A Machine Learning Approach for the Prediction of Formability and Thermodynamic Stability of Single and Double Perovskite Oxides

Masrat Rasool, Samir Belhouari, Fedwa El Mellouhi

May, 2022

1 Motivation

We focus on single and double oxide perovskites in this work, and we explore the large chemical space to identify compositions that are likely to form stable perovskites. We investigated the formability and thermodynamic stability of oxide perovskites using a multifaceted approach. To accomplish this, we used an initial training data set [11] using a combination of first-principles calculations and literature data. These training data sets are then used to build machine learning (ML) classification models. The models are then used to search an exhaustive chemical space of single and double perovskites for new oxide perovskites that are formable and stable.

2 Methods

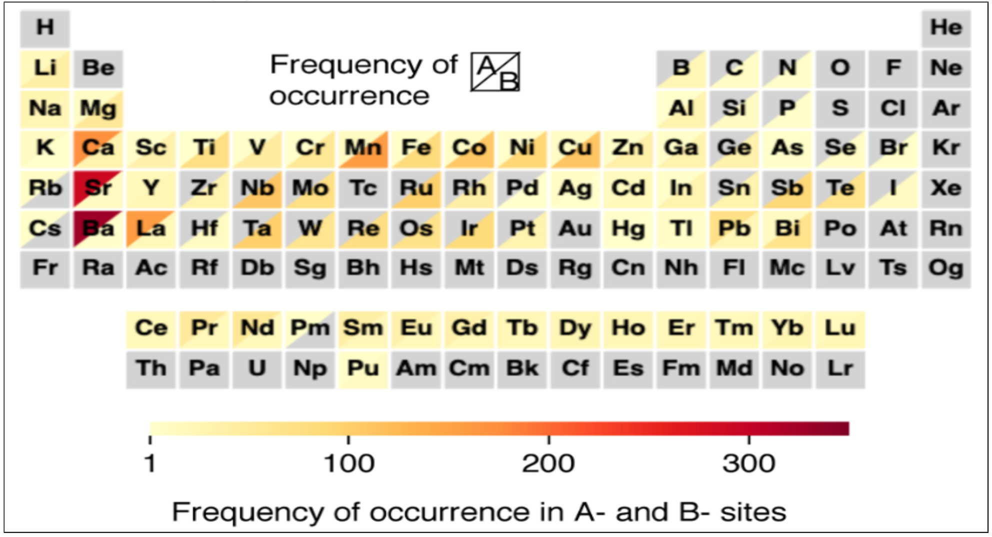
For predicting formability and thermodynamic stability of perovskites, our machine learning models went through five stages of development and validation. (i) The development of a feature set capable of capturing the thermodynamic properties of perovskite oxides. (ii) Using feature selection to identify significant features that have a high correlation with stability. (iii) From the set of possible machine learning algorithms, choose the best machine learning model. (iv) Validation of the model for various perovskite composition spaces, depending on the frequency with which each element appears in the training dataset. Each of the above processes required to build our machine learning models is described in detail in the sections that follow.
For all machine learning models, feature selection methods, and model evaluations, we used the python library scikit-learn [5], which is a free machine learning library.

2.1 Dataset and Feature Generation

2.1.1 Literature-Based Formability database ($\mathbf{D}_F$)

As a part of this work, we used two databases created by [11] one for formability and the other for stability predictions. The training data set for perovskite formability D_F was collected from literature which included 1505 experimentally known oxides of the form ABO3 and AABBO6. Non-perovskites constitute 318 of the chemicals while 1187 are perovskites. Data which covers both single and double perovskites obtained from the Inorganic Crystal Structure Database (ICSD) and databases prepared by [13] [1] are also included in the data collection. For all compounds in figure 1 the periodic table depicts the distribution of the elements filling the A- and B-sites. The presence of a large set of known perovskite oxides containing elements Sr and Ba at the A-site, for example, may be seen.

Figure 1: The number of occurrences of each element in the perovskite formability database D_F is shown in a periodic table [11]



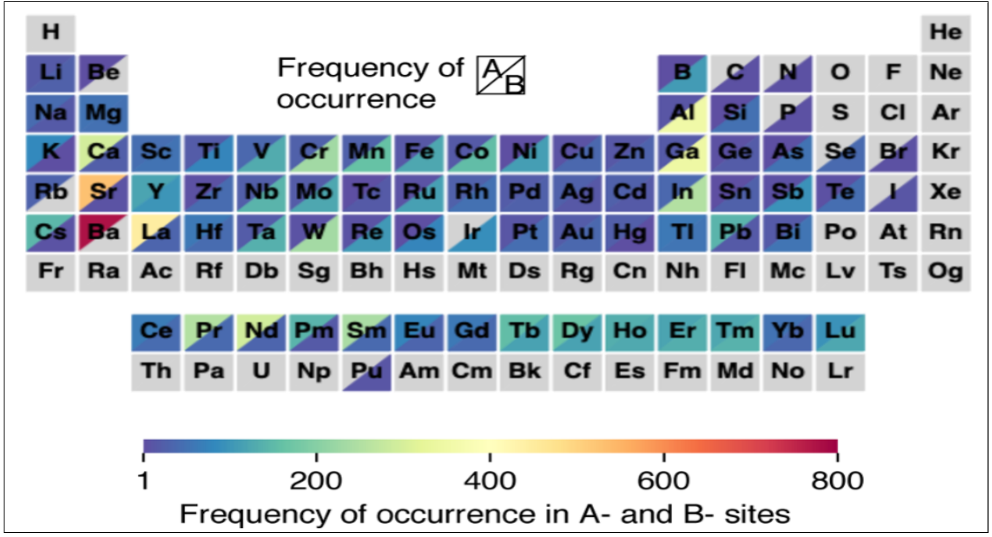
2.1.2 Stability Database(D$_s$)

We begin by listing all potential $ABO3$, $A2BBO6$, $AAB2O6$, and $AABBO6$ compounds for the second training data set, which is used to forecast the thermodynamic stability of perovskites. With the exclusion of noble gases, halogens, and C, N, S, and P, all elements up to Bi are considered.As illustrated in Figure 2, this equates to 68 elements, resulting in $^{68}P_2 = \frac{68!}{(68-2)!} = 4556$ different ABO3 combinations. We have $^{68}P_3 = \frac{68!}{(68-3)!} = 150,843$ distinct combinations for the A2BBO6 and AAB2O6, which account for equivalent A and A sites and B and B sites, respectively. There are $^{68}P_4 = \frac{68!}{(68-4)!} = 19,545,240$ combinations for two A cations and two B cations, that is, the AABBO6 structures, which reduces to 4,886,310 combinations when analogous A and A sites and B and B sites are taken into account.To keep the data set manageable, we only considered equimolar cation compositions at the A- and/or B-site sublattices for the double perovskites. The A, A, B, B cations were assigned valences based on the most usually displayed oxidation states, and only those combinations with the total of the valences of 6 for single perovskites and 12 for double perovskites were maintained for further investigation. This enumeration eliminated double perovskite compositions with a shared cation chemistry between the A and B sublattices. As a result, a number of compounds having different valence states were discovered, and they were all included in the data set. As a result, we have a data collection comprising 946, 292 distinct compounds (some of which have numerous valence combinations), which we name D_C (chemically compatible data set)[11].

The training data set for thermodynamic stability, D_S, of perovskites was calculated using the list of chemicals that correspond to D_C. Compounds were chosen for calculation to maintain a high level of overlap with the formability database D_F, as well as to ensure maximal diversity in the chemistries represented.

To distinguish between stable and unstable perovskites, we used an E_{hull} value of 50 meV/atom as a threshold (i.e., $E_{hull} = 50meV/atom$ is a stable compound). Any compound that is not on

Figure 2: The number of occurrences of each element in the estimated thermodynamic stability data set D_S is shown in a periodic table [11]



the convex hull is considered technically unstable. However, earlier research utilizing chemicals from the Materials Project database has revealed that a cutoff of 50 meV/atom is a fair threshold for distinguishing between materials that are most likely stable and those that are metastable or unstable [9].

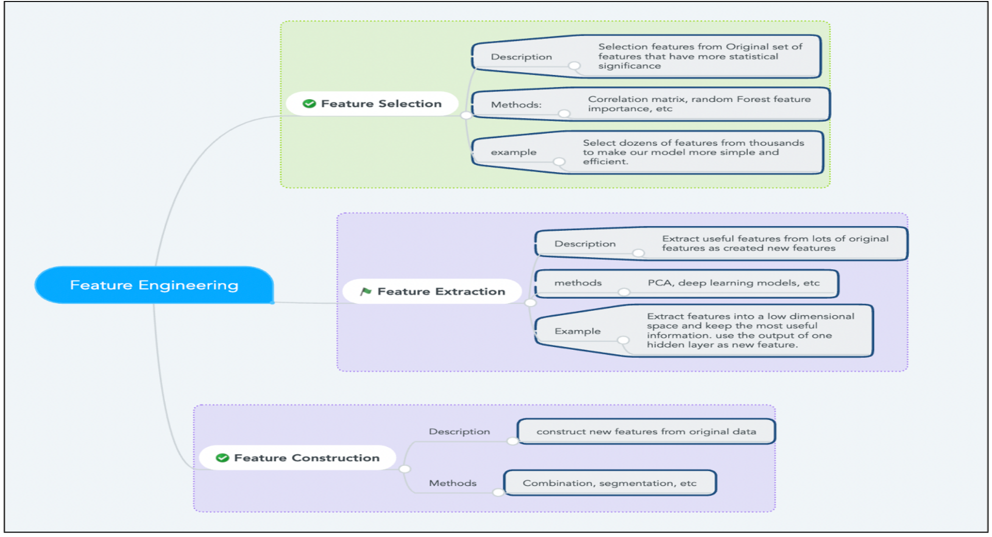
2.1.3 Feature Generation and Selection

The ontology of process of Feature Engineering in this work is depicted in figure 3. The databases include values for properties associated with the A and B atoms in single perovskites and A, A, B, Batoms in double perovskites for all compounds. Researchers have employed symmetric and antisymmetric compound characteristics for the double perovskites, which is a technique that has previously been studied [6] . For a given property P in a double perovskite $AABBO6$, the symmetric compound feature $P^{A+} = \frac{(P_A+P_A^{'})}{2}$ and the antisymmetric compound feature $P_{A-} = \frac{|P_A-P_A^{'}|}{2}$ can be calculated for the A-site, where P_A and $P_A^{'}$ are the elemental properties of A and A, respectively; similar features were defined for the B-site.

To train the categorization models, a vast number of structural and chemical features were first added. We used Recursive Feature Elimination (RFE) [3], a feature selection approach to remove redundant or irrelevant features. The number of times that the randomized approach picks a certain feature by repeating random subsamples of the data and fitting to a logistic regression model in a classification task (or lasso model for regression) [4] is used to select features in stability selection. RFE picks the most important features by iteratively deleting those with the smallest weight supplied by an extra trees classifier in a classification task (or an extra trees regressor in a regression task) [2]. The quantity of mutual information between each feature and the target classification value is used to rank all features in univariate feature selection. Based on entropy measured by a nearest neighbor, mutual information assesses the degree of reliance between features and the goal value.

Prior to performing feature selection, all features are standardized to have a mean of 0 and a standard deviation of 1. As standardization of the feature set is a frequent need for many machine learning models, such as artificial neural networks [12] and support vector machines [10], which

Figure 3: Ontology of the Fetaure Engineering in this work.


Feature Engineering
Feature Selection
Description
Selection features from Original set of features that have more statistical significance
Methods:
Correlation matrix, random Forest feature importance, etc
example
Select dozens of features from thousands to make our model more simple and efficient.
Feature Extraction
Description
Extract useful features from lots of original features as created new features
methods
PCA, deep learning models, etc
Example
Extract features into a low dimensional space and keep the most useful information. use the output of one hidden layer as new feature.
Feature Construction
Description
construct new features from original data
Methods
Combination, segmentation, etc

are both sensitive to feature scaling, this normalization is employed to ensure that all features are scaled in the same way. Only a subset of the initially evaluated properties are truly relevant in describing the formability and the stability of the oxide perovskite, according to the features that were retained.

The highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO) energies, ionization energy, electro- negativity, Zunger's pseudopotential radius (sum of the radii for the s and p orbitals), and electron affinity, as well as geometric features of the compound, the tolerance factor (t), octahedral factor (μ), and mismatch factor ($\bar{\mu}_B$), are among them. We incorporate an additional mismatch factor for the A-site since we include compounds with two cations in both the A- and B-sites: $\bar{\mu}_A = \frac{|r_A - r_{A'}|}{2r_X}$. For the first six of these characteristics, symmetric and antisymmetric compound features for the A- and B-sites are used, resulting in 24 compound features. We now have a total of 20 features, including the four geometric features.

2.2 ML Classification: Technical Details

The process of categorizing a set of data into categories is known as classification. In machine learning terms, classification is a type of supervised learning in which a variable (or class) is predicted based on the values of other variables or features. For both formability and stability classification models, we employed the random forest classifier(RFC) [8], support vector machines [10], Gradient Boosting, extra trees classifier and Naive Bayes classifiers in this study.

In this work, we use the Scikit-learn package to create random forest classifiers. For both the formability and stability classification models, we adopt an 80/20 training/test split. The training and test data sets were stratified according to perovskite/non-perovskite candidates or stable/unstable chemistries in each example. As a result, the models were trained on 80% of the data sets before being evaluated on the remaining 20%. The details of the implemented classifiers is provided below:

1. **Random Forest:** A bagging-based ensemble learning method, the random forest (also known as random decision trees), is a bagging-based ensemble learning method. In comparison to using a single decision tree, tree-based ensemble approaches integrate numerous decision trees to generate higher predicting performance. Bagging is a technique for growing ensemble decision trees in which a random subset of the examples in the training set is chosen to create

each tree [8].

Random Forest is a variation on bagging in that it grows trees using several randomly selected subsets of the input feature sets rather than the complete feature set at once. As a result, many random decision trees are built utilizing randomly selected subsets of training data and feature sets. Better variance-bias trade-offs increase prediction, and the technique is intrinsically resistant to overfitting. The number of decision trees in the forest and the tree depth examined by each tree when splitting a node are both hyperparameters in a random forest. The mean predicted class probabilities of the trees in the forest are used to compute the predicted class probabilities of an input sample. These class probabilities are used to find the best perovskite oxide candidates in this study.

For formability, the maximum tree depth was set at 22 and for stability, it was set at 23. In both classification models, the number of estimators or trees was 28. The hyper-parameters were calculated using a 5-fold cross-validation method, with the goal of maximizing accuracy while lowering the standard deviation on unknown data. A subset of 80 percent of the training data was used for cross-validation. The accuracy for formability was 0.897 while the accuracy for stability was 0.9265.

2. **Extra Trees:** The accuracy for formability was 0.910 while the accuracy for stability was 0.9265.

3. **Gradient Boosting:** The mean accuracy for formability was 0.897 while the accuracy for stability was 0.9193.

4. **Support Vector Machine:** The accuracy for formability was 0.807 while the accuracy for stability was 0.7190.

5. **Naive Bayes:** The accuracy for formability was 0.817 while the accuracy for stability was 0.7536.

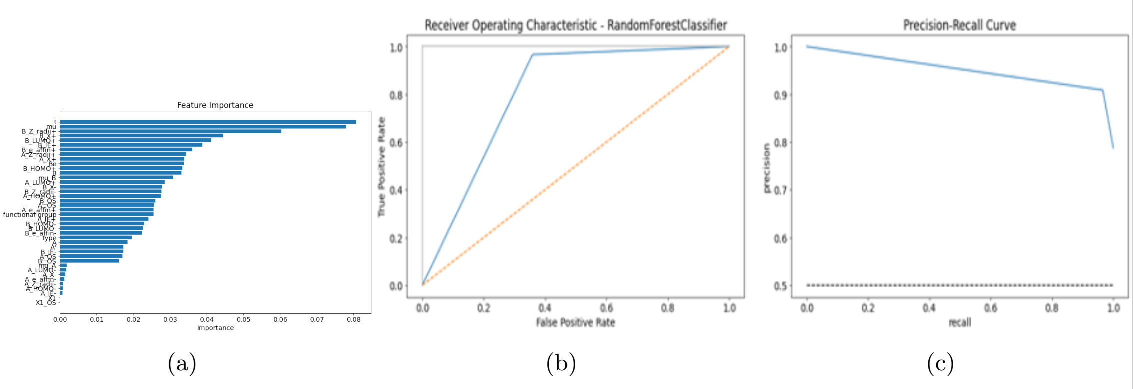
!h

3 Results and Discussion

The 20 features and the related training data sets D_F and D_S were used to create the two RFC models for formability and stability prediction stated above. Both models were also put to the test with random characteristics in order to determine their resilience and determine the genuine relevance values for the features. These random traits ranked near the bottom of the relative feature relevance rankings, as expected. The models were then put to the test on the test data sets that were constructed with the 80/20 shuffle split indicated before. Figures 4 and 8 illustrate the formability and cubic stability classification models' comprehensive feature importances and performance metrics, respectively.

3.1 Perovskite Formability Classification:

The feature importances for the perovskite formability classification problem are shown in Figure 4a. The classical geometric tolerance factor (t) and octahedral factor (μ) are heavily featured on this list, which is anticipated. Furthermore, a variety of B-site symmetric properties such as the Zunger pseudopotential radius (Z radius), electronegativity (X), and LUMO are the most crucial



(a) (b) (c)

Figure 4: Random forest classification results for perovskite formability. (a) Feature importance plot for all the features with non-zero values, (b) receiver operating characteristic (ROC) curves, and (c) precision-recall curves of the cross-validated random forest classification on test data.

Table 1: Comparison of Classification Accuracy, Precision, Recall and F1 Score among five classifiers with 20 features.

Model	Random Forest	Extra Trees	Gradient Boosting	Support Vector	Naive Bayes
Accuracy	0.897	0.910	0.897	0.90	0.90
Precision	0.89	0.91	0.89	0.89	0.89
Recall	0.89	0.91	0.89	0.89	0.89
F1 Score	0.88	0.91	0.89	0.89	0.89

in distinguishing between compounds that form perovskites and those that do not.

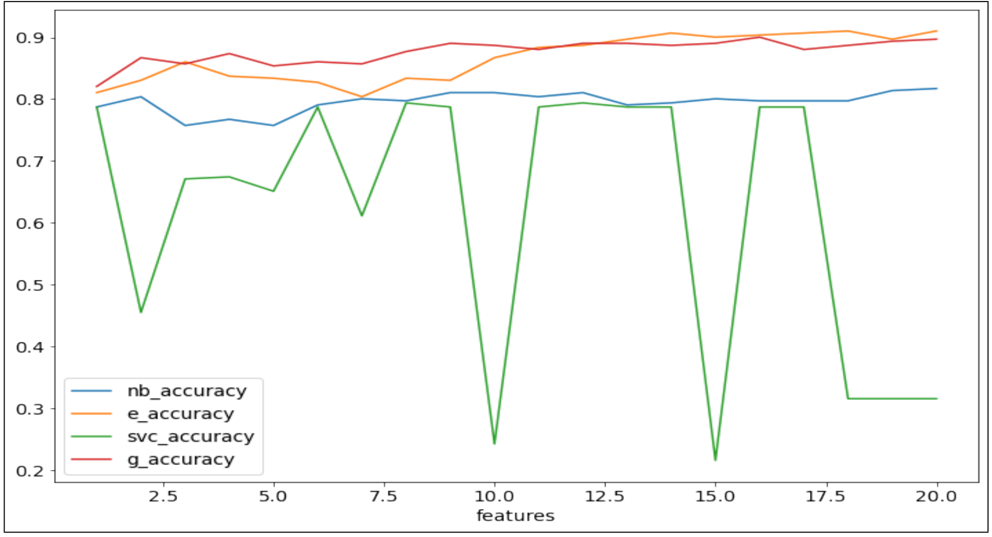
The cross-validated random forest classification's receiver operating characteristic (ROC) curves for the test data are shown in Figure 4b. It shows how a binary classifier system performs while the discrimination threshold is changed. On the y axis is the genuine positive rate, and on the x axis is the false positive rate. The ideal point on the curve is at the top left corner, where the true positive rate is 1 and the false positive rate is 0; as a result, the larger the area under the curve (AUC), the better the classifier's performance.With an AUC of 0.87, the classifier performs admirably.

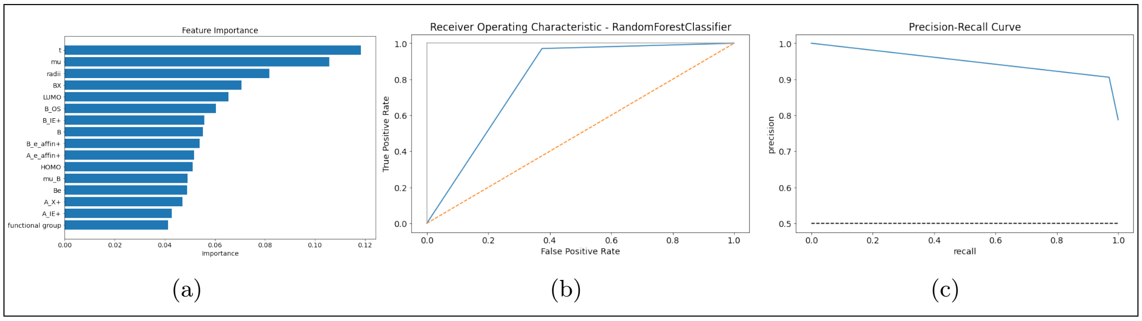
The precision and recall curves for the classifier are shown in Figure 4c. A precision-recall (PR) curve is a plot of the precision rate on the y axis and the recall rate on the x axis for various threshold values, similar to the ROC curve. As a result, a model with flawless prediction abilities is shown as a point at a coordinate of (1,1). A skilled model is represented as a curve that converges to a coordinate of (1,1).

We tested five classifiers, including: Gradient Boosting, support vector machines, Extra Trees, Naive Bayesian and Random Forest.The F1 score from cross-validation was used to optimize the parameters of the five models. Table 1 presents the accuracy, precision, recall, and F1 score of these five models.

Further, the classification accuracy of these five classifiers against the number of features is graphed for comparison purposes. The results obtained were graphed and is depicted in figure 5. As can be seen, extra trees classifier provides the highest accuracy with less than 20 features.

Figure 5: The classification accuracy of test dataset and the number of features for five classifiers





(a) (b) (c)

Figure 6: Random forest classification results for perovskite formability with newly generated features. (a) Feature importance plot for all the features with non-zero values, (b) receiver operating characteristic (ROC) curves, and (c) precision-recall curves of the cross-validated random forest classification on test data.

3.1.1 Generating New Features

We manually produced new features and used our classification models to forecast the accuracy of the new feature set as a further test of our model validation. The new features were produced by combining the features with similar chemical properties and almost same importance. Four new features BX,HOMO,LUMO and radii are generated by combining their A-site and B-site ions. The combination is as

$$BX = B\_X^{+} + B\_X^{-}$$

$$HOMO = A\_HOMO^{+} + B\_HOMO^{+}$$

$$LUMO = B\_LUMO^{+} + A\_LUMO^{+}$$

$$radii = A\_Z\_radii^{+} + B\_Z\_radii^{+}$$

We add these new descriptors to the initial feature set and rerun the feature selection to ensure their dependability.This features list is used to build the Classification models using the five classifiers defined above in section 2.2.Figure 6 depicts the RFC model for formability predictions and detailed feature importance.

The feature importances for the perovskite formability classification problem with the new dataset are shown in Figure 6a. The classical geometric tolerance factor (t) and octahedral factor (μ) are heavily featured on this list as well, which is typical. Furthermore, a variety of B-site symmetric properties such as the Zunger pseudopotential radius (Z radius), electronegativity (X), and LUMO are the most crucial in distinguishing between compounds that form perovskites and those that do not. Therefore, the importance of features is almost similar to the old feature set. This build the curiosity to know about the classification accuracy, precision, recall and area under the curve values. Table 2 summarizes the performance of these five models for accuracy, precision, recall and F1 score.

Table 2: Comparison of Classification Accuracy, Precision, Recall and F1 Score among five classifiers with reduced feature set-16 features.

Model	**Random Forest**	**Extra Trees**	**Gradient Boosting**	**Support Vector**	**Naive Bayes**
Accuracy	0.897	0.903	0.897	0.81	0.81
Precision	0.89	0.90	0.89	0.85	0.81
Recall	0.90	0.90	0.89	0.81	0.82
F1 Score	0.80	0.90	0.88	0.74	0.81

3.1.2 Principal Component Analysis

Principal Component Analysis, or PCA, is a dimensionality-reduction approach for reducing the dimensionality of large data sets by transforming a large collection of variables into a smaller one that retains the majority of the information in the large set.

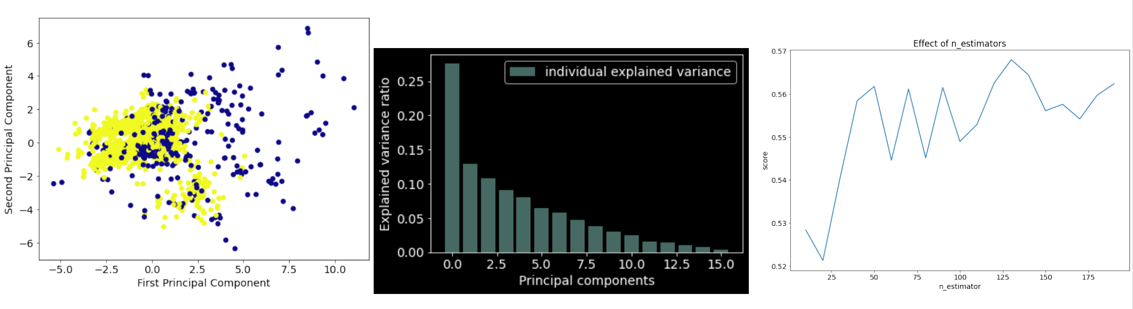
Naturally, reducing the number of variables in a data set reduces accuracy; nevertheless, the answer to dimensionality reduction is to exchange some accuracy for simplicity. Because smaller data sets are easier to study and display, and because machine learning algorithms can analyze data more easily and quickly without having to deal with superfluous factors. The PCA for the formability dataset is shown in figure 7.

3.2 Perovskite Stability Classification:

Figure 8 shows similar results for the perovskite stability categorization. Figure 8a shows the feature importances, which differ significantly from those reported in the formability classification. The tolerance factor t is the only geometric property that matters, but the B-site symmetric features of HOMO, ionization energy, LUMO, and pseudo-potential radius are also crucial stability markers.The ROC curves and precision-recall curves for the stability classification model are shown in Figure 8b-c.

Figure 8c shows that, with a PR-AUC of 0.922, the classifier is suitably close to the ideal (1,1) position on the plot with a threshold of 0.5, indicating a near-perfect model. As a result, all measures show that the classifier is reliable and capable of distinguishing between stable and unstable perovskites. The huge training data sets that were painstakingly generated, assuring the inclusion of the largest possible number of chemistries, are credited with the exceptional performance of both the perovskite formability and thermodynamic stability classification models.

The graphical results of the stability classification using five different classifiers is depicted in 9



(a) (b) (c)

Figure 7: Principal Component analysis.

Table 3: Comparison of Classification Accuracy, Precision, Recall and F1 Score among five classifiers for Pervoskite stability.

Model	**Random Forest**	**Extra Trees**	**Gradient Boosting**	**Support Vector**	**Naive Bayes**
Accuracy	0.92	0.93	0.92	0.92	0.89
Precision	0.93	0.92	0.91	0.91	0.88
Recall	0.93	0.92	0.91	0.91	0.85
F1 Score	0.92	0.92	0.91	0.91	0.81

3.2.1 Generating New Features

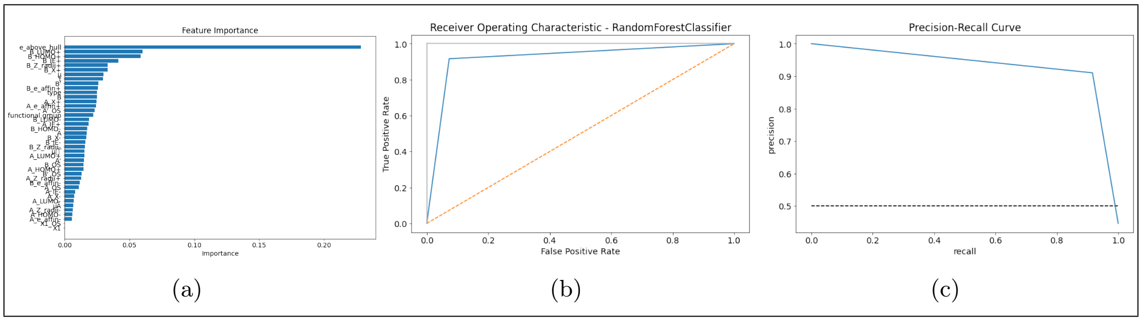
Similar to Formability dataset, new features were created in Stability dataset manually. The resulting dataset was tested for accuracy and the importance of each feature generated using ExtraTrees classifier, Random Forest, Naive bayes, Gradient Boosting and Support Vector. The feature importance plot for the thermodynamic stability for the new dataset is shown in figure 10.

The benefits would be:

1. Reduces Overfitting: Because there is less redundant data, there is less opportunity to make decisions based on noise.

2. Improves Accuracy: Less misleading data means better modeling accuracy.

3. Reduces Training Time: With less data, algorithms can train faster.



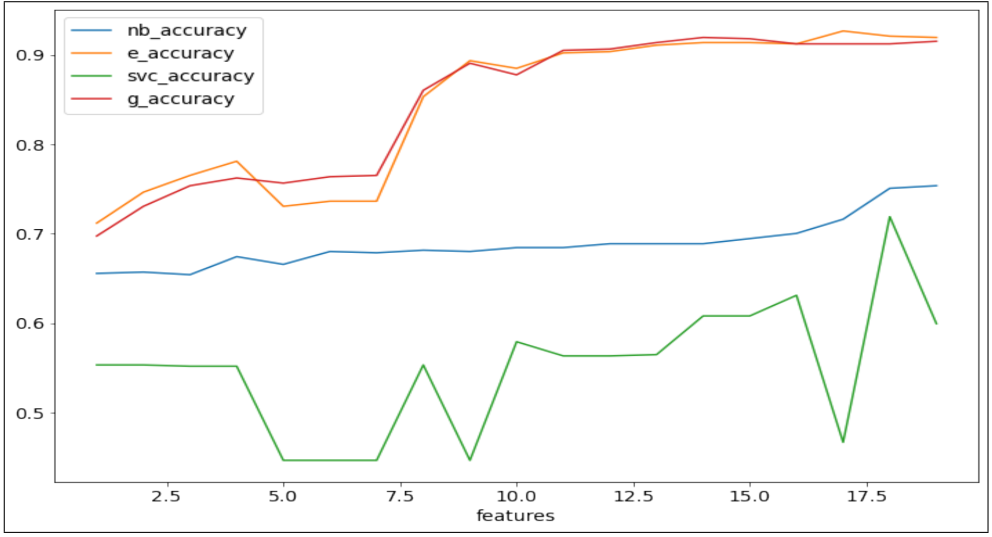
(a)

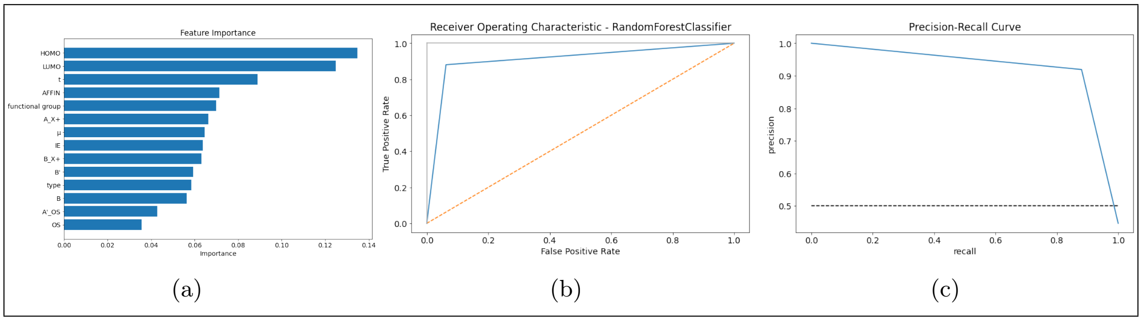
(b)

(c)

Figure 8: Random forest classification results for cubis stability. (a) Feature importance plot for all the features with non-zero values, (b) receiver operating characteristic (ROC) curves, and (c) precision-recall curves of the cross-validated random forest classification on test data.

Figure 9: The classification accuracy of test dataset and the number of features for five classifiers





(a)

(b)

(c)

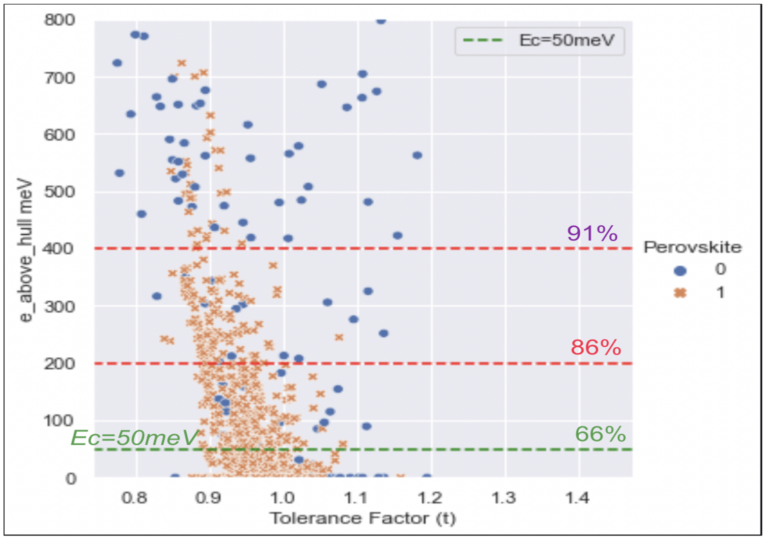
Figure 10: Random forest classification results for thermodynamic stability with newly generated features. (a) Feature importance plot for all the features with non-zero values, (b) receiver operating characteristic (ROC) curves, and (c) precision-recall curves of the cross-validated random forest classification on test data.

3.3 Comparison of Perovskite Formability and Cubic Stability

It is unknown whether a formable perovskite is necessarily thermodynamically stable and vice versa. Is it necessary to have both formability and thermodynamic stability to ensure a composition's viability as a perovskite candidate, or is one a more robust metric than the other? The intersection of the compounds in the formability and stability data sets was extracted to form a new data set $D_{F\bigcap S}$, which includes chemistries that are both experimentally known to form a perovskite or not and have energies above hull from DFT.Nonperovskites account for approximately 10% of these, with the remainder (1067) being perovskites. Using an energy greater than the hull threshold of 50 meV/atom, 816 of the 1237 are classified as stable, while the remaining 421 are classified as unstable. It should be noted that we are only comparing chemistries here. We assume cubic symmetry for the DFT calculations; however, many of the known perovskites have been experimentally synthesized at lower non-cubic symmetries. Further, that we fixed a 50 meV/atom stability threshold for our stability data set to account for stability at lower symmetries.

Figure 11 depicts the energy above the hull (Eh) in meV versus the tolerance factor (t) for this data set $D_{F\cap S}$ Nonperovskite compounds are indicated in blue, while perovskite compounds are indicated in orange. Close inspection reveals that, for an Ec threshold of 50 meV/atom, 56 percent of the compounds in the data set are perovskites and thermodynamically stable, while 10 percent are not perovskites and are not stable, implying that there is agreement between the formability and stability metrics for 66 percent of the data set.

Figure 11: Energy above hull (ΔEh) in meV is plotted against the tolerance factor (t) for the data set $D_{F\cap S}$ which is the intersection of the perovskite formability and stability data sets (D_F and D_S).



3.4 Ensemble Methods

Ensemble methods are widely regarded as the cutting-edge solution to many machine learning problems. By training multiple models and combining their predictions, such methods improve the predictive performance of a single model. Ensemble methods mimic human nature by soliciting multiple opinions before making a major decision. The main idea behind such methods is to weigh and combine several individual models in order to improve predictive performance [7].

The application of Ensemble methods to the given dataset improved the predictive performance to 96.6%.

4 Conclusion

To summarize, a systematic computational screening strategy for identifying novel single and double oxide perovskites is presented, which efficiently explores a large fraction of the double perovskite chemical space. To predict the formability and thermodynamic stability of single and double oxide perovskites, ML classification models based on random forests and Extra Trees were developed. The ML models were trained to be as accurate as possible while minimizing variance. To begin, a large number of elemental features were considered by researchers using domain knowledge, intuition, and prior work. These were eventually reduced to 28 features, and then to 16 features in the second step of generating new features. The perovskite formability model was trained using a data set of experimentally known perovskites and nonperovskites obtained from the literature, referred to as D_F.

The thermodynamic cubic stability metric was discovered to be more conservative than perovskite formability, possibly due to the $E_c = 50meV$ threshold.The ML approach employed in this produced similar results with less features.

References

[1] Prasanna V Balachandran et al. "Predictions of new AB O 3 perovskite compounds by combining machine learning and density functional theory". In: *Physical Review Materials* 2.4 (2018), p. 043802.

[2] Pierre Geurts, Damien Ernst, and Louis Wehenkel. "Extremely randomized trees". In: *Machine learning* 63.1 (2006), pp. 3–42.

[3] Isabelle Guyon et al. "Gene selection for cancer classification using support vector machines". In: *Machine learning* 46.1 (2002), pp. 389–422.

[4] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression.* Vol. 398. John Wiley & Sons, 2013.

[5] Fabian Pedregosa et al. "Scikit-learn: Machine learning in Python". In: *the Journal of machine Learning research* 12 (2011), pp. 2825–2830.

[6] Ghanshyam Pilania et al. "Machine learning bandgaps of double perovskites". In: *Scientific reports* 6.1 (2016), pp. 1–10.

[7] Omer Sagi and Lior Rokach. "Ensemble learning: A survey". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8.4 (2018), e1249.

[8] Erwan Scornet, Gérard Biau, and Jean-Philippe Vert. "Consistency of random forests". In: *The Annals of Statistics* 43.4 (2015), pp. 1716–1741.

[9] Qingde Sun and Wan-Jian Yin. "Thermodynamic stability trend of cubic perovskites". In: *Journal of the American Chemical Society* 139.42 (2017), pp. 14905–14908.

[10] Johan AK Suykens et al. *Least squares support vector machines.* World scientific, 2002.

[11] Anjana Talapatra et al. "A Machine Learning Approach for the Prediction of Formability and Thermodynamic Stability of Single and Double Perovskite Oxides". In: *Chemistry of Materials* 33.3 (2021), pp. 845–858.

[12] Bayya Yegnanarayana. *Artificial neural networks*. PHI Learning Pvt. Ltd., 2009.

[13] Guo-Xu Zhang et al. “Performance of various density-functional approximations for cohesive properties of 64 bulk solids”. In: *New Journal of Physics* 20.6 (2018), p. 063020.