Materials and Design 196 (2020) 109194

Contents lists available at ScienceDirect

Materials and Design

journal homepage: www.elsevier.com/locate/matdes

Method construction of structure-property relationships from data by machine learning assisted mining for materials design applications

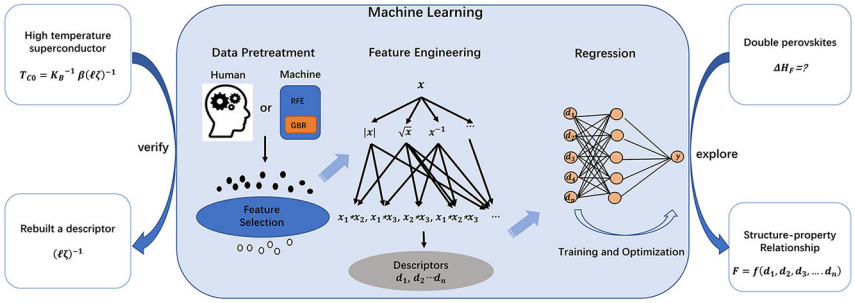
Dongbo Dai [a], Qing Liu [a], Rui Hu [a], Xiao Wei [a,b], Guangtai Ding [a,b], Baoyu Xu [a], Tao Xu [b], Jincang Zhang [b], Yan Xu [c], Huiran Zhang [a,b,*]

[a] School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China
[b] Materials Genome Institute of Shanghai University, Shanghai 200444, China
[c] College of Mathematics and Physics, Shanghai University of Electric Power, Shanghai 200090, China

HIGHLIGHTS

- Construct appropriate descriptors with feature preprocessing and feature engineering.
- Non-linear combination of descriptors can help build structure-property relationship.
- The linear regression model combined with descriptors can be used to fit the structure-property relationship of materials.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 27 July 2020
Received in revised form 8 September 2020
Accepted 26 September 2020
Available online 30 September 2020

Keywords:
Structure-property relationship
Descriptor
Machine learning
Feature engineering
Linear regression

ABSTRACT

Data driven material research is a hot topic in the cross field of artificial intelligence and materials science. The core of new material prediction is to find the relationship between material structure and properties. In this research, machine learning will have important advantages and play an important role for materials data. In this paper,we put forward a framework combining feature engineering and linear regression to find the correlation between structure and properties from materials data. High temperature superconductor and double perovskites for solar cells were employed to test the feasibility of the method. In the former, we successfully rebuilt a descriptor $(\ell\zeta)^{-1}$ from data mining which is consistent with the theoretical formula. In the latter, as an exploration, we obtain a new descriptor $\left(\chi_{b2} r s_x{}^2 e^{r s_x}\right)^{-1}$ from data mining which expresses the heat of formation (ΔH_F) in the double perovskite. By our experiment, the method can obtain related expressions of structure-property relationship for material.The results show that the method is a simple yet efficient paradigm to construct the structure-property relationship and provides valuable hints to accelerate the process of materials design.

1. Introduction

Discoveries of the relationship between materials structure and properties come from the experiments, empirical or observational data [1]. Scientists got hypothesis and theories via correlations revealed by statistical analysis when they have enough data. With the integration of information science and materials science, machine learning (ML) is seen as an alternative to theoretical work [2,3]. In some researches, it shows that ML can prove original conclusions and discover new theories. For example, Kim, C. et al. used machine learning to get phenomenological theory about dielectric breakdown from organized High-Throughput data [4]. Li, Z. et al. used ML to verify the solid solubility theory of binary alloys and study other influencing factors [5], which shows that ML can indeed be used as a tool to help material theory

* Corresponding author at: School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China.
E-mail address: hrzhangsh@shu.edu.cn (H. Zhang).

https://doi.org/10.1016/j.matdes.2020.109194

research. The use of ML in materials science, however, has been hindered by the accuracy and interpretability of predictive models. Lack of interpretability of most ML predictive models prevents further mechanistic understanding, such as finding key ingredients for target properties [6]. It is difficult for us to obtain a clear relationship or expression to understand the association between structure and properties in this way. Thus, finding the appropriate ML algorithms that can achieve both accurate prediction and interpretability is crucial to the further advance of data-driven materials research [7,8].

In recent years, so as to make it easier to understand and find the structure-property relationship of a certain material, many researchers try to find descriptors to intuitively predict material properties through basic parameter combinations [9,10]. Many methods are specifically designed to find descriptors, like Symbol Regression [11], LASSO algorithm [12], and SISSO algorithm [13]. Materials property prediction based on descriptors is becoming a new approach in materials science. Although these methods can achieve good results, they need to rely on many conditions, such as large-scale data, suitable algorithms, sufficient features [14], and it is relatively difficult for materials researchers who are not familiar with computer algorithms to understand. So it is not easy and efficient to achieve ideal result.

In this paper, we put forward a framework combining feature engineering and linear regression to find the correlation between structure and properties. In the ML process, feature engineering is needed in the first place to remove redundant features and construct new descriptors, then the constructed descriptors are used to perform the regression analysis by linear regression (LR). The coefficient of determination (R^2) of the LR is used as the criterion for seleof the descriptors. Here, we aimed at the superconducting transition temperature (T_{C0}) for high temperature superconductors [15,16]. Comparing the expressions obtained by the method, a descriptor $(\ell\zeta)^{-1}$ which is the correlation with the transition temperature T_{C0}, with the expressions derived from theoretical derivation. It was found that this method could well verify the structure-property relationship about T_{C0}. As an exploration, this approach was also used to find the structure-property relationship about the heat of formation in double perovskite materials [17,18]. As a result, several relevant expressions were obtained by the method, for example, a new descriptor $\left(\chi_{b2} r s_x^{\,2} e^{r s_x}\right)^{-1}$, which indicates that this approach could be applied to other materials.

2. Methods

The whole framework and workflow schematically illustrated shows in Fig. 1, and it contains several steps as follow:

Step 1, collect available material data sets from different ways.

Step 2, construct the machine learning methods to assist target materials. In the ML process, the feature engineering method is used to construct new descriptors from the datasets. The regression model is used to train and fit the algebraic expression.

Step 3, obtain corresponding structure-property relationships, which should be verified by experiments or domain knowledge.

2.1. Material dataset

Here, we collected two batches of datasets. One dataset comes from Ref. [19]. It includes 36 compounds (high temperature superconductors), which were used in the first experiment, with T_{C0} is the value obtained through experimental measurement. Here, 2 original features(ζ,ℓ) and T_{C0} of each instance(Table.1) are included.

The other dataset is about the heat of formation (ΔH_F), which is calculated by electronic structures of the double perovskite AB_1B_2X refer to Ref [20]. It includes 540 perovskite materials, with the heat of formation computed by density functional theory (DFT), as called the ΔH_F of perovskite. Here, 33 original features and ΔH_F of each instance (Table.2) are selected, including chemical information and geometric information, such as bond length and crystal symmetry etc. For the convenience of the experiment, the abbreviations are used for the index of the dataset, and the detailed information of index is following: *a*, *b1*, *b2*, *x*: cations at A^-, B^{1+}, B^{3+}site, and anions at X^-site.

2.2. Data pretreatment

For any ML method that targets toward a prespecified material property, it usually depends on a certain amount of features. Although there may be many factors that affect the targeted property of materials, the number of features must be reasonable. The best strategy is to choose features that perfectly represent the materials' property [21]. Both the use of existing literature knowledge and the use of machine learning techniques can be used for feature selection [22,23]. On the one hand,

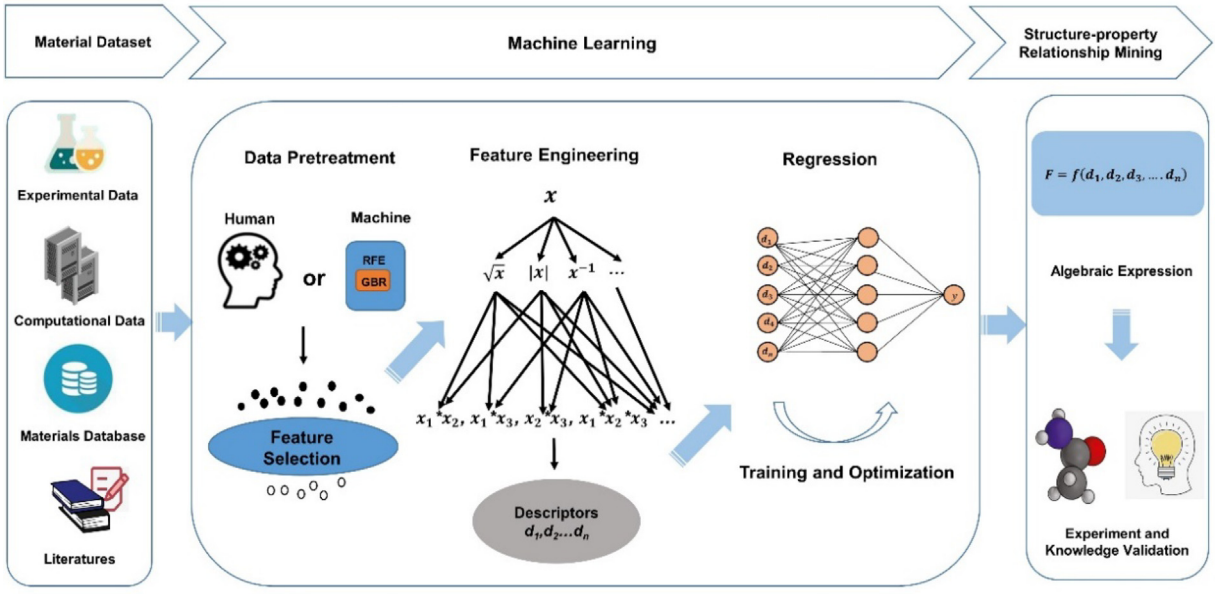


Fig. 1. Overall workflow of the structure-property relationship mining by machine learning assisting.

Table 1
A snapshot of the first 2 instances of the dataset in this work. The 2 original features and T_{c0} of each instance.

Feature	Feature description	Compound 1 $YBa_2Cu_3O_{6.92}$	Compound 2 $Tl_2Ba_2CuO_6$
ζ	the distance between interacting layers	2.2677	1.9291
ℓ	the calculated spacing between interacting charges within layers	5.7085	8.0965
T_{c0}	The value of measured	93.7	80

machine learning can help to obtain features whose importance cannot be determined from the existing literature knowledge. On the other hand, the features obtained from the literature knowledge can also help verify the correctness of machine learning methods. In this work, since the first batch of data has few features and it can be determined from the literature that the features are important. Considering time and efficiency, it is not necessary to use technique for feature selection. For the second batch of data, since there are more data features, and it is difficult to determine which are the important features from the literature. For more feature combinations, it is necessary to choose technique for feature selection. There are many methods for feature selection, such as LASSO [24], model-based feature selection [25], etc. LASSO reduces dimension by retaining one of several features that are equally relevant to the target value, it is useful when you need to reduce the number of features, but not so good for data comprehension and unselected but userful features [26]. However, in this work, we need to consider the importance of each feature and more feature combinations, Recursive Feature Elimination (RFE) is a better choice. RFE is a wrapper method of feature selection strategy, it is designed to select features by recursively minimizing the features with an external estimator, and the external estimator could be replaced by kinds of machine learning (ML) methods [27]. The main idea of RFE is to use a base model for multiple rounds of training. After each round of training, the features of some weight coefficients are eliminated, and then the next round of training is conducted based on the new feature set, then repeat the process over the remaining features until all the features are traversed [28].

In order to find a suitable estimator for RFE, several different types of regression estimators were tested:support vector regressor (SVR), random forests regressor (RFR), gradient boosting regressor (GBR), linear regressor (LR). SVR is suitable for handling nonlinear problems with the aid of nonlinear mapping features to high-dimensional feature space where the linear regression is conducted [29]. RFR uses a random method to construct multiple unrelated decision trees to form a forest. For a new sample, each decision tree will be judged separately, and finally the mean of the k models will be calculated as the final result [30]. GBR is an integrated learning algorithm, the learning principle of this method is to improve the accuracy of the final regression results by gradually reducing the error of the training process [31]. LR can be used for variable prediction, the purpose is to obtain the linear relationship between the output vector and the input feature, and find the linear regression coefficient [32].

In order to test the prediction effect of the model for material systems, as we all known, R^2(coefficient of determination) is a frequently used measure of the differences between predicted value and the target value, which has the advantage of providing some balanced measurement indexes that can be used as a criterion. MAE (mean absolute error) is to consider the mean error of the predicted value and the experimental value. At the same time, MSE (mean square error) assesses the quality of a target variables from a predictor [33]. Hence, three measures were used to evaluate the performance: R^2, MAE and MSE. The three computational formulas are shown in Table.3.

For observing the relationship between the structure features and the target property, some measures should be chosen. The most straightforward way is the value of the Pearson correlation coefficient (r) [34], which is a measure of the linear correlation between predicted and experimental output. The Pearson correlation coefficient(r) is defined as below:

$$r_{xy} = \frac{1}{n-1}\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{s_x s_y} \quad (1)$$

Table 2
A snapshot of the first 2 instances of the dataset in this work. The 33 original features and ΔH_F of each instance.

Feature	Feature description	Compound 1 Cs.Ag.As.Br	Compound 2 Cs.Ag.Al.Br
d_a	distance between cations at A^- site and anions at X^- site	3.925	3.870
d_{b1}	distance between cations at $B1^+$ site and anions at X^- site	2.858	2.912
d_{b2}	distance between cations at B^{3+} site and anions at X^- site	2.692	2.555
cubic	cubic crystal structures	1	1
ortho	orthorhombic crystal structures	0	0
χ_a	electronegativity of cations at A^- site	0.79	0.79
χ_{b1}	electronegativity of cations at B^{1+} site	1.93	1.93
χ_{b2}	electronegativity of cations at B^{3+} site	2.18	1.61
χ_x	electronegativity of anions at X^- site	2.96	2.96
h_a	highest occupied energy level of cations at A^- site	−2.086	−2.086
h_{b1}	highest occupied energy level of cations at B^{1+} site	−4.404	−4.404
h_{b2}	highest occupied energy level of cations at B^{3+} site	−5.194	−2.712
h_x	highest occupied energy level of anions at X^- site	−7.859	−7.859
i_a	ionization energy of cations at A^- site	3.893	3.893
i_{b1}	ionization energy of cations at B^{1+} site	7.576	7.576
i_{b2}	ionization energy of cations at B^{3+} site	9.788	5.985
i_x	ionization energy of anions at X^- site	11.813	11.813
l_a	lowest unoccupied energy level of cations at A^-site	−2.086	−2.086
l_{b1}	lowest unoccupied energy level of cations at B^{1+} site	−4.404	−4.404
l_{b2}	lowest unoccupied energy level of cations at B^{3+} site	−5.194	−2.712
l_x	lowest unoccupied energy level of anions at X^- site	−7.859	−7.859
rd_a	radius of d-orbital of cations at A^- site	0	0
rd_{b1}	radius of d-orbital of cations at B^{1+} site	0.385	0.385
rd_{b2}	radius of d-orbital of cations at B^{3+} site	0.155	0
rd_x	radius of d-orbital of anions at X^- site	0.143	0.143
rp_a	radius of p-orbital of cations at A^- site	2.6	2.6
rp_{b1}	radius of p-orbital of cations at B^{1+} site	1.33	1.33
rp_{b2}	radius of p-orbital of cations at B^{3+} site	0.745	0.905
rp_x	radius of p-orbital of anions at X^- site	0.62	0.62
rs_a	radius of s-orbital of cations at A^- site	1.71	1.71
rs_{b1}	radius of s-orbital of cations at B^{1+} site	1.045	1.045
rs_{b2}	radius of s-orbital of cations at B^{3+} site	0.67	0.77
rs_x	radius of s-orbital of anions at X^- site	0.58	0.58
ΔH_F	heat of formation in eV	−1.012	−1.240

Table 3
Three measure variables and statistical error measures applied for model comparison.

Measure variables	Expression
coefficient of determination (R^2)	$R^2 = 1-\frac{\{\sum_i(\hat{y}_i-y_i)\}^2}{\sum_i(\hat{y}_i-y_i)^2}$
mean squared error (MSE)	$MSE = \frac{1}{m}\sum_{i=1}^{m}(y_i-\hat{y}_i)^2$
mean absolute error (MAE)	$MAE = \frac{1}{m}\sum_{i=1}^{m}\mid(y_i-\hat{y}_i)\mid$

Table 4
A Pearson diagram of superconducting transition temperature and selected structural features.

Pearson correlation coefficient	T_{c0}	ζ	ℓ
T_{c0}	1	0.2	−0.86
ζ	0.2	1	−0.19
ℓ	−0.86	−0.19	1

Where x_i and y_i are the sample value of feature x and feature y, $\overline{x}$ and $\overline{y}$ are the mean value of feature x and feature y, s_x and $s_y$are the standard deviations of corresponding features , n represents the number of samples. The value of the coefficient is always between −1.0 and 1.0. If the value is close to 0, it means there is non-correlated. On the contrary, if the value is close to 1 or − 1, it means there is strong correlation. Table.4 gives the degree of association between the two structural features of the superconductors and the T_{c0}. It can be seen that the linear correlation between features and target property is not strong. Relying on these two features only, it is difficult to obtain the structure-property relationship.

2.3. Feature engineering

In order to quickly find the relationship between structure and property, we visualized the distribution of the features of the two batches of datasets. As shown in Fig. 2 and Fig. 3, the raw dataset of observed features is positive skewed and the range of data is very wide. Therefore, dataset transformation methods are necessary in this situation [35]. In the work, new descriptors are created by the feature engineering (Fig. 4). Feature engineering is essentially an engineering activity, which is designed to maximize the extraction of features from raw data by algorithms and models [36]. Therefore, feature engineering method could construct new descriptors and evaluated the best performance with selected subset.

In the feature engineering process, the selected features, 3 interactions,8 prototypical functions are employed to construct descriptors, and the descriptors were constructed by non-linear combinations [37].

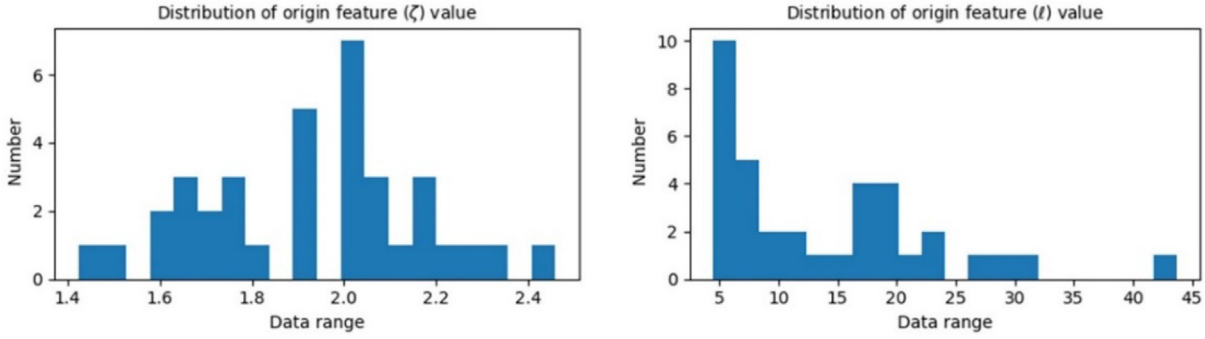


Fig. 2. Distribution of two features for T_{c0}.

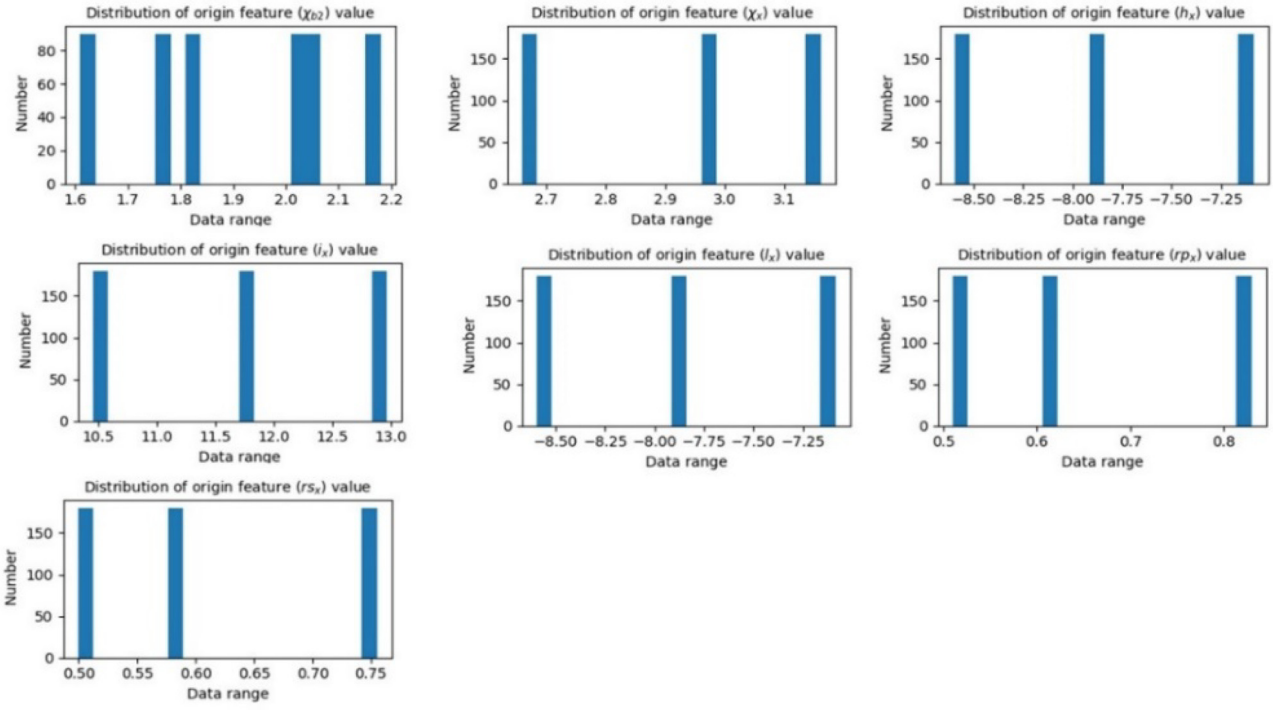


Fig. 3. Distribution of seven selected features for ΔH_f.

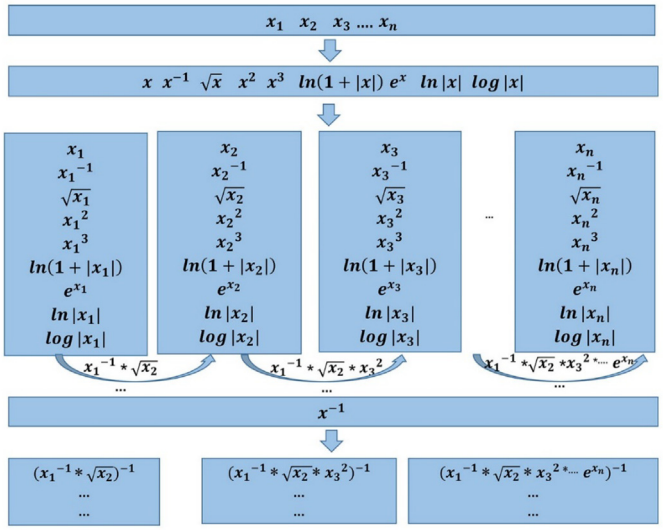


Fig. 4. The process of dimensionality augmentation based on feature engineering. The x_i represent the selected features. The parameter behind the arrow represents the generated new descriptors. There are 8 prototypical functions:x^{-1},$\sqrt{x}$, x^2, x^3,$ln(1+|x|)$, $ln|x|$, $e^x$and $log|x|$ to expand feature dimension, the descriptors are all nonlinearly combined in order to get more descriptors.

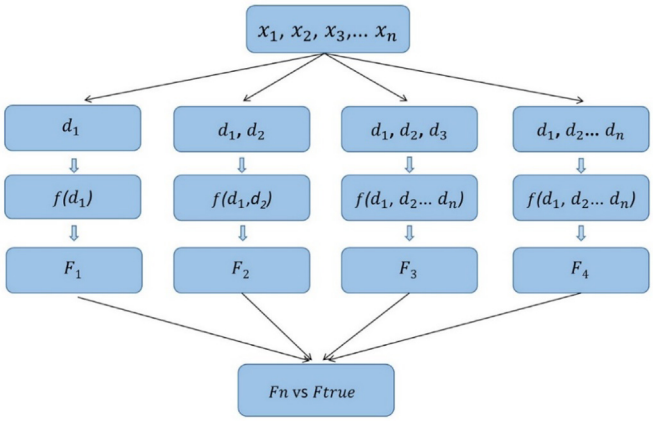


Fig. 5. The frame diagram of the linear model. The x_i represents the feature selected by the model, d_i represents the descriptor obtained after feature engineering, $f(d_1,d_2,d_3,\ldots. d_n)$ represents the process of fitting the descriptors, F_n can be roughly understood as the value of structure-property relationship, F_{true} can be roughly understood as the true value.

It was worth mentioning that the non-linear descriptors which represent meaningless combinations were not considered. Through the feature engineering method, a large number of descriptors have been added and the feature dimension have been expanded. From the large set of descriptors, looking for the suitable descriptors is the goal. The primary features constructed in the following way:

Step 1. Input selected important features into the 8 prototypical functions, namely: x^{-1}, $\sqrt{x}$, x^2, x^3, $ln(1+|x|)$, $ln|x|$,e^x and $log|x|$, with x being one of the primary features selected, which immediately lead to new descriptors.

Step 2. Multiply two or three descriptors obtained in the first step each time to get more descriptors.

Step 3. Put the new descriptors obtained in *Step 2* into the prototypical function x^{-1} to add more descriptors.

2.4. Regression

Machine learning algorithm is effective to conduct regression, and linear regression is also a type of machine learning. It is worth mentioning that linear regression is more suitable for fitting expression. This is because linear regression has many applications in some related materials research and got good results [38]. At the same time, the process of getting the expression is relatively simple and easy to understand, while for other machine learning algorithms, the formula parameters are complex and the process is not easy to understand [39]. Here, the linear regression(Fig. 5) is suitable to select the most suitable descriptor and fit the structure-property expression. In the experiment, a 60%/40% dataset split was found to be a reasonable compromise between the accuracy of the machine learning model and the tendency of the model to be over-fit. Finally, the most important descriptor was selected by comparing the effects of the model with the DFT or measured value.

2.5. Structure-property relationship mining – Experiment and knowledge validation

In this paper, we first collected the dataset of materials, and then used data preprocessing to select important features. Feature engineering constructed descriptors based on the selected important features, and linear regression was employed to select the most appropriate descriptors and fitted related structure-property relationship. Finally structure-property relationship was mined by comparing the effect of the model with measured value or the DFT value. At the same time, related domain knowledge is used to valid the accuracy and practicality of the model.

3. Result and discussion

3.1. Feature selection about ΔH_F

A detailed description of R^2 for testing set is shown in Fig. 6. It can be seen that the result of GBR shows better performance than other regressors (the orange line in Fig. 6). Since different algorithms have different dependence on data, in this section, GBR is a better estimation model for the present dataset compared with the others. Usually, GBR is a flexible non-parametric statistical machine leaning algorithm, and it is good at dealing with high-dimensional data. GBR algorithm evolves from the combination of boosting methods and regression trees, which makes it suitable for effectively mining features and feature engineering [40]. RFE builds the GBR model repeatedly through features, then selects the best features based on the coefficients, and then repeats this process on the remaining features until all features are traversed. In the end, there are 7 most important features for the heat of formation sorted out by GBR and constitute as an optimal feature set. The new feature set contains 7 features (the electronegativity of b2, the electronegativity of x, the highest occupied energy level of x, the ionization energy of x, the lowest unoccupied energy level of x, the radius of p-orbital for x, the radius of s-orbital for x).

3.2. Case study: Superconducting transition temperature (T_{C0})

Here, we take the superconducting transition temperature (T_{C0}) as an example to study the performance of model [41]. Based on the above work, the goal is to find the most suitable descriptor through the linear regression model. It is assumed that eventually some suitable descriptors will be fitted. As a result, it is expected to get a concrete expression which could express the structure-property relationship, which could be conceived as follow:

$$F = f(d_1, d_2, d_3, \ldots .d_n) \tag{2}$$

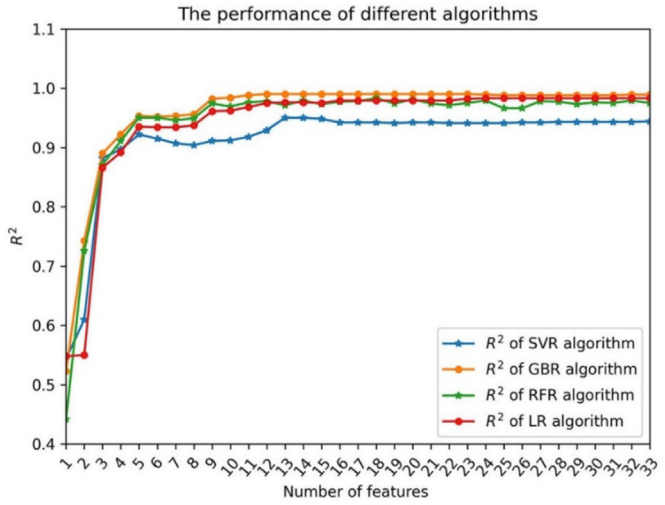


Fig. 6. Evaluations of machine learning algorithms by estimating the value of R^2 for different regression algorithms. The horizontal axis represents the features and the vertical axis represents the R^2 for different machine learning algorithms.

Where d_i represents the ultimate descriptors, F represents the target property, and $f(d_1, d_2, d_3, \ldots. d_n)$ represents an expression which can explain the relationship of structure-property.

As a data-driven research, we should consider the objectivity of the results and explain more information through domain knowledge. Some studies have shown that the superconducting transition temperature (T_{C0}) depends on crystal structure, cell parameters, ionic valences, and coulomb coupling between electronic bands in adjacent, spatially separated layers of high temperature superconductors, and T_{C0} can be given by the following algebraic expression [42]:

$$T_{C0} = {K_B}^{-1}\,\beta(\ell\zeta)^{-1} \tag{3}$$

Here, ℓ is related to the mean spacing between interacting charges in the layers, ζ is the distance between interacting electronic layers, β is a universal constant, and k_B is Boltzmann's constant. Here, for our generalization expression (2), F stands for T_{C0}, and d_i represents $(\ell\zeta)^{-1}$.

We first chose a single descriptor or multiple descriptors to describe the relationship. The Fig. 7.a gives a trend between the prediction effect of the model and the number of descriptors. It is clear that the prediction effect of the model decreases with the number of descriptors superimposed. This is the consequence of overfitting caused by feature engineering. When the low-dimensional descriptor is converted into a high-dimensional descriptor, it will increase the complexity of the model and cause the model to learn too much detail. As a result, it performs well on training set, but has a bad generalization ability on unseen sample. Thus, a single descriptor training linear model is better than multiple descriptors input together. As shown in Fig. 7.b, it displays the performances of each descriptor for the model. It can be seen that the prediction effect of the models constructed by different descriptors is different. Those descriptors that perform well have a strong correlation with the target property. In other words, looking for the most suitable descriptor is the key to understand the structure-property relationship.

Table 5
The descriptors selected by the model and three evaluation indicators with the measured values.

Method	No.	Descriptors	R^2	MAE	MSE
LR	1	$(\ell\zeta)^{-1}$	0.9985	1.1729	2.3285
	2	$(\ell\zeta^{1/2})^{-1}$	0.9934	2.8611	10.1529
	3	$\{\ell\ log\ (1+\zeta)\}^{-1}$	0.9954	2.4165	7.0045
	4	$(\ell\zeta^2)^{-1}$	0.9782	4.8977	33.4343
	5	$(\ell e^{\zeta})^{-1}$	0.9844	4.0728	23.8150

Then, we will choose the most suitable descriptor to train the linear regression model. As shown in Table.5, the model eventually led to five predictive expression for the superconducting transition temperature (T_{C0}). It can be seen that the best performing descriptor is $(\ell\zeta)^{-1}$, the scores of MAE and MSE is smaller than others, and the value of R^2 is 0.9985, which shows that the effect of this descriptor is very good.

A comparison between ML-predicted and measured results is presented in in Fig. 8. Obviously, the best-fit descriptor $(\ell\zeta)^{-1}$ is very consistent with the formula 3, and it is shown that T_{C0} may be obtained from an average of the coulomb interaction forces between the two layers, which also proves that this method can verify the theoretically formula. Furthermore, the R^2 can reach 0.9985, which means that the ML model is consistent with calculating method. In summary, our framework provides a possibility of achieving theory accuracy.

3.3. An exploration: The heat of formation (ΔH_F)

Recently, the ΔH_F is also the hot point in halide double perovskites research [43,44], in order to verify the applicability of this method, we give an exploration on a kind of more complex materials(halide double perovskites). In the study, we got six descriptors for the ΔH_F (Table.6), whereas all six descriptors are indistinguishable having the same value of MSE and almost same MAE. A little improved value in fourth digit is arguable (comparing No. 2, No.5, and No.6), so it is necessary to take liberty and provide some domain knowledge(obtained from the literature knowledge) expertise to determine the best descriptor.

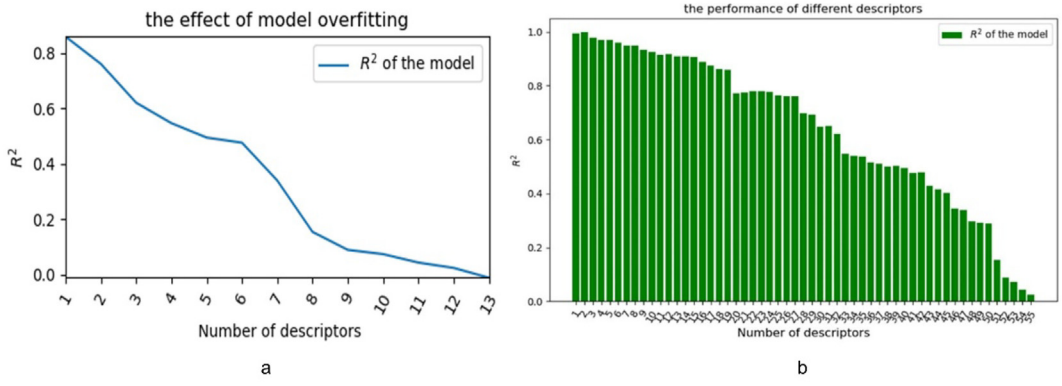


Fig. 7. a. The effect of model overfitting with the increasing of descriptor numbers. **b.** The prediction effect of a single descriptor with different sequence number. The abscissa represents the sequence number of the descriptor (a part of the descriptors selected), the ordinate represents R^2.

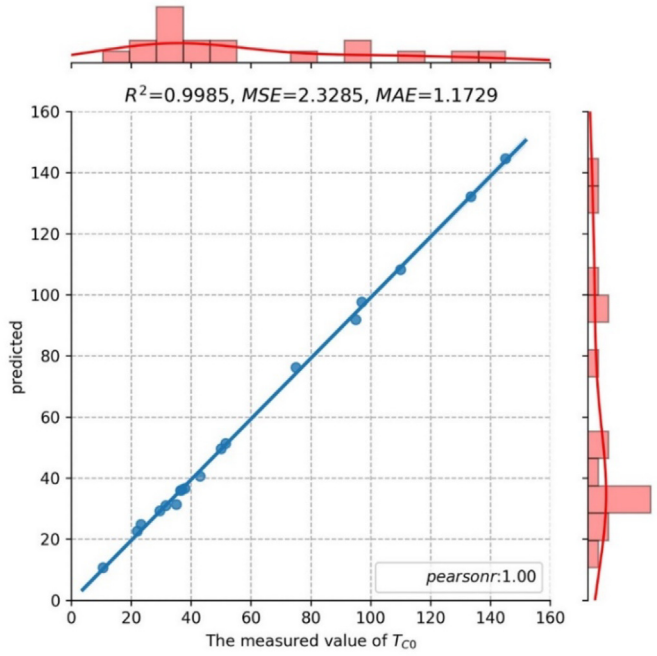


Fig. 8. Comparison of the measured value of T_{C0} with the predicted value of the model on the test set. Horizontal axis represents the true value of the target property. Vertical axis represents the model predictive value. The red part represents the data distribution.

Since the prediction effects of No. 1, No. 3, No. 4 are not good, these descriptors can be ignored.

At the same time, it shows that differences between electronegativity of bonding partners are good indicators of the bonding character and bonding strength, which strongly affect ΔH_F [45]. The high electronegativity of the compounds leads to tightly bound electronic distribution around the atoms and reflects strong hybridization via small bonding length [46]. In addition to the atomic levels, the electronegativity can play a crucial role in determining ΔH_F by controlling energy splitting between the bonding and antibonding states. Along with electronegativity (χ_{b2}), the radius (rs_x) is also a good indicator of bonding strength. Usually, strong bonding is accompanied by a shorter bond length. In that case, the radius of cation can affect the distances between cation and anion,and it can also affect the ΔH_F [47]. Therefore, for the selection of (No. 2, No.5, and No.6), we can only make a qualitative explanation about two parameters (χ_{b2}, rs_x) for selecting the appropriate descriptor. In the end, No.5 $(\chi_{b2} rs_x{}^2 e^{rs_x})^{-1}$ is the most suitable descriptor, which is consistent with previous known scientific knowledge well. Then we choose a simple linear expression:

$$\Delta H_F = a\left(\chi_{b2} rs_x{}^2 e^{rs_x}\right)^{-1} + b \tag{4}$$

Here, $a = -0.8067$ and $b = -0.4186$ are the values of coefficients, which were fit by the linear regression.

Compared with the DFT value, the reliability of the produced analytical expression for the heat of formation from the machine learning is tested. As shown in Fig. 9, this expression gives a coefficient of determination value $R^2 = 0.8854$ on the test, and the Pearson correlation coefficient between the calculated value of DFT and the predicted value is 0.94,which indicates the DFT calculations are consistent with the predictions. According to the analytical expression and the mechanism analysis, the ΔH_F in halide double perovskites has a stronger dependence on the electronegativity of *b2* and the orbital radius of *x* [48]. It reveals that the effect of electronegativity and radius is much more dominant than the other effects. This analytical expression reveals a clear relationship between ΔH_F and its structure. The presented model can be used to investigate the structure effects on the heat of formation in halide double perovskites. Based on the above analysis, it further validates the availability of the model and find the relevant expression for ΔH_F.

Table 6
The descriptors selected by the model and the comparison of the three evaluation indicators with the DFT values.

Method	No.	Descriptors	R^2	MAE	MSE
LR	1	$(\chi_{b2} i_x rs_x{}^3)^{-1}$	0.8845	0.067	0.006
	2	$(\chi_{b2} l_x rs_x{}^3)^{-1}$	0.8853	0.067	0.006
	3	$\left(\chi_{b2} l_x{}^3 rp_x{}^3\right)^{-1}$	0.8852	0.066	0.006
	4	$(\chi_{b2} rs_x{}^{5/2})^{-1}$	0.8846	0.066	0.006
	5	$(\chi_{b2} rs_x{}^2 e^{rs_x})^{-1}$	0.8854	0.067	0.006
	6	$(\chi_{b2}{}^{-1} e^{\chi_{b2}} rs_x{}^2)^{-1}$	0.8853	0.067	0.006

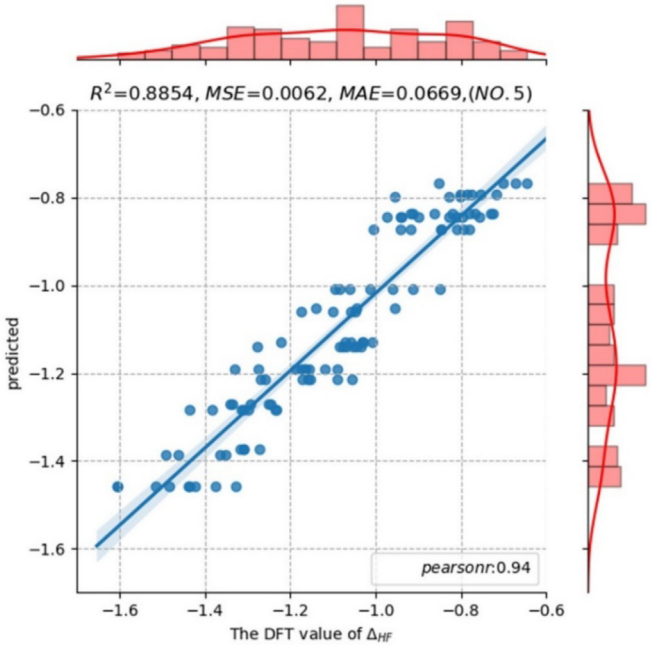


Fig. 9. Comparison of the DFT value of ΔH_F with the predicted value of the model on the test set. Horizontal axis represents the DFT value of the target property. Vertical axis represents the model predictive value. The red part represents the data distribution.

4. Conclusion

In this paper, based on the basic structure dataset, feature engineering is applied to construct new descriptors and linear regression to help us fit relationships between structure and properties in the field of materials. We put forward a framework combining feature engineering and linear regression to find the correlation between structure and properties from materials data by ML methods. New descriptors are constructed and used to perform the regression analysis and selected by LR. The results show a valuable ML framework on structure-property relationships. In order to check the present results, two materials systems were employed in high temperature superconductor and

double perovskites for solar cells. In high temperature superconducting system, we obtain the expression $(\ell\zeta)^{-1}$of the superconducting transition temperature. And in double perovskites for solar cells, we derive the expression $\left(\chi_{b2} r s_x{}^2 e^{r s_x}\right)^{-1}$ of the heat of formation ΔH_F.

The results show these experiment results validated the theoretically derived formula, at the same time, it supplemented structure-property relationship of the ΔH_F in perovskite materials. The present performance shows that this method can be combined with domain knowledge to effectively find a relationship or expression that can describe the association between structure and property. This provides us with new insights, and also provides some help for us to find some structure-property laws in material without prior knowledge. It also suggests that the current framework can be introduced to extract meaningful information from datasets and applied to solve inverse problems of material designs, which is very useful in discovering new materials and material manufacturing.

Data availability

The data that support the findings of this study are available from the corresponding author upon request.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by both the National Key Research and Development Program of China (No. 2018YFB0704400) and (No. 2016YFB0700502).

References

[1] K.T. Butler, D.W. Davies, C. Hugh, I. Olexandr, W. Aron, Machine learning for molecular and materials science, Nature. 559 (2018) 547–555.
[2] R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, C. Kim, Machine learning in materials informatics: recent applications and prospects, Npj Comput. Mater. 3 (2017) 54.
[3] P. Raccuglia, K.C. Elbert, P.D.F. Adler, C. Falk, M.B. Wenny, A. Mollo, M. Zeller, S.A. Friedler, J. Schrier, A.J. Norquist, Machine-learning-assisted materials discovery using failed experiments, Nature. 533 (2016) 73–76.
[4] Kim Chiho, Ghanshyam, Pilania, Rampi, Ramprasad, machine learning assisted predictions of intrinsic dielectric breakdown strength of ABX3 Perovskites, J. Phys. Chem. C (2016) 14575–14580.
[5] S. Li, H. Zhang, D. Dai, G. Ding, X. Wei, Y. Guo, Study on the factors affecting solid solubility in binary alloys: an exploration by machine learning, J. Alloy. Compd. 782 (2019) 110–118.
[6] D. Xue, P.V. Balachandran, J. Hogden, J. Theiler, D. Xue, T. Lookman, Accelerated search for materials with targeted properties by adaptive design, Nat. Commun. 7 (2016) 11241.
[7] P.V. Balachandran, Machine learning guided design of functional materials with targeted properties, Comput. Mater. Sci. 164 (2019) 82–90.
[8] A. Agrawal, A. Choudhary, Perspective: materials informatics and big data: realization of the “fourth paradigm” of science in materials science, Apl Mater. 4 (2016) 53208.
[9] L. Ward, C. Wolverton, Atomistic calculations and materials informatics: a review, Curr. Opin. Solid State Mater. Sci. 21 (2017) 167–176.
[10] L.M. Ghiringhelli, J. Vybiral, S.V. Levchenko, C. Draxl, M. Scheffler, Big Data of Materials Science: Critical Role of the Descriptor, Phys. Rev. Lett. 114 (2015) 105503.
[11] A. Lino, Á. Rocha, A. Sizo, Á. Rocha, Virtual teaching and learning environments: automatic evaluation with symbolic regression, J. Intell. Fuzzy Syst. 31 (2016) 2061–2072.
[12] S. Yuan, Z. Jiao, N. Quddus, S.I. Kwon, C. V Mashuga, Developing Quantitative Structure–Property Relationship Models To Predict the Upper Flammability Limit Using Machine Learning, Ind. Eng. Chem. Res. (2019) 3531–3537.
[13] R. Ouyang, S. Curtarolo, E. Ahmetcik, M. Scheffler, L.M. Ghiringhelli, SISSO: a compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates, Phys. Rev. Materials 2 (2018), 083802, .
[14] S.P. Ong, Accelerating materials science with high-throughput computations and machine learning, Comput. Mater. Sci. 161 (2019) 143–150.
[15] D.R. Harshman, A.T. Fiory, J.D. Dow, Theory of high-Tc superconductivity: transition temperature, Phys.Rev.Lett. 58 (2011) 2794.
[16] D.R. Harshman, A.T. Fiory, Superconducting interaction charge in thallium-based high-T_C cuprates: roles of cation oxidation state and electronegativity, J. Phys. Chem. Solids 85 (2015) 106–116.
[17] S. Lu, Q. Zhou, Y. Ouyang, Y. Guo, Q. Li, J. Wang, Accelerated discovery of stable lead-free hybrid organic-inorganic perovskites via machine learning, Nat. Commun. 9 (2018) 3405.
[18] D. Yao, X. Mao, X. Wang, Y. Yang, M.T. Hoang, A. Du, E.R. Waclawik, G.J. Wilson, H. Wang, The effect of ethylene-amine ligands enhancing performance and stability of perovskite solar cells, J. Power Sources 463 (2020) 228210.
[19] H. Zhang, Y. Zhang, D. Dai, M. Cao, W. Shen, Modelling and optimization of the superconducting transition temperature, Mater. Des. 92 (2016) 371–377.
[20] J. Im, S. Lee, T.-W. Ko, H.W. Kim, Y. Hyon, H. Chang, Identifying Pb-free perovskites for solar cells by machine learning, Npj Comput. Mater. 5 (2019) 37.
[21] M. Ankita, E.A. Holm, A comparative study of feature selection methods for stress hotspot classification in materials, Integr Mater Manuf Innov. 7 (2018) 87–95.
[22] G. Pilania, K.R. Whittle, C. Jiang, R.W. Grimes, C.R. Stanek, K.E. Sickafus, B.P. Uberuaga, Using machine learning to identify factors that govern Amorphization of irradiated Pyrochlores, Chem. Mater. 29 (2017) 2574–2583.
[23] C. Suh, K. Rajan, Invited review: data mining and informatics for crystal chemistry: establishing measurement techniques for mapping structure–property relationships, Met. Sci. J. 25 (2009) 466–471.
[24] Z. Jiao, S. Yuan, Z. Zhang, Q. Wang, Machine learning prediction of hydrocarbon mixture lower flammability limits using quantitative structure-property relationship models, Process. Saf. Prog. 39 (2020), e12103, .
[25] M. Toğaçar, B. Ergen, Z. Cömert, Classification of flower species by using features extracted from the intersection of feature selection methods in convolutional neural network models, Measurement. 158 (2020) 107703.
[26] R. Liu, M. Yuan, H. Xu, P. Chen, X.S. Xu, Y. Yang, Adaptive weighted sum tests via LASSO method in multi-locus family-based association analysis, Comput. Biol. Chem. 88 (2020) 107320.
[27] S. Sahran, D. Albashish, A. Abdullah, N.A. Shukor, S. Hayati Md Pauzi, Absolute cosine-based SVM-RFE feature selection method for prostate histopathological grading, Artif. Intell. Med 87 (2018) 78–90.
[28] Su Ran, Xinyi Liu, Wei Leyi, MinE-RFE: determine the optimal subset from RFE by minimizing the subset-accuracy-defined energy, Brief. Bioinform. 21 (2) (2020) 687–698.
[29] M.S. Ahmad, S.M. Adnan, S. Zaidi, P. Bhargava, A novel support vector regression (SVR) model for the prediction of splice strength of the unconfined beam specimens, Constr. Build. Mater. 248 (2020) 118475.
[30] K. Liou, S.-Y. Ooi, Resting full-cycle ratio (RFR) in the assessment of left Main coronary disease: caution required, Hear. Lung Circ. 29 (2020) 1256–1259.
[31] G.L. Abe, J.-I. Sasaki, C. Katata, T. Kohno, R. Tsuboi, H. Kitagawa, S. Imazato, Fabrication of novel poly(lactic acid/caprolactone) bilayer membrane for GBR application, Dent. Mater. 36 (2020) 626–634.
[32] A. Rahbari, T.R. Josephson, Y. Sun, O.A. Moultos, D. Dubbeldam, J.I. Siepmann, T.J.H. Vlugt, Multiple linear regression and thermodynamic fluctuations are equivalent for computing thermodynamic derivatives from molecular simulation, Fluid Phase Equilib. 523 (2020) 112785.
[33] D. Dai, T. Xu, X. Wei, G. Ding, Y. Xu, J. Zhang, H. Zhang, Using machine learning and feature engineering to characterize limited material datasets of high-entropy alloys, Comput. Mater. Sci. 175 (2020) 109618.
[34] J.M. Rickman, T. Lookman, S.V. Kalinin, Materials Informatics: From the Atomic-Level to the Continuum, Acta Mater. 168 (2019) 473–510.
[35] S. Yuan, Z. Zhang, Y. Sun, J. Kwon, C. Mashuga, Liquid flammability ratings predicted by machine learning considering Aerosolization, J. Hazard. Mater. 386 (2019) 121640.
[36] O. Isayev, C. Oses, C. Toher, E. Gossett, S. Curtarolo, A. Tropsha, Universal fragment descriptors for predicting electronic properties of inorganic crystals, Nat. Commun. 8 (2017) 15679.
[37] C. Wang, C. Shen, Q. Cui, C. Zhang, W. Xu, Tensile property prediction by feature engineering guided machine learning in reduced activation ferritic/martensitic steels, J. Nucl. Mater. 529 (2020) 151823.
[38] Shen Zhong-Hui, Wang Jian-Jun, Jiang Jian-Yong, X. Huang Sharon, Yuan-Hua, Phase-field modeling and machine learning of electric-thermal-mechanical breakdown of polymer-based dielectrics, Nat. Commun 10 (2019) 1843.
[39] M. Zeng, S. Yuan, D. Huang, Z. Cheng, Accelerated Design of Catalytic Water-Cleaning Nanomotors via machine learning, ACS Appl. Mater. Interfaces 11 (2019) 40099–40106.
[40] A.D. Sendek, E.D. Cubuk, E.R. Antoniuk, G. Cheon, Y. Cui, E.J. Reed, Machine learning-assisted discovery of solid Li-ion conducting materials, Chem. Mater. 31 (2) (2018) 342–352.
[41] Y. Zhang, X. Xu, Yttrium barium copper oxide superconducting transition temperature modeling through gaussian process regression, Comput. Mater. Sci. 179 (2020) 109583.
[42] D.R. Harshman, A.P. Mills, Concerning the nature of high-Tc superconductivity: Survey of experimental properties and implications for interlayer coupling, Phys. Rev. B 45 (1992) 10684.
[43] P.V. Balachandran, B. Kowalski, A. Sehirlioglu, T. Lookman, Experimental search for high-temperature ferroelectric perovskites guided by two-step machine learning, Nat. Commun. 9 (2018) 1668.
[44] H. Weici, F. Peter, C. Paulette, P. Matthias, Efficient search of compositional space for hybrid organic–inorganic perovskites via Bayesian optimization, Npj Comput. Mater. 4 (2018) 51.

[45] G. Volonakis, M.R. Filip, A.A. Haghighirad, N. Sakai, B. Wenger, H.J. Snaith, F. Giustino, Lead-free halide double perovskites via heterovalent substitution of noble metals, J. Phys. Chem. Lett. 7 (2016).

[46] W. Chen, Y. Wu, Y. Yue, J. Liu, W. Zhang, X. Yang, H. Chen, E. Bi, I. Ashraful, M. Gratzel, Efficient and stable large-area perovskite solar cells with inorganic charge extraction layers, Sci. 350 (2015) 944–948.

[47] A.A. Emery, C. Wolverton, High-throughput DFT calculations of formation energy, stability and oxygen vacancy formation energy of ABO3 perovskites, Sci. Data. 4 (2017) 170153.

[48] G. Hautier, S. Ong, A. Jain, C. Moore, G. Ceder, Accuracy of density functional theory in predicting formation energies of ternary oxides from binary oxides and its implication on phase stability, Phys. Rev. B 85 (2012) 155208.