemical space by mapping the attributes of the material structure to its property for a subset of materials[12]. The material's attributes (features or profile) should be such that they ensure unique representation of every structure. Once the mapping is established and appropriately validated, we can accurately predict properties for all the materials in the chemical space at negligible computational cost, thereby bypassing the cumbersome time-consuming computations and experimental evaluations for all the structures. Such modelling methods are termed as Quantitative Structure Property Relationships[34] or

QSPR. In this contribution we utilized such modelling methods to find materials with optimum efficiency for replacing lead-based materials in solar cells[2,4].

Conventionally lead-based solar cells are in use as they have high power conversion efficiency (Figure 1) and low-cost manufacturing [4], but the lead element poses a significant concern to the environment and health. Previous studies have suggested exploring the vast combinatorial chemical space of the double perovskites class (chemical formula $AA'BB'O_6$) to replace lead-based materials in solar cells [18,19]. To summarize, perovskites are crystal structures (Figure 2) of the type $ABX_{3,}$ where A and B are cations and X is the anion bounded to both, the X anion here is usually oxygen or a halide. The A cation can take up charge form +1, +2, +3 and B cation occupies the position within the X anion octahedra. The A atom occupies every hole left vacant by the eight BX_6 octahedra. In comparison double perovskites have the same geometry as that of a perovskite but twice the unit cell. The perovskite family in general is an excellent choice for material discovery as it provides good chemical flexibility, and perovskite frameworks open the window to a broad spectrum of diverse compositions[12,10]. Material property such as bandgap is instrumental in classifying a material based on its solar absorption efficiency[21]. Thus, using existing or computationally generated structure–bandgap dataset for a subset of the materials in the material space and combining the data with ML modelling methods we can accurately draw high fidelity bandgap predictions for all the structures within the chemical space. This accelerates our screening process for the candidate materials with desired solar absorption efficiency.

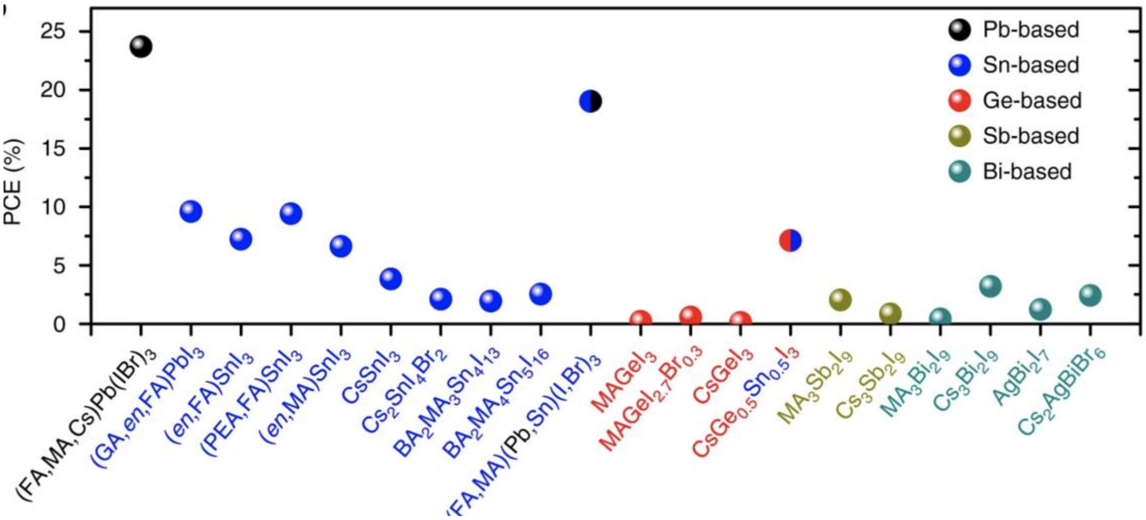


Figure 1: Power conversion efficiencies (PCE) of solar cells using Pb, Sn, Ge, Sb and Bi based absorbers. [1]

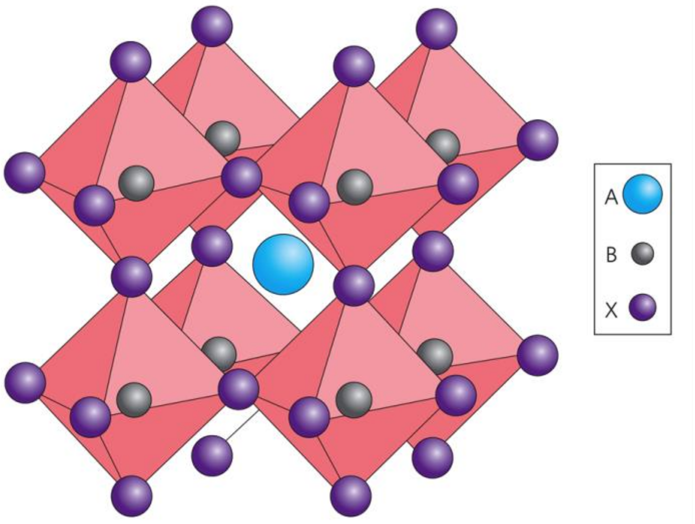


Figure 2: Crystal structure of perovskite of the type ABX$_3$.

For the machine learning model to work effectively we need to provide the model with an accurate structure-bandgap dataset. For this purpose, we computed our own perovskite oxide structure dataset using density functional theory implemented through an open-source quantum chemistry software package called CP2k[27]. Specifically, we randomly generated a material dataset of 205 structures for perovskite oxides by making substitutions over the A and B sites with varying degrees of concentration of substituents.

Density Functional Theory (DFT) is a computational quantum mechanical modelling method that computes the ground state electronic structure of chemical compounds. DFT is successful in the computational domain due to its high accuracy description of the electronic structure with a moderate computational cost[32] and hence is the workhorse for such high throughput computational screening. As such it has enabled the development of databases covering calculated properties for number of materials either known or hypothetical[37,2]. Using DFT we can generate a high-quality subset of calculated bandgaps for perovskite oxides, which can further be used to train Machine Learning models that learn and accurately predict bandgaps for all the structures within the perovskite oxide composition space (Figure 3).

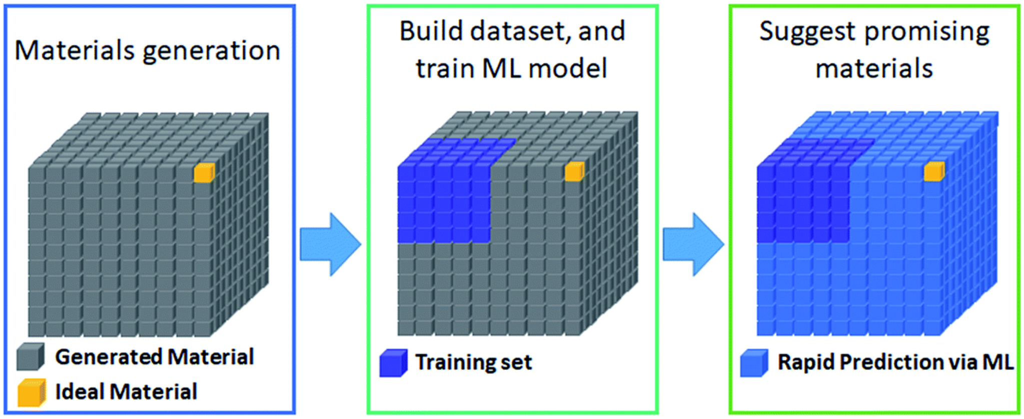


Figure 3: Workflow for Machine Learning based material screening and property prediction.

Thus, a combination of high throughput DFT computations and Machine Learning can provide a practical, high fidelity, rapid screening of the combinatorial chemical space occupied by the perovskite oxide class. Though machine learning techniques have made significant breakthrough in the domain of material science, such modelling methods suffer from lack of transparency misinterpretation of results and low accuracy[5]. Most machine learning models fail to show the true relationship between attributes and properties which hinders our interpretation about the structures. In our study we have used gradient-boost decision trees to tackle lack of transparency as it provides quantification to what degree a structure's attribute is related to its bandgap by means of importance scores.

CHAPTER TWO

COMPUTATIONAL METHODS

Structure Library

Here we explain how the dataset was conceptualized for the training of the ML prediction model. The choice of elements is referred to from the study conducted by Pilania, et.al using the Computational Materials Repository dataset [12,11]. The perovskite oxides crystal structure in this study are of the form ABO_3 as shown in Figure 4c which represents one of the randomly generated structures in the dataset. The element list of A sites involves alkali and alkaline earth metals, while the list for the B sites majorly involves transition metals. In Figure 4c, Nb occupies the B-sites and Ba, Cs and Rb occupy the A-sites. Figure 4b sheds some light on the geometrical arrangement of our perovskite oxides structures at the unit cell level. The B site cations represented by blue atoms in Figure 4b have a 6-fold oxygen coordination as they occupy positions within the oxygen octahedra (red atoms), this leaves a hole at the center of the lattice which is filled up the A cation represented by a green atom in Figure 4b, having a 12-fold oxygen coordination. Since these structures are periodic in nature, the unit cell could be expanded to incorporate more A sites (8 in this study) thus allowing us to implement more substitutions per structure. Multiple element substitutions allow us to explore the perovskite oxide chemical space more thoroughly, but this could also very quickly make our computational resources over encumbered as the possible structures rise steeply. To tackle this, we conduct the sampling (substitutions) of structures for the machine learning model non-uniformly, as described in detail below.

The element space for A cations is Ag, Ba, Ca, Cs, K, La, Li, Mg, Na, Pb, Rb, Sr, Tl and Y and the element space for the B-site is Al, Hf, Nb, Sb, Sc, Si, Ta, Ti, V, Zr. A primitive structure was first generated for every B-site cation list (viz. TiO3, TaO3, SiO3...), and for every such oxide, we randomly conducted the substitutions over the A-cations elements space.

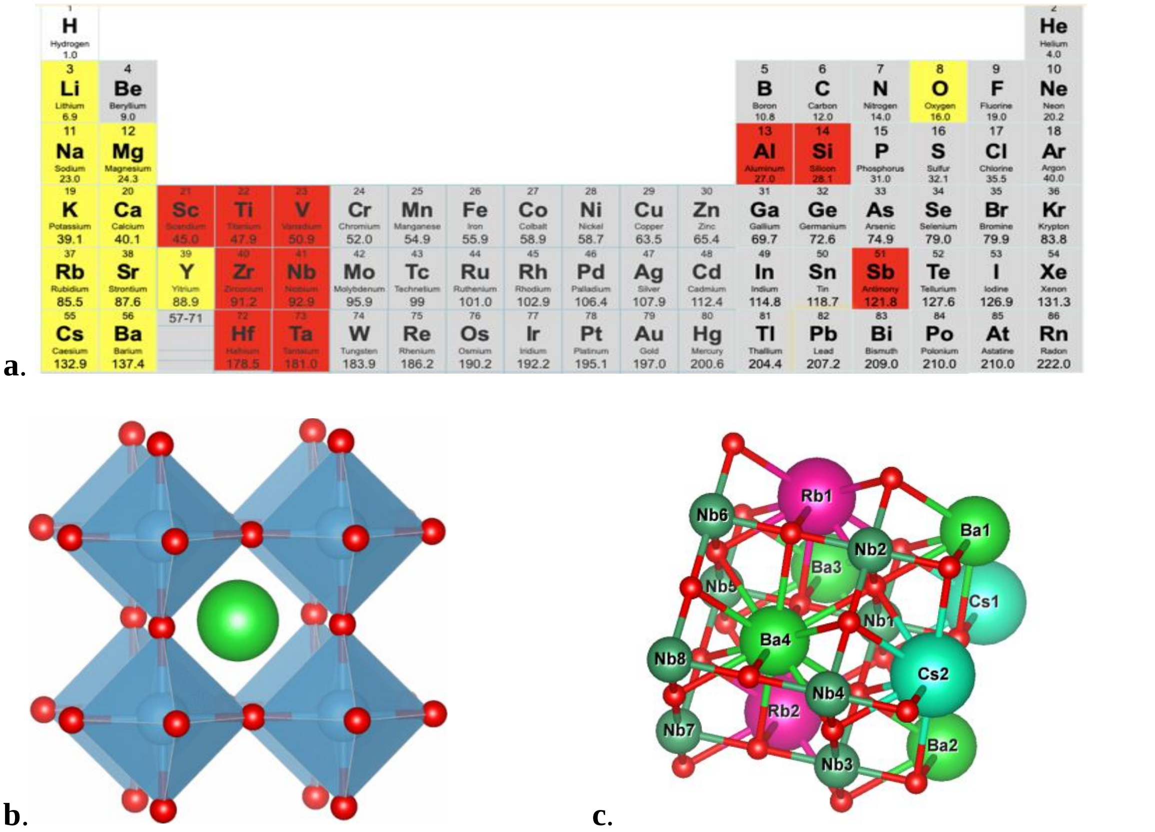


Figure 4: **a** Candidate elements featuring in the chemical space, the yellow elements occupy the A sites and the red elements occupy the B sites. **b** visual representation of unit cell of the perovskite oxide class [29]. **c** sample structure of Cs2-Rb2-Ba4-Nb8-O24 [23].

Making every possible substitution in this way would result in roughly 227,000 compounds. The colossal dataset arises due to the different combinations or orderings the A cations can take up within the octahedral arrangement of B oxide. Furthermore, the permutations for selecting A-site cations can further widen up the chemical space possible. Machine Learning plays a vital role in screening the entire chemical space and providing high fidelity predictions but computing even a paltry subset of around 2000 structures using DFT to train the ML model will be a time-consuming step, compounded further by the fact that not all DFT calculations complete straightforwardly. This gives rise to a tradeoff between capturing structural complexity and training examples or computational expense. Here we employ a strategy that guides selecting a reasonable number of structures for the training set while utilizing our chemical space's element diversity.

First, many structural configurations mirror each other and provide no additional information to the Machine Learning model; eliminating symmetrical structures will ensure that we have fewer but unique structures and can achieve similar proportions of all the elements in our training set. This elimination is achieved by employing the Enumeration [13] library implementation in the Materials Project's pymatgen python package. Before the enumeration step, we randomly automate selecting the elements and concentrations for the A-site elements within the B-site sublattice. The enumeration transformation is then conducted to remove symmetrically equivalent structures from the dataset. Second, we set the same reference frame to identify sites for every structure depending on their XYZ coordinates to capture the relative positions of elements concerning each other in lattice space.

Density Functional Theory Calculations

Density functional theory (DFT) was used to compute the bandgap for the structure library. The HOMO-LUMO gap as extracted from CP2k is used to estimate the bandgap (E_g) in this study. Bandgap is the metric to assess the ability of materials to absorb solar energy, previous studies have shown that materials having bandgap values between 1.1 eV to 1.8 eV have optimum solar absorption efficiency[35].

DFT calculations were performed using the PBE (Perdew-Burke-Ernzehof) exchange correlational functional with DFT-D3 [26] dispersion correction. Here we use the Quickstep [27] code from CP2k [14] which is an open-source quantum chemistry and molecular dynamics software based on Fortran, employed with GTH pseudopotentials and DZVP-MOLOPT-SR Gaussian basis sets, and supporting basis sets such as Auxiliary Density Matrix (ADMM) were employed[24]. The plane wave cutoff is 900 Ry with relative cutoff of 60 Ry. More info on the parameters can be found in Appendix B. A sample input file for CP2k is provided in Appendix C.

The energy functional minimization scheme used here was the Orbital Transformation (OT) [28] method, which has guaranteed convergence coupled with a full single inverse preconditioner to achieve superior convergence speeds. Along with OT, the Pulay Method optimizer was also used, and the minimum bandgap parameter was set at 0.01 eV. However, E_g here calculated via DFT is underestimated than the experimental values due to its limitations. Specifically, the semi local DFT functional (GGA)[30, 41] employed in this study cannot fully capture the self-interaction of the electrons in the HOMO, which results in their energies getting pushed upwards and in turn reduces the bandgap. Using a more accurate functional such as the GLLB-SC being further optimized for solids can help recover optimal values from the materials' underestimated bandgap values but the computational cost (twice as costly[12] compared to our current functional) for the functional hinders the development of the structure dataset in reasonable timeframe.

Machine Learning Model

Machine Learning methods can be roughly described as algorithms that iteratively search for the optimum equation having the lowest possible average mean squared error between the predicted and calculated values. The equation should accurately describe the relation between the value of interest and features (values that help explain the output value). For the Machine Learning model to be successful, it requires that we gather a good quality dataset and select an appropriate learning algorithm. Learning algorithms such as kernel ridge regression[12], LASSO[36], support vector machines etc. are some of the popular choices. We in this study use gradient boosted regression trees (GBRT) as our learning algorithm as it provides high accuracy as well as good interpretability of the features.

A virtual decision-making tree could be explained as system of nodes and leaves, where a node represents classification based on one of the explanatory variables or features as shown in Fig 5, and a leaf is the prediction of the target variable. In a regular regression problem, the model updates its parameters or coefficients of the explanatory variables to achieve the best fit. But in a decision tree the model updates decisions or functions instead of parameters. For example, to check whether a patient is suffering from hypertension from their blood pressure report, a basic or weak learner decision tree generates a function such that patients with blood pressure above a certain limit will be classified as positive and vice versa. A weak learner or tree stump is simply referred to decision trees that make their decision based on just one feature. They are termed weak as their predictions are not highly accurate (slightly above random guessing).

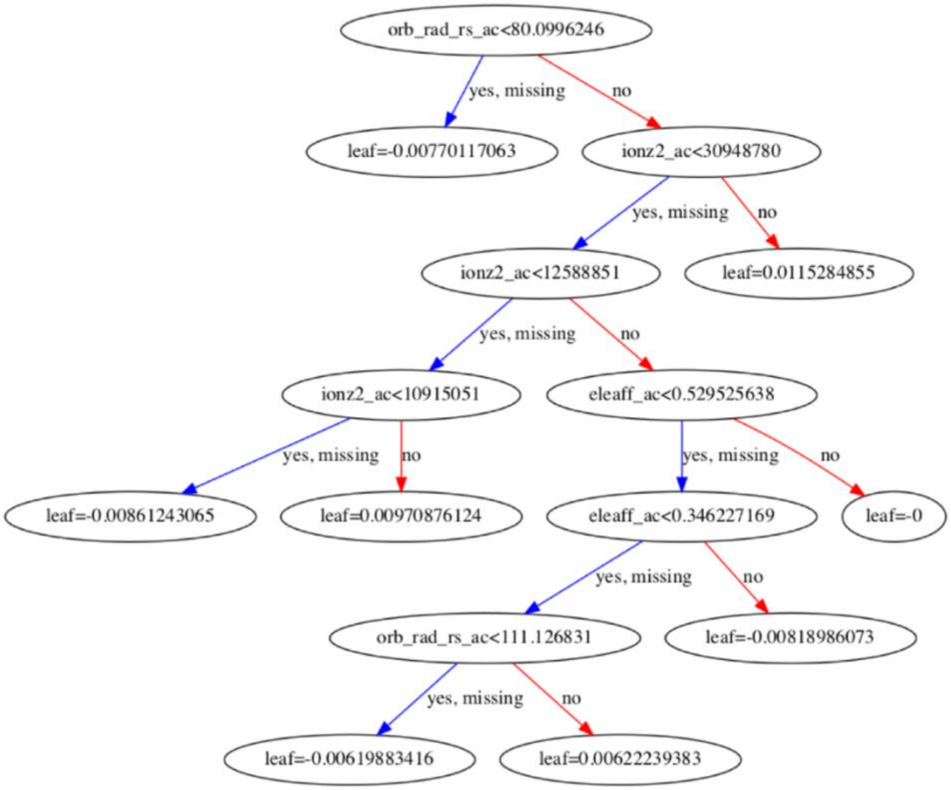


Figure 5: One of the many decision trees that were generated by the regression machine learning model. To summarize the process generally, the structure and its attributes goes through the decision tree and based on tree's decision points it branches into leaf or prediction. The values in the bubble or node represent decision point based on the feature and based on this decision the bubble branches into two paths ('yes, missing' if the data point matches the condition, else 'no' path is followed). Names in the bubble denote feature names, orb_rad_rs_ac = orbital radius for s orbital, eleaff_ac = electron affinity, ionz2_ac = ionization potential 2. Leaf represents the value needed to be added to the base value to make the prediction for the given structure. Base value here is the average of all bandgap values.

Gradient boosting tree adopts gradient boosting method that combine weak learners into one strong learner (classification model with better accuracy than random guessing). This method combines features into functions (for example, the feature atomic radius could be transformed to a new feature (atomic radii)^2 , (atomic radii)^3 , or (electronegativity * atomic radii) to draw better correlations) thus opening doors to accurately capture non-linear relationships as the explanatory variables or features can be explored to a higher degree as they can be used again as nodes further down the decision tree. This approach eliminates the need for manually experimenting to identify the optimal relationship between features.

From a mathematical point of view the model's optimization objective is the loss function shown in equation 1,

$$\mathcal{L} = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \tag{1}$$

l is the function that measures the difference between the predicted response $\hat{y}_i$ and the actual value y_i. The second term Ω is called the regularization term. The regularization terms help improve the model's performance by preventing features from influencing the output due to their heavier (higher order of magnitude) weights. The model learns by adding new decision tree iteratively that reduces the loss $\mathcal{L}$. A squared error or logarithmic loss could be used for regression or classification models, respectively. However, such decision tree-based models cannot be optimized using conventional optimization methods as shown in equation 1 as it includes 'functions' (decision tree f_k) as parameters which cannot be optimized within the Cartesian space spanned by n-tuples of real numbers, (x_1, x_2, ..., x_n)[5].

$$\mathcal{L} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i)\right) + \Omega(f_t) \tag{2}$$

As shown in equation 2 for sample i and iteration t, $y^{(t-1)}$ is the prediction with (t-1) tree setting and to minimize the loss function; tree f_t needs to be added and n is the number of data points. The model searches for the point to split (if else decision) within every feature column that causes the maximum decrease in the loss function. The model further decides to either stop with the current prediction or further add another tree or split that significantly reduces the loss function. The model learns by adding a decision tree $f_t(\mathbf{x}_i)$ for a sample i, to the existing set of trees $\hat{y}_i^{(t-1)}$ if there is a maximum decrease in the loss function. It greedily (iterates until it reaches the best value possible) adds a function or split to the tree that most improves the model or that records the maximum decrease in the loss function. The model keeps track of how each feature or explanatory variable influences the decision for splitting the leaves and their contribution in reducing the loss function[5]. The evaluation metric is based on the number of times a feature was used to split the leaves or the mean of gains achieved, in context of the loss function, from splitting a feature or adding a tree.

The machine learning model is assessed on how well its predictions are for the structures that it wasn't trained on. The model performing well on the structures that it was trained on but performing extremely poor on newer structures is unacceptable. Such a phenomena is termed as overfitting. To counter this, we randomly split our dataset into two parts, 80 per cent is used for training the model and 20 per cent is kept as the test or new/untrained structures set. Out of the 80 percent we further randomly sample 20 per cent data as cross-validates set which tests the model for very iteration. Cross validated procedure is like the concept of test set, where the model evaluates itself on every iteration during the training process. We couple cross-validation with parameter grid search which finds the best possible combination of model parameters.

A 7-fold cross-validated hyperparameter grid search was conducted for the machine learning model. k-fold cross-validation is the method of validating a Machine learning model's prediction performance. In this method we divide the data into k samples and select one sample as validation set and k-1 samples as training set, this process is then repeated for all the k samples. The model is then trained and validated k number of times using any metric such as least squares, mean absolute error etc., the result is the average of all k tests. Hypermeter grid is the set of parameters used when the model is initialized, which affects the model's performance. The objective (metric) or cost function was set to be negative median absolute error for the regression model and accuracy score for the classification model. A list of hyperparameters for the models and feature name key can be found in Appendix A and B respectively.

To understand the Machine Learning model's interpretability, we can access the weights the model assigns to every feature in deciding the final value of the bandgap. Hence these weights shed light on the importance of every feature in the model, which further guides the extraction of powerful features. To establish a structure-property relationship, we need to choose appropriate features that would describe or relate with the target properties. The performance of the model accuracy is heavily influenced by choice of input features.

Feature Engineering

The most critical step to setting up a high-fidelity machine learning model is finding the right set of input features. The selection is based on characteristics that completely describe the elements in lattice space. The features selected here are Pauling Electronegativity, electron affinity, atomic radii, ionization potentials, highest occupied energy level, lowest unoccupied energy level and orbital radii for s, p, d and f orbitals of the neutral isolated elements A and A’[12,3].

The transparency of the Gradient Boosting algorithm's decision-making process is achieved through its feature importance scores. The evaluation of scores helps in assessing qualitatively to what degree a feature contributes to the prediction. Furthermore, it also helps extract critical features out of the entire set to avoid model overfitting, familiar with the gradient boosting setup. One downside of feature importance scores is that they do not accurately evaluate scores for highly correlated features.

Removing one of such features, also known as dimensionality reduction, is a common practice to reduce computational cost and prevent overfitting. The correlation between two features is determined by the Pearson Correlation coefficient, the covariance ratio between two features to the product of their standard deviations. The equation normalizes the covariances to lie between the values –1 and 1. Figure 6 shows a Pearson covariance matrix for our input features.

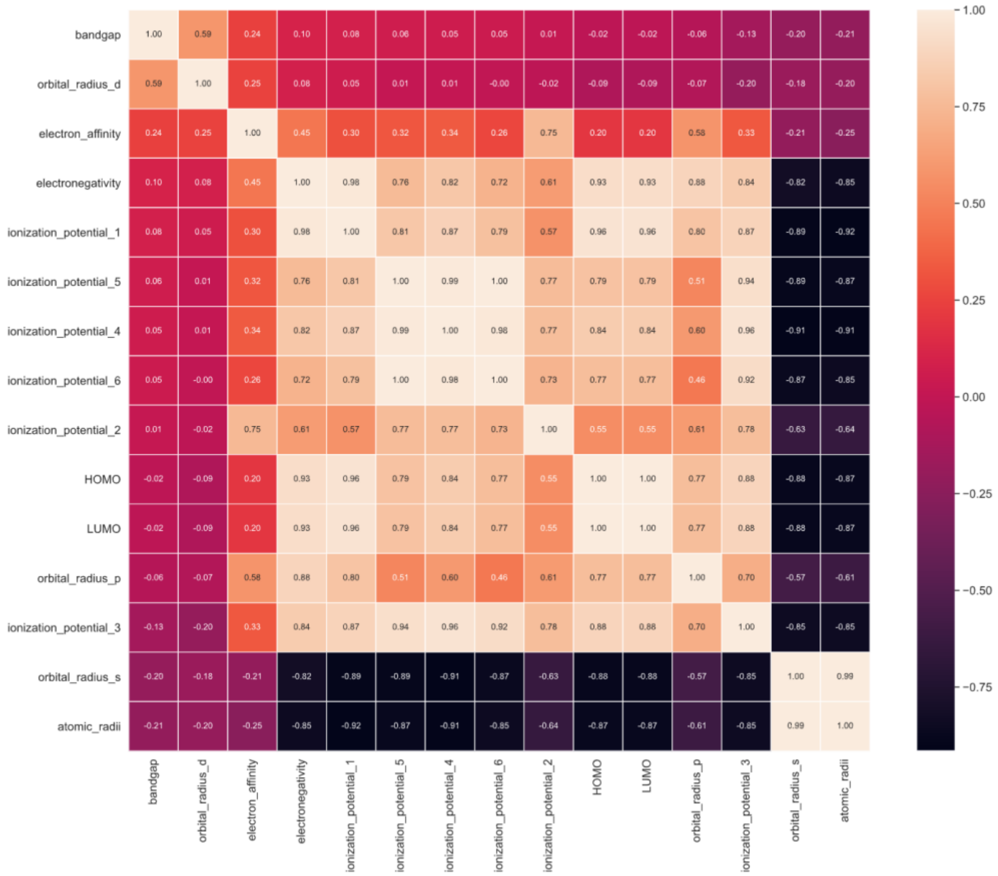


Figure 6: Pearson correlation matrix for the input features after the auto-correlation function transformation. The values [1, -1] denotes correlation between two feature columns.

The Pearson covariance matrix displays strong correlations between multiple features, most prominently the highest occupied level and lowest unoccupied level, and the ionization potential energies. Though eliminating standard features helps reduce the computational cost, especially if the dataset is scaled, it does not offer much change in the performance of the model. We also conduct a dimensionality reduction on our feature set. The rationale here is that for a given structure n with the total number of descriptors equal to 15, we would end up having 15 * 3 (for each A, A' and B) = 45 feature columns which can drive up the computational cost and the information from relative positions or correlations of elements also needs to reflect in the input features. To overcome these issues, we constructed fingerprints from the primary features.

Fingerprinting

It is very crucial that features clearly and exclusively describe a single given material, to ensure this we make use of the unique arrangement of the substituent elements within the unit cell. Instead of taking 8 different values per feature for 8 sites, we correlate the 8 values with the distance spanning between them using a fingerprint equation. So, in case we have two structures having the same substituent elements but occupying different positions, they will have a unique value per feature after they are operated by the fingerprint equation, which is calculated as the sum of the ratio of the product features to the distance spanning between them.

$$\mathrm{F}_{\mathrm{p,i}} = \sum_{\mathrm{j=1}}^{8} \sum_{\mathrm{j \neq i}}^{7} \frac{\mathrm{p_i} * \mathrm{p_j}}{\mathrm{D_{ij}}} \tag{3}$$

Where i and j represent the substitutions sites in the lattice space, and $\mathrm{D_{ij}}$ is the distance between them. In this notation, p_i and p_j are the feature values for substituents at sites i and j respectively. Though the equation doesn’t work fairly for features that have a zero value, an addition or subtraction-based numerator would resolve such an issue. This deficiency was brought to light when the feature importance scores were examined. The features such as orbital radii for p, d and f suffer from this equation since for many elements such as barium, calcium, strontium etc., the values for p, d and f orbitals are zero.

CHAPTER THREE

RESULTS & DISCUSSION

Now that we have a dataset that is diverse with respect to composition (Figure 7) and bandgap (Figure 8), we can conduct regression using the gradient boosted regression tree ensemble implemented within the XGBoost python library to predict bandgap values for our perovskite oxide structures. We set the test set size to 20 per cent of the entire structure dataset. The average root mean squared error (RMSE) value for the prediction model using all input features is 0.1327 eV (12.8036 kJ/mol), compared to the range of bandgap values from 0.3 eV to 3.2 eV. A parity plot for this model as shown in Figure 9 serves as a visualization to this result. Though the error value is not highly accurate (order of 10^{-2}), it is acceptable because the range of bandgap values for solar cell materials is extensive. The test error value of 0.1327 eV for full feature model, may look impressive when compared with the span of bandgap values ranging from 0.3 eV to 3.2 eV, it should be noted that majority of bandgap values within the dataset lies between 0.3 eV to 1.0 eV as shown in Figure 8. For our application with solar cells discovery, the effective range of bandgap values is 1.1 eV to 1.8 eV (Figures 8 and 9). Hence the error of 0.1327 eV is reasonable for large scale screening and identifying candidate structures; however, the most promising structures should be verified with DFT before synthesis.

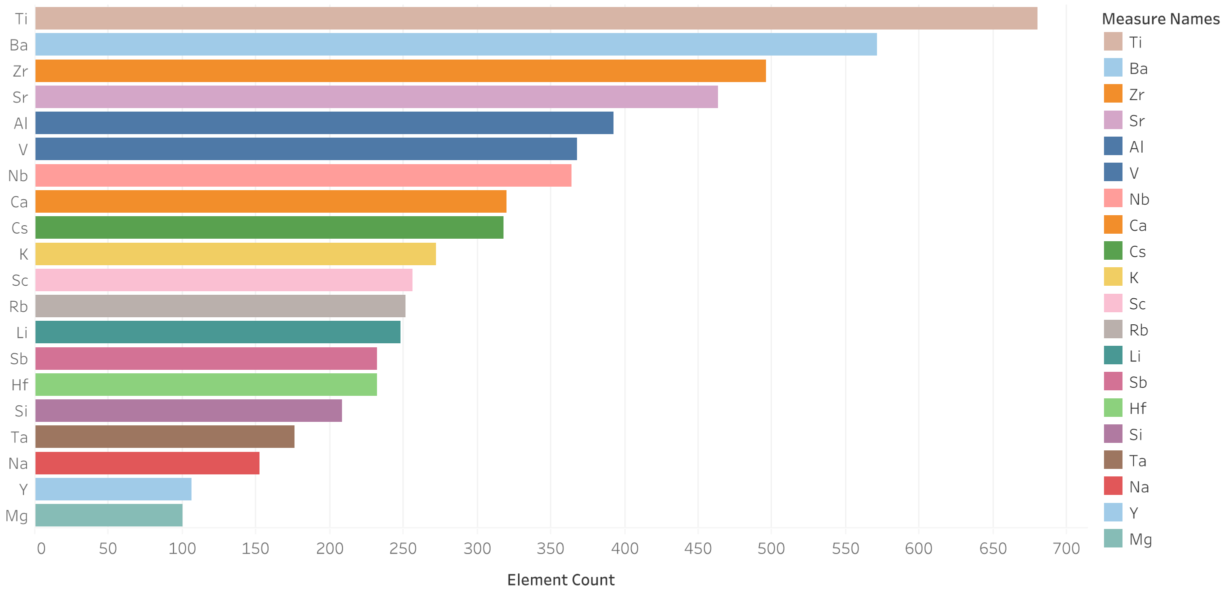


Figure 7: Distribution of elements within the structural dataset.

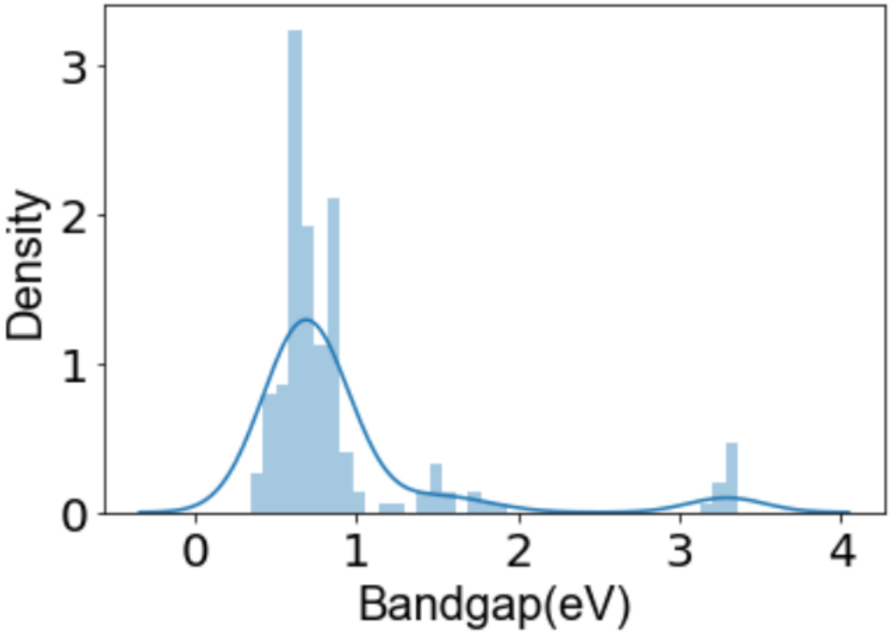


Figure 8 Distribution of the bandgap values for the DFT calculated 191 structures. The blue line here represents probability density function of the distribution.

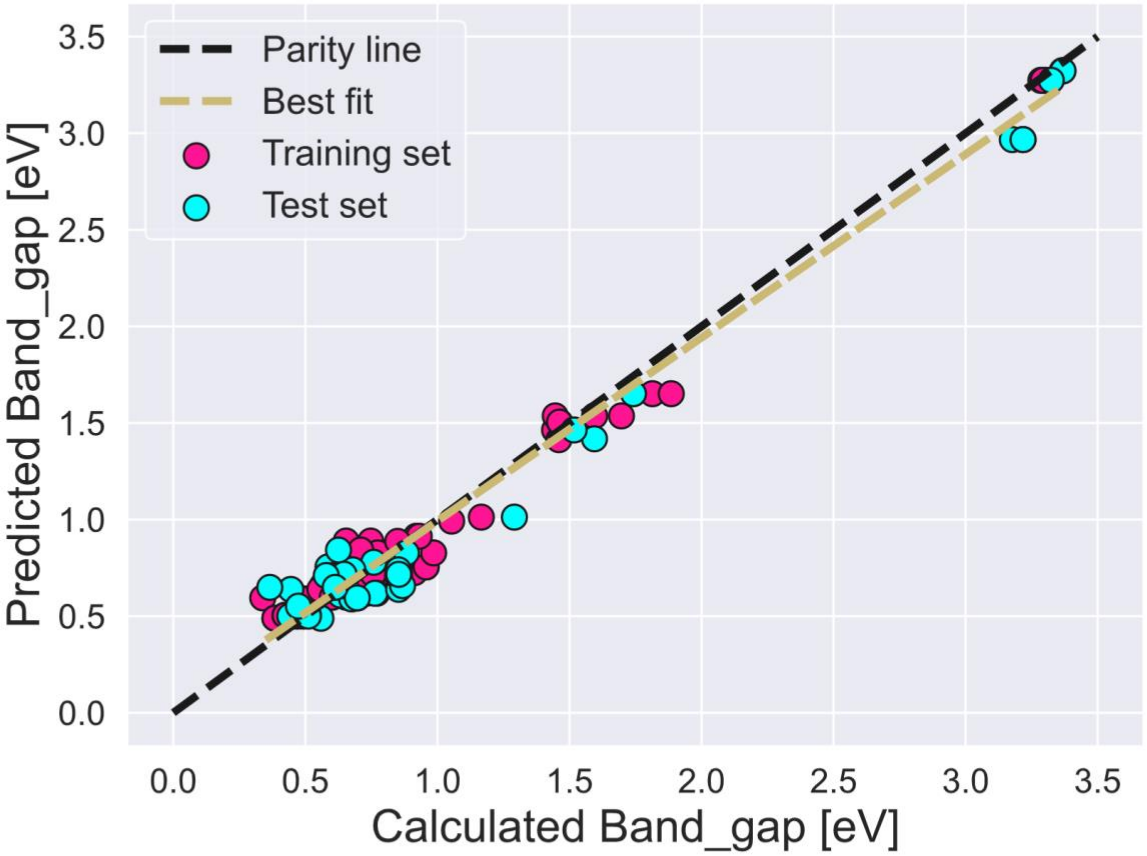


Figure 9: Parity plot for the trained model predictions (from machine learning) vs calculated bandgap testing values (from DFT) for the model with all selected features

As mentioned earlier, the importance scores (Fig 10) demystify the underlying relations within the machine learning model's decision-making process. They quantify how much a given feature influenced the model’s decision. In comparing the importance scores in Figure 10, ionization potential, electronegativity and atomic radius are essential features in predicting the bandgap. Apart from the orbital radii for p, d, and f, all the features have a significant score which says that the three critical features cannot just explain the model, and there is a highly complex and non-linear relationship involved between features and the bandgap.

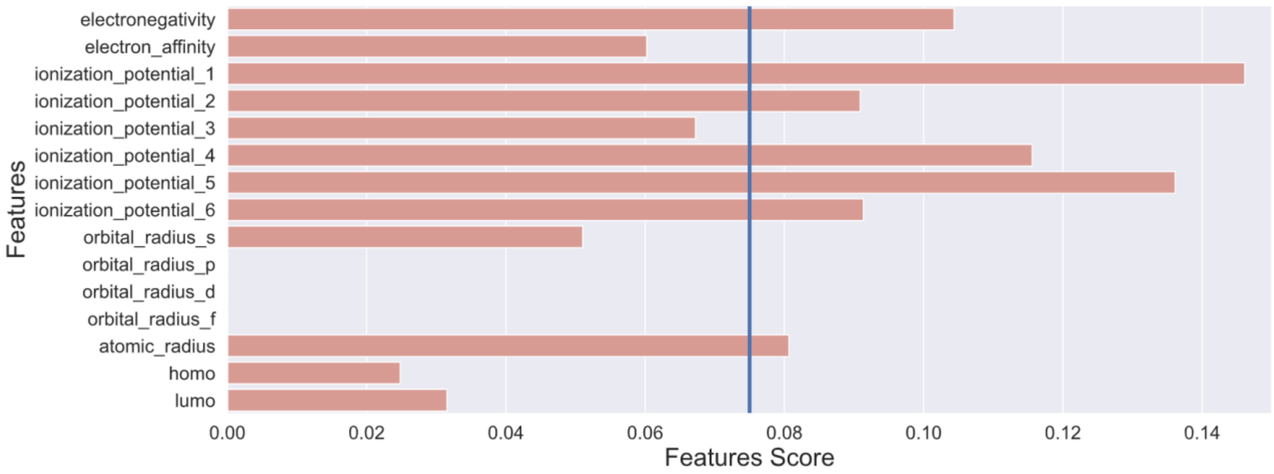


Figure 10: Bar plot of importance scores for the XGBoost model with all selected features. The vertical blue line is the cutoff line for influential features.

Another use of feature importance scores is that we could eliminate less important features which would further bring down the computing cost for running the machine learning model. For this purpose, we use the feature importance scores plot to recursively remove least important feature from the model and assess the change in metric such as average RMSE. This process is also called recursive feature elimination.

This process also helps in countering the phenomenon called overfitting, which is observed in a machine learning model. Overfitting occurs when the model performs well on the training data and severely underperforms on test data. High number of features has a chance to make the model overfit.

The blue line in Figure 10 is the cutoff line for influential features, it's based on rule of thumb where a vertical line is drawn from half of the ceiling value of plot, e.g., ionization_potential_1 has value greater than 0.14 and less than 0.15, so here ceiling value is 0.15 and line is drawn at 0.15/2 or 0.075. Based on the feature extraction process, we found the essential features are electronegativity, ionization potential 1, 2, 4, 5 and 6 and the atomic radius.

Here in Figure 11, we display how the average training and test RMSE values vary for the models with recursively eliminated features. Note that the selection of features represents the top features based on the importance scores. The objective here is to minimize the number of features in the model without taking a hit to the performance. The performance here is assessed by the difference between training and testing errors represented by orange and blue line in Fig 11. Based on the above plot the model with 7 features shown in Fig 12 (features: electronegativity, ionization potential [1,2,4,5,6] and atomic radius) has decent performance with test error RMSE value of 0.1516 eV and can be used in scenarios where the scale of data is enormous, to cut the computational cost.

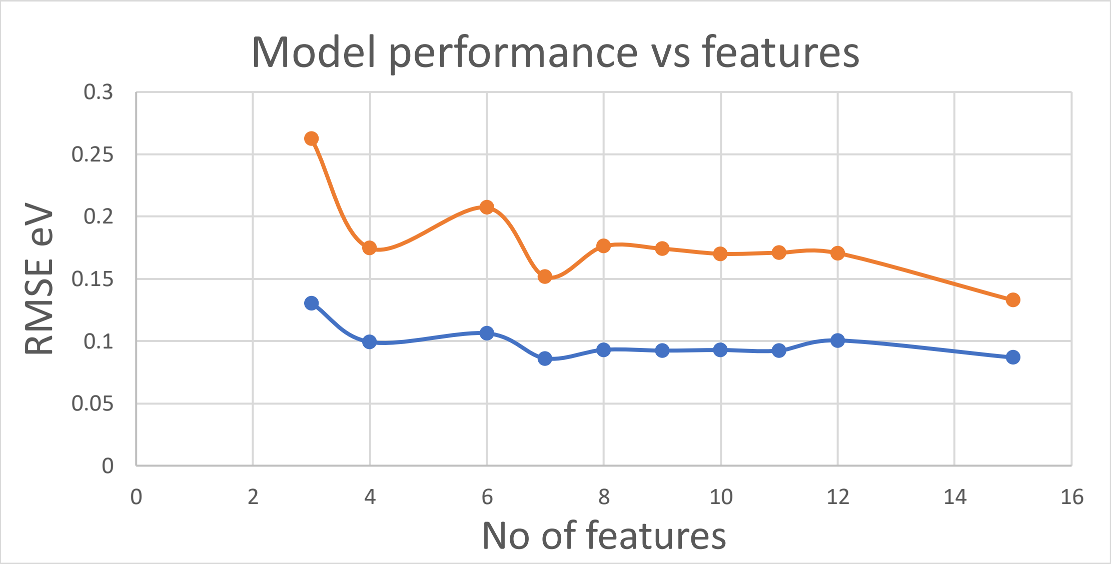


Figure 11: Averaged root mean squared error of bandgap of perovskite oxide ML models as a function of number of features. The orange line corresponds to test errors and blue line correspond to training error.

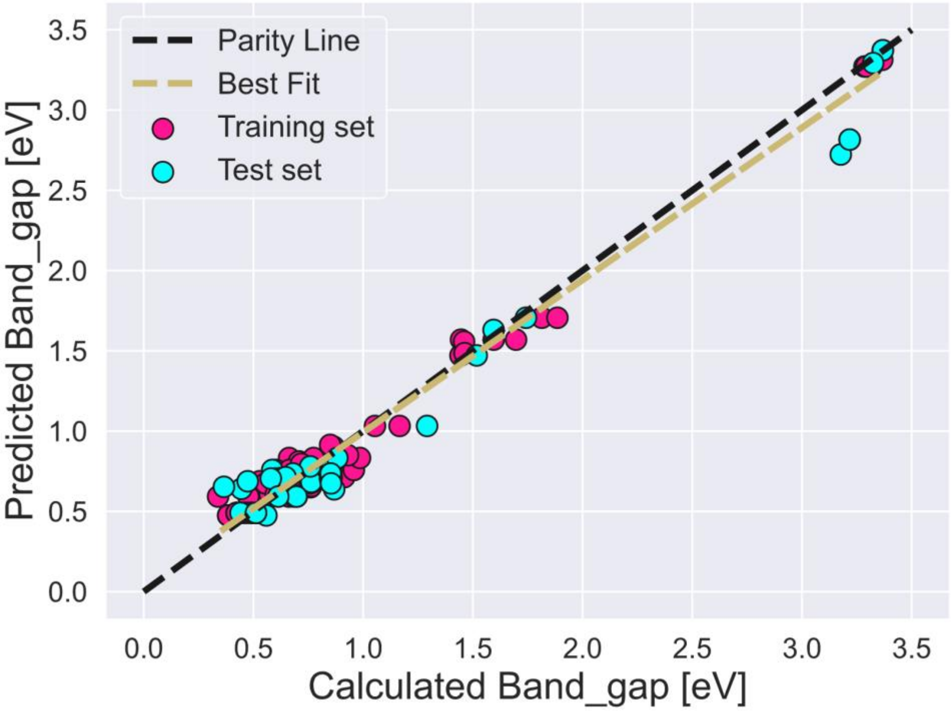


Figure 12: Parity plot for the trained model predictions (from machine learning) vs calculated bandgap testing values (from DFT) for the model with best seven extracted features.

We transform our model to a classification model that predicts what structure arrangement will have a bandgap value between 0.3 eV and 0.6 eV and vice versa. While training the structures, we convert the bandgap values into zeroes if greater than or equal to 0.6 eV and ones if between 0.3 eV than 0.6 eV. The rationale here is to utilize the optimal range of bandgap values that correspond to higher solar absorption efficiency. In practice high solar adsorption efficiency is considered to be a bandgap of 1.1 eV to 1.8 eV. Since our DFT computed bandgap values are underestimated by ~ 0.8 eV (for example, BaTiO3$_{(DFT)}$ = 2.61 eV, BaTiO3[39]$_{(literature)}$ = 3.4 eV; other examples in our dataset show the same trend), we assume that the range between 0.3 to 0.6 eV can approximate to the optimal range of 1.1 to 1.8 eV. Since we have a disproportionate dataset selecting the upper limit near 1.0 eV can lead to unreliable biased predictions as we have comparatively fewer data points above 1.0 eV.

The diagnostic for a classification model is the ratio of correct predictions to total predictions or accuracy score. To understand a better picture of accuracy score, we further dissect the accuracy scores in two parts: false positives and false negatives. False positives would mean the model incorrectly classifies a higher bandgap material among the lower ones and vice versa. The confusion matrix (Figure 13) is an advanced form of accuracy data which sheds light on correct and incorrect classifications for every class. The best cross-validated accuracy score for the classification model with the seven extracted features is 90.13 %, which is approximately 5 points better than the accuracy score for all features' accuracy score at 84.95%.

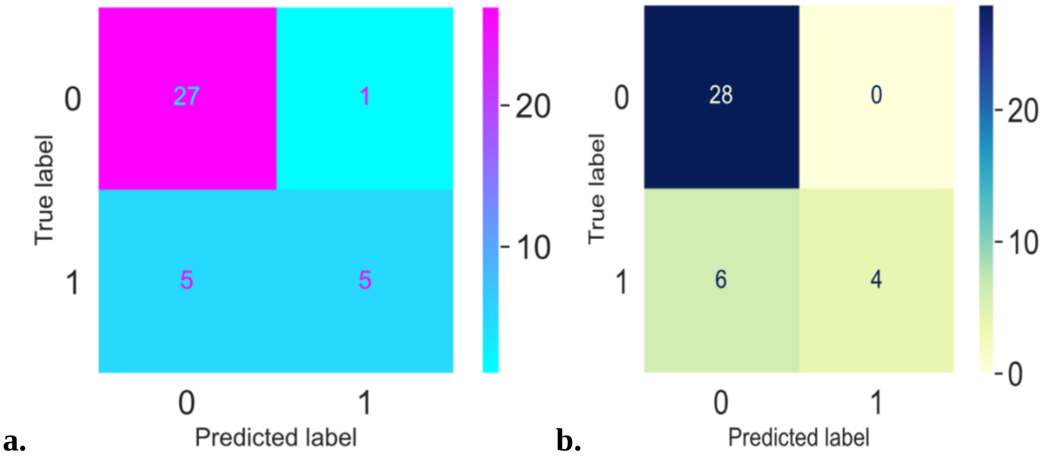


Figure 13 **a.** Confusion matrix for the model with seven key features. **b.** Confusion matrix for the model with full features. Label 1 or Class I is for predicting values less than 0.6eV and Label 0 or Class II is for predicting values equal to or above 0.6eV. Top lefts are true negatives, top right are false negatives, bottom left are false positives and bottom right are true positives. The color intensity is directly proportional to values in the block.

However, both the predictive models perform similarly against the same test data. The full feature model performs better with true negatives, whereas the eight feature model fares better in false positives. To assess a classification model's performance, it is important to know how many predictions were classified correctly or incorrectly for each class for varying degree of acceptance threshold. An ideal classification model should perform well even when the threshold is low i.e., it is prone to accept false entries for a given class and when the threshold is high i.e., it doesn't accept entries easily. The confusion matrix and the receiver operating characteristic (ROC) curve helps us in the diagnosis.

To compute the ROC curve, which is a diagnostic plot that illustrates the ability of classifier by varying its discrimination threshold values, we derive two variables from the confusion matrix. They are the true positive rate (TPR) shown in equation 4 and false positive rate (FPR) shown in equation 5. FPR or fall out is the ratio of false-positive values to the sum of false-positives (and true negatives). TPR is the ratio of true positives to the sum of true positives and false negatives.

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{P}} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} = 1 - \mathrm{FNR}. \tag{4}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{N}} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} = 1 - \mathrm{TNR} \tag{5}$$

Here 'TP' is True Positive, 'P' equals total positives (true positives plus false positives), 'FNR' implies false negative rate, 'FP' means false positive, 'N' means total negatives, ‘TN’ equals true negative, and ‘TNR’ is true negative rate. True positive and false positive rates for a binary classification model are assessed against multiple threshold values. Such an ROC curve helps analyze the classifier’s performance under different discrimination threshold values between 0 and 1.

This comparison is further evident when we look at the ROC curves for both models in Fig 14. The 8-feature model performs better when the tolerance of accepting values is low, whereas the full feature model performs better when threshold values are high. We expect the 8-feature model performance will get better when more training structures are added further. The training data set that we used to train the Gradient Boosting models comprised 191 structures in total.

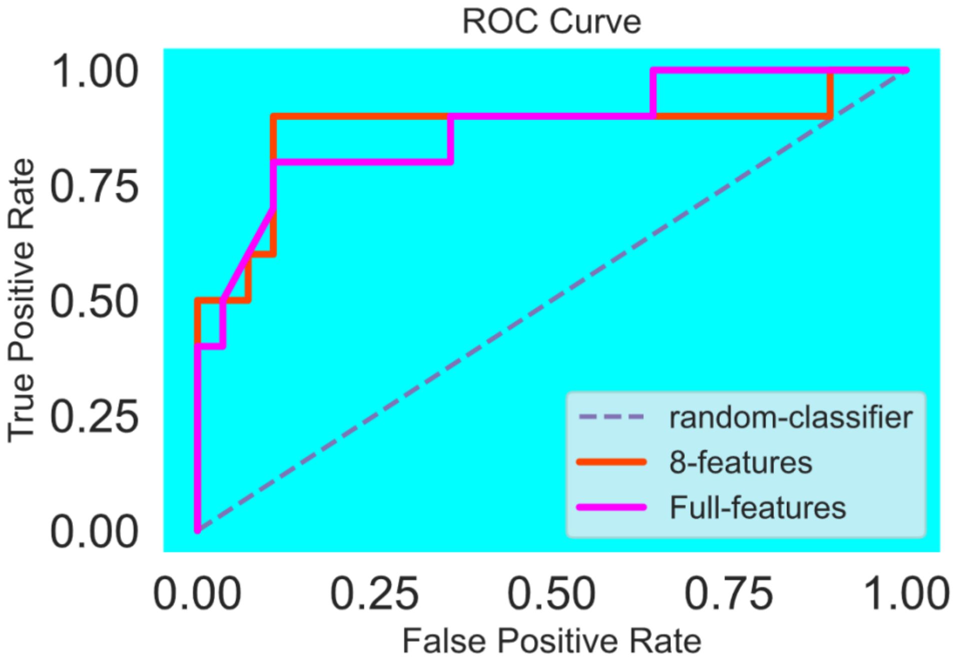


Figure 14: ROC area under curve diagnostic plot comparing the classification models with full features, model with eight key extracted features and a random classifier.

Though the structures’ configurations capture the element diversity, the training data is few for the model to effectively screen the material space. The machine learning model always gets better when more training examples are used.

CHAPTER FOUR

CONCLUSION

To summarize, we created a gradient boosting-based Machine Learning model that predicts bandgap values for perovskite oxide structure class with high fidelity by taking input from the structural formula and what positions they occupy within the unit cell. We also created a classification model that could sort the structures against the determined bandgap threshold of 0.6 eV. Based on the feature importance scores, we extracted critical features from the entire feature set, which helped counter the model overfitting and reduce the average RMSE for the prediction model. The model error of 0.168eV is acceptable when compared against the range of bandgaps that are considered optimal for solar absorption; however, predicted structures should be checked with DFT before synthesis. Though the model's training dataset has significantly less (determined by the high ratio of total structures to sample structures, for an ideal model factor should be of the order 10^2) structures, the results further indicate that model performance will only get better as more structures are added to the training dataset.

Despite fewer training examples, the regression model does perform well in predicting the values in the range where training examples were few. As mentioned earlier, the bandgap values obtained in the dataset through the DFT method are underestimated compared to the experimental values. The classification model developed here provides the solution to this problem since its ability to identify which set of materials and their configurations could have a very low bandgap. After this coarse screening, the candidate materials bandgaps could further be recovered to their optimal values using high fidelity DFT methods incorporating hybrid functionals. Our model with a limited dataset and inaccurate bandgap values could be termed surrogate-optimization or active-learning model since we use such limited resources to arrive at satisfactory results.

Although the machine learning model performs very well, its applicability only lies within the perovskite oxides chemical space. Another reason to explain the excellent fit is that the training structures are similar in class. The model could suffer in accuracy when predicting materials with different orderings. It will be fascinating to see how the selected features could generally apply to distinct classes from the perovskite's family.

CHAPTER FIVE

FUTURE STUDIES

Deeper understanding of structure-property relationships demands evaluation of effect of substituent sites on the structure's property. Though our study takes internal orderings and connections within the unit cell as input, it does not explain their dependence and their interplay on the bandgap's final value. Hence quantitatively understanding how neighboring elements influence a site's contribution to the property will make our model more robust. Such strategies bypass the need entirely for training the model for different properties, and thus one single model should be good enough for predicting all properties for the material. Thereby reducing computational resources [15]. Such machine learning techniques are in use to analyze social network behavior.

Another exciting strategy on material discovery involved collaborated filtering to narrow down the material search. Theoretical intuition is used to generate the chemical space and filtering of unstable compounds. The process involves that the selected candidates are validated experimentally and polled among industrial experts to select the best material. Recently research utilizing this method showed close to a 22% jump in efficiency [17].

Recently researchers at the Ulsan National Institute of Science and Technology, in collaboration with the Swiss Federal Institute of Technology Lausanne, have achieved a new record conversion efficiency of 25.6 per cent in a single junction perovskite solar cell. The material is a metal halide perovskite (FAPbI3). The rationale was to use anion formate to nullify vacancy defects at the film's surface to support the structure's crystallinity [16].

Compared to the total possible structures in the chemical space the total training structures in the study are very few. Structures having bandgap values in the range 1 to 3 eV are required to be added in the training dataset to even out the distribution of bandgaps, which would further ensure that the ML model is more robust and could deliver higher quality predictions.

APPENDICES

Appendix A

Machine Learning model hyperparameters and model metrics

Gradient Boosted Regression tree models

The best hyperparameters were selected from the cross- validated grid search to find the lowest negative median absolute error. They are as follows:

learning rate = 0.1, *colsample_bytree* = 0.7, *gamma* = 0.03, *max_depth* = 8, *min_child_weight* = 6, *reg_alpha* = 0.1, *subsample* = 0.52

Gradient Boosted Classification tree models

Like the regression model we conduct a grid search for classification models as well with accuracy as cost objective. The parameters are as follows:

learning rate = 0.1, *colsample_bytree* = 0.5, *gamma* = 0.03, *max_depth* = 8, *min_child_weight* = 4, *reg_alpha* = 0.1, *subsample* = 0.52

Hyperparameter Terminology

Learning rate - step size to update each boosting step

Gamma - minimum loss required to make a split to the leaf

Max depth – Limit to which the decision tree is allowed to grow

Sub sample – subsample ratio of training instances.

Colsample by tree – subsample ratio of features when constructing each tree.

Alpha – Regularization term

Model validation data

Table1. Root mean squared error (RMSE) and mean absolute error (MAE) values for the models

No of Features	RMSE Train(eV)	RMSE Test(eV)	MAE Train(eV)	MAE Test(eV)
3	0.1304	0.2624	0.0858	0.1578
4	0.0992	0.1745	0.0733	0.1211
6	0.1062	0.2072	0.0712	0.1292
7	0.0861	0.1516	0.0643	0.1095
8	0.0930	0.1761	0.0653	0.1202
9	0.0923	0.1741	0.0649	0.1210
10	0.0928	0.1700	0.0661	0.1185
11	0.0923	0.1709	0.0639	0.1182
12	0.1005	0.1706	0.0729	0.1236
15	0.0868	0.1327	0.066	0.1072

Appendix B

Feature Abbreviations & Terminologies

The atomic features used in the study weren't computed for any elements. The data was gathered from information available online. The aim was to generalize model's applicability, such that the only input required to the is the information about the atom arrangement within the unit cells and element properties that are available online.

Feature Abbreviations

ionz – ionization potential

orb_rad_rs – orbital radius for s orbital

orb_rad_rp – orbital radius for p orbital

orb_rad_rd – orbital radius for d orbital

orb_rad_rf – orbital radius for f orbital

elenega – Electronegativity Pauling

eleaff – Electron affinity

homo – Highest occupied molecular orbital

lumo – Lowest unoccupied molecular orbital

DFT Terminologies

Orbital transformation – Seeks to find energy functional minimum with respect to the molecular orbital coefficients, with the constraint that molecular orbitals are normalized.

Preconditioner – A metric that effects transformation of internal structure coordinates to new coordinates where the optimization problem is better conditioned, hence allowing rapid convergence of the algorithms.[31]

Full Single Inverse – Preconditioner based on Cholesky decomposition.

Cholesky Decomposition – Decomposition of positive definite matrix into lower triangular matrix and its conjugate transpose.

Pulay Mixing – Pulay mixing is used to accelerate the convergence of Hartree Fock self-consistent field method. It attempts to find good approximation through linear combination of set of trial vectors generated during the iteration.

Machine Learning Terminologies

Training set – The dataset of structural attributes as explanatory variables and bandgap as target variable provided to the model for regression.

Test set – The dataset where the model applies its learned parameters from the training set, on newer unseen data to compare its predictions with the actual data.

Gini impurity – Gini impurity quantifies model classification at every iterative step. The model improves by optimizing Gini impurity score.

Ensemble/Bagging – Bagging is a process of improving decision tree

performance. In bagging we create multiple datasets either same or less the size of original. The new datasets sample data from the original where repeated entries are allowed. Multiple decision trees are constructed for each new dataset and these individual trees try to classify a test sample. The class of the sample is then selected by a majority.

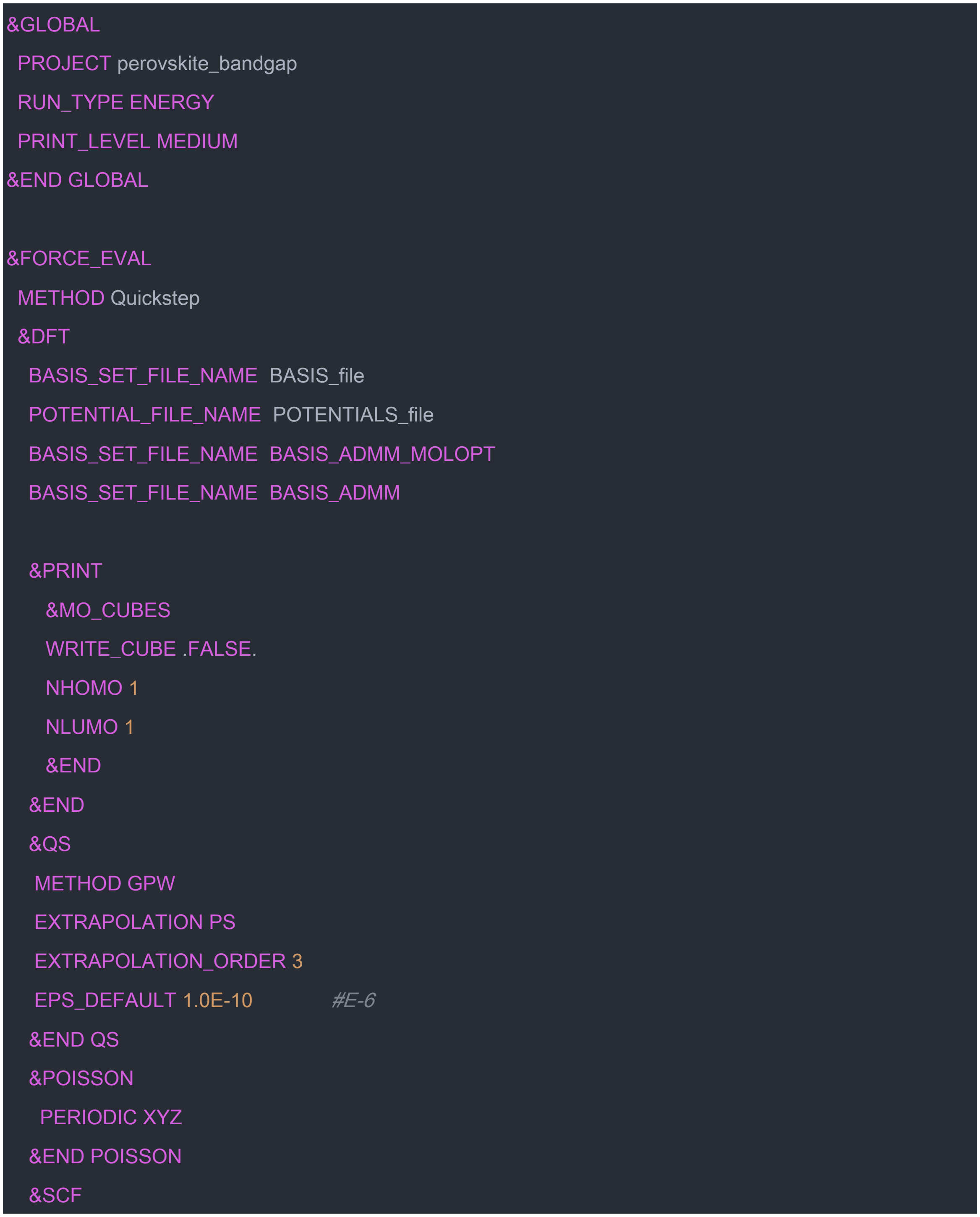
Supplementary information

The DFT calculation input file, the structure dataset and the Machine Learning script is provided through portal called GitHub following is the link.

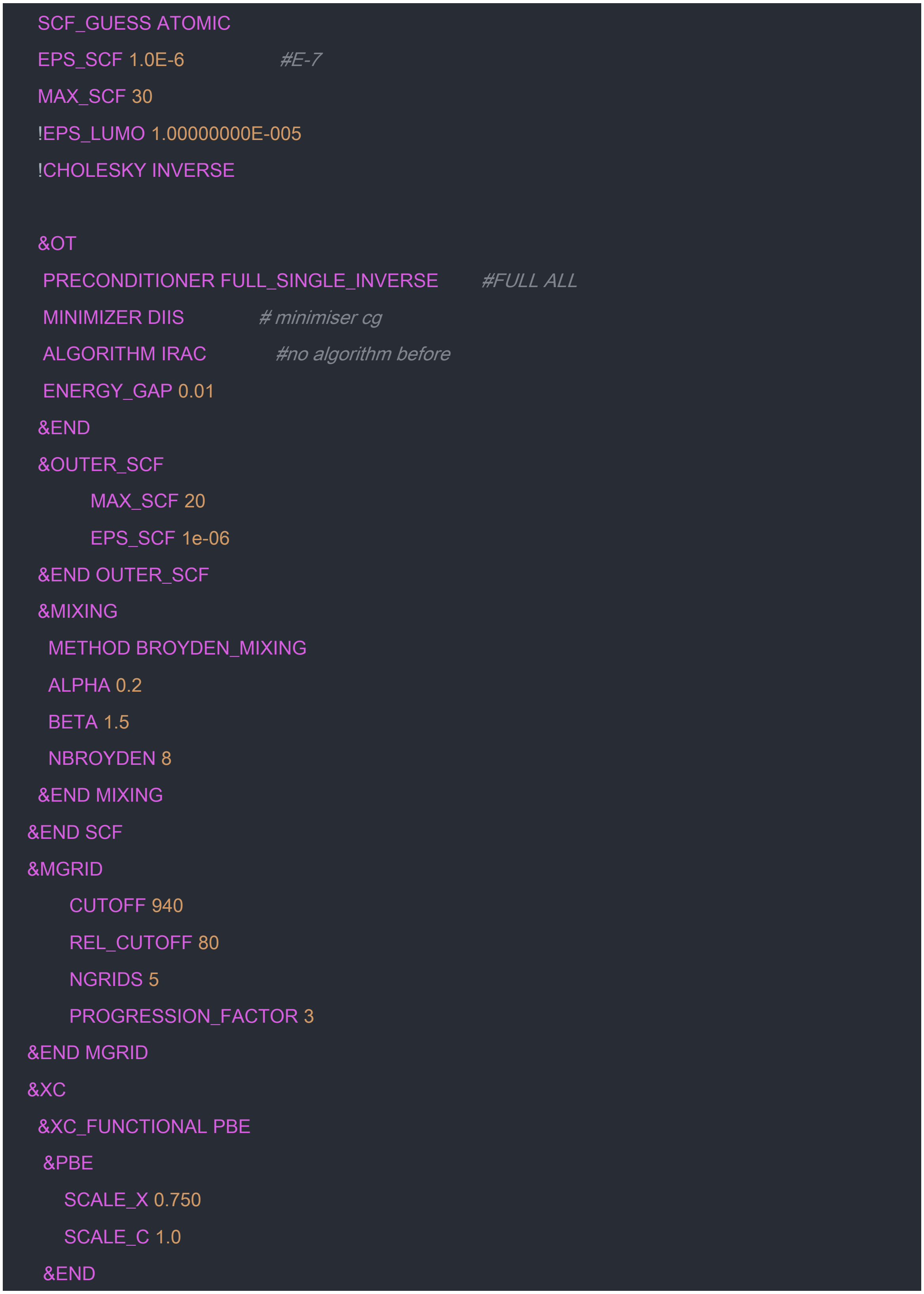
https://github.com/afrid341/Solar-cell-Material-discovery

Appendix C

DFT calculations input template

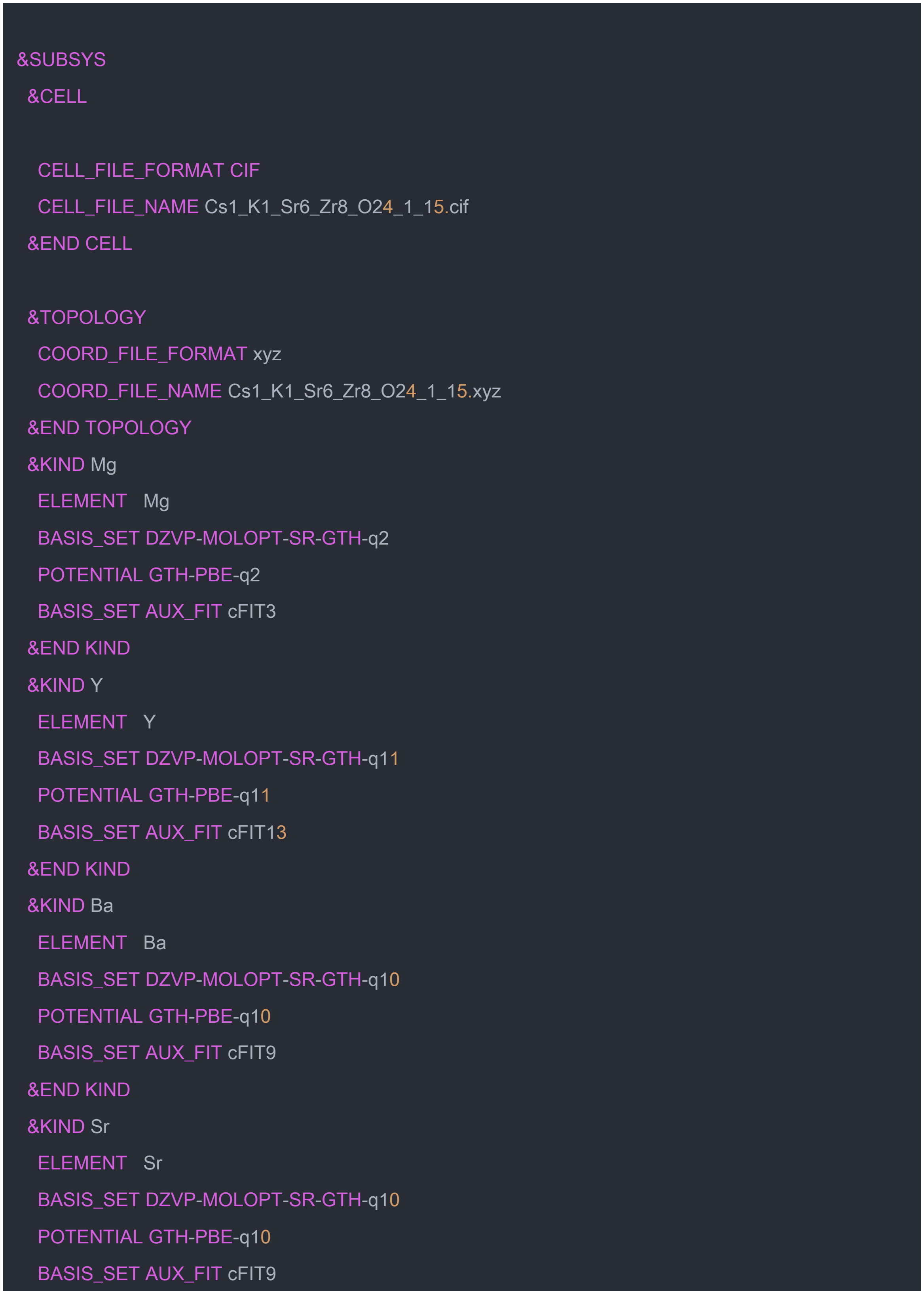


```
&GLOBAL
  PROJECT perovskite_bandgap
  RUN_TYPE ENERGY
  PRINT_LEVEL MEDIUM
&END GLOBAL

&FORCE_EVAL
  METHOD Quickstep
  &DFT
    BASIS_SET_FILE_NAME  BASIS_file
    POTENTIAL_FILE_NAME  POTENTIALS_file
    BASIS_SET_FILE_NAME  BASIS_ADMM_MOLOPT
    BASIS_SET_FILE_NAME  BASIS_ADMM

    &PRINT
       &MO_CUBES
       WRITE_CUBE .FALSE.
       NHOMO 1
       NLUMO 1
       &END
    &END
    &QS
     METHOD GPW
     EXTRAPOLATION PS
     EXTRAPOLATION_ORDER 3
     EPS_DEFAULT 1.0E-10             #E-6
    &END QS
    &POISSON
      PERIODIC XYZ
    &END POISSON
    &SCF
```

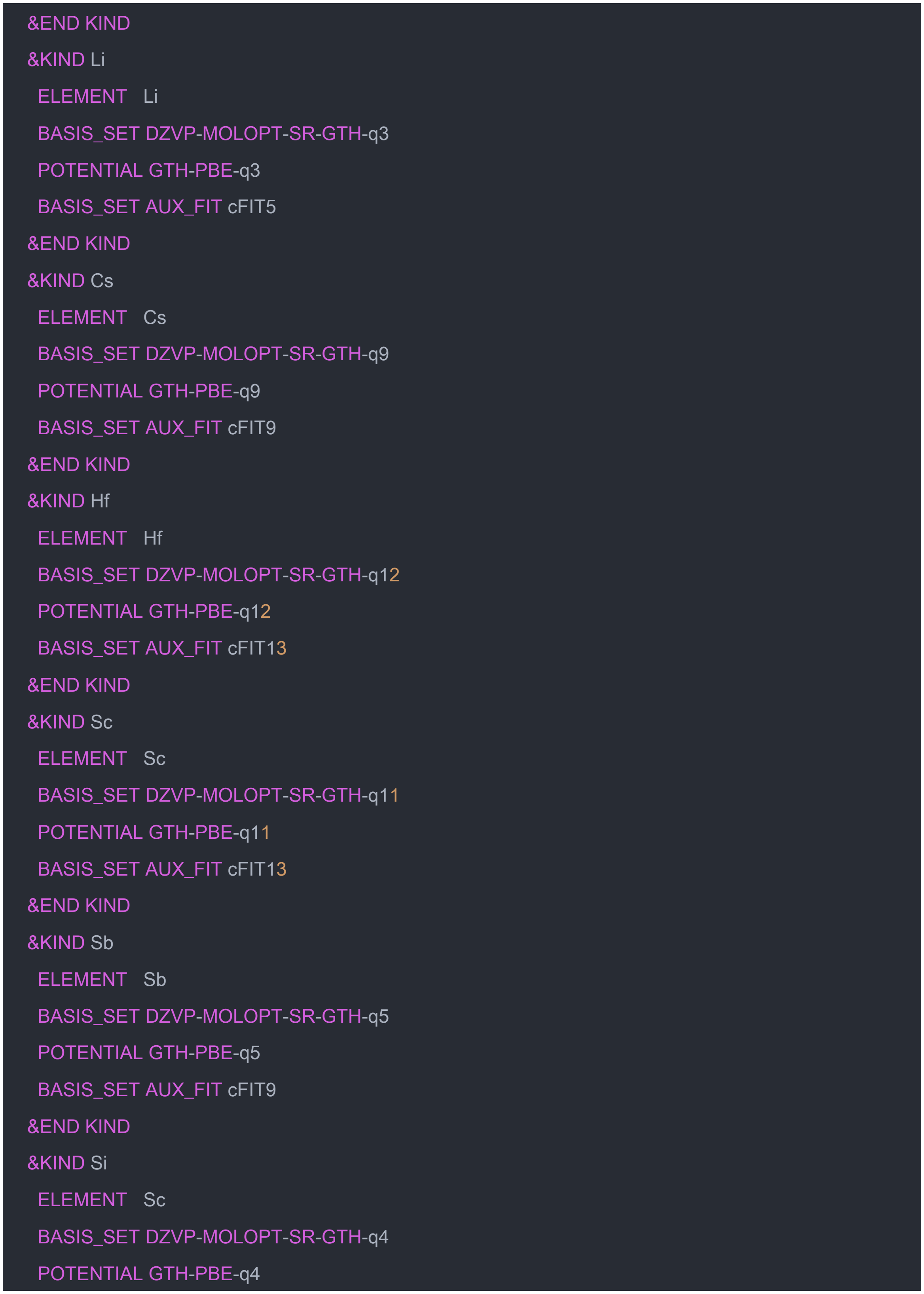


```
   SCF_GUESS ATOMIC
   EPS_SCF 1.0E-6            #E-7
   MAX_SCF 30
   !EPS_LUMO 1.00000000E-005
   !CHOLESKY INVERSE

   &OT
    PRECONDITIONER FULL_SINGLE_INVERSE    #FULL ALL
    MINIMIZER DIIS            # minimiser cg
    ALGORITHM IRAC            #no algorithm before
    ENERGY_GAP 0.01
   &END
   &OUTER_SCF
        MAX_SCF 20
        EPS_SCF 1e-06
   &END OUTER_SCF
   &MIXING
    METHOD BROYDEN_MIXING
    ALPHA 0.2
    BETA 1.5
    NBROYDEN 8
   &END MIXING
  &END SCF
  &MGRID
      CUTOFF 940
      REL_CUTOFF 80
      NGRIDS 5
      PROGRESSION_FACTOR 3
  &END MGRID
  &XC
   &XC_FUNCTIONAL PBE
    &PBE
      SCALE_X 0.750
      SCALE_C 1.0
    &END
```

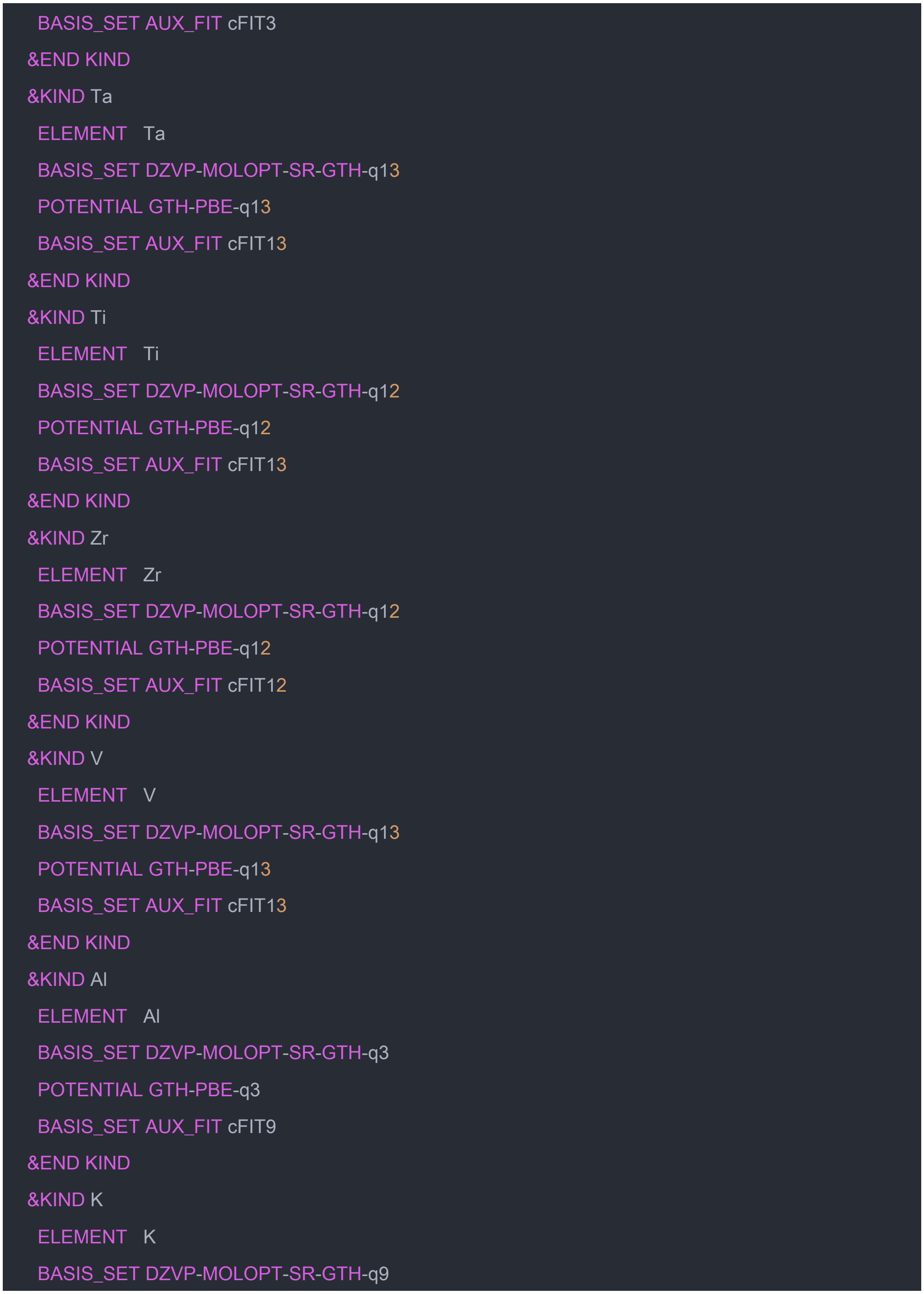
```
       &PBE_HOLE_T_C_LR
           SCALE_X 0.25       ! + 25% of truncated PBE0 functional - that includes exact hfx
           CUTOFF_RADIUS 4.19759400000000000000  ! that has interaction truncated at 3.5 A from the atomic core
       &END
      &END XC_FUNCTIONAL
      &VDW_POTENTIAL
        POTENTIAL_TYPE pair_potential
        &PAIR_POTENTIAL
          TYPE DFTD3(BJ)
          PARAMETER_FILE_NAME dftd3.dat
          REFERENCE_FUNCTIONAL PBE
        &END PAIR_POTENTIAL
      &END VDW_POTENTIAL
      &HF
        FRACTION 0.25
        &SCREENING
          EPS_SCHWARZ 1.0E-7
          !SCREEN_ON_INITIAL_P TRUE
        &END
        &MEMORY
          MAX_MEMORY 7500
        &END
        &INTERACTION_POTENTIAL
          POTENTIAL_TYPE TRUNCATED
          CUTOFF_RADIUS 4.19759400000000000000
          T_C_G_DATA ./t_c_g.dat
        &END
      &END
    &END XC
    &AUXILIARY_DENSITY_MATRIX_METHOD
            ADMM_PURIFICATION_METHOD NONE
            METHOD BASIS_PROJECTION
      &END AUXILIARY_DENSITY_MATRIX_METHOD
  &END DFT
```



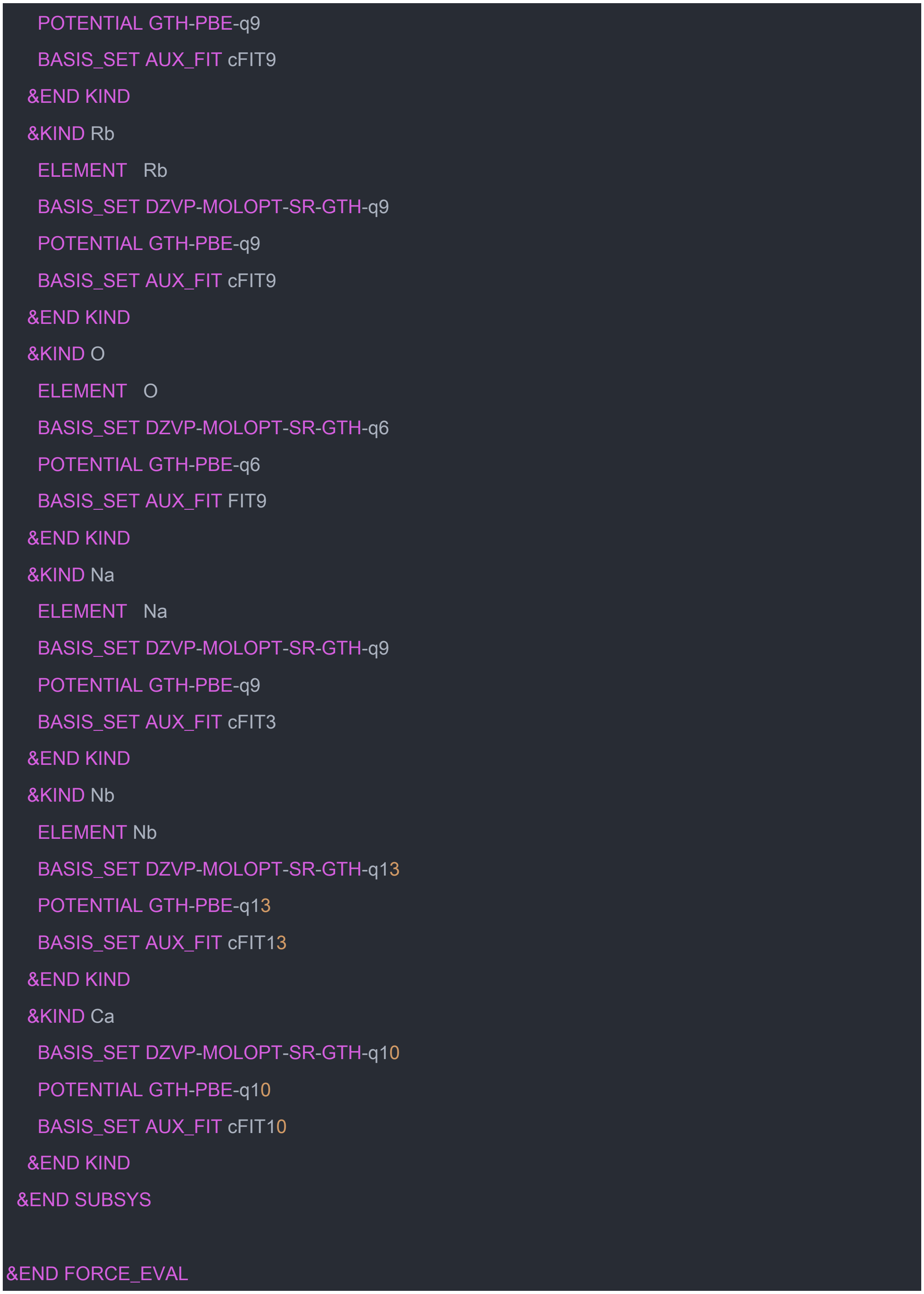
```
&SUBSYS
 &CELL

  CELL_FILE_FORMAT CIF
  CELL_FILE_NAME Cs1_K1_Sr6_Zr8_O24_1_15.cif
 &END CELL

 &TOPOLOGY
  COORD_FILE_FORMAT xyz
  COORD_FILE_NAME Cs1_K1_Sr6_Zr8_O24_1_15.xyz
 &END TOPOLOGY
 &KIND Mg
  ELEMENT   Mg
  BASIS_SET DZVP-MOLOPT-SR-GTH-q2
  POTENTIAL GTH-PBE-q2
  BASIS_SET AUX_FIT cFIT3
 &END KIND
 &KIND Y
  ELEMENT   Y
  BASIS_SET DZVP-MOLOPT-SR-GTH-q11
  POTENTIAL GTH-PBE-q11
  BASIS_SET AUX_FIT cFIT13
 &END KIND
 &KIND Ba
  ELEMENT   Ba
  BASIS_SET DZVP-MOLOPT-SR-GTH-q10
  POTENTIAL GTH-PBE-q10
  BASIS_SET AUX_FIT cFIT9
 &END KIND
 &KIND Sr
  ELEMENT   Sr
  BASIS_SET DZVP-MOLOPT-SR-GTH-q10
  POTENTIAL GTH-PBE-q10
  BASIS_SET AUX_FIT cFIT9
```



```
&END KIND
&KIND Li
  ELEMENT   Li
  BASIS_SET DZVP-MOLOPT-SR-GTH-q3
  POTENTIAL GTH-PBE-q3
  BASIS_SET AUX_FIT cFIT5
&END KIND
&KIND Cs
  ELEMENT   Cs
  BASIS_SET DZVP-MOLOPT-SR-GTH-q9
  POTENTIAL GTH-PBE-q9
  BASIS_SET AUX_FIT cFIT9
&END KIND
&KIND Hf
  ELEMENT   Hf
  BASIS_SET DZVP-MOLOPT-SR-GTH-q12
  POTENTIAL GTH-PBE-q12
  BASIS_SET AUX_FIT cFIT13
&END KIND
&KIND Sc
  ELEMENT   Sc
  BASIS_SET DZVP-MOLOPT-SR-GTH-q11
  POTENTIAL GTH-PBE-q11
  BASIS_SET AUX_FIT cFIT13
&END KIND
&KIND Sb
  ELEMENT   Sb
  BASIS_SET DZVP-MOLOPT-SR-GTH-q5
  POTENTIAL GTH-PBE-q5
  BASIS_SET AUX_FIT cFIT9
&END KIND
&KIND Si
  ELEMENT   Sc
  BASIS_SET DZVP-MOLOPT-SR-GTH-q4
  POTENTIAL GTH-PBE-q4
```



```
  BASIS_SET AUX_FIT cFIT3
 &END KIND
 &KIND Ta
  ELEMENT   Ta
  BASIS_SET DZVP-MOLOPT-SR-GTH-q13
  POTENTIAL GTH-PBE-q13
  BASIS_SET AUX_FIT cFIT13
 &END KIND
 &KIND Ti
  ELEMENT   Ti
  BASIS_SET DZVP-MOLOPT-SR-GTH-q12
  POTENTIAL GTH-PBE-q12
  BASIS_SET AUX_FIT cFIT13
 &END KIND
 &KIND Zr
  ELEMENT   Zr
  BASIS_SET DZVP-MOLOPT-SR-GTH-q12
  POTENTIAL GTH-PBE-q12
  BASIS_SET AUX_FIT cFIT12
 &END KIND
 &KIND V
  ELEMENT   V
  BASIS_SET DZVP-MOLOPT-SR-GTH-q13
  POTENTIAL GTH-PBE-q13
  BASIS_SET AUX_FIT cFIT13
 &END KIND
 &KIND Al
  ELEMENT   Al
  BASIS_SET DZVP-MOLOPT-SR-GTH-q3
  POTENTIAL GTH-PBE-q3
  BASIS_SET AUX_FIT cFIT9
 &END KIND
 &KIND K
  ELEMENT   K
  BASIS_SET DZVP-MOLOPT-SR-GTH-q9
```



```
   POTENTIAL GTH-PBE-q9
   BASIS_SET AUX_FIT cFIT9
  &END KIND
  &KIND Rb
   ELEMENT   Rb
   BASIS_SET DZVP-MOLOPT-SR-GTH-q9
   POTENTIAL GTH-PBE-q9
   BASIS_SET AUX_FIT cFIT9
  &END KIND
  &KIND O
   ELEMENT   O
   BASIS_SET DZVP-MOLOPT-SR-GTH-q6
   POTENTIAL GTH-PBE-q6
   BASIS_SET AUX_FIT FIT9
  &END KIND
  &KIND Na
   ELEMENT   Na
   BASIS_SET DZVP-MOLOPT-SR-GTH-q9
   POTENTIAL GTH-PBE-q9
   BASIS_SET AUX_FIT cFIT3
  &END KIND
  &KIND Nb
   ELEMENT Nb
   BASIS_SET DZVP-MOLOPT-SR-GTH-q13
   POTENTIAL GTH-PBE-q13
   BASIS_SET AUX_FIT cFIT13
  &END KIND
  &KIND Ca
   BASIS_SET DZVP-MOLOPT-SR-GTH-q10
   POTENTIAL GTH-PBE-q10
   BASIS_SET AUX_FIT cFIT10
  &END KIND
 &END SUBSYS

&END FORCE_EVAL
```

REFERENCES

1. Ke, W., Kanatzidis, M.G. Prospects for low-toxicity lead-free perovskite solar cells. Nat Commun 10, 965 (2019).
2. G. H. Gu, J. Noh, I. Kim and Y. Jung, Machine learning for renewable energy materials, *J. Mater. Chem. A*, 2019, **7**, 17096 —17117
3. Im, J., Lee, S., Ko, TW. et al. Identifying Pb-free perovskites for solar cells by machine learning. npj Comput Mater 5, 37 (2019).
4. Im, J., Lee, S., Ko, TW. et al. Identifying Pb-free perovskites for solar cells by machine learning. npj Comput Mater 5, 37 (2019).
5. Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. arXiv:1603.02754 (2016).
6. Pymatgen Package: Shyue Ping Ong, William Davidson Richards, Anubhav Jain, Geoffroy Hautier, Michael Kocher, Shreyas Cholia, Dan Gunter, Vincent Chevrier, Kristin A. Persson, Gerbrand Ceder. Python Materials Genomics (pymatgen) : A Robust, Open-Source Python Library for Materials Analysis. Computational Materials Science, 2013, 68, 314–319.
7. Xie, Tian and Jeffrey C. Grossman. "Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties." Physical Review Letters 120, 14 (April 2018): 145301 © 2018 American Physical Society.
8. Guidon, Manuel & Hutter, Jürg & VandeVondele, Joost. (2010). Auxiliary Density Matrix Methods for Hartree−Fock Exchange Calculations. Journal of

Chemical Theory and Computation - J CHEM THEORY COMPUT. 6. 10.1021/ct1002225.

9. Fischer, C. C. et al. Predicting crystal structure by merging data mining with quantum mechanics. Nat. Mater. 5, 641 (2006).
10. Mitchell, R. H. Perovskites: Modern and Ancient (Almaz Press, Ontario, Canada, 2002).
11. Computational Materials Repository https://wiki.fysik.dtu.dk/cmr/ (Documentation) and https://cmr.fysik.dtu.dk/
12. Pilania, G., Mannodi-Kanakkithodi, A., Uberuaga, B. et al. Machine learning bandgaps of double perovskites. Sci Rep 6, 19375 (2016).
13. Gus L. W. Hart and Rodney W. Forcade, "Algorithm for generating derivative structures," Phys. Rev. B 77 224115, (26 June 2008)
14. T. D. Kühne, M. Iannuzzi, M. Del Ben, V. V. Rybkin, P. Seewald, F. Stein, T. Laino, R. Z. Khaliullin, O. Schütt, F. Schiffmann et al., “CP2K: An electronic structure and molecular dynamics software package—Quickstep: Efficient and accurate electronic structure calculations,” J. Chem. Phys. 152, 194103 (2020)
15. Xie, T. & Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. Phys. Rev. Lett. 120, 145301 (2018).
16. Jeong, J., Kim, M., Seo, J. et al. Pseudo-halide anion engineering for α-$FAPbI_3$ perovskite solar cells. Nature (2021).

17. R. Gómez-Bombarelli, J. Aguilera-Iparraguirre, T. D. Hirzel, D. Duvenaud, D. Maclaurin, M. A. Blood-Forsythe, H. S. Chae, M. Einzinger, D.-G. Ha, T. Wu, G. Markopoulos, S. Jeon, H. Kang, H. Miyazaki, M. Numata, S. Kim, W. Huang, S. I. Hong, M. Baldo, R. P. Adams and A. Aspuru-Guzik, Nat. Mater., 2016, 15, 1120.

18. Volonakis, G. et al. Lead-free halide double perovskites via heterovalent substitution of noble metals. J. Phys. Chem. Lett. 7, 1254 (2016).

19. Philip, M. R. et al. Band gaps of the lead-free halide double perovskites Cs2BiAgCl6 and Cs2BiAgBr6 from theory and experiment. J. Phys. Chem. Lett. 7, 2579 (2016).

20. Rajan, K. in Informatics for Materials Science and Engineering: Data-driven Discovery for Accelerated Experimentation and Application (ed. Rajan, K.), Ch. 1, 1–16 (Butterworth-Heinemann, Oxford, 2013).

21. P., Dey et al. Informatics-aided bandgap engineering for solar materials. Com. Mat. Sci. 83, 185–195 (2014).

22. K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev and A. Walsh, Nature, 2018, 559, 547–555

23. Visualization for electronics and structure software (VESTA): https://jp-minerals.org/vesta/en/

24. Manuel Guidon, Jürg Hutter, and Joost VandeVondele. Auxiliary Density Matrix Methods for Hartree−Fock Exchange Calculations Journal of Chemical Theory and Computation 2010 6 (8), 2348-2364 DOI: 10.1021/ct1002225

25. Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized gradient approximation made

simple. Physical review letters 1996, 77, 3865.

26. Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. J.Comput.Chem. 2011, 32, 1456– 1465, DOI: 10.1002/jcc.21759
27. VandeVondele, J.; Krack, M.; Mohamed, F.; Parrinello, M.; Chassaing, T.; Hutter, J. Quickstep: fast and accurate density functional calculations using a mixed Gaussian and plane waves approach. Comput. Phys. Commun. 2005, 167, 103– 128, DOI: 10.1016/j.cpc.2004.12.014
28. Joost VandeVondele and Jürg Hutter, "An efficient orbital transformation method for electronic structure calculations", The Journal of Chemical Physics 118, 4365-4369 (2003)
29. Predicting the Thermodynamic Stability of Solids Combining Density Functional Theory and Machine Learning Jonathan Schmidt, Jingming Shi, Pedro Borlido, Liming Chen, Silvana Botti, and Miguel A. L. Marques Chemistry of Materials 2017 29 (12), 5090-5103 DOI: 10.1021/acs.chemmater.7b00156
30. Gregor Michalicek (https://physics.stackexchange.com/users/166799/gregor-michalicek), Why does Density Functional Theory (DFT) underestimate bandgaps?, URL (version: 2021-03-28): https://physics.stackexchange.com/q/360454
31. Mones, L., Ortner, C. & Csányi, G. Preconditioners for the geometry optimisation and saddle point search of molecular systems. Sci Rep 8, 13991 (2018). https://doi.org/10.1038/s41598-018-32105-x

32. Lee, J., Seko, A., Shitara, K. & Tanaka, I. Prediction model of band-gap for AX binary compounds by combination of density functional theory calculations and machine learning techniques. arXiv preprint arXiv:1509.00973 (2015).

33. Pozun, Z. *et al.* Optimizing transition states via kernel-based machine learning. *Chem. Phys.* **136,** 174101 (2012).

34. Roy K, Kar S, Das RN (2015). "Chapter 1.2: What is QSAR? Definitions and Formulism". A primer on QSAR/QSPR modeling: Fundamental Concepts. New York: Springer-Verlag Inc. pp. 2–6. ISBN 978-3-319-17281-1.

35. Shockley, W. & Queisser, H. J. Detailed balance limit of efficiency of p-n junction solar cells. J. Appl. Phys. 32, 510–519 (1961).

36. Shounak Datta, Vikrant A. Dev, Mario R. Eden, Developing QSPR for Predicting DNA Drug Binding Affinity of 9-Anilinoacridine Derivatives Using Correlation-Based Adaptive LASSO Algorithm, Editor(s): Antonio Espuña, Moisès Graells, Luis Puigjaner, Computer Aided Chemical Engineering, Elsevier, Volume 40,2017, Pages 2767-2772, ISSN 1570-7946, ISBN 9780444639653,https://doi.org/10.1016/B978-0-444-63965-3.50463-3.

37. Castelli, I. E., Thygesen, K. S. & Jacobsen, K. W. Bandgap engineering of double perovskites for one-and two-photon water splitting. MRS Proceedings 1523, mrsf12-1523-qq07-06 (2013), doi: 10.1557/opl.2013.450.

38. Schneider, G. et al. Voyages to the (un)known: adaptive design of bioactive compounds. Trends Biotechnol. 27, 18–26 (2009).

39. AIP Conference Proceedings 1875, 020017 (2017); https://doi.org/10.1063/1.4998371 Published Online: 08 August 2017

40. Yusuke Noda, Masanari Otake & Masanobu Nakayama (2020) Descriptors for dielectric constants of perovskite-type oxides by materials informatics with first-principles density functional theory, Science and Technology of Advanced Materials, 21:1, 92-99, DOI: 10.1080/14686996.2020.1724824

41. Perdew, J. P., & Levy, M. (1983). Physical Content of the Exact Kohn-Sham Orbital Energies: Band Gaps and Derivative Discontinuities. Physical Review Letters, 51(20), 1884–1887. doi:10.1103/physrevlett.51.1884