Supplementary Information

Machine learning and Feature Importance in data-driven search for Pb-free Perovskite Solar Cell Materials

Jino Im\textsuperscript{1*}, Seongwon Lee\textsuperscript{2*}, Tae-Wook Ko\textsuperscript{2}, Hyun Woo Kim\textsuperscript{1}, YunKyong Hyon\textsuperscript{2†}, and Hyunju Chang\textsuperscript{1†}

\textsuperscript{1}Korea Research Institute of Chemical Technology (KRICT), Daejeon 34114, Republic of Korea
\textsuperscript{2}National Institute for Mathematical Sciences (NIMS), Daejeon 34047, Republic of Korea

*These authors contributed equally to this work
†Correspondence should be addressed to hjchang@kRICT.re.kr and hyon@nims.re.kr

1. Machine learning on oxide double perovskites

The dataset was initially developed by Castelli et al. for water splitting applications\textsuperscript{1}. The dataset included 1307 unique cubic oxide double perovskites, \textit{AA’BB’O\textsubscript{6}}, within a combination of various chemical elements for \textit{A}, \textit{A’}, \textit{B}, and \textit{B’} cations. In this dataset, \textit{E\textsubscript{g}} was calculated using density functional theory (DFT) with the “Gritsenko-Leeuwen-Lenthe-Baerends functionals for solid and correction” (GLLB-SC) functional, which is known to provide a reasonable estimate of actual \textit{E\textsubscript{g}} values with a low computational cost\textsuperscript{2-4}.

We utilized the 28 primary atomic features and additional structural features, such as the bond length between cations and anions. A total of 32 features were examined:

Pauling’s electronegativity ($\chi$), ionization potential (IP), highest occupied atomic level (h), lowest unoccupied atomic level (l), and s-, p- and d- valence orbital radii $r_s$, $r_p$ and $r_d$ of isolated neutral atoms of \textit{A}, \textit{A’}, \textit{B}, and \textit{B’}, and atomic distance (d) between cations and the
nearest oxygen atom.

Like the previous study by Pilania et al., we symmetrized and anti-symmetrized all features over $A$ and $A'$ (or $B$ and $B'$), denoted $f_{A+A'} := f_A + f_{A'}$ and $f_{A-A'} := |f_A - f_{A'}|$. The symmetry of the double perovskite structure was invariant under the structural transformation by exchanging $A$ and $A'$ (or $B$ and $B'$), which shared the same pool of chemical species.

Here, we compare four different machine learning algorithms, including the linear regression (LR), the $k$-nearest neighbors ($k$-NN), the support vector regression (SVR), and the gradient-boosted regression trees (GBRT) for the same dataset. See the computation details below. Among these methods, LR only captured the linear behavior between the target properties and the features, whereas $k$-NN, SVR, and GBRT performed nonlinear regression. $E_g$ prediction via LR provided the worst accuracy (the averaged root-mean-square-error (RMSE) of the test sets = 0.926 eV). On the other hand, nonlinear methods showed better accuracy. Here, we note that GBRT (RMSE of test sets = 0.270 eV) outperformed other methods (the RMSE of the test sets = 0.359 eV for SVR and 0.524 eV for $k$-NN). These results strongly indicated that $E_g$ and the features were nonlinearly related, and GBRT well captured the nonlinearity in the prediction of $E_g$. (see Figure S1)

![Figure S1. Prediction of the bandgap of the oxide double perovskites through a) linear regression (LR), b) $k$-nearest neighbors ($k$-NN), c) support vector regression (SVR), and d) gradient-boosted regression trees (GBRT).](image)
regression, b) k-nearest neighbor, c) support vector machine, and d) gradient boosted regression tree. Orange filled circles correspond to the training dataset, and blue circles correspond to the test dataset. Red solid lines indicate the reference line for a perfect fit.

**Computation details**

*Hyperparameters for ML*

The hyperparameters of each model were optimized by using a cross-validated grid search or a randomized search over the parameter settings. The root-mean-square error (RMSE) of the target variable was used as a cost function for all models.

1) Linear regression

We performed a first-order linear regression.

2) k-nearest neighbors regression

The k-nearest neighbors (k-NN) algorithm is an instance-based learning approach. The prediction of an object is the (weighted) average of the values of its k nearest neighbors. We identified an optimal number of the hyperparameter k=3 using cross-validation, and we used the inverse distance-weighted average with nearest neighbors, calculated by Euclidean distance.

3) Support vector machines

The support vector machine (SVM) is one of the most well-known machine learning algorithms. We used a version of SVM for regression, the support vector regression (SVR). We performed SVR using a radial basis function (RBF) kernel that was nonlinear, along with a penalty parameter C=80 and threshold $\epsilon=0.03$.

4) Gradient boosted regression tree

Several hyperparameters exist for this model. A parameter subsample for bagging (subsample ratio of the training instance) and colsample_bytree (subsample ratio of features) for the random forest were optimized to prevent overfitting. The regularization parameters were also optimized. The values of the hyperparameters used here are:

for the oxide double perovskite dataset, max_depth = 4, min_child_weight = 2, colsample_bytree = 0.3, subsample = 0.9, reg_alpha = 0.11, learning_rate = 0.1, and

for the halide double perovskite dataset, max_depth = 6, min_child_weight = 1,
colsample_bytree = 0.5, subsample = 0.7, reg_alpha = 0.1, learning_rate = 0.03.

References


2. A dataset of halide double perovskite.

The dataset of halide double perovskite used in machine learning is included in the excel file “Dataset_halide_doubleperovskite.csv”. Detailed information of index is following:

- distance: Distance between cations at A-, B\textsuperscript{1+}, and B\textsuperscript{3+} site, and anions at X-site
- eleneg: Electronegativity
- hoe: Highest occupied energy level
- ionenergy: Ionization energy
- luep: Lowest unoccupied energy level
- rs: radius of s-orbital
- rp: radius of p-orbital
- rd: radius of d-orbital
- ind-gap: band gap in eV
- heat_of_formation: heat of formation in eV