Supporting Information

Identification of high-reliability regions of machine learning predictions based on materials chemistry

Evan M. Askenazi[1], Emanuel A. Lazar[2], and Ilya Grinberg[1*]

[1]Department of Chemistry, Bar-Ilan University, Ramat Gan 52900,

Israel

[2]Department of Mathematics, Bar-Ilan University, Ramat Gan 52900,

Israel

[2] Email: ilya.grinberg@biu.ac.il

S1. Convex hull construction

As a first approximation of for identifying a high-reliability region in feature space, we take the points with sufficiently low prediction error, for example the 5% of data points with the lowest error, for the convex hull construction. However, as shown for the ABO_3 lattice parameter data set in Fig. S1, this approach does not provide good separation between the high-reliability region containing only well-predicted systems and the rest of the feature space that contains both well- and poorly-predicted systems. As can be seen in Fig. S1a,b the error distributions for the systems inside the constructed CH and for the systems outside the constructed CH are the same for both the training set (80% of the systems included in CH construction) and the test set (20% of the systems not included in the CH construction. Thus, no enhancement in reliability is obtained inside the CH. Similarly, the error distributions as a function of the distance from the CH boundaries for the training and test set show that high prediction errors are often obtained inside the CH. Thus, it is clear that a method that constructs the CH while explicitly excluding the poorly predicted systems is necessary.

Therefore, we introduce a cost function for including systems with higher error and choose the points for CH construction such that the cost function is minimized as described by Eqs. 1-3 in the main text. The minimization of the introduced cost function and the MAE for the lattice parameter data set during the CH construction procedure as a function of iteration is shown in Fig. S2. It can be seen that the optimal convex hull has the cost function below 0 and the MAE is approximately 0.014 Å.

S2. Analysis of charge balance and tolerance factors of systems in the high-reliability CH region

To identify the common properties of the systems in the perovskite lattice parameter dataset identified to be located in the high-reliability region in feature space, we plot the total charge and tolerance factor distributions of the systems inside and outside the CH boundary in Fig. S3a,b. It can be seen from figure S3a that the systems inside the CH boundary are close to charge balanced (with total cation charge of 5or 6) with only a few systems with larger (±2) deviations from charge balance. By contrast, the systems outside the CH boundary show a much broader distribution of total cation charge. Fig. S3b shows that most of the systems inside the CH boundary have tolerance factors in the range of 0.7-0.95, while the systems outside the CH boundary have a wide range of tolerance factors.

S3. Analytical models and previous ML models for prediction of ABO_3 lattice parameters

Prediction of perovskite lattice parameters has been studied both using analytical (33,41-42) and statistical and artificial intelligence based methods (43-48). As shown in Fig. S4, the analytical methods capture the rough trend of the lattice parameter changes between different systems but show large errors with MAE of approximately 0.14 Å. Artificial neural networks (ANN), Support Vector Machines (SVR), Convolutional Neural Networks (CNN) and General Regression Neural Networks have been the primary techniques for this analysis. The Crystal Graph Convolutional Neural Network (CGCNN) has also been used. In the CGCNN method, a graph network consisting of nodes containing ionic information is used to predict various structure properties. Hirshfeld surfaces, due to their use in the analysis of the structure of atoms bound in molecules (49-52) have been used in conjunction with AI methods as well. CNNs have been combined with these surfaces (30) which provide molecular shapes for compounds by summing the average electron densities of the atoms within the compound and within a reasonable short distance away from it. To complement the Hirshfeld surface, the contact distances from each point on the Hirshfeld surface to the closest atoms inside and outside the surface can be converted to a set of two dimensional fingerprint plots. This leads to a set of input data on which the CNN can then be trained.

A survey of the performance of these methods shows that the mean absolute error (MAE) of ML methods is between 0.026 and 0.04 Å while the MAE of analytical methods is approximately 0.14 Å.

S4. Our XGboost model for prediction of ABO_3 lattice parameters

To obtain a high-quality ML model, in this study, we use an ensemble-based XGboost machine learning method trained on a set of input features based on the key properties of the individual A and B site ions of cubic oxides. We use element labels, ionic radii and valence charge values, electronegativity and the periodic table block of the A and B site ions as the regression features. Since ionic size depend on the valence of the ion and its coordination, we use ionic radii for the two most common ionic valences in the 12-fold coordination for the A-site and in the 6-fold coordination for the B-site. If only one ionic valence/charge value is possible, only that value is

used. The valence charge numbers and the periodic table block of the ions (s,p,d) are used as numerical label classifiers. The set of oxides used for model training is obtained from OQMD (34) and includes 5250 ABO_3 systems. For all data sets used in model fitting, the data is split such that 90% are used for training and the remaining 10% form the test set. Additionally, for the 5250 oxides, the testing data sets are randomly split into 10 subsets such that each oxide is in only one of the ten subsets. For each subset, an Xgboost model is trained, with the aforementioned hyperparameter optimization through 10 fold cross validation, on the other 9 subsets and tested on the selected subset. A total of ten models are thus trained, one for each subset. In this manner, a lattice parameter prediction for each of the 5250 oxides is made.

The ability of the Xgboost model to predict the lattice parameters for oxides under various conditions is shown in Fig. S5a. As shown in Table 1, when trained on the full ABO_3 data set and used for experimentally viable oxides, the model outperforms the analytical formulas and obtains a substantially smaller MAE than a recent support vector based model and deep learning model, despite using a small set of features. For mean square error (MSE), an order of magnitude reduction in the error is obtained. Displaying the residuals, as shown in Fig. S5b for a test set of computational oxides, illustrates the models' effectiveness in fitting for experimentally feasible and unfeasible oxides. Its MAE is 0.025 Å and its MSE is 0.0015 $Å^2$. Its overall error is lower than previous attempts at AI based lattice parameter predictions, albeit with greater likelihood of outliers. More than 95% of the lattice parameter predictions have an error less than 0.1 Å.

The feature importance and plot of error relative to number of features are presented in Figs. S5c and S5d, and the list of the most important features for each number of features is presented in Table 2. Feature importance was computed using Gini selection while the optimal features for each specific number of features were determine using backwards recursive feature selection. The greater importance of the B site properties relative to those of the A site is evident from the feature importance plot when all features are utilized in the selection process. This may be due to the B site features having substantially greater variance than the A site features. Alternatively, this may be due because the B-O bonds are stronger than A-O bonds and therefore are more important for determining the ABO_3 lattice parameter. However, for optimizing the feature set for a given number of features, A and B site properties are both necessary.

The first major drop in the error is observed when two features are used. In this case, the optimal feature set consists of the B site label and the smaller value for the A-site ionic radius. With three features, the B-site label, the A site electronegativity, and the smaller value of the two values of the A-site ionic radius are included. When the desired number of features is set to five, the second major drop in the error is recorded, and the higher of the two valence charges of the A site is added as a feature. Therefore, the usefulness of features in both the A site and B site is evident in terms of obtaining the optimal fit. It is clear that no more than nine features are necessary to reach the limit of ML model accuracy. The excellent performance of our method compared to the previous ML studies, including the deep learning model based on the Hirshfield surface indicates that our feature set is well-suited for modeling the lattice parameters of perovskite oxides, and provides another example of the crucial role of feature selection in the application of ML to small materials science data sets.

Comparison of the effect of the choice of training data for the prediction of the systems in the high-reliability region of ABO_3 lattice parameter data set feature space

We now investigate whether the ABO_3 systems lying outside this region of reliability contribute to the prediction accuracy of the data points in the reliability region of physically reasonable ABO_3 oxides. Intuitively, based on the illustration shown in Figure 1, we expect that some systems outside the limits of the region of reliability must be included for the accurate interpolation of the ABO_3 systems close to the edge of the region of reliability. First, in Table S3 we present the MAE, stdError and MaxAE metrics for the full data set, the subset of perovskite oxide systems in the region in feature space enclosed by CH, the manually selected subset of systems chosen based on meeting the experimental viability criteria ($0.9<t<1.1$ and charge-balance) and the joint subset containing both the CH and experimentally viable system. In all cases the XGboost model was trained on the full data set. A reduction in all error metrics is observed going from the full data set to the high-reliability subsets.

Then, we also train three Xgboost models using only the systems in the CH subset, the experimentally subset or the joint subset, respectively, and then test for the accuracy of model prediction for each of the full set and the three data set, thus obtaining a 4x4 matrix of MAE results for lattice parameter prediction as shown in Table S4. Examination of the MAE data presented in Table S4 shows the following. First, the high-reliability regions identified by the CH procedure and manually contain systems that follow different relationships between the features and the lattice parameter. This can be seen from the very high MAE obtained when training on the systems in the CH region and testing on the experimentally viable systems (0.096 Å) or vice versa (0.088 Å). Second, when the model is trained on systems from both high-reliability regions a fairly low MAE of 0.017 Å is obtained showing that when given the information about both subsets, the Xgboost model can distinguish between the two different regions in feature space and provide appropriate predictions for both subsets. As expected, training only on the systems in the high-reliability regions leads to poor prediction ability for the rest of the dataset as shown by the high MAE value of 0.061 Å obtained when training on both CH and experimentally viable systems and testing on all oxides. Third, confirming the conclusions from the trends presented in Fig. 8f and 9c-d, we find that some systems outside the high-reliability region must be included in the training for maximum accuracy. Therefore, training using the full set leads to better prediction accuracy for the experimentally viable high-reliability region (MAE of 0.018 Å) than when only the data in the experimentally viable high-reliability region are used for training (MAE of 0.022 Å). Finally, the high-reliability region identified by the CH construction procedure gives slightly smaller MAE than the manually identified high-reliability region, with MAE values of 0.018 Å and 0.015 Å, respectively, obtained when using the full oxides data set for training, and MAE values of 0.014 Å and 0.022 Å, respectively, when using both the CH and experimentally viable data subsets for training. This is not surprising due to the explicit focus of the CH construction procedure on excluding poorly predicted systems.

Relationship between bond length and bond strength

To illustrate composition-structure-property relationships as discussed in the main text, the relationship between bond length and strength for carbon-carbon, carbon nitrogen and nitrogen bonds is shown in Figure S6. It shows that strength-bond length relationship is only valid for individual classes of bonds, but cannot be used a general predictor of bond strength of all bonds and all elements.

Effects of number of input features and percentage tolerance on convex hull predictive ability

Effects of varying the number of input features, for the TCO Eg dataset, are in Fig S7. It can be observed that for 5 input features the convex hull is most effective at removing the number of poorly predicted points while keeping overall prediction error reasonably low. Effects of varying the tolerance factor, for TCO Eg, Dilute Solute and TCO Ef data set, are in Figure S8. General trends of improved ability of convex hull to remove poorly predicted points for lower tolerance factors are observed.

Convex hull data split

The 80/20 split for construction of convex hull training and test data points is performed randomly. The results of different such splits are given in S9 and S10 for the TCO E_g dataset and show similar results for all five splits.

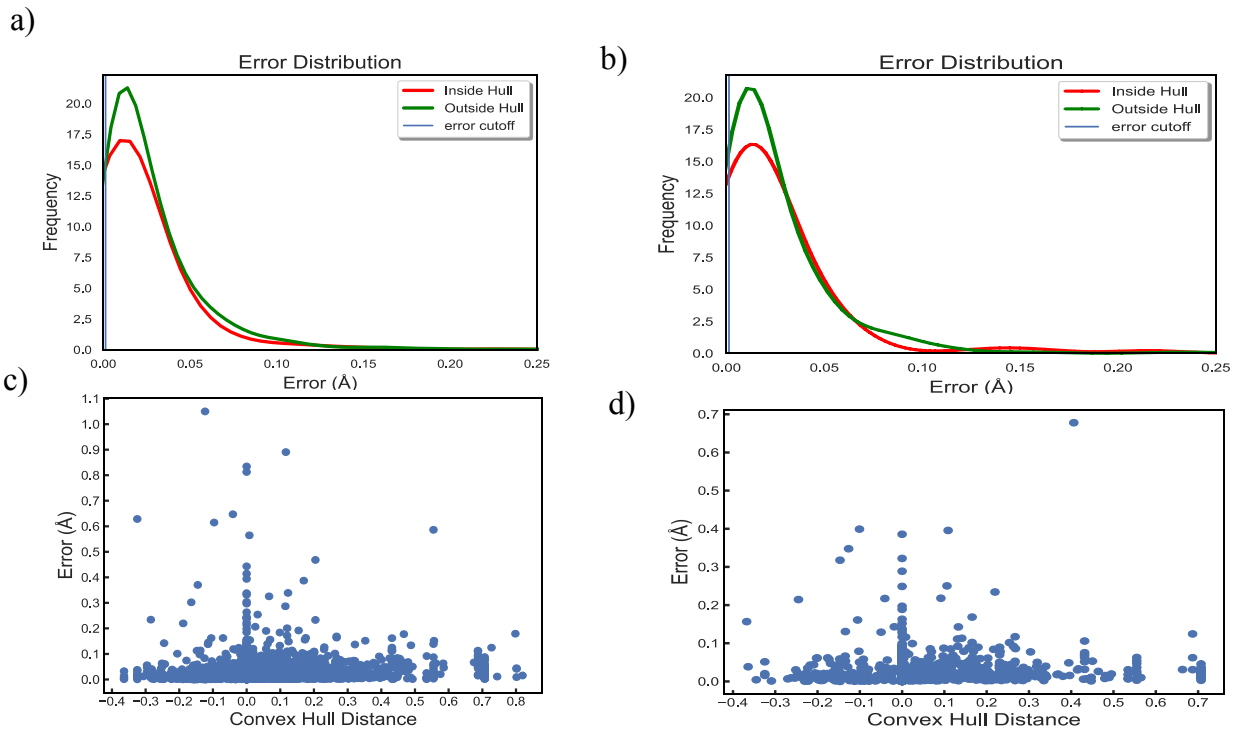

Figure S1 Results for convex hull construction using the 5% of the data points with the lowest error for the perovskite oxide lattice parameters a) Error distributions for systems inside and outside the convex hull for the training set used in CH construction b) Error distributions for systems inside and outside the convex hull for the test set not used in CH construction. c) Errors plotted versus the distance from the CH boundary for the training set used in CH construction d) Errors plotted versus the distance from the CH boundary for the test set not used in CH construction

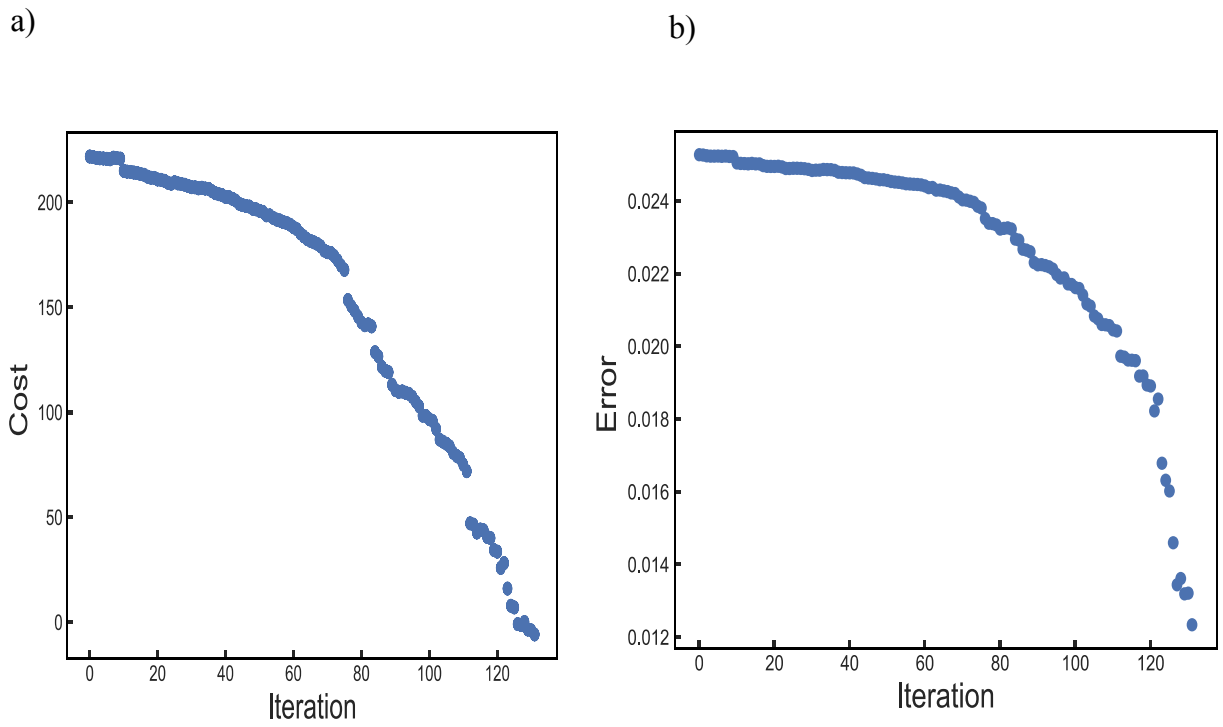

Figure S2 a) Cost function as a function of iteration b) MAE of the ABO_3 systems chosen by convex hull construction as a function of iteration

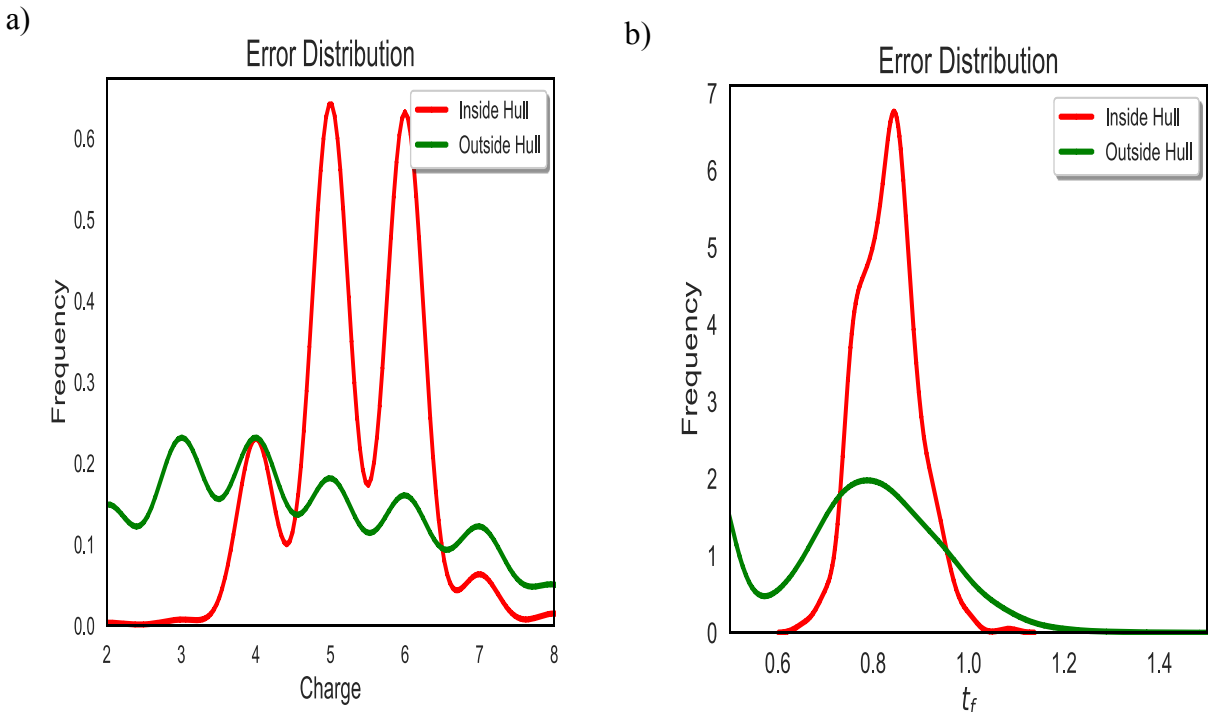

Figure S3 Distributions of the total cation charge and tolerance factor for the perovskite oxides systems inside and outside the convex constructed hull a) Distributions of the total cation charge b) distributions of the tolerance factor

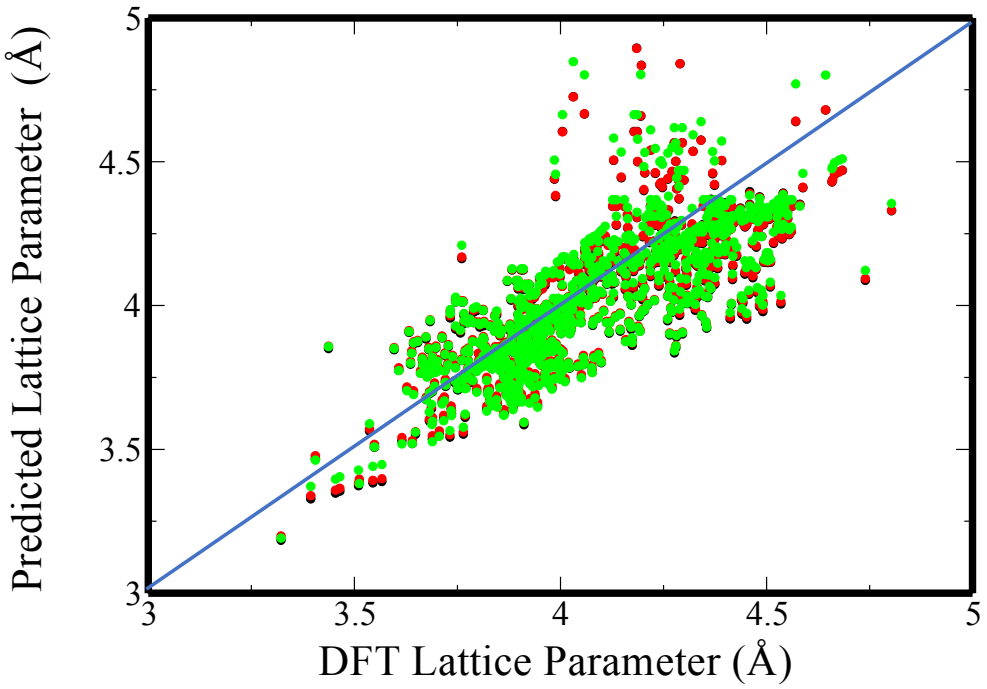

Figure S4 Predictions of analytical models for perovskite lattice parameters a) Predictions of Analytical models for ABO_3 lattice parameter compared to DFT calculate values. The data for the Sidey [20], Ubic [19] and Jiang [21] models are shown in black, red and green

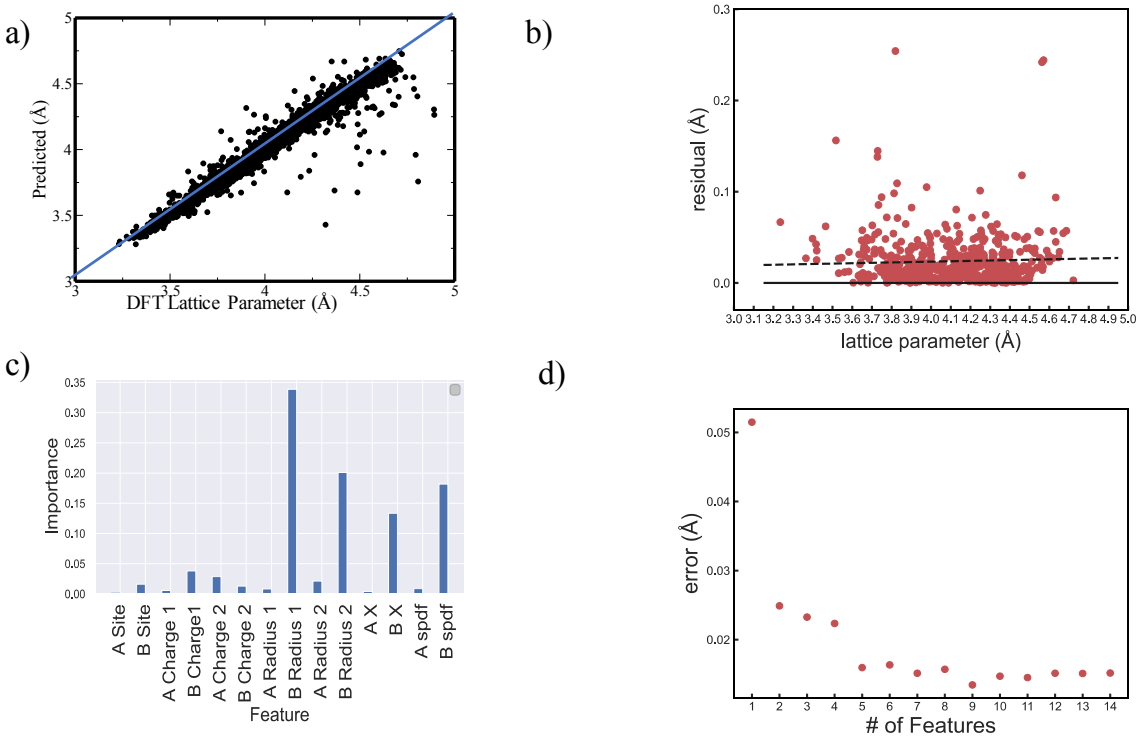

Figure S5 Characteristics of the XGboost model used for perovskite oxide lattice parameter prediction a) Predicted versus DFT lattice parameter for our XGboost ML model trained on all data points b) Residuals for our XGboost model as a function of lattice parameter c) Feature importance plot for our XGboost model d Error as a function of number of features for our XGboost model

Table S1 Errors for lattice parameter predictions on ABO_3 systems for analytical models and machine learning methods.

Method	MAE (Å)	MSE (Å^2)	stdevError (Å)	MaxAE (Å)
Sidey, Analytical, all points	0.142	0.038	0.11043	0.71078
Jiang, Analytical, all points	0.146	0.044	0.11561	0.87062
Ubic, Analytical, all points	0.140	0.037	0.10948	0.70892
Olowabi, SVRA-PSO, all points	0.029	0.00336	0.0503	1.08432
Fingerprint CNN, all points	0.037	0.00282	N/A	N/A
Xgboost, all points	0.026	0.00289	0.04690	1.04991

Table S2: Best Features for each number of optimal features

# of Features	Best Features
1	B Site Label
2	B Site Label and smaller A Radius
3	B Site Label, smaller A Radius, A X
4	B Site Label, smaller A Radius, smaller B Radius, A X
5	B Site Label, larger A Site Charge, smaller A Radius, smaller B Radius, A X
6	B Site Label, larger A Site Charge, smaller B Radius, smaller A radius, A X, B X
7	B Site Label, smaller B Site Charge, larger A Site Charge, smaller B Radius value, smaller A Radius , A X, B X
8	B Site Label, smaller B Site Charge, larger A Site Charge, smaller B Radius , smaller A Radius , A X, B X and A Site s,p,d,f block
9	B Site Label, smaller B Site Charge, larger A Site Charge, smaller and larger B radius, smaller A Radius, A X, B X and A Site s,p,d,f block

Table S3: ML accuracy metrics for the predicted ABO_3 lattice parameter for various data sets for the XGboost model trained on the full set of ABO_3 systems.

Set of Data Points	MAE (Å)	stdError (Å)	MaxAE (Å)

All Oxide Points	0.026	0.0469	1.0499
Manually Chosen Experimentally Viable Points	0.018	0.0187	0.1135
Convex Hull Defined Points	0.015	0.0179	0.1159
Manually Chosen and Convex Hull Defined Points	0.017	0.0184	0.1159

Table S4: ML accuracy metrics for the full, experimentally viable, and convex-hull data set for the XGboost models trained on the full, experimentally viable, and convex-hull data sets of ABO3 systems.

	All Oxides Test (Å)	Experimentally viable Test (Å)	Convex Hull Defined Test (Å)	Human and Convex Hull Defined Test (Å)
All Oxides Train	0.026	0.018	0.015	0.017
Experimentally viable Train	0.11	0.021	**0.088**	0.062
Convex Hull Defined Train	0.074	**0.096**	0.018	0.031
Human and Convex Hull Defined Train	0.061	0.022	0.014	0.017

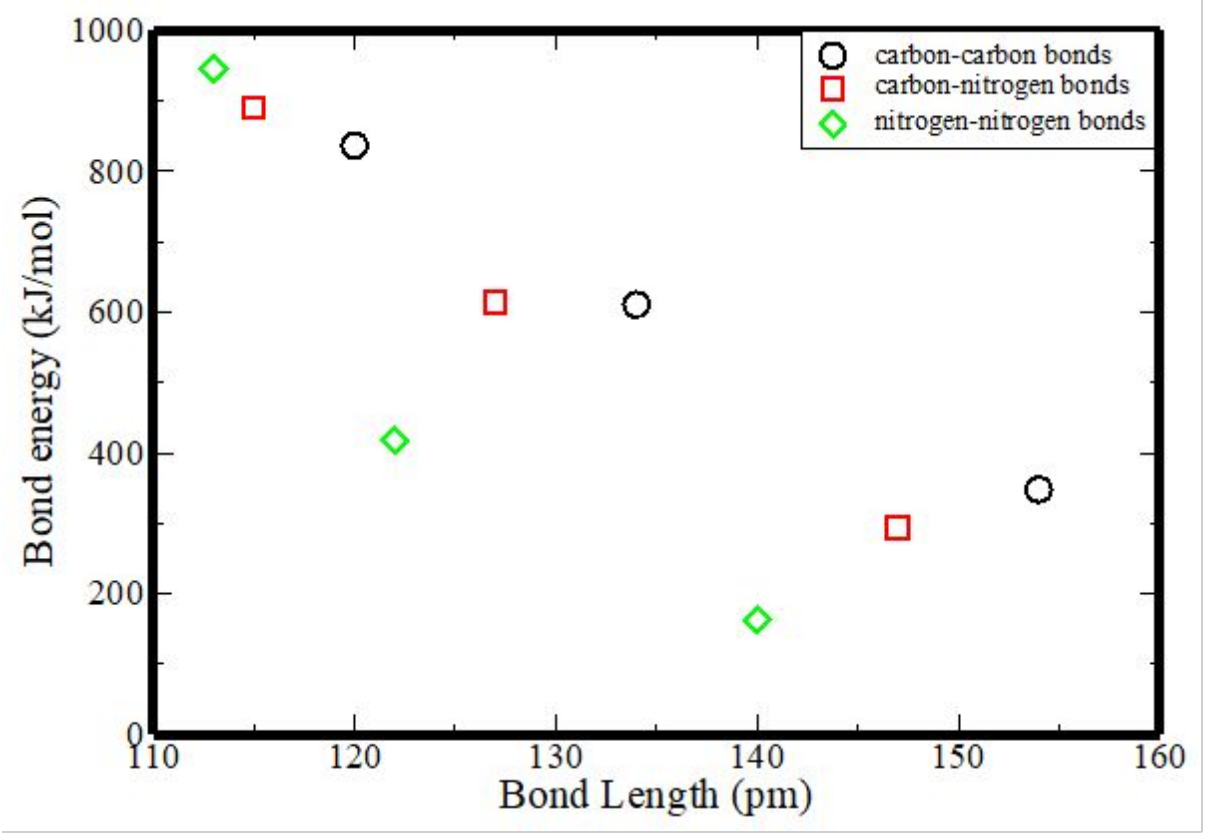

Figure S6. Relationship between bond energy and bond length.

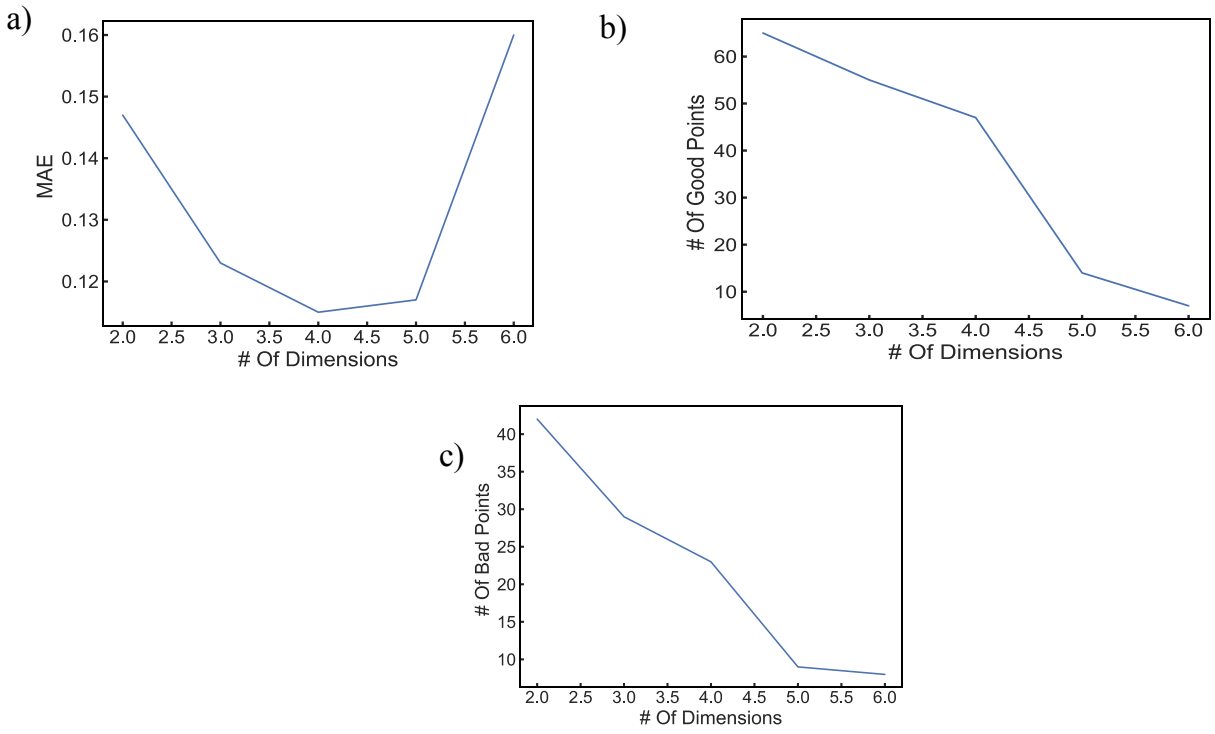

Figure S7 Characterization of the CH accuracy for different number of features (dimensions) used in CH construction using the TCO E_g dataset as an example. (a) MAE of the validation data points (not used in CH construction) inside the CH as a function of the number of features used in CH construction. (b) Number of well-predicted validation data points (not used in CH construction) inside the CH as a function of the number of features used in CH construction. (b) Number of poorly-predicted validation data points (not used in CH construction) inside the CH as a function of the number of features used in CH construction.

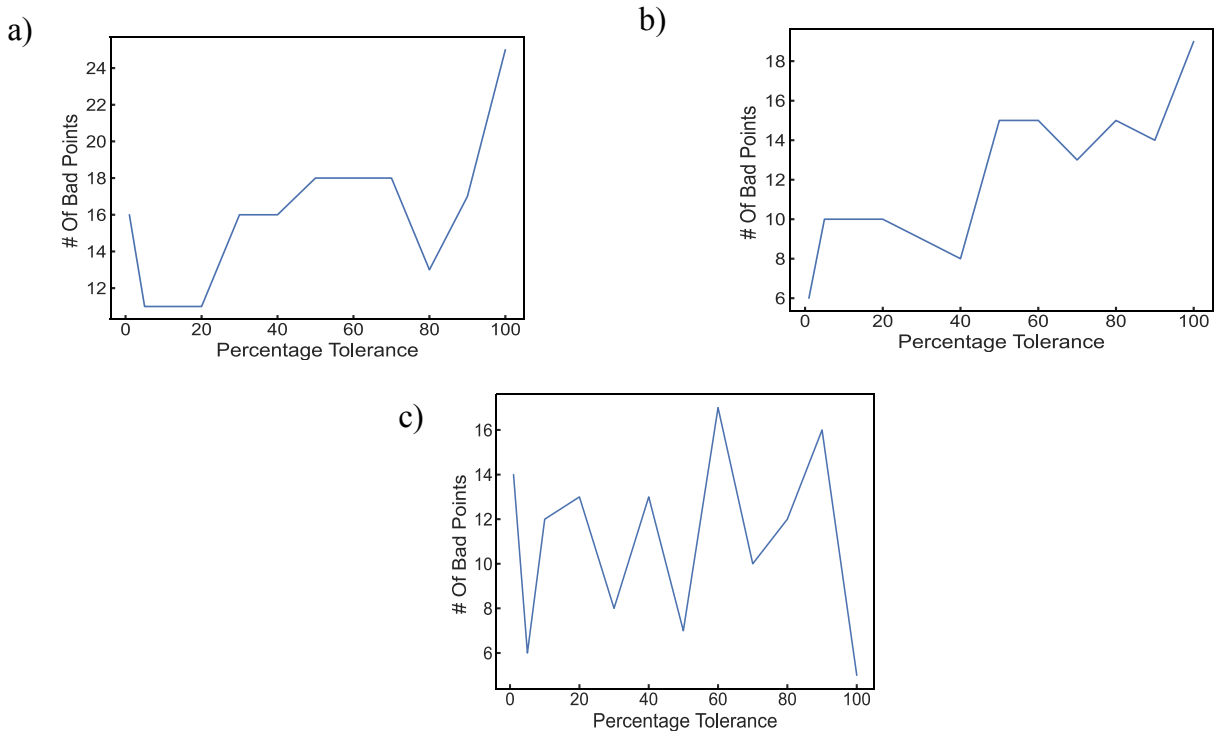

Figure S8. Number of poorly-predicted validation (not used in CH construction) data points of the validation data points inside the CH as a function of the tolerance magnitude used in CH construction for (a) TCO E_g dataset, (b) Dilute solute dataset (c) TCO E_f dataset.

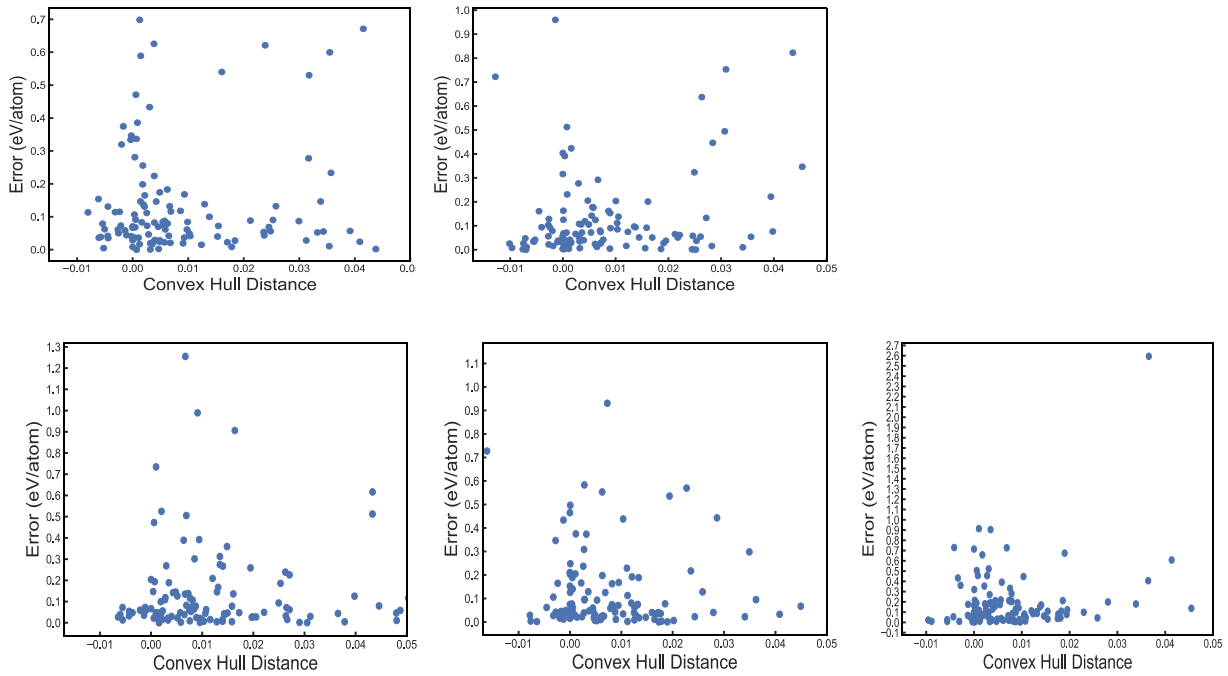

Figure S9. Error distributions for the TCO E_g data sets for the 20% of the dataset not used in CH construction for five different 80/20 splits.

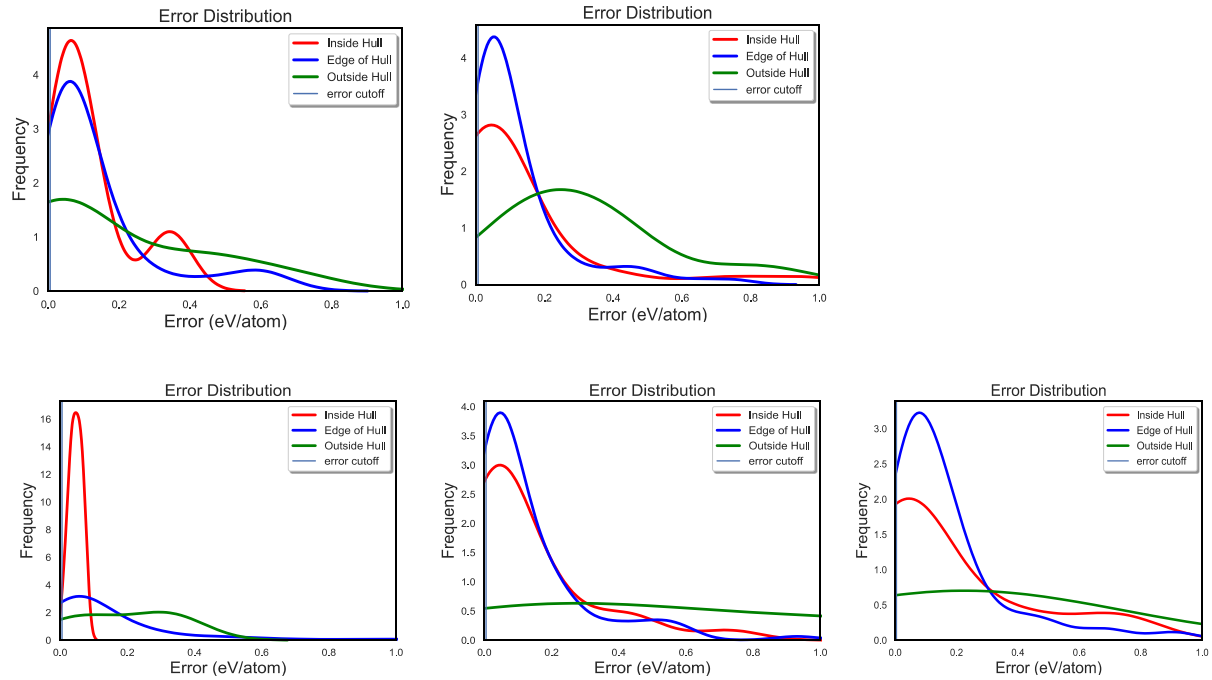

Figure S10. Error for the TCO E_g data sets for the 20% of the dataset not used in CH construction for five different 80/20 splits plotted as a function of the distance from the CH boundary.