Deng *et al*. *Energy Mater* 2021;1:100006
DOI: 10.20517/energymater.2021.10

Energy Materials

Article

Open Access

Automated machine learning structure-composition-property relationships of perovskite materials for energy conversion and storage

Qin Deng, Bin Lin

Yangtze Delta Region Institute (HuZhou), School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Huzhou 313001, Zhejiang, China.

Correspondence to: Prof. Bin Lin, School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Huzhou 313001, Zhejiang, China. E-mail: bin@uestc.edu.cn

How to cite this article: Deng Q, Lin B. Automated machine learning structure-composition-property relationships of perovskite materials for energy conversion and storage. *Energy Mater* 2021;1:100006. http://dx.doi.org/10.20517/energymater.2021.10

Received: 12 Sep 2021 **First Decision**: 27 Sep 2021 **Revised**: 3 Oct 2021 **Accepted**: 9 Oct 2021 **Published**: 10 Oct 2021

Academic Editors: Yuping Wu, Bin Zhu **Copy Editor**: Yue-Yue Zhang **Production Editor**: Yue-Yue Zhang

Abstract

Perovskite materials are central to the fields of energy conversion and storage, especially for fuel cells. However, they are challenged by overcomplexity, coupled with a strong desire for new materials discovery at high speed and high precision. Herein, we propose a new approach involving a combination of extreme feature engineering and automated machine learning to adaptively learn the structure-composition-property relationships of perovskite oxide materials for energy conversion and storage. Structure-composition-property relationships between stability and other features of perovskites are investigated. Extreme feature engineering is used to construct a great quantity of fresh descriptors, and a crucial subset of 23 descriptors is acquired by sequential forward selection algorithm. The best descriptor $[ln(1+|r_A|)r_B^{-1}c^2\alpha^{-1}]^{-1}$ for stability of perovskites is determined with linear regression. The results demonstrate a high-efficient and non-priori-knowledge investigation of structure-composition-property relationships for perovskite materials, providing a new road to discover advanced energy materials.

Keywords: Perovskites, structure-composition-property relationships, stability, descriptors, automated machine learning

INTRODUCTION

To discover materials, the investigation of structure-composition-property relationship of inorganic materials is essential, and a huge number of material composition pose a big challenge to investigate the

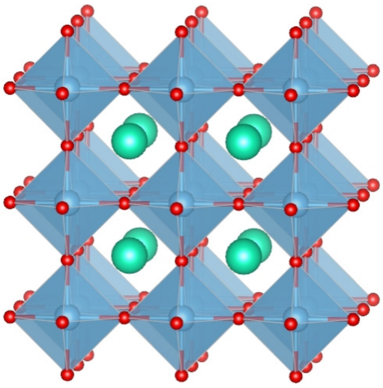


Figure 1. A-site 12-coordinated cations (green), B-site 6-coordinated cations (blue), and O-site oxygen anions in the crystal structure of perovskite oxide ABO_3 (red).

hidden structure-composition-property relationships[1,2]. It is usually supported by magnanimous lab experiments that are demanding both in terms of time and technology. Accordingly, the exploration of the structure-composition-property relationship is very difficult[3-5]. Machine learning is intensively applied in the field of advanced materials exploration and discovery for almost a decade[6-9], becoming a high-efficient approach to investigate inorganic materials[5]. However, the complicacy of machine-learning processes and the inability to comprehend models make it hard to obtain good rules for describing connections between structure, composition and property of materials, which impedes their deeper comprehension[10]. Consequently, it is particularly significant to improve the approach of exploring the structure-composition-properties relationship of inorganic materials[7,11]. So far, several approaches for discovering important descriptors have been published, such as the symbolic regressionr algorithm[12], the least absolute shrinkage and selection operator algorithm algorithm[13], and the sure independence screening and sparsifying operator (SISSO) algorithm[14]. The purpose of these approaches is to find some vital descriptors describing the target variables or some hidden mathematical formulas from the given feature space so that these vital descriptors can be used to predict the target variables[4,9,10]. Although these methods achieved good results, they need to rely on many conditions, such as a large amount of data, suitable algorithms, *etc*., which are obviously tough for material scientists who are not familiar with computer algorithms[15]. Therefore, these algorithms are extremely low efficient[14-16].

Perovskite materials are essential for energy storage and conversion, due to their excellent electrocatalytic properties[11]. The stability of perovskite compounds is the focus and challenging dimension in perovskite-based fuel cells, and is a key material property whose value may determine the use of perovskite oxides[17]. When considering numerous different A- and B-site elements [Figure 1], as well as various conventional doping ratios and combinations, the amount of perovskite components should be huge. The full compositional flexibility of perovskite structure gives it a complex set of functional properties. In addition, the flexibility poses the big challenge for predicting stability[18]. A recent research paper obtained a subset of nine important descriptors by constructing a large number of new descriptors and using recursive feature elimination method. Furthermore, the optimal descriptor of lattice constant was obtained by linear regression algorithm, and the simple linear expression of lattice constant was obtained successfully[19]. It helps to explore structure-composition relationships of materials without prior knowledge. In this work, the approach was further improved.

In this paper, the structure-composition-property connections between stability and other features of perovskite compounds was investigated via a high-effective approach of extreme feature engineering and automated machine learning[19-27]. The feature engineering approach was used to remove redundant features while generating many fresh descriptors[28]. The subset of significant descriptors was obtained

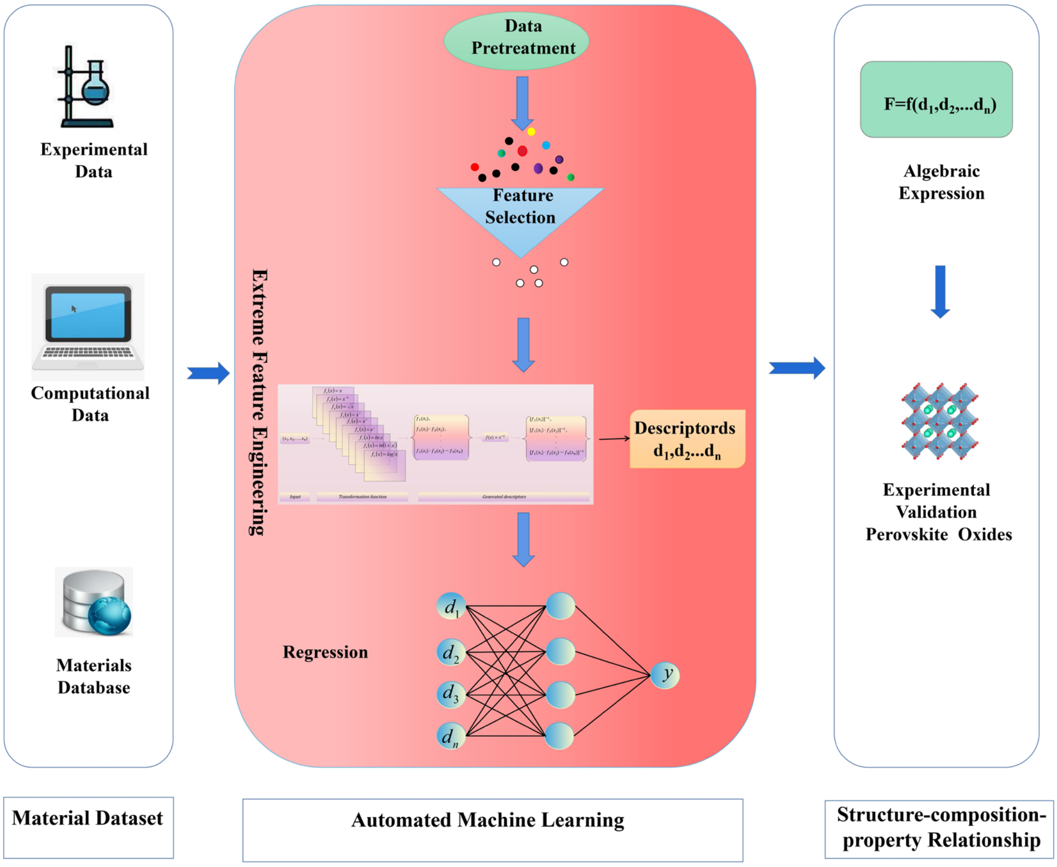


Figure 2. Whole process of adaptively learning structure-composition-property connections of ABO_3 perovskite compounds via automated machine learning.

by sequence forward selection algorithm, the best descriptor was obtained via linear regression analysis to obtain expression of stability. Instead of trying to model all the feature combinations, the sequential forward selection algorithm aborts the search by finding an optimal solution, which greatly reduces the computational effort. This new approach combining feature engineering with linear regression algorithms does not demand researchers to have an in-depth understanding of computer algorithms and does not depend on advanced knowledge or model[29]. Compared with symbolic regression algorithm and SISSO algorithm, this algorithm has obvious advantages[9]. The acquired structure-composition-property relationships will speed up the design and optimization of perovskite materials, and offer a new way for the exploration and research of inorganic materials.

EXPERIMENTAL

The whole process of adaptively learning structure-composition-property connections of ABO_3 perovskite compounds shows in Figure 2, and it contains several steps as follows:

Step 1: Collect the material dataset from different ways;
Step 2: Perform pretreatment on the material dataset;
Step 3: Extreme feature engineering is used to generate a large number of new descriptors;
Step 4: Apply the feature selection on a significant number of new descriptors, discover the subset of important descriptors, and then apply regression fitting on the subset of important descriptors. This enables the discovery of the optimal descriptor as well as the gain of the related structure-composition-property relationships.

Table 1. Explanation for ABO_3 perovskite compounds datasheet

No.	Property	Unite	Description
1	r_A	Å	Ionic radius of A site
2	r_B	Å	Ionic radius of B site
3	ΔH_f	eV/atom	Formation energy as calculated by equation of the distortion with the lowest energy
4	ΔH_s	eV/atom	Stability as calculated by equation of the distortion with the lowest energy
5	V	$Å^3$/atom	Volume per atom of the relaxed structure
6	ΔE	eV	PBE band gap obtained from the relaxed structure
7	a	Å	Lattice parameter a of the perovskite structure
8	b	Å	Lattice parameter b of the perovskite structure
9	c	Å	Lattice parameter c of the perovskite structure
10	α		α angle of the relaxed structure. α = 90 for the cubic, tetragonal and orthorhombic distortion
11	β		β angle of the relaxed structure. β = 90 for the cubic, tetragonal and orthorhombic distortion
12	γ		γ angle of the relaxed structure. γ = 90 for the cubic, tetragonal and orthorhombic distortion
13	ΔE_V^O	eV per O atom	Oxygen vacancy formation energy

Dataset of perovskite materials

The dataset of ABO_3 perovskite oxide utilized in this paper originates from DFT high-throughput compute, including the ion radius of the A-site, the ion radius of the B-sit, formation energy, crystal volume, band gap, lattice parameters (a, b, c, α, β, γ), oxygen vacancy formation energy, stability, *etc.*[30]. Different radii of ions were used in A sites and B sites, including those of A ∈ [Al, As, Ag, Be, B, Bi, Ba, Ca, Co, Cu, Cr, Cd, Ce, Zr, Zn, Dy, Er, Fe, Ge, Gd, Ga, Hf, Ho, In, Ir, K, La, Lu, Mn, Mg, Mo, Ni, Nb, Nd, Pr, Pd, Ru, Pb, Rb, Re, Rh, Si, Sc, Sr, Sb, Sn, Sm, Ta, Th, Tb, Te, Ti, Tm, U, V, W, Y, Yb, *etc.*] and B ∈ [Al, Ag, As, Li, B, Bi, Be, Ca, Zn, Co, Cu, Cd, Cr, Ce, Zr, Dy, Eu, Er, Fe, Ga, Ge, Gd, Hf, Ho, In, Ir, Lu, Mn, Mg, Mo, Ni, Nd, Nb, Pr, Pb, Pd, Ru, Rb, Rh, Re, Si, Sb, Sc, Sn, Sm, Sr, Ti, Ta, Th, Tb, Te, Tm, U, V, W, Y, Yb, *etc.*], for perovskite ABO_3 compounds. Through a cursory examination of the data, a total of 4912 sets of ABO_3 perovskite compound high-throughput data were chosen. The values of stability were in the -0.729~3.927 eV/atom range. Data sets are indexed via abbreviations to make experimentation easier, using the following details: r_A for ionic radius at the A-site, r_B for ionic radius at the B-site, ΔH_f for formation energy, V for crystal volume, ΔE for band gap, a for lattice parameter a, b for lattice parameter b, c for lattice parameter c, α for the lattice parameter of the α phase, β for the lattice parameter of the β phase, γ for the lattice parameter of the γ phase, ΔE_V^O for oxygen vacancy formation energy and ΔH_s for stability. For ABO_3 perovskite oxide, more detailed descriptions are shown in Table 1.

Data pretreatment

Data pretreatment is used to process the missing and repeated values in the data, raising the data's accuracy and helping to raise the precision and efficiency of the subsequent learning procedure. The common processes of data pretreatment include missing value processing, attribute coding, feature selection, *etc.*[31,32]. There are three common ways to deal with missing values: use the feature that contains the missing value directly, delete the feature that contains the missing value (this only works if the feature contains blank values in a big number), and complete the missing value[33,34]. Because there are a small number of blank values in the raw Dataset, the features containing blank values are employed in this paper to process blank values.

Feature selection refers to the procedure of picking a subset of relevant features from a given feature collection[35]. Although a variety of factors influence the target characteristics of perovskite oxides, the amount of features must be appropriate, the features must be uneven for the category of interest, and certain non-essential information must be removed[36]. Correlation is a term that describes the degree and direction of the link between these two measurable features. Pearson correlation analysis is often adopted for analyzing the connections between two measurable features[37]. In this paper, Pearson correlation coefficient was used to examine the link between composition, structure and property, and the linear correlation among composition, structure and property is measured[38]. Pearson correlation coefficient can be defined easily as follows:

$$cor(x,y)=\frac{1}{n-1}\cdot\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\sigma_x\sigma_y} \qquad (1)$$

where x_i represents the value of feature x, y_i represents the value of feature y, $\bar{x}$ represents the average value of value of feature x, $\bar{y}$ represents the average value of feature y, σ_x represents the standard deviation of feature x, σ_y represents the standard deviation of feature y and n represents the sample size[38]. Correlation coefficient of 1 indicates strongly positively correlated, whereas a correlation value of -1 indicates a strong negative correlation. The correlation coefficient near to 0 implies that there is no association[37,38].

Pearson correlation coefficients were utilized to choose the raw dataset in this paper. Table 2 depicts the degree of connection between the 12 properties of perovskite oxide ABO_3 and their stability. Figure 3 depicts Pearson correlation map for different features.

Extreme feature engineering

For the sake of rapidly discovering the connections between structure, composition, and properties, we displayed dataset's feature distribution. Figure 4 depicts the distribution of raw features and stability. The distribution of the raw data set of observation feature and the predicted variable stability is positively biased, and the range of data is quite broad. Therefore, data transformation methods must be employed to generate new descriptors through feature engineering[39]. Essentially, the data provided to the algorithm should be compatible with the required structure or characteristics of the underlying data. Feature engineering is the process of turning data attributes into data features and extracting features from raw data through algorithms and models to the greatest extent possible[40]. Therefore, the feature engineering approach may generate a large number of new descriptors and assess their performance with a subset of them.

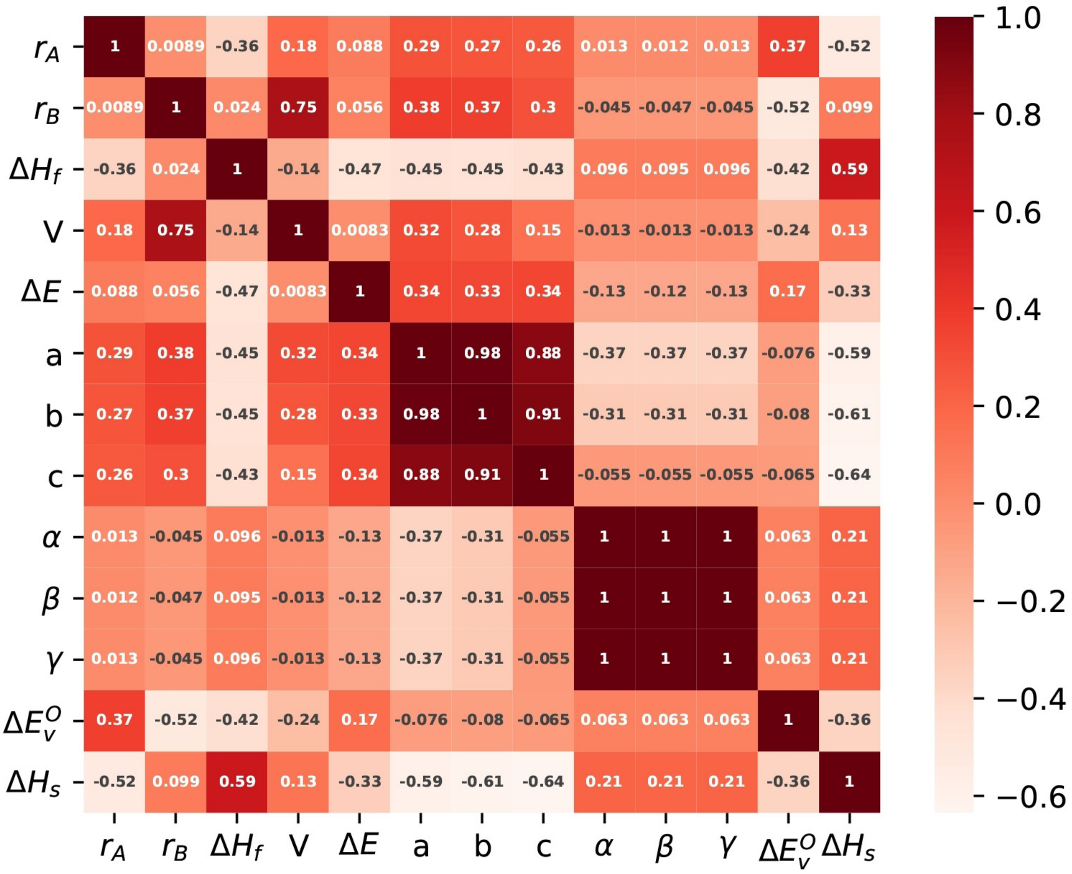
In machine learning, feature engineering is a critical data preparation activity that creates suitable descriptors from a given feature to improve prediction performance[41]. Feature engineering is adding some functions of conversion, such as arithmetic and aggregation operators, into a given attribute to create a huge number of new descriptors[42]. The transformation functions contribute to increase the dimensions of features or to turn the nonlinear connection between features and stability into a more understandable linear one[40,43]. Feature combination is a highly important method in feature engineering to integrate features from several categories into a single feature[44]. This is a beneficial method when a combination of features outperforms a single feature. The feature combination is the cross multiplication of all conceivable eigenvalues in mathematics. The features of each combination really constitute the information synergy.

A huge number of brand-new descriptors were gained through extreme feature engineering, where the dimensionality of features was also expanded. Figure 5 shows the construction process of the descriptors by extreme feature engineering. In their midst, x_i (i = 1, 2, ... n) indicates the selected feature. The parameters following the yellow arrows reflect a significant number of new descriptors that were produced. These 9 functions of x, x^{-1}, $\sqrt{x}$, x^2, x^3, e^x, $ln|x|$, $ln(1+|x|)$ and $log|x|$ are utilized for nonlinear transformation of features. In order to generate additional descriptors, these descriptors are merged non-linearly[45]. The primary descriptors were generated in the following way:

Step 1: Import the chosen vital features into these 9 functions of x, x^{-1}, $\sqrt{x}$, x^2, x^3, e^x, $ln|x|$, $ln(1+|x|)$ and $log|x|$, where x is one of the vital features chosen from the raw features, and it can directly generate brand-

Table 2. Pearson correlation coefficients of built-up features and stability

PCC	r_A	r_B	ΔH_f	V	ΔE	a	b	c	α	β	γ	ΔE_V^O	ΔH_s
r_A	1.000000	0.008855	-0.364382	0.183594	0.088459	0.286041	0.272757	0.257065	0.012843	0.012120	0.013025	0.371450	-0.522209
r_B	0.008855	1.000000	0.023527	0.754276	0.055515	0.384344	0.370947	0.295416	-0.045394	-0.047375	-0.045085	-0.517963	0.098718
ΔH_f	-0.364382	0.023527	1.000000	-0.140423	-0.472760	-0.448701	-0.450275	-0.431603	0.095894	0.094795	0.095947	-0.424975	0.594845
V	0.183594	0.754276	-0.140423	1.000000	0.008301	0.317787	0.277428	0.150485	-0.012753	-0.012908	-0.012622	-0.242627	0.127233
ΔE	0.088459	0.055515	-0.472760	0.008301	1.000000	0.337769	0.332856	0.340382	-0.126428	-0.124734	-0.126561	0.169527	-0.332567
a	0.286041	0.384344	-0.448701	0.317787	0.337769	1.000000	0.982619	0.879740	-0.366276	-0.365131	-0.366207	-0.076172	-0.591562
b	0.272757	0.370947	-0.450275	0.277428	0.332856	0.982619	1.000000	0.913307	-0.305305	-0.305172	-0.305283	-0.079756	-0.610395
c	0.257065	0.295416	-0.431603	0.150485	0.340382	0.879740	0.913307	1.000000	-0.054803	-0.055040	-0.054817	-0.065271	-0.635102
α	0.012843	-0.045394	0.095894	-0.012753	-0.126428	-0.366276	-0.305305	-0.054803	1.000000	0.996639	0.999847	0.062728	0.208962
β	0.012120	-0.047375	0.094795	-0.012908	-0.124734	-0.365131	-0.305172	-0.055040	0.996639	1.000000	0.995354	0.062613	0.207678
γ	0.013025	-0.045085	0.095947	-0.012622	-0.126561	-0.366207	-0.305283	-0.054817	0.999847	0.995354	1.000000	0.062694	0.209099
ΔE_V^O	0.371450	-0.517963	-0.424975	-0.242627	0.169527	-0.076172	0.079756	-0.065271	0.062728	0.062613	0.062694	1.000000	-0.359896
ΔH_s	-0.522209	-0.522209	0.594845	0.127233	-0.332567	-0.591562	-0.610395	-0.635102	0.208962	0.207678	0.209099	-0.359896	1.000000



	r_A	r_B	ΔH_f	V	ΔE	a	b	c	α	β	γ	ΔE_v^O	ΔH_s
r_A	1	0.0089	-0.36	0.18	0.088	0.29	0.27	0.26	0.013	0.012	0.013	0.37	-0.52
r_B	0.0089	1	0.024	0.75	0.056	0.38	0.37	0.3	-0.045	-0.047	-0.045	-0.52	0.099
ΔH_f	-0.36	0.024	1	-0.14	-0.47	-0.45	-0.45	-0.43	0.096	0.095	0.096	-0.42	0.59
V	0.18	0.75	-0.14	1	0.0083	0.32	0.28	0.15	-0.013	-0.013	-0.013	-0.24	0.13
ΔE	0.088	0.056	-0.47	0.0083	1	0.34	0.33	0.34	-0.13	-0.12	-0.13	0.17	-0.33
a	0.29	0.38	-0.45	0.32	0.34	1	0.98	0.88	-0.37	-0.37	-0.37	-0.076	-0.59
b	0.27	0.37	-0.45	0.28	0.33	0.98	1	0.91	-0.31	-0.31	-0.31	-0.08	-0.61
c	0.26	0.3	-0.43	0.15	0.34	0.88	0.91	1	-0.055	-0.055	-0.055	-0.065	-0.64
α	0.013	-0.045	0.096	-0.013	-0.13	-0.37	-0.31	-0.055	1	1	1	0.063	0.21
β	0.012	-0.047	0.095	-0.013	-0.12	-0.37	-0.31	-0.055	1	1	1	0.063	0.21
γ	0.013	-0.045	0.096	-0.013	-0.13	-0.37	-0.31	-0.055	1	1	1	0.063	0.21
ΔE_v^O	0.37	-0.52	-0.42	-0.24	0.17	-0.076	-0.08	-0.065	0.063	0.063	0.063	1	-0.36
ΔH_s	-0.52	0.099	0.59	0.13	-0.33	-0.59	-0.61	-0.64	0.21	0.21	0.21	-0.36	1

Figure 3. This is a Pearson correlation map for raw data. The correlation coefficient is shown by the color bar: red indicates strongly positive correlations, white denotes strongly negative correlations. The worth of the related Pearson correlation coefficient is represented by the filled fraction in each tiny square.

new descriptors;
Step 2: Feature combination combines the brand-new descriptors from Step 1. Increase the number of descriptive words by multiplying them by two or more and then combining them into a brand-new one;
Step 3: Substituting brand-new descriptors gained in step 2 into the function x^{-1} for nonlinear conversion, and the number of descriptors acquired has increased.

Regression

The term "regression" refers to the process of determining the quantitative connection between two or more variables using a group of data, the establishment of simulations from mathematics, and the estimation of unidentified factors[46]. Machine learning is an efficient way of performing regression. The capacity to do linear regression is defined as to properly depict the connection between data using a straight line, which is more suited to fitting the expression[47]. The modeling speed of linear regression is rapid, it does not need sophisticated calculation, and it may even run quickly when dealing with huge amounts of data[48]. The gained linear expression can be understood and interpreted according to the coefficient of each variable, and the influence of each feature on the result can be directly seen from the weight, which is much easier to grasp[43,49]. Nonlinear expressions are more complex than other machine learning methods, and the related process is difficult to learn[48]. Clearly, linear regression is appropriate for selecting the most appropriate descriptor. In this paper, we gained 55%/45% of the optimized data sets, which nicely balanced the accuracies and overfitting of the machine learning model. In the end, the important descriptor was gained by comparing the effectiveness of models for various descriptors.

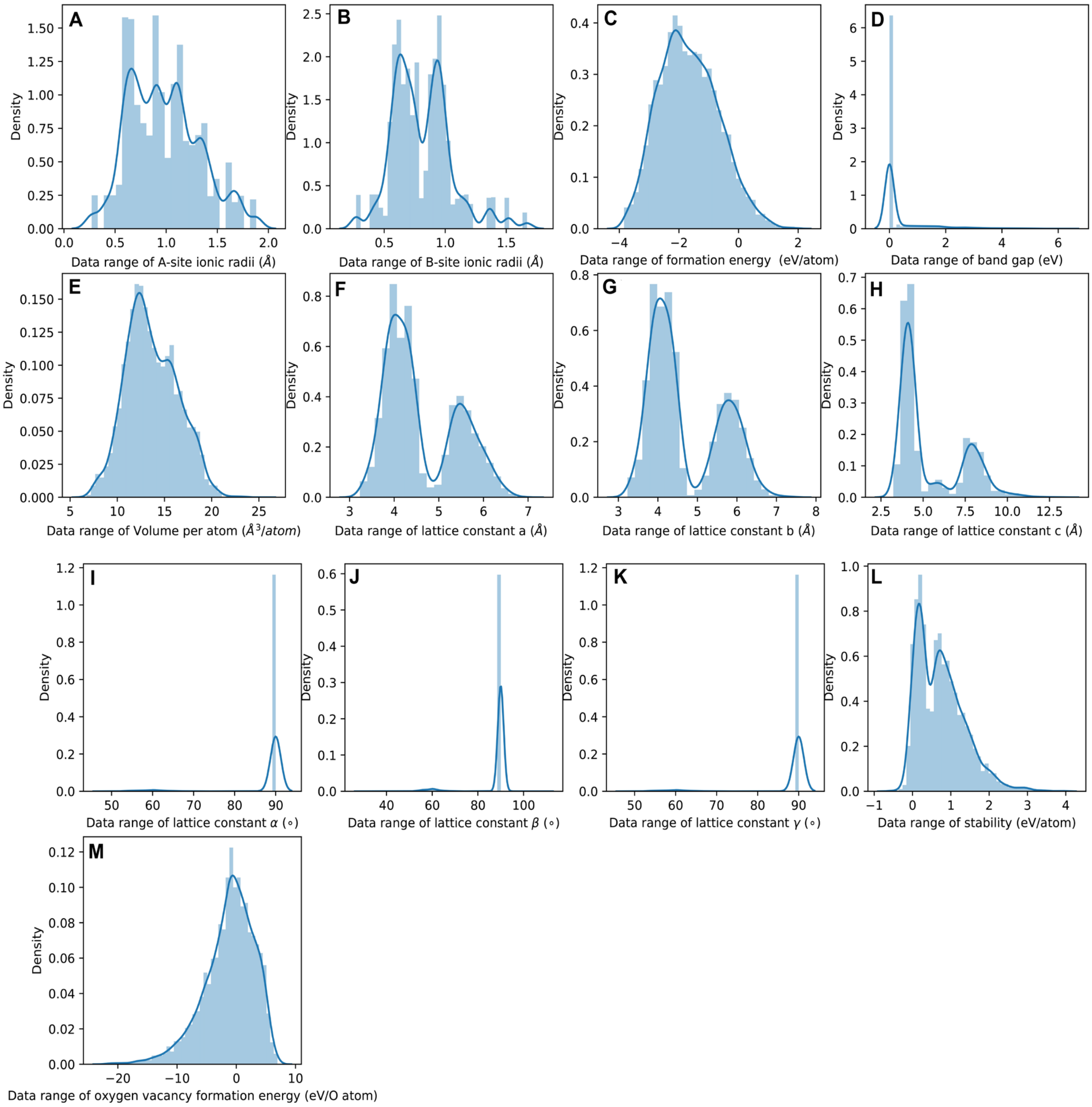


Figure 4. (A) The radius distribution of A-site ions; (B) the radius distribution of B-site ions; (C) distribution of formation energy; (D) distribution of band gap feature; (E) distribution of volume; (F) distribution of lattice constant a; (G) distribution of lattice constant b; (H) distribution of lattice constant c; (I) distribution of α angle of the crystal structure; (J) distribution of β angle of the crystal structure; (K) distribution of γ angle of the crystal structure; (L) distribution of stability; and (M) distribution of oxygen vacancy formation energy.

Performance evaluation

In order to assess the prediction accuracy and model performance, we employed the mean absolute error (MAE), mean square error (MSE) and coefficient of determination (R^2). Simply, the smaller MAE and MSE values to 0 and the bigger R^2 values to 1 suggest the higher prediction accuracy and better model performance. The corresponding equations can be summarized:

$$MAE = \frac{1}{n}\sum_{j=1}^{n}\left|\hat{y}_j - y_j\right| \tag{2}$$

$$MSE = \frac{1}{n}\sum_{j=1}^{n}\left(\hat{y}_j - y_j\right)^2 \tag{3}$$

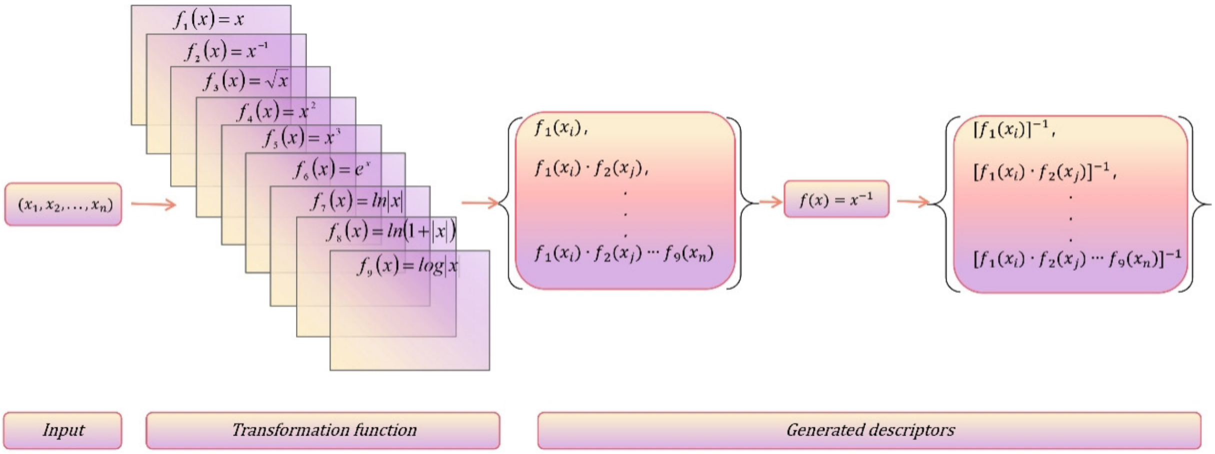


Figure 5. The process of descriptor construction by extreme feature engineering. The x_i (i = 1, 2, ... n) indicates the chosen feature. The parameters following the arrows denote the constructed a large number of brand-new descriptors. These 9 functions of x, x^{-1}, $\sqrt{x}$, x^2, x^3, e^x, $ln|x|$, $ln(1 + |x|)$ and $log|x|$ are employed in the nonlinear conversion of features. All of the descriptors mixed in a nonlinear way to construct more descriptors.

$$R^2 = 1 - \frac{\sum_{j=0}^{n-1}\left(y_j - \hat{y}_j\right)}{\sum_{j=0}^{n-1}\left(y_j - \overline{y}_j\right)^2} \tag{4}$$

Where, n indicates the sample size, $\hat{y}_j$ indicates experimental value and y_j indicates predicted value, and $\overline{y}$ is the average value.

RESULTS AND DISCUSSION

Extreme feature engineering

Following data pretreatment and feature transformation, the amount and quality of the description dataset must be checked further. Feature processing plays an important role in feature engineering and is also the most time-consuming aspect of data analysis. Because feature processing lacks a defined phase, such as algorithms and models with greater technical knowledge and compromises, there is no unified feature processing way. Fortunately, scikit-learn offers a more comprehensive feature processing approach, which includes data preparation, feature selection, dimension reduction, and so on[50]. Scikit-learn is a free and open-source machine learning library licensed under the Berkeley Software Distribution license[51]. Thus, in this paper, the python package scikit-learn was used for data pretreatment, feature transformation, feature processing, machine-learning model training and model performance evaluation[51-53]. Feature selection is the process of eliminating duplicate and unnecessary characteristics from a data collection, determining the important features in the data set, and eventually obtaining the feature subset[51]. Wrapper methods are common methods for feature selection[54,55]. The basic description of wrapper methods is:

Step 1: A subset of features is chosen to train the model. The model here usually refers to a machine learning algorithm, also called an objective function;
Step 2: Evaluate the model with a validation dataset;
Step 3: Perform the above operations on different feature subsets based on some search algorithm;
Step 4: Based on the evaluation results, the best feature subset is selected.

Clearly, the method for finding the optimal descriptors subset belongs to the family of greedy search algorithms. Wrapper methods include three common selection methods, such as sequential feature selection (SFS)[56], exhaustive feature selection[57] and recursive feature elimination[58]. Among them, SFS

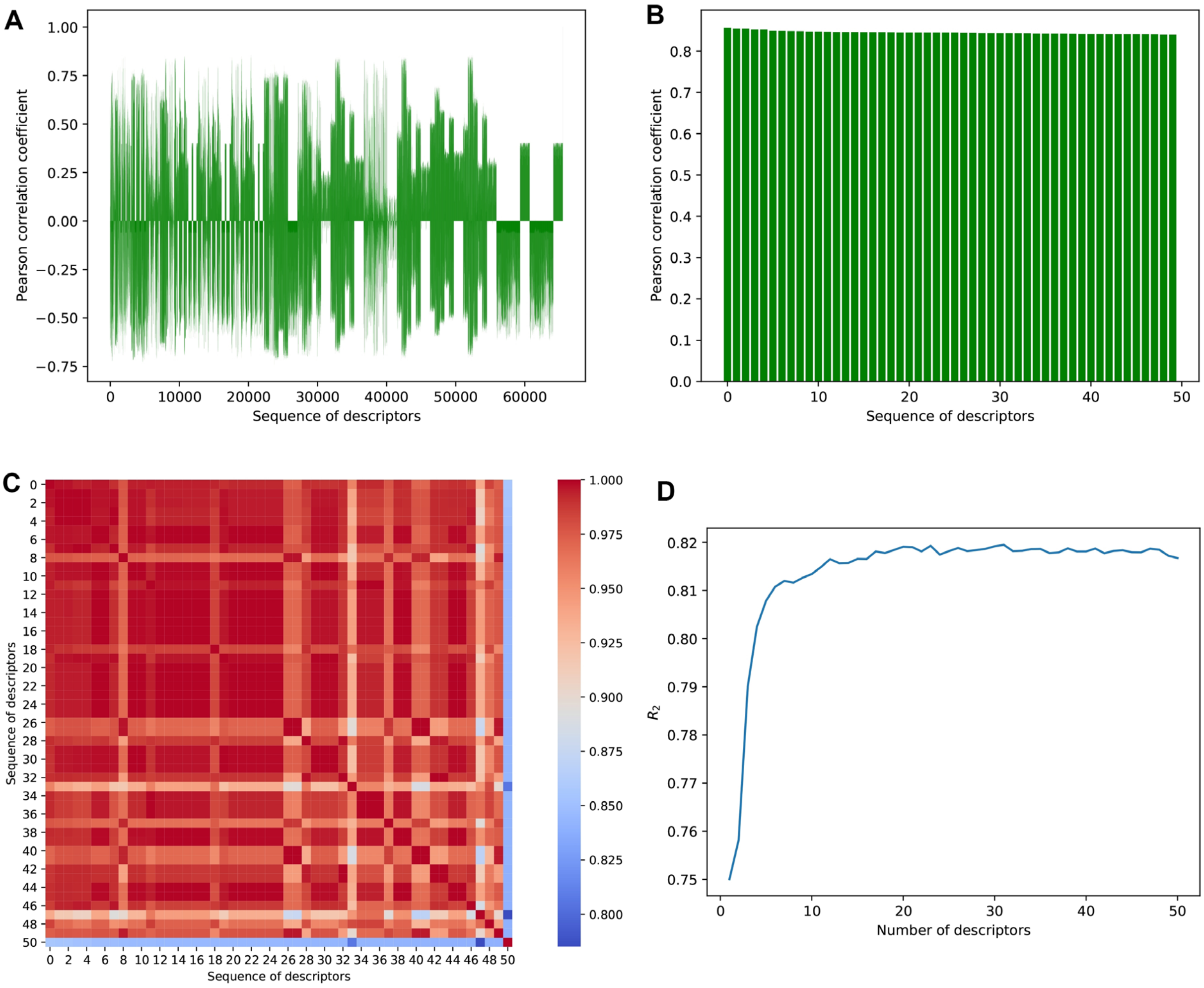


Figure 6. (A) Indicators of Pearson correlation coefficients for distinct descriptors with varying sequence numbers and stability. The horizontal axis displays the sequence number for the descriptors while the vertical axis is a reference to the relative Pearson correlation coefficient. (B) Indicators of Pearson correlation coefficients for the selected 50 distinct descriptors with varying sequence numbers and stability. The horizontal axis displays the sequence number for the descriptors while the vertical axis is a reference to the relative Pearson correlation coefficient. (C) Pearson correlation map for the selected 50 descriptors and the stability. The color bar on the right represents the correlation coefficient. (D) R^2 values of GBR models are used to evaluate machine learning algorithms. There are values of descriptors on the horizontal axis, and R^2 values for GBR models on the vertical axis.

includes two algorithms, such as sequential forward feature selection algorithm and sequential backward feature selection algorithm. Sequential forward selection algorithm is about execution of the following steps to search the most appropriate features out of N features to fit in K-features subset. Instead of trying to model all the feature combinations, the sequential forward feature selection algorithm aborts the search by finding an optimal solution, which greatly reduces the computational effort[56]. Therefore, we adopted the sequential forward feature selection algorithm to perform feature selection. In this work, gradient augmented regression (GBR) was used as the objective function.

The extreme feature engineering created many descriptors, and was followed by a preliminary screening of these descriptors. By analyzing the Pearson correlation coefficient, the top 50 descriptors with the highest Pearson correlation coefficient were successfully selected. Figure 6A shows the Pearson correlation coefficients for different new-constructed descriptors. Figure 6B shows the Pearson correlation coefficients for the as-selected 50 descriptors. Figure 6C shows the Pearson correlation map of the as-selected 50 descriptors and stability. Figure 6D shows the trend between the prediction effect of GBR models and the descriptor number.

GBR is an enhancement to the Boosting algorithm[59]. Boosting is a type of integrated machine-learning algorithm that transforms the poor learner into the strong learner. Each sample was initially allocated an equal weight value in the Boosting algorithm[60,61]. Because each training produced a significant change in the values of data points, the weight values were processed by adding mis-splitting points at the end of each step, and then N iterations were done to obtain N simple base classifiers[62,63]. Finally, the N basic classifiers acquired were weighted together to form a final model.

The distinction between GBR and Boosting is that each GBR computation is designed to minimize the last residual. To reduce the associated residuals, a new model must be created in the gradient's orientation to reduce the residuals. As a result, in GBR, each new modeling aimed to reduce the previous model residuals in the gradient orientation[64], the associated loss-function negative gradient was employed as the estimated value of the residual in the GBR algorithm, and then the regression tree was fitted[65,66]. As a weak classifier, GBR typically employs a fixed size regression tree. The capacity to analyze mixed data and create models with complicated functions are two properties of the regression tree that make it more accurate in the promotion process[64]. The GBR model is as follows:

$$f_M(x) = \sum_{m=1}^{M} r_m T(x; \theta_m) \tag{5}$$

Here, r_m is the weight, $T(x; \theta_m)$ is the regression tree, θ_m is the parameter of the regression tree and m is the number of trees[67]. The GBR models were built iteratively in this paper. The best descriptors were chosen based on the coefficients of descriptor significance, and this process was repeated for the other descriptors until all descriptors had been explored. Finally, the optimal descriptor subset consists of 23 most important descriptors, as shown in Table 3.

Figure 6D depicts the relationship between GBR model prediction effect and descriptor numbers. It is clear that when the number of descriptors was raised, the prediction effect of the GBR models grew and eventually stabilized[68,69]. Clearly, the best effect of the GBR models was obtained when the optimum subset of 23 descriptors was employed. Figure 6B depicts the Pearson correlation coefficients for the 50 descriptors chosen and the stability, which ranges from 0.845766 to 0.839595 with minor variations. These descriptors are strongly correlated with the stability of perovskite compounds. The key to understanding the structure-composition-property relationship is to choose the best relevant description.

Following the selection of 23 key descriptors through SFS, the ideal subset of descriptors was chosen based on the three evaluation indices of the GBR model to train the linear regression model, as shown in Table 3. After a large number of experiments, the results showed that these fluctuations were within the range of 3%. It is apparent that the descriptor $(ln(1+|r_A|)r_B^{-1}c^2\alpha^{-1})^{-1}$ exhibited the highest R^2 of 0.716913, the lowest MAE of 0.230453 and the lowest MSE of 0.230453, respectively, indicating the best model performance. The R^2, MSE and MAE values for descriptors $[ln(1+|r_A|)r_B^{-1}c^2\alpha^{-1}]^{-1}$, $[ln(1+|r_A|)r_B^{-1}\sqrt{b}c^2\alpha^{-1}]^{-1}$, $[ln(1+|r_A|)r_B^{-1}ln(1+|b|)c^2\alpha^{-1}]^{-1}$, $[ln(1+|r_A|)r_B^{-1}ln|b|\,c^2\alpha^{-1}]^{-1}$, $[ln(1+|r_A|)r_B^{-1}bc^2\alpha^{-1}]^{-1}$, $(r_Ar_B^{-1}c^2\alpha^{-1})^{-1}$ and $[ln(1+|r_A|)r_B^{-1}bc^2\,ln|\alpha|\,]^{-1}$ are exceptionally small. The other 16 descriptors exhibited poor prediction effects and were not considered in the subsequent work. For these 7 different descriptors, the performance of the model was similar, so we chose 7 descriptors, $[ln(1+|r_A|)r_B^{-1}c^2\alpha^{-1}]^{-1}$, $[ln(1+|r_A|)r_B^{-1}\sqrt{b}c^2\alpha^{-1}]^{-1}$, $[ln(1+|r_A|)r_B^{-1}ln(1+|b|)c^2\alpha^{-1}]^{-1}$, $[ln(1+|r_A|)r_B^{-1}ln|b|\,c^2\alpha^{-1}]^{-1}$, $[ln(1+|r_A|)r_B^{-1}bc^2\alpha^{-1}]^{-1}$, $(r_Ar_B^{-1}c^2\alpha^{-1})^{-1}$ and $[ln(1+|r_A|)r_B^{-1}bc^2\,ln|\alpha|\,]^{-1}$, to train the linear regression model.

Due to the limited capabilities of experimental and theoretical tools, traditional material discovery has always been a process of trial and error. The widely used tolerance factor (t for short) to measure the stability of perovskite was proposed by Goldschmidt in 1926. t has become a popular descriptor of stability and has accelerated stability screening of perovskite over the past century. It is worth noting that Goldschmidt tolerance factor t has been widely used to predict the stability of perovskite structures based

Table 3. Comparison of three evaluation indicators and brand-new descriptors chosen by the GBR model

Method	No.	Descriptors	Evaluation index		
			R^2	MAE	MSE
GBR	1	$(ln(1+\lvert r_A\rvert)r_B^{-1}c^2\alpha^{-1})^{-1}$	0.716913	0.230453	0.110132
	2	$(ln(1+\lvert r_A\rvert)r_B^{-1}\sqrt{b}c^2\alpha^{-1})$	0.713565	0.231627	0.111434
	3	$(ln(1+\lvert r_A\rvert)r_B^{-1}\,ln(1+\lvert b\rvert)c^2\quad{}^{-1})^{-1}$	0.715694	0.230980	0.110606
	4	$(ln(1+\lvert r_A\rvert)r_B^{-1}\,ln\lvert b\rvert c^2\alpha^{-1})^{-1}$	0.712077	0.234185	0.112013
	5	$(ln(1\quad\lvert r_A\rvert)r_B^{-1}bc^2\quad{}^{-1})^{-1}$	0.711042	0.235473	0.112416
	6	$(ln(1+\lvert r_A\rvert)r_B^{-1}\sqrt{b}c^2\,ln\lvert\alpha\rvert)^{-1}$	0.699977	0.240292	0.116721
	7	$(ln(1+\lvert r_A\rvert)r_B^{-1}bc\alpha^{-1})^{-1}$	0.707636	0.235522	0.113741
	8	$(ln(1+\lvert r_A\rvert)r_B^{-1}bc^2)^{-1}$	0.707648	0.238117	0.113736
	9	$(\sqrt{r_A}r_B^{-1}\sqrt{b}c^2\alpha^{-1})^{-1}$	0.687836	0.241530	0.121444
	10	$(r_Ar_B^{-1}c^2\alpha^{-1})^{-1}$	0.711809	0.233404	0.112117
	11	$(ln(1+\lvert r_A\rvert)r_B^{-1}bc^2\,ln(1+\lvert\alpha\rvert))^{-1}$	0.703372	0.239687	0.115400
	12	$(ln(1+\lvert r_A\rvert)r_B^{-1}bc^2\,ln\lvert\alpha\rvert)^{-1}$	0.714631	0.234160	0.111020
	13	$(r_Ar_B{}^{-1}\,ln(1+\lvert b\rvert)c^2\alpha^{-1})^{-1}$	0.673522	0.248478	0.127012
	14	$(\sqrt{r_A}r_B^{-1}bc\alpha^{-1})^{-1}$	0.645613	0.247656	0.137870
	15	$(ln(1+\lvert r_A\rvert)r_B^{-1}ln(1+\lvert b\rvert)c^2\sqrt{\alpha})^{-1}$	0.691656	0.243092	0.119958
	16	$(\sqrt{r_A}r_B^{-1}\,ln\lvert b\rvert c^2\alpha^{-1})^{-1}$	0.691806	0.242466	0.119899
	17	$(\sqrt{r_A}r_B^{-1}ln\lvert b\rvert c^2\alpha^{-1})^{-1}$	0.669111	0.246630	0.128729
	18	$(r_Ar_B^{-1}\,ln\lvert b\rvert c^2\alpha^{-1})^{-1}$	0.708067	0.235591	0.113573
	19	$(r_Ar_B^{-1}\,log\lvert b\rvert c^2\alpha^{-1})^{-1}$	0.708067	0.235591	0.113573
	20	$(ln(1+\lvert r_A\rvert)r_B^{-1}b^2c\alpha^{-1})^{-1}$	0.703448	0.238094	0.115370
	21	$(ln(1+\lvert r_A\rvert)r_B^{-1}bc^2\sqrt{\alpha})^{-1}$	0.585948	0.274807	0.161083
	22	$(ln(1+\lvert r_A\rvert)r_B^{-1}b^{-1}c^2\alpha^{-1})^{-1}$	0.683018	0.244765	0.123318
	23	$(\sqrt{r_A}r_B^{-1}\sqrt{b}c^2)^{-1}$	0.683018	0.244765	0.123318

GBR: Gradient augmented regression; MAE: mean absolute error; MSE: mean square error.

only on a universal formula of ABX_3 with matching ionic sizes of A-site, B-site and X-site[70]. Its expression is:

$$t=\frac{(r_A+r_X)}{\sqrt{2}(r_B+r_X)} \tag{6}$$

Here, r_A is the A-site ionic radius, r_B is the B-site ionic radius, and r_X is the X-site ionic radius. This is a semi-empirical formula with an accuracy of only 70% that gives a rough indication of the stability of perovskite materials. The descriptors constructed in this work are not only related to the A-site ion radius, B-site ion radius, but also related to the lattice parameters, which are considered to be key features related to the stability of perovskite materials.

Automated machine learning

By automated machine learning, we discovered the quantitative relationships between various variables based on a collection of data, which resulted in the construction of a mathematical model and the estimation of unknown parameters. The linear regression algorithm, as an effective machine-learning algorithm, accurately depicted the connection of data via the straight line and was better for fitting expressions in this paper[71]. Table 4 shows the 7 descriptors $\{[ln(1+|r_A|)r_B^{-1}c^2\alpha^{-1}]^{-1}$, $[ln(1+|r_A|)r_B^{-1}\sqrt{b}c^2\alpha^{-1}]^{-1}$, $[ln(1+|r_A|)r_B^{-1}\,ln(1+|b|)c^2\alpha^{-1}]^{-1}$, $[ln(1+|r_A|)r_B^{-1}\,ln|b|\,c^2\alpha^{-1}]^{-1}$, $[ln(1+|r_A|)r_B^{-1}bc^2\alpha^{-1}]^{-1}$, $(r_Ar_B^{-1}c^2\alpha^{-1})^{-1}$, $[ln(1+|r_A|)r_B^{-1}bc^2\,ln|\alpha|]^{-1}\}$ and the corresponding evaluation indexes selected by the linear regression model. It is easy to see that the last descriptor of $[ln(1+|r_A|)r_B^{-1}c^2\alpha^{-1}]^{-1}$ achieved the greatest R^2 value of 0.735529, the lowest MAE value of 0.224526 and the lowest MSE value of 0.102889. As a result, the descriptor of $[ln(1+|r_A|)r_B^{-1}c^2\alpha^{-1}]^{-1}$ was chosen as the best descriptor for investigating structure-composition-property relationships in perovskite compounds (ABO_3), which was only related with A-site ion radius, B-site ion radius, lattice constant b, and α angle of the crystal structure.

Table 4. The linear regression model's specified descriptors and assessment indices

Method	No.	Descriptors		Evaluation index	
			R^2	MAE	MSE
Linear Regression	1	$(ln(1+\|r_A\|)r_B^{-1}c^2\alpha^{-1})^{-1}$	0.735529	0.224526	0.102889
	2	$(ln(1+\|r_A\|)r_B^{-1}\sqrt{b}c^2\alpha^{-1})^{-1}$	0.732250	0.226233	0.104165
	3	$(ln(1+\|r_A\|)r_B^{-1}ln(1+\|b\|)c^2\alpha^{-1})^{-1}$	0.732245	0.226250	0.104167
	4	$(ln(1+\|r_A\|)r_B^{-1}ln\|b\|c^2\alpha^{-1})^{-1}$	0.728364	0.228253	0.105677
	5	$(ln(1+\|r_A\|)r_B^{-1}bc^2\alpha^{-1})^{-1}$	0.721646	0.231526	0.108291
	6	$(r_A r_B^{-1}c^2\alpha^{-1})^{-1}$	0.718913	0.109353	0.235988
	7	$(ln(1+\|r_A\|)r_B^{-1}bc^2 ln\|\alpha\|)^{-1}$	0.718752	0.235184	0.109416

MAE: Mean absolute error; MSE: mean square error.

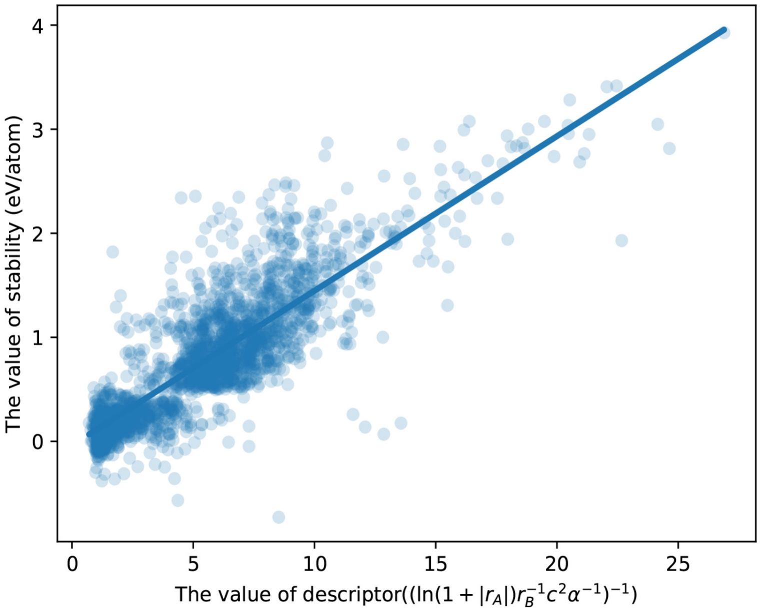


Figure 7. Scatter plots showing correlations between the best descriptor and stability for ABO_3 perovskite compounds. The blue line showed anticipated stability values, whereas the scatter points reflect actual stability values.

The straightforward linear equation is intended to represent the structure-composition-property relationships of ABO_3 perovskite compounds after obtaining the optimal descriptor using a linear regression model. The following is the equivalent formula:

$$F = f(d_1, d_2, \ldots d_n) \tag{7}$$

Where d_i is the final descriptors, F is the stability, $f(d_1, d_2, \ldots d_n)$ is a linear representation of the structure-composition-property connection. A simple linear expression was produced using linear regression analysis as follows:

$$\Delta H_S = k(ln(1+|r_A|)r_B^{-1}c^2\alpha^{-1})^{-1} + z \tag{8}$$

where k = 0.1485 and z = -0.0380 are the coefficient values. Following linear regression fit and comparison with DFT calculation value, as shown in Figure 7, the dependability of the automated-machine-learning stability expression was validated. The results showed that the effects of A-site ion radius, B-site ion radius, lattice constant b, and α angle of the crystal structure are more significant than that of other variables. The equation of $\Delta H_S = 0.1485(ln(1+|r_A|)r_B^{-1}c^2\alpha^{-1})^{-1} - 0.0380$ showed the relationship between structure, composition and property in perovskite oxides. Our technique produces a more accurate expression than the semi-empirical formula. In a nutshell, the novel approach may be utilized to investigate the structure-composition-property relationships of ABO_3 perovskite oxides.

CONCLUSIONS

For the sake of conquering the huge complexity of structure-composition-property in ABO_3 perovskite materials for energy conversion and storage, we presented a new way to combine extreme feature engineering and automated machine learning for investigating structure-composition-property connections in perovskite oxides. A great number of brand-new descriptors were generated via extreme feature engineering and a subset of 23 significant descriptors was gained via SFS. Furthermore, by linear regression algorithm, the optimal descriptor of $[ln(1+|r_A|)r_B^{-1}c^2\alpha^{-1}]^{-1}$ was found, and the straightforward linear equation of $\Delta H_S = 0.1485(ln(1+|r_A|)r_B^{-1}c^2\alpha^{-1})^{-1} - 0.0380$ for the stability was achieved. It has been shown that the influence of radius of A-site ions, radius of B-site ions, lattice constant b, and α angle of the crystal structure on the stability of ABO_3 perovskites are more significant than others. In this way, we can obtain expression with higher accuracy than a semi-empirical formula. The results demonstrate a high-efficient and non-priori-knowledge investigation of structure-composition-property relationships for perovskite materials, providing a new road to discover advanced energy materials.

DECLARATIONS

Authors' contributions

Made substantial contributions to conception and design of the study and performed data analysis and interpretation: Deng Q, Lin B

Performed data acquisition, as well as provided administrative, technical, and material support: Deng Q, Lin B

Wrote and reviewed the manuscript: Deng Q, Lin B

Availability of data and materials

Not applicable.

Financial support and sponsorship

The authors are grateful for the support offered by the UESTCYDRI (U04210029), the University of Electronic Science and Technology of China (No. A03018023601020) and the Thousand Talents Program of Sichuan Province.

Conflicts of interest

Both authors declared that there are no conflicts of interest.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Copyright

REFERENCES

1. Hong Y, Hou B, Jiang H, Zhang J. Machine learning and artificial neural network accelerated computational discoveries in materials science. *WIREs Comput Mol Sci* 2020;10:e1450.
2. Sparks TD, Kauwe SK, Parry ME, Tehrani AM, Brgoch J. Machine learning for structural materials. *Annu Rev Mater Res* 2020;50:27-48.
3. Hwang J, Rao RR, Giordano L, Katayama Y, Yu Y, Shao-Horn Y. Perovskites in catalysis and electrocatalysis. *Science* 2017;358:751-6.
4. Butler K T, Davies D W, Cartwright H, et al. Machine learning for molecular and materials science. *Nature* 2018;559:547-55.
5. Juan Y, Dai Y, Yang Y, Zhang J. Accelerating materials discovery using machine learning. *J Mater Sci Mater Med* 2021;79:178-90.

6. Fischer CC, Tibbetts KJ, Morgan D, Ceder G. Predicting crystal structure by merging data mining with quantum mechanics. *Nat Mater* 2006;5:641-6.
7. Raccuglia P, Elbert KC, Adler PD, et al. Machine-learning-assisted materials discovery using failed experiments. *Nature* 2016;533:73-6.
8. Sanchez-Lengeling B, Aspuru-Guzik A. Inverse molecular design using machine learning: generative models for matter engineering. *Science* 2018;361:360-5.
9. Ong SP. Accelerating materials science with high-throughput computations and machine learning. *Computational Materials Science* 2019;161:143-50.
10. Balachandran PV. Machine learning guided design of functional materials with targeted properties. *Computational Materials Science* 2019;164:82-90.
11. Peña MA, Fierro JL. Chemical structures and performance of perovskite oxides. *Chem Rev* 2001;101:1981-2017.
12. Lino A, Rocha Á, Sizo A, Rocha Á. Virtual teaching and learning environments: automatic evaluation with symbolic regression. *IFS* 2016;31:2061-72.
13. Yuan S, Jiao Z, Quddus N, Kwon JS, Mashuga CV. Developing quantitative structure-property relationship models to predict the upper flammability limit using machine learning. *Ind Eng Chem Res* 2019;58:3531-7.
14. Ouyang R, Curtarolo S, Ahmetcik E, Scheffler M, Ghiringhelli LM. SISSO: a compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates. *Phys Rev Materials* 2018;2:083802.
15. Zhang Y, Xu X. Machine learning lattice constants for cubic perovskite compounds. *Chemistry Select* 2020;5:9999-10009.
16. de Franca FO, de Lima MZ. Interaction-transformation symbolic regression with extreme learning machine. *Neurocomputing* 2021;423:609-19.
17. Li Z, Xu Q, Sun Q, et al. Stability engineering of halide perovskite via machine learning. *arXiv preprint arXiv* 2018;1803.06042.
18. Li W, Jacobs R, Morgan D. Predicting the thermodynamic stability of perovskite oxides using machine learning models. *Computational Materials Science* 2018;150:454-63.
19. Deng Q, Lin B. Exploring structure-composition relationships of cubic perovskite oxides via extreme feature engineering and automated machine learning. *Materials Today Communications* 2021;28:102590.
20. Gardner S, Golovidov O, Griffin J, et al. Constrained multi-objective optimization for automated machine learning. 2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA); 2019 Oct 5-8; Washington, DC, USA. IEEE; 2019. p. 364-73.
21. Masrom S, Mohd T, Jamil N S, et al. Automated machine learning based on genetic programming: a case study on a real house pricing dataset. 2019 1st International Conference on Artificial Intelligence and Data Sciences (AiDAS); 2019 Sep 19; Ipoh, Malaysia. IEEE; 2019. p. 48-52.
22. Chauhan K, Jani S, Thakkar D, et al. Automated machine learning: the new wave of machine learning. 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA); 2020 Mar 5-7; Bangalore, India. IEEE; 2020. p. 205-12.
23. Ge P. Analysis on approaches and structures of automated machine learning frameworks. 2020 International Conference on Communications, Information System and Computer Engineering (CISCE); 2020 Jul 3-5; Kuala Lumpur, Malaysia. IEEE; 2020. p. 474-7.
24. Han J, Park KS, Lee KM. An automated machine learning platform for non-experts. Proceedings of the International Conference on Research in Adaptive and Convergent Systems. Association for Computing Machinery, New York, NY, USA; 2020. p. 84-6.
25. Umamahesan A, Babu DMI. From zero to AI hero with automated machine learning. Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Association for Computing Machinery, New York, NY, USA; 2020. p. 3495.
26. Waring J, Lindvall C, Umeton R. Automated machine learning: review of the state-of-the-art and opportunities for healthcare. *Artif Intell Med* 2020;104:101822.
27. Zeineddine H, Braendle U, Farah A. Enhancing prediction of student success: automated machine learning approach. *Computers & Electrical Engineering* 2021;89:106903.
28. Sun Y, Yang G. Feature engineering for search advertising recognition. 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC); 2019 Mar 15-17; Chengdu, China. IEEE; 2019. p. 1859-64.
29. Li Z, Ma X, Xin H. Feature engineering of machine-learning chemisorption models for catalyst design. *Catalysis Today* 2017;280:232-8.
30. Emery AA, Wolverton C. High-throughput DFT calculations of formation energy, stability and oxygen vacancy formation energy of ABO_3 perovskites. *Sci Data* 2017;4:170153.
31. Kotsiantis SB, Kanellopoulos D, Pintelas PE. Data preprocessing for supervised leaning. *International journal of computer science* 2006;1:111-7.
32. Liu N, Gao G, Liu G. Data Preprocessing based on partially supervised learning. Proceedings of the 6th International Conference on Information Engineering for Mechanics and Materials; 2016 Nov. Atlantis Press; 2016. p. 678-83.
33. Zainuddin Z, Lim E A. A comparative study of missing value estimation methods: which method performs better? 2008 International Conference on Electronic Design; 2008 Dec 1-3; Penang, Malaysia. IEEE; 2008, p. 1-5.
34. Kang H. The prevention and handling of the missing data. *Korean J Anesthesiol* 2013;64:402-6.
35. Li J, Cheng K, Wang S, et al. Feature Selection: A Data Perspective. *ACM Comput Surv* 2018;50:1-45.
36. Mangal A, Holm E A. A comparative study of feature selection methods for stress hotspot classification in materials. *Integrating Materials and Manufacturing Innovation* 2018;7:87-95.
37. Zhou H, Deng Z, Xia Y, Fu M. A new sampling method in particle filter based on Pearson correlation coefficient. *Neurocomputing*

2016;216:208-15.

38. Hauke J, Kossowski T. Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data. *Quaestiones Geographicae* 2011;30:87-93.
39. Khurana U, Samulowitz H, Turaga D. Feature engineering for predictive modeling using reinforcement learning. Proceedings of the AAAI Conference on Artificial Intelligence. *Thirty-Second AAAI Conference on Artificial Intelligence* 2018;32:3407-14.
40. Heaton J. An empirical analysis of feature engineering for predictive modeling. SoutheastCon 2016; 2016 Mar 30-Apr 3; Norfolk, VA, USA. IEEE; 2016. p. 1-6.
41. Zheng A, Casari A. Feature engineering for machine learning: principles and techniques for data scientists. 'O'Reilly Media, Inc.; 2018.
42. Dai D, Xu T, Wei X, et al. Using machine learning and feature engineering to characterize limited material datasets of high-entropy alloys. *Computational Materials Science* 2020;175:109618.
43. Nargesian F, Samulowitz H, Khurana U, et al. Learning feature engineering for classification. Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence Main track. IJCAI; 2017. p. 2529-35.
44. Hou J, Pelillo M. A simple feature combination method based on dominant sets. *Pattern Recognition* 2013;46:3129-39.
45. Ghiringhelli LM, Vybiral J, Levchenko SV, Draxl C, Scheffler M. Big data of materials science: critical role of the descriptor. *Phys Rev Lett* 2015;114:105503.
46. Fox J. Regression diagnostics: an introduction. Sage publications; 2019.
47. Montgomery DC, Peck EA, Vining GG. Introduction to linear regression analysis. John Wiley & Sons; 2021.
48. Weisberg S. Applied linear regression. John Wiley & Sons; 2005.
49. Dai D, Liu Q, Hu R, et al. Method construction of structure-property relationships from data by machine learning assisted mining for materials design applications. *Materials & Design* 2020;196:109194.
50. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *The Journal of machine Learning research* 2011;12:2825-30.
51. Abraham A, Pedregosa F, Eickenberg M, et al. Machine learning for neuroimaging with scikit-learn. *Front Neuroinform* 2014;8:14.
52. Jović A, Brkić K, Bogunović N. A review of feature selection methods with applications. 2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO); 2015 May 25-29; Opatija, Croatia. IEEE; 2015. p. 1200-5.
53. Remeseiro B, Bolon-Canedo V. A review of feature selection methods in medical applications. *Comput Biol Med* 2019;112:103375.
54. Dash M, Liu H. Feature selection for classification. *Intelligent Data Analysis* 1997;1:131-56.
55. Chandrashekar G, Sahin F. A survey on feature selection methods. *Computers & Electrical Engineering* 2014;40:16-28.
56. Rückstieß T, Osendorfer C, van der Smagt P. Sequential feature selection for classification. In: Wang D, Reynolds M, editors. AI 2011: advances in artificial intelligence. Berlin: Springer Berlin Heidelberg; 2011. p. 132-41.
57. Lee CY, Chen BS. Mutually-exclusive-and-collectively-exhaustive feature selection scheme. *Applied Soft Computing* 2018;68:961-71.
58. Su R, Liu X, Wei L. MinE-RFE: determine the optimal subset from RFE by minimizing the subset-accuracy-defined energy. *Brief Bioinform* 2020;21:687-98.
59. Yang F, Wang D, Xu F, Huang Z, Tsui K. Lifespan prediction of lithium-ion batteries based on various extracted features and gradient boosting regression tree model. *J Power Sources* 2020;476:228654.
60. Sun X, Zhou H. Experiments with two new boosting algorithms. *IIM* 2010;02:386-90.
61. Schapire RE. The boosting approach to machine learning: an overview. In: Denison DD, Hansen MH, Holmes CC, Mallick B, Yu B, editors. Nonlinear estimation and classification. New York: Springer; 2003. p. 149-71.
62. Oza N C. Online bagging and boosting. 2005 IEEE International Conference on Systems, Man and Cybernetics; 2005 Oct 12; Waikoloa, HI, USA. IEEE; 2005, p. 2340-5.
63. Elith J, Leathwick JR, Hastie T. A working guide to boosted regression trees. *J Anim Ecol* 2008;77:802-13.
64. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Statist* 2001;29.
65. Abe GL, Sasaki JI, Katata C, et al. Fabrication of novel poly(lactic acid/caprolactone) bilayer membrane for GBR application. *Dent Mater* 2020;36:626-34.
66. Sharafati A, Asadollah SBHS, Hosseinzadeh M. The potential of new ensemble machine learning models for effluent quality parameters prediction and related uncertainty. *Process Safety and Environmental Protection* 2020;140:68-78.
67. Friedman JH. Stochastic gradient boosting. *Computational Statistics & Data Analysis* 2002;38:367-78.
68. Domingos P. A few useful things to know about machine learning. *Commun ACM* 2012;55:78-87.
69. Lu S, Zhou Q, Guo Y, Zhang Y, Wu Y, Wang J. Coupling a crystal graph multilayer descriptor to active learning for rapid discovery of 2D Ferromagnetic Semiconductors/Half-Metals/Metals. *Adv Mater* 2020;32:e2002658.
70. Bartel CJ, Sutton C, Goldsmith BR, et al. New tolerance factor to predict the stability of perovskite oxides and halides. *Sci Adv* 2019;5:eaav0693.
71. Uyanık GK, Güler N. A study on multiple linear regression analysis. *Procedia - Social and Behavioral Sciences* 2013;106:234-40.