Computational Materials Science 151 (2018) 41–48

Contents lists available at ScienceDirect

Computational Materials Science

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

Accelerated search for perovskite materials with higher Curie temperature based on the machine learning methods

Xiuyun Zhai[a,b], Mingtong Chen[c], Wencong Lu[d,*]

[a] College of Materials Science and Engineering, Shanghai University, Shanghai 200444, China
[b] School of Mechanical Engineering, Panzhihua University, Panzhihua 617000, China
[c] Material Engineering School, Panzhihua University, Panzhihua 617000, China
[d] Department of Chemistry, College of Sciences, Shanghai University, Shanghai 200444, China

ARTICLE INFO

Keywords:
Perovskite materials
Curie temperature
Machine learning
Support vector machine
Relevance vector machine
Random forest

ABSTRACT

Curie temperature (Tc), the second order phase transition temperature, is also one of the important physical properties of perovskite materials. It is a meaningful work to quickly and efficiently predict Tc of new perovskite materials before doing a considerable amount of experimental work. In the work, SVM (support vector machine), RVM (relevance vector machine) and RF (random forest) were employed to establish the prediction models of Tc with the physicochemical parameters, respectively. The results reveal that the three models all have high precision and reliability. According to K-fold cross validation, the SVR model had better prediction performance than the RVM and RF models. Meanwhile, the potential perovskite material with higher Tc was found by using the SVR model integrated with the search strategy of genetic algorithm from the virtual samples. The methods outlined here can provide valuable hints into the exploration of materials with desired property and can accelerate the process of materials design.

1. Introduction

Perovskite-type oxides, commonly represented by ABO_3, have been considered as one of the most promising materials due to the application of electronic and magnetic components such as multilayer capacitors and sensors [1,2]. In the properties of perovskite materials, Curie temperature (Tc), also called Curie point, is the phase transition temperature of ferroelectrics from ferroelectric phase to paraelectric phase. So, it has an important influence on many applications of perovskite materials such as erasing and writing new data of magneto-optical storage medium, temperature control of soldering irons, and stabilizing the magnetic field of tachometer generators against temperature variation. In recent years, many researchers attempt to synthesize perovskite materials with high Tc or higher Tc than room temperature. Therefore, the effects of different doping elements on Tc of perovskite materials have been widely reported [3,4]. Yu et al. [5] synthesized a very complex perovskite material ($Pb_{0.6}Bi_{0.4}Ti_{0.75}Zn_{0.15}Fe_{0.1}O_3$) with Tc of 978 K. However, it is a challenge to break through existing Tc in that the compositions of perovskite materials and different doping ratios of elements are highly complex.

At present, materials design with assistance of machine learning methods, promoted by efforts such as the Materials Genome Initiative, has become a research hotspot and an alternative approach to trial-and-error experiments. Pilania et al. [6] constructed a model to predict the bandgaps of double perovskites with the help of the machine learning methods. Raccuglia et al. [7] used the resulting data of failed experiments to train a machine-learning model to predict reaction success. Xue et al. [8] provided an adaptive approach and employed the machine learning regression algorithms to find very low thermal hysteresis (ΔT) NiTi-based shape memory alloys. Accordingly, it is no doubt that machine learning methods can shorten the cycle of materials design and realize controllable synthesis of materials.

In this work, a slew of machine learning methods was employed to forage for the model with the optimal regression performance to predict Tc of perovskite materials. To develop a really useful machine-learning predictor for a material or biological system as reported in a series of recent publications [9–23], one should observe the Chou's 5-step rule [24]; i.e., making the following five steps very clear: (i) how to construct or select a valid benchmark dataset to train and test the predictor; (ii) how to formulate the samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted; (iii) how to introduce or develop a powerful algorithm (or engine) to operate the prediction; (iv) how to properly perform cross-validation tests to objectively evaluate the anticipated accuracy of

* Corresponding author.
E-mail address: wclu@shu.edu.cn (W. Lu).

https://doi.org/10.1016/j.commatsci.2018.04.031
Received 31 March 2018; Received in revised form 13 April 2018; Accepted 15 April 2018

Table 1
The twenty-one descriptors of perovskite materials.

No.	Meanings	Features
1	Weighted ionic radii of A-site (Å)	R_a
2	Weighted ionic radii of B-site (Å)	R_b
3	Weighted electronegativity Pauling of A-site	χ_{pa}
4	Weighted electronegativity Pauling of B-site	χ_{pb}
5	Tolerance factor	t
6	Unit cell lattice edge (Å)	$\alpha_O{}^3$
7	Critical radii (Å)	r_c
8	Weighted ionization energy of A-site (kJ/mol)	I_{1a}
9	Weighted ionization energy of B-site (kJ/mol)	I_{1b}
10	Molecular mass (g/mol)	M
11	Ratio of ionic radii of A-site to B-site	R_a/R_b
12	Weighted electron affinity of A-site (eV)	EA_a
13	Weighted electron affinity of B-site (eV)	EA_b
14	The melt point of A-site metal (°C)	t_{ma}
15	The melt point of B-site metal (°C)	t_{mb}
16	The boil point of A-site metal (°C)	t_a
17	The boil point of B-site metal (°C)	t_b
18	The enthalpy of fusion of A-site (kJ/mol)	$\Delta_{fus}H_a$
19	The enthalpy of fusion of B-site (kJ/mol)	$\Delta_{fus}H_b$
20	The density of A-site metal (g/cm^3)	ρ_a
21	The density of B-site metal (g/cm^3)	ρ_b

the predictor; (v) how to establish a user-friendly web-server for the predictor that is accessible to the public. The fifth step is the direction of our future work that we will provide a web-server for the prediction tools presented in this paper. Below, we are to describe how to deal with these steps one-by-one.

The main outcomes of the paper are: the support vector regression (SVR), relevance vector machine (RVM) and random forest (RF) models were constructed with the better regression and generalization performances according to the K-fold cross validation; meanwhile, the SVR model has been verified that it has the stunning performance; finally, by the means of the SVR model and genetic algorithm (GA), the candidate perovskite material with perhaps higher Tc was provided to guide future researches and experiments, and then accelerate search for perovskite materials with higher Tc.

2. Methods

2.1. Dataset

We collected forty-seven perovskite materials from nine references [25–33] as the dataset in Tab S1 of supplementary information. There are several types of elements with different doping ratio both in A-site and B-site of the samples in the dataset to provide the conditions for following screening perovskite materials. The range of the target value (Tc) is from 170 K to 380 K. Besides, there are twenty-one physico-chemical parameters [34] (in Table 1) as the descriptors of perovskite materials and the candidate inputs of the models.

Table 2
The results of GA feature selection.

Algorithm	The selected features
SVR	χ_{pb}, r_c, R_a/R_b, EA_a, t_{mb}, t_a
RVM	R_a, χ_{pa}, t, r_c, I_{1a}, R_a/R_b, EA_a, t_a, $\Delta_{fus}H_a$, ρ_b

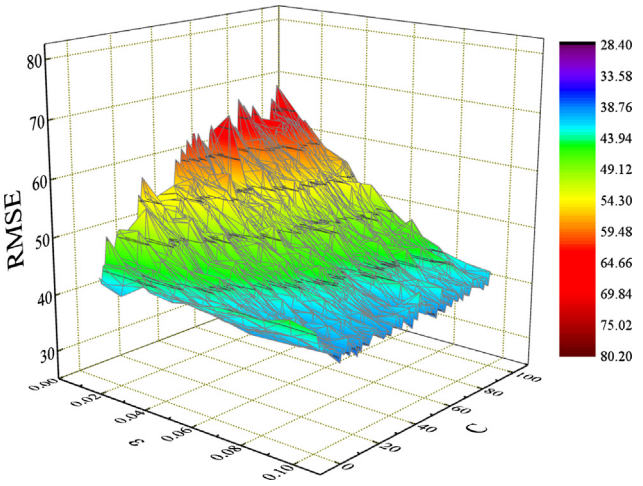


Fig. 2. The optimization process of hyper-parameter of the SVR model in GA.

Table 3
The list of hyper-parameters of SVR and RVM.

Algorithm	Hyper-parameters
SVR	C = 2; σ = 0.2; ε = 0.01
RVM	γ = 0.171

2.2. SVR

SVR [35–37], a powerful methodology for solving problems in nonlinear classification and regression, is also a supervised learning algorithm that has been widely applied to various fields. It considers the balance between empirical risk and expected risk, and then makes computational model have the good prediction and generalization performances.

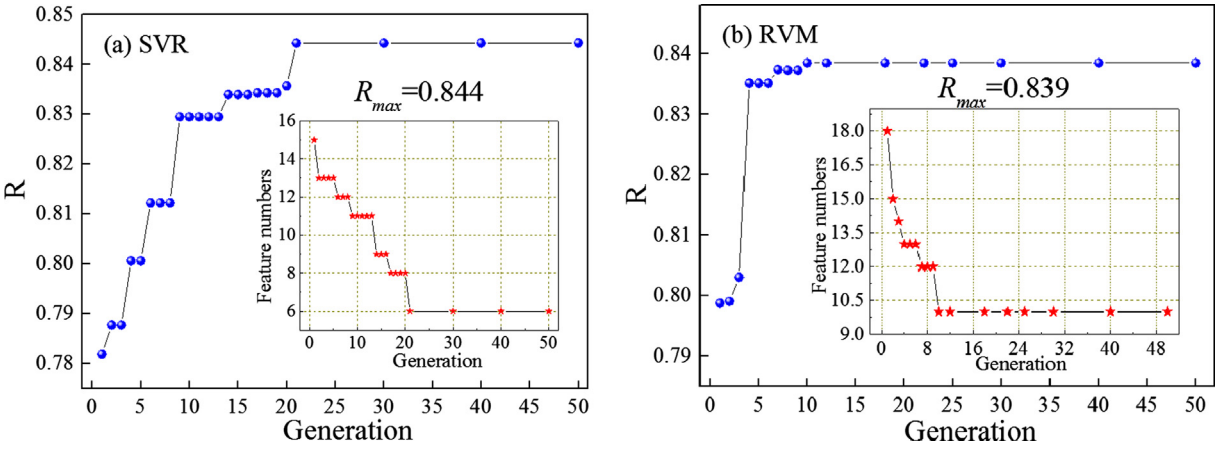


Fig. 1. The R versus generation of the evolution process in GA.

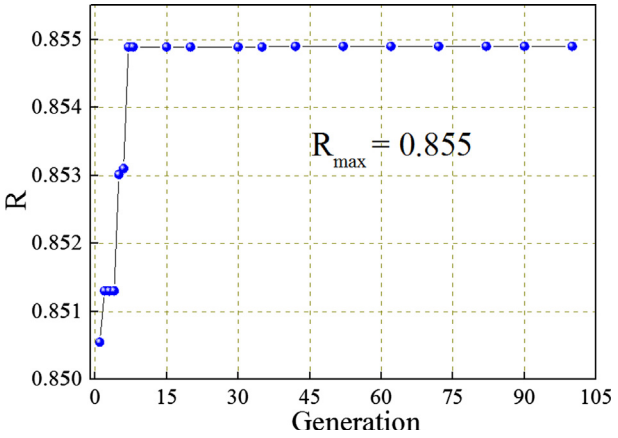


Fig. 3. The optimization process of hyper-parameter of the RVM model in GA.

A flat function, $f(x)$, is derived and has most ε deviation for each training input in SVR. In the linear case, $f(x)$ is defined as:

$$f(x) = \langle \boldsymbol{w},\boldsymbol{x} \rangle + b \tag{1}$$

where $\boldsymbol{w}$ is the input pattern space; $\langle \boldsymbol{w},\boldsymbol{x} \rangle$ is the dot product of vectors $\boldsymbol{w}$ and $\boldsymbol{x}$. The following SVR approximation with ε precision is constructed:

$$\text{Minimize } \frac{1}{2}|\boldsymbol{w}|^2 \tag{2}$$

$$\text{Subject to } \begin{cases} y_i - \langle \boldsymbol{w},x_i \rangle - b \leqslant \varepsilon \\ \langle \boldsymbol{w},x_i \rangle + b - y_i \leqslant \varepsilon' \end{cases} \tag{3}$$

here $\boldsymbol{i}$ is the number of sample points.

The optimization problem in dual form via Lagrange theory can be written as the Lagrange function as follows:

$$\text{Maximize } \frac{1}{2}\sum_{i,j=1}^{N}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\langle x_i,x_j^* \rangle - \varepsilon\sum_{i=1}^{N}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{N} y_i(\alpha_i - \alpha_i^*) \tag{4}$$

$$\text{Subject to } \sum_{i=1}^{N}(\alpha_i - \alpha_i^*) = 0 \ and \ \alpha_i,\alpha_i^* \in [0,C] \tag{5}$$

where N is the number of samples. α_i and α_i^* are the optimized Lagrange multipliers. Parameter C is a regularized constant determining the trade-off between the training error and the model flatness.

By introducing the kernel function (K), the original input is first nonlinearly mapped into the feature space, and then the SVR can be used for complicated nonlinear regression problems. The SVR mathematical form is as follows:

$$f(x) = \sum_{i=1}^{N}(\alpha_i + \alpha_i)K\langle x_i,x \rangle + b \tag{6}$$

$$\text{b} = y_i - \sum_{i=1}^{N}(\alpha_i + \alpha_i)K\langle x_i,x \rangle + \varepsilon \tag{7}$$

In the study, the SVR model applies Gaussian kernel function (RBF) shown as following:

$$\text{K}(x_i,x) = \exp\left(\frac{-\|x_i - x_j\|^2}{\sigma^2}\right) \tag{8}$$

2.3. RVM

RVM [38,39], firstly proposed by Tipping in 2001, has been successfully applied in the field of machine learning. It uses Bayesian inference to obtain parsimonious solutions for regression and probabilistic classification [40–42]. Moreover, the most compelling feature of the RVM is that it typically utilizes significantly fewer kernel functions compared to the SVM, while providing a similar performance.

Supervised learning techniques make use of a training set that consists of a set of sample input vectors $\{x_n\}_{n=1}^N$ together with the corresponding targets $\{t_n\}_{n=1}^N$, and t_n is expressed as follows:

$$t_n = y(x_n,w) + \varepsilon_n \tag{9}$$

where $\text{w} = (w_1,w_2,...,w_M)^T$ is the weight vector; ε_n is the error which is assumed to be mean-zero Gaussian noise with variance σ^2.

The regression function of RVM with a linear combination of the weighted kernel functions is presented as follows:

$$\text{y(x,w)} = \sum_{n=1}^{N} w_n K(x,x_n) + w_0 \tag{10}$$

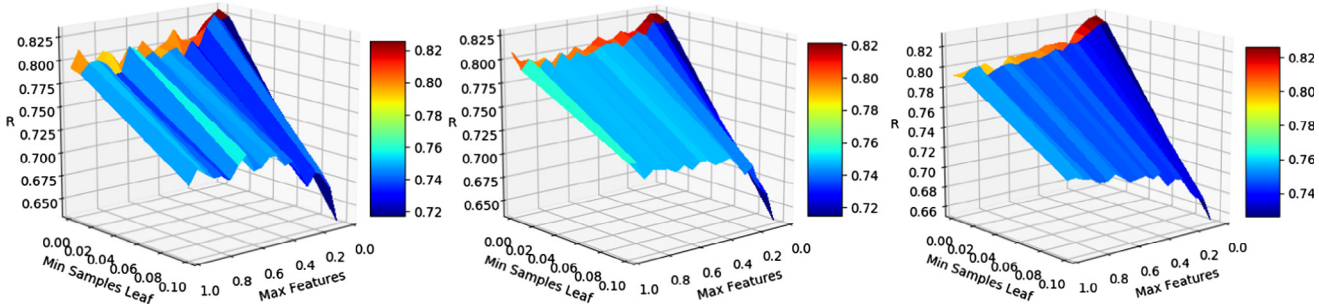
where $K(x,x_n)$ is the kernel function; w_0 is the bias. In the work, Gaussian RBF kernel was used and expressed as follows:

$$K_{Gaussian}(x_i,x_j) = \exp\left(\frac{-\|x_i - x_j\|^2}{\gamma^2}\right) \tag{11}$$

here γ is the kernel parameter that can be adjusted to optimize the RVM model. The target values obey the Gaussian distribution with mean $y(x_n)$ and the variance (σ^2), $p(t_n|x) \sim N(t_n|y(x_n),\sigma^2)$. The similar function of the data set is described as follow:

$$\text{p(t|w},\sigma^2) = \prod_{n=1}^{N} p(t_n|\text{w},\sigma^2) = (2\pi\sigma^2)^{-\frac{N}{2}}\exp\left\{-\frac{1}{2\sigma^2}\|t - \varnothing w\|^2\right\} \tag{12}$$

where $\Phi = [\varnothing(x_1),\varnothing(x_2),...,\varnothing(x_n)]^T$; $\varnothing(x_i) = [1,K(x_i,x_1),...,K(x_i,x_n)]^T$. It is assumed that w satisfies the Gaussian distribution with zero-mean and variance α_i^{-1}:



(a) Estimators=100 (b) Estimators=250 (c) Estimators=500

Fig. 4. Hyper-parameters optimization in RF.

Table 4
The hyper-parameters of RF.

Algorithm	Hyper-parameters
RF	Estimators = 100
	Max features = 0.05
	Min samples leaf = 0.001

$$p(w|\alpha) = \prod_{n=0}^{N} N(w_n|0,\alpha_n^{-1}) \tag{13}$$

here $\alpha = [\alpha_0,\alpha_1,...,\alpha_n]^T$ is a vector of $n + 1$ hyper-parameters that determines how far can deviate from each zero weight. For more details of RVM, see the Ref. [38,39].

2.4. RF

Based on decision tree algorithms [43], random forest (RF) [44–46] is an ensemble learning approach adopted in classification and regression analysis with preferable generalization performance. RF outperforms decision tree algorithms because it can correct for decision trees' habit of overfitting to the training set.

RF is an ensemble of B subtrees $\{T_1(X),...,T_2(X)\}$, where $X = \{x_1,...,x_p\}$ is a p-dimensional vector features associated with a target variable. The ensemble produces B outputs, $\{\hat{Y}_1 = T_1(X),...,\hat{Y}_B = T_B(X)\}$, where $\hat{Y}_b$, $b = 1,...,B$, is the prediction for a target variable by the bth tree. Outputs of all trees, $\hat{Y}$, are aggregated to produce one final prediction. $D = \{(X_1,Y_1),...,(X_n,Y_n)\}$ is a training set, where X_i is a vector of features and Y_i is a value of target variable; $i = 1,...,n$. The training process of RF [47] is described as follows.

(1) The bootstrap samples are drawn from the training data.
(2) Each bootstrap sample grows a tree with the following modification: at each node, the best split is selected among a randomly selected subset of $M_{feature}$ (rather than all) features. The tree is grown to the maximum size and not pruned back.
(3) Repeat the above steps until such B (a sufficiently large number) trees are grown.

So, RF is an ensemble of unpruned classification or regression trees created by using bootstrap samples of the training data and random feature selection in tree induction, and then outputs the class during classification or mean prediction of the individual trees during regression. Since it was proposed, RF has become a well-known data analysis method that has been applied to a wide variety of scientific areas.

The correlation coefficient (R), root mean square error (RMSE), mean squared error (MSE), mean absolute error (MAE) and mean relative error (MRE) of K-fold cross validation (K = 20) are used as the model evaluation metrics. Obviously, the model with higher R and lower RMSE, MSE, MAE and MRE is better.

2.5. Implementation

The calculations were performed on Online Computational Platform of Material Data Mining (OCPMDM) developed by us. It can be freely used on the website of the Laboratory of Materials Data Mining in Shanghai University (http://materialdata.shu.edu.cn). Its predecessor is HyperMiner software package [46,48] written by us. Its free version can be downloaded from URL: http://chemdata.shu.edu.cn:8080/MyLab/Lab/download.jsp.

3. Result and discussion

3.1. Feature selection

Before developing the SVM and RVM models, feature selection, a key factor to determine a success model, can reduce the dimension of feature space to further decrease the risk of over fitting, and can better remove features unrelated to target value and noise interference. Meanwhile, it can also make the training time shorten, and further promote the prediction ability and generalization performance of the models. Apparently, the performances of the models can be significantly impacted if irrelevant features are not removed prior to training. In the work, GA [49,50] and greedy feature selection [51] are adopted to select independent variables that are used to form the optimal feature set. GA is a meta-heuristic algorithm inspired by the process of natural selection and belongs to the larger class of evolutionary algorithms (EA). Compared with other optimization algorithms, GA has an ability to move from local optima present on the response surface and can work out a wide variety of optimization with the requirement of no knowledge or gradient present about the response surface.

According to K-fold cross validation results, GA outperformed greedy feature selection. Fig. 1 illustrates how GA was used to search for the optimal feature sets. From Fig. 1, the biggest R occurs after the evolutions of 20 and 9 generations for SVR and RVM algorithms respectively, where the two optimal features sets were found and included six and ten parameters (in Table 2) as inputs of the SVR and RVM models, respectively.

Decision Tree's performance is customarily insensitive to the presence of irrelevant features in that feature selection is intrinsic to the tree growing process (embedded variable selection [52]). Ensembles of trees should be even more capable of avoiding the influence of irrelevant descriptors. Instead, the relationship between the features and the target is hidden inside a "black box". So, RF cannot gain much in accuracy if feature reduction is implemented [47], and then it needn't implement feature selection.

3.2. Optimizing hyper-parameters

Most modeling tools require a modest amount of parameter tuning to lead the model to achieve optimal performance. In the work, GA and

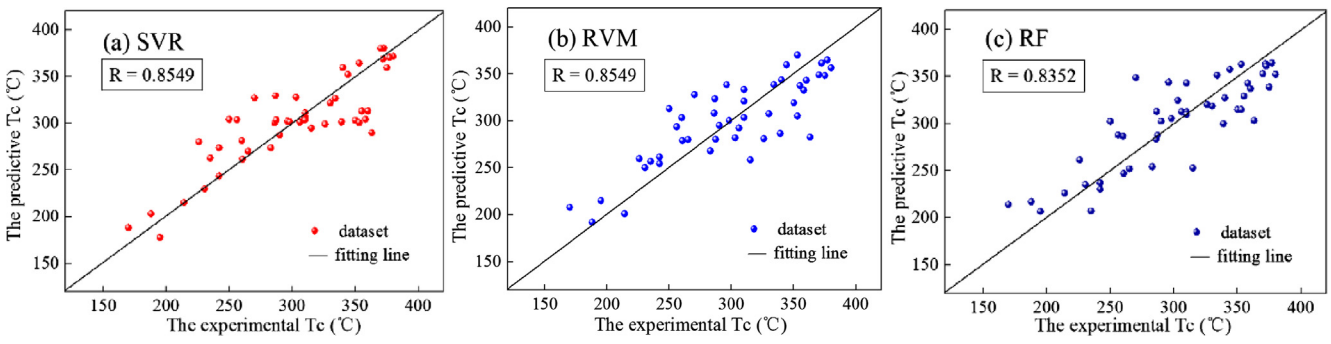


Fig. 5. The experimental versus the predictive Tc of the (a) SVR, (b) RVM and (c) RF models for K-fold cross validation.

Table 5
The results of the three models by K-fold validation.

Algorithm	R	RMSE	MSE	MAE	MRE
SVR	0.8549	28.6659	821.74	21.2125	0.0725
RVM	0.8549	28.6691	821.92	22.7540	0.0796
RF	0.8352	30.3969	923.97	24.4089	0.0842

grid search method were employed to optimize hyper-parameters of SVR and RVM algorithms to further improve the performance of the models. The RMSE and R between experimental values and predictive values of K-fold cross validation are defined as the fitness function evaluations of hyper-parameters optimization of the SVR and RVM models, respectively. For the nonlinear SVR-RBF model, its generalization performance depends on the suitable setting of meta-parameters C, σ and ε. Parameter C is a constant that determines regularized penalty to estimation errors and was set from 0 to 100 with step 1. Parameter σ of RBF kernel function changed from 0.5 to 1.4 with step 0.1. Parameter ε is an important parameter that can prevent the entire training set from meeting boundary conditions and was set from 0.01 to 1.0 with step 0.01. The optimization processes and results of the hyper-parameters are shown in Fig. 2 and in Table 3, respectively.

For RVM model, only γ of RBF kernel function need be optimized. The optimization processes and results of hyper-parameters are shown in Fig. 3 and in Table 3, respectively. The SVM and RVM models with highest accuracy can be achieved when $RMSE_{min}$ = 28.666 and R_{max} = 0.855, respectively.

In the work, the grid search is used as the strategy to optimize the three hyper-parameters (namely, estimators, max features and min samples leaf) of the RF model. 'Estimators' presents the number of the subtrees in RF. 'Max features' indicates the maximum number of features allowed to be used by a subtree of RF. 'Max features' is equal to 0.2, meaning that the features that can be used by a subtree account for 20% of the total features. 'Min samples leaf' is used to restrict the percent that minimum sample number of leaf nodes accounts for the total. In the optimization process, 'Estimators' was set to select from one of 100, 250 and 500; 'Max features' changes from 0.0 to 1.0 with step 0.05; 'Min sample leaf' makes a choice from 0.1, 0.01 and 0.001. The processes and results of the optimization of the hyper-parameters are shown in Fig. 4 and in Table 4, respectively.

3.3. Developing the models

The three models were evaluated by K-fold cross validation as shown in Fig. 5 and Table 5. Comparative analyses show that the SVR model has the best performance because of highest R and lowest RMSE, MSE, MAE and MRE.

To further analyze the models, we plotted the distribution maps of the prediction residuals for each model as shown in Fig. 6. It can be found that the residuals basically fit the normal distributions. In the three models, the continuities of residual distributions of the RVM and

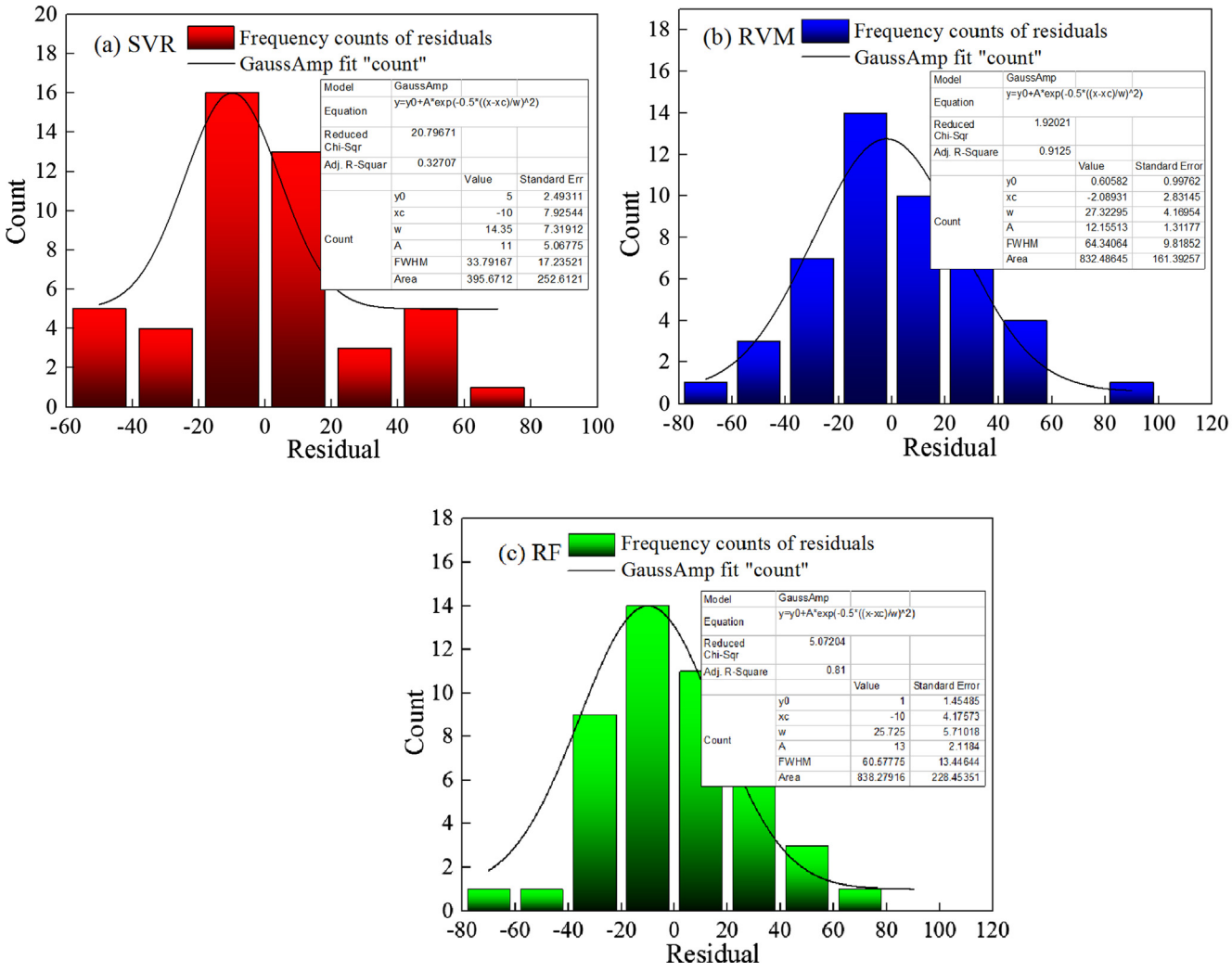


Fig. 6. The prediction residual distributions of the (a) SVR, (b) RVM and (c) RF models.

Table 6
The predictive results of the samples of the testing set.

No.	Tc	$Tc_{(SVR)}$	$e_{(SVR)}$	$Tc_{(RVM)}$	$e_{(RVM)}$	$Tc_{(RF)}$	$e_{(FR)}$
1	380	370.124	−9.876	345.971	−34.029	359.159	−20.841
2	377	360.919	−16.081	351.626	−25.374	349.503	−27.497
3	375	360.033	−14.967	334.755	−40.245	351.786	−23.214
4	372.5	368.314	−4.186	350.114	−22.386	352.234	−20.266
5	372	356.356	−15.644	351.438	−20.562	347.942	−24.058

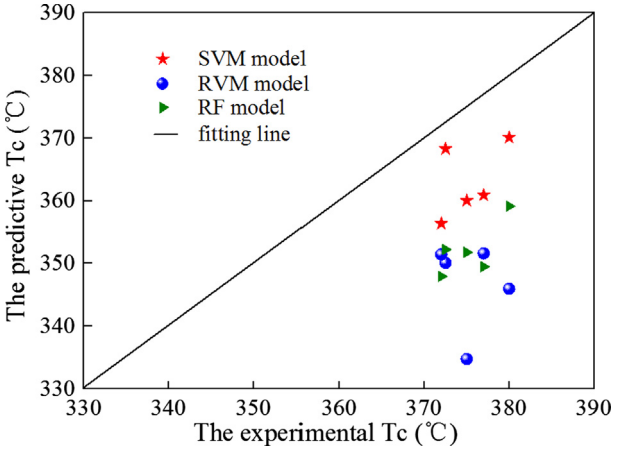


Fig. 7. The experimental versus the predictive Tc of the three models with external validation.

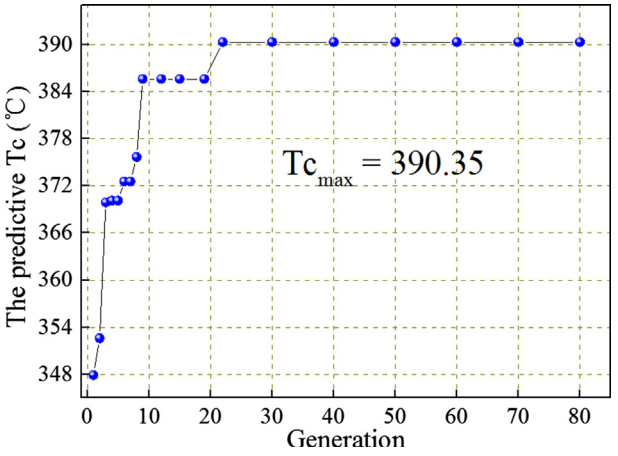


Fig. 8. The process of finding the higher Tc by using GA.

Table 7
The perovskite material with higher Tc.

Perovskite material	SVR	RVM	RF
$La_{0.66}Sr_{0.3}Ba_{0.04}MnO_3$	390.35	373.07	366.77

RF models are better than the SVR model. The residuals of the RVM and RF models mainly vary from −40 to 40. While it is obvious that the residual distribution of the SVR model is the most concentrated, and changes from −20 to 20. From the above analyses, the SVR model has the best prediction performance among all the models.

3.4. Testing the models

The data set was split into two parts, the training set for modeling and the testing set for validation. The testing set was made up of the samples with top five Tc. Obviously, the remainder of the dataset constructs the training set. The three models were established by using the training set and predicted the samples in the testing set. The results are listed in Table 6 and Fig. 7. In Table 6, "*e*" represents the residuals of predictive values.

From Table 6, all the models have the less prediction values than the experimental values, and the RVM model is most obvious. That may be because the values of the samples in the testing set are all higher than those in the training set. Fig. 7 illustrates the comparisons of the three models with external validation. From Fig. 7, all points are mapped on the right of the fitting line, and the predictive points of the RVM and the SVR models are farthest away from and nearest to the fitting line, respectively. To sum up, SVM performs the best in prediction beyond the range of the training set, whereas RVM is the worst performer. Furthermore, the low prediction values of the three models correspond to the high experimental values because the predictive values of the samples with higher Tc than the highest Tc in the training set generally have negative residuals.

3.5. Virtual screening

On the whole, the SVR model exhibits stable performance in predicting Tc. Hence, the SVR model was used to virtual screening the candidate perovskite materials with probably higher Tc than the highest Tc in the dataset. According to the characters of the samples in the dataset, the following restrictions were obeyed when generating the new samples:

(1) The A-site and B-site contain no more than 3 and 2 different doping ions, respectively.
(2) The first element in the A-site space is La with the doping ratio from 0.6 to 1.0 with step 0.02. The second element in A-site space is Ca, Sr, Ag, Pb or Ba with the doping ratio from 0.0 to 0.4 with step 0.02. The third element in the A-site space is Ag, Ba, Sm, Nd, Ca or Pr with the rest doping ratio.
(3) The first element in the B-site space is Mn with the doping ratio from 0.9 to 1.0 with step 0.02. The second element in the B-site space may be Cu, Fe, V, Al or Cr with the rest doping ratio.

GA was employed to generate the virtual samples, and its population was set to 80. Fig. 8 illustrates how GA was used to find the higher Tc. From Fig. 8, the higher Tc occurs when generation is 22, where a perovskite material was found as shown in Table 7. The predictive value of Tc of the perovskite material ($La_{0.66}Sr_{0.3}Ba_{0.04}MnO_3$) using the SVR model has exceeded the highest Tc in the dataset, and its experimental Tc will very likely exceed more because the predictive values of the samples with higher Tc than highest Tc of the training set typically have negative residuals. Although the RVM and RF models were also used to predict the Tc of the virtual samples, their results were unsatisfactory because the predictive Tc don't exceed the highest Tc in the dataset.

4. Conclusion

This work focuses on the prediction problem for Tc of perovskite materials with the physicochemical parameters. The SVR, RVM and RF models constructed to predict Tc were used search for perovskite materials with higher Tc. By K-fold cross validation, the SVR model had the higher predictive performance than the two models. The model is a very effective way to predict Tc of perovskite materials because of owing R of 0.8549, RMSE of 28.6659 and MRE of 0.0725. Finally, the perovskite material ($La_{0.66}Sr_{0.3}Ba_{0.04}MnO_3$) has been found by virtual screening and using the SVR model and will possibly have the higher Tc than the highest Tc in the dataset. The method exhibited in the study can be generalized in materials design and controllable synthesis of other compounds, and further improves the study about machine learning to assist the material design.

As pointed out in [53] and demonstrated in a series of recent publications (see, e.g., [10–18,20,54–62]), user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful prediction methods and computational tools. Actually, many practically useful web-servers have increasing impacts on medical science [63], driving medicinal chemistry into an unprecedented revolution [64]. We shall make efforts in our future work to provide a web-server for the prediction methods presented in this paper.

Acknowledgements

The authors acknowledge the financial support from the National Key Research and Development Program of China (No. 2016YFB0700504).

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.commatsci.2018.04.031.

References

[1] T. Shi, G. Li, J. Zhu, Compositional design strategy for high performance ferroelectric oxides with perovskite structure, Ceram. Int. 43 (3) (2017) 2910–2917, http://dx.doi.org/10.1016/j.ceramint. 2016.11.085.
[2] K. Abe, S. Komatsu, Ferroelectric properties in epitaxially grown Ba_xSr_{1-x}TiO3thin films, J. Appl. Phys. 77 (12) (1995) 6461–6465, http://dx.doi.org/10.1063/1.359120.
[3] P.T. Phong, L.T.T. Ngan, L.V. Bau, et al., Study of critical behavior using the field dependence of magnetic entropy change in $La_{0.7}Sr_{0.3}Mn_{1-x}Cu_xO_3$ (x = 0.02 and 0.04), Ceram. Int. 43 (18) (2017) 16859–16865, http://dx.doi.org/10.1016/j.ceramint.2017.09.085.
[4] D.C. Linh, T.D. Thanh, L.H. Anh, et al., Critical properties around the ferromagnetic-paramagnetic phase transition in $La_{0.7}Ca_{0.3-x}A_xMnO_3$ compounds (A = Sr, Ba and x = 0, 0.15, 0.3), J. Alloy. Compd. 725 (2017) 484–495, http://dx.doi.org/10.1016/j.jallcom.2017.07.168.
[5] F.F. An, F. Cao, J. Yu, Piezoelectric properties of Ca-modified $Pb_{0.6}Bi_{0.4}(Ti_{0.75}Zn_{0.15}Fe_{0.10})O_3$ ceramics, Ceram. Int. 38 (2012) S211–S214, http://dx.doi.org/10.1016/j.ceramint.2011.04.085.
[6] G. Pilania, A. Mannodi Kanakkithodi, B.P. Uberuaga, et al., Machine learning bandgaps of double perovskites, Sci. Rep. 6 (2016) 19375, http://dx.doi.org/10.1038/srep19375.
[7] P. Raccuglia, K.C. Elbert, P.D. Adler, et al., Machine-learning-assisted materials discovery using failed experiments, Nature 533 (7601) (2016) 73–76, http://dx.doi.org/10.1038/nature17439.
[8] D. Xue, P.V. Balachandran, J. Hogden, et al., Accelerated search for materials with targeted properties by adaptive design, Nat. Commun. 7 (2016) 1–9, http://dx.doi.org/10.1038/ncomms11241.
[9] J. Jia, Z. Liu, X. Xiao, et al., iPPI-Esml: an ensemble classifier for identifying the interactions of proteins by incorporating their physicochemical properties and wavelet transforms into PseAAC, J. Theor. Biol. 377 (2015) 47–56, http://dx.doi.org/10.1016/j.jtbi.2015.04.011.
[10] W. Chen, P. Feng, H. Yang, et al., iRNA-AI_ identifying the adenosine to inosine editing sites in RNA sequences, Oncotarget 8 (2017) 4208–4217, http://dx.doi.org/10.18632/oncotarget.13758.
[11] B. Liu, L. Fang, F. Liu, et al., Identification of real microRNA precursors with a pseudo structure status composition approach, PLoS One 10 (3) (2015) e0121501, http://dx.doi.org/10.1371/journal.pone. 0121501.
[12] X. Cheng, S.G. Zhao, X. Xiao, et al., iATC-mISF_ a multi-label classifier for predicting the classes of anatomical therapeutic chemicals, Bioinformatics 33 (3) (2017) 341–346, http://dx.doi.org/10.1093/bioinformatics/btw644.
[13] P. Feng, H. Ding, H. Yang, et al., iRNA-PseColl: identifying the occurrence sites of different RNA modifications by incorporating collective effects of nucleotides into PseKNC, Mol. Ther. Nucleic Acids 7 (2017) 155–163, http://dx.doi.org/10.1016/j.omtn.2017.03.006.
[14] B. Liu, S. Wang, R. Long, et al., iRSpot-EL: identify recombination spots with an ensemble learning approach, Bioinformatics 33 (1) (2017) 35–41, http://dx.doi.org/10.1093/bioinformatics/btw539.
[15] B. Liu, F. Yang, K.C. Chou, 2L-piRNA: a two-layer ensemble classifier for identifying Piwi-interacting RNAs and their function, Mol. Ther. Nucleic Acids 7 (2017) 267–277, http://dx.doi.org/10.1016/j.omtn.2017.04.008.
[16] L.M. Liu, Y. Xu, K.C. Chou, iPGK-PseAAC: identify lysine phosphoglycerylation sites in proteins by incorporating four different tiers of amino acid pairwise coupling information into the general PseAAC, Med. Chem. 13 (999) (2017) 552–559, http://dx.doi.org/10.2174/1573406413666170511 5120507.
[17] W.R. Qiu, S.Y. Jiang, Z.C. Xu, et al., iRNAm 5C-PseDNC_ identifying RNA 5-methylcytosine sites by incorporating physical-chemical properties into pseudo dinucleotide composition, Oncotarget 8 (25) (2017) 41178–41188, http://dx.doi.org/10.18632/oncotarget.17104.
[18] W.R. Qiu, B.Q. Sun, X. Xiao, et al., iPhos-PseEvo: identifying human phosphorylated proteins by incorporating evolutionary information into general PseAAC via grey system theory, Mol. Inf. 36 (5–6) (2017) 1–9, http://dx.doi.org/10.1002/minf.201600010.
[19] Q. Su, W. Lu, D. Du, et al., Prediction of the aquatic toxicity of aromatic compounds to tetrahymena pyriformis through support vector regression, Oncotarget 8 (2017) 49359–49369, http://dx.doi.org/10.18632/oncotarget.17210.
[20] Y. Xu, Z. Wang, C. Li, et al., iPreny-PseAAC: identify C-terminal cysteine prenylation sites in proteins by incorporating two tiers of sequence couplings into PseAAC, Med. Chem. 13 (6) (2017) 544–551, http://dx.doi.org/10.2174/1573406413666170419150052.
[21] W. Chen, P. Feng, H. Yang, et al., iRNA-3typeA: identifying 3-types of modification at RNA's adenosine sites, Mol. Ther. Nucleic Acids (2018), http://dx.doi.org/10.1016/j.omtn.2018.03.012.
[22] Y.W.J. Song, F. Li, T. Akutsu, N.D. Rawlings, G.I. Webb, iProt-Sub: a comprehensive package for accurately mapping and predicting protease-specific substrates and cleavage sites, Briefings Bioinf. (2018), http://dx.doi.org/10.1093/bib/bby028.
[23] H. Yang, W.R. Qiu, G. Liu, et al., iRSpot-Pse6NC: Identifying recombination spots in Saccharomyces cerevisiae by incorporating hexamer composition into general PseKNC, Int. J. Biol. Sci. (2018), http://dx.doi.org/10.7150/ijbs.246.
[24] K.C. Chou, Some remarks on protein attribute prediction and pseudo amino acid composition, J. Theor. Biol. 273 (1) (2011) 236–247, http://dx.doi.org/10.1016/j.jtbi.2010.12.024.
[25] N. Kallel, S. Kallel, A. Hagaza, et al., Magnetocaloric properties in the Cr-doped $La_{0.7}Sr_{0.3}MnO_3$ manganites, Phys. B: Condens. Matter 404 (2) (2009) 285–288, http://dx.doi.org/10.1016/j.physb. 2008.10.049.
[26] C.R. Koubaa, M. Koubaa, A. Cheikhrouhou, Structural, magnetotransport, and magnetocaloric properties of $La_{0.7}Sr_{0.3-x}Ag_xMnO_3$ perovskite manganites, J. Alloy. Compd. 453 (1–2) (2008) 42–48, http://dx.doi.org/10.1016/j.jallcom.2006.11.185.
[27] M. Koubaa, C.R. Koubaa, A. Cheikhrouhou, Magnetocaloric effect and magnetic properties of $La_{0.75}Ba_{0.1}M_{0.15}MnO_3$ (M = Na, Ag and K) perovskite manganites, J. Alloy. Compd. 479 (1–2) (2009) 65–70, http://dx.doi.org/10.1016/j.jallcom.2009.01.030.
[28] J.C. Debnath, R. Zeng, J.H. Kim, et al., Large magnetic entropy change near room temperature in $La_{0.7}(Ca_{0.27}Ag_{0.03})MnO_3$ perovskite, J. Alloy. Compd. 509 (8) (2011) 3699–3704, http://dx.doi.org/10.1016/j.jallcom.2010.12.169.
[29] S. Sankarrajan, K. Sakthipandi, P. Manivasakan, et al., On-line phase transition in $La_{1-x}Sr_xMnO_3$ ($0.28 \le x \le 0.36$) perovskites through ultrasonic studies, Phase Trans. 84 (7) (2011) 657–672, http://dx.doi.org/10.1080/01411594.2011.556915.
[30] M.S. Anwar, F. Ahmed, B. Heun Koo, Influence of Ce addition on the structural, magnetic, and magnetocaloric properties in $La_{0.7-x}Ce_xSr_{0.3}MnO_3$ ($0 \le x \le 0.3$) ceramic compound, Ceram. Int. 41 (4) (2015) 5821–5829, http://dx.doi.org/10.1016/j.ceramint.2015.01.011.
[31] J. Dhahri, A. Dhahri, M. Oummezzine, et al., Effect of substitution of Fe for Mn on the structural, magnetic properties and magnetocaloric effect of $LaNdSrCaMnO_3$, J. Magn. Magn. Mater. 378 (2015) 353–357, http://dx.doi.org/10.1016/j.jmmm.2014.10.163.
[32] N. Chau, H.N. Nhat, N.H. Luong, et al., Structure, magnetic, magnetocaloric and magnetoresistance properties of $La_{1-x}Pb_xMnO_3$ perovskite, Phys. B: Condens. Matter 327 (2–4) (2003) 270–278, http://dx.doi.org/10.1016/s0921-4526(02)01759-3.
[33] M.S. Anwar, F. Ahmed, B.H. Koo, Structural distortion effect on the magnetization and magnetocaloric effect in Pr modified $La_{0.65}Sr_{0.35}MnO_3$ manganite, J. Alloy. Compd. 617 (2014) 893–898, http://dx.doi.org/10.1016/j.jallcom.2014.08.105.
[34] D.R. Lide, CRC Handbook of Chemistry and Physics, Internet Version, CRC Press, Boca Raton, FL, 2005.
[35] M. Nazemi, A. Heidaripanah, Support vector machine to predict the indirect tensile strength of foamed bitumen-stabilised base course materials, Road Mater. Pavement Des. 17 (3) (2016) 768–778, http://dx.doi.org/10.1080/14680629.2015.1119712.
[36] X. Zhang, J. Liu, W. Hou, et al., Preparation and properties of pesticide/cyclodextrin complex intercalated into ZnAl-layered double hydroxide, Ind. Eng. Chem. Res. 55 (6) (2016) 1550–1558, http://dx.doi.org/10.1021/acs.iecr.5b04001.
[37] T.O. Owolabi, K.O. Akande, S.O. Olatunji, Application of computational intelligence technique for estimating superconducting transition temperature of YBCO superconductors, Appl. Soft Comput. 43 (2016) 143–149, http://dx.doi.org/10.1016/j.asoc.2016.02.005.
[38] B. Demir, S. Erturk, Hyperspectral image classification using relevance vector machines, IEEE Geosci. Remote Sens. Lett. 4 (4) (2007) 586–590, http://dx.doi.org/10.1109/LGRS.2007.903069.
[39] M.E. Tipping, Sparse Bayesian learning and the relevance vector machine, J. Mach. Learn. Res. 1 (2001) 211–244.
[40] D. Liu, J. Zhou, D. Pan, et al., Lithium-ion battery remaining useful life estimation with an optimized Relevance Vector Machine algorithm with incremental learning, Measurement 63 (2015) 143–151, http://dx.doi.org/10.1016/j.measurement.2014.11.031.
[41] C. Hu, G. Jain, C. Schmidt, et al., Online estimation of lithium-ion battery capacity using sparse Bayesian learning, J. Power Sources 289 (2015) 105–113, http://dx.doi.org/10.1016/j.jpowsour. 2015.04.166.
[42] P. Spyridonos, G. Gaitanis, I.D. Bassukas, et al., Evaluation of vermillion border descriptors and relevance vector machines discrimination model for making probabilistic predictions of solar cheilosis on digital lip photographs, Comput. Biol. Med. 63 (2015) 11–18, http://dx.doi.org/10.1016/j.compbiomed.2015.04.024.

[43] J.R. Quinlan, Induction on decision tree, Mach. Learn. 1 (1) (1986) 81–106.
[44] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32, http://dx.doi.org/10.1023/A:1010933 404324.
[45] Y. Liu, S. Tang, C. Fernandez-Lozano, et al., Experimental study and Random Forest prediction model of microbiome cell surface hydrophobicity, Expert Syst. Appl. 72 (2017) 306–316, http://dx.doi.org/10.1016/j.eswa.2016.10.058.
[46] B. Hu, K. Lu, Q. Zhang, et al., Data mining assisted materials design of layered double hydroxide with desired specific surface area, Comput. Mater. Sci. 136 (2017) 29–35, http://dx.doi.org/10.1016/j.commatsci.2017.03.027.
[47] V. Svetnik, A. Liaw, C. Tong, et al., Random Forest_ A classification and regression tool for compound classification and QSAR modeling, J. Chem. Inf. Comput. Sci. 43 (2003) 1947–1958.
[48] P. Xiong, X. Ji, X. Zhao, et al., Materials design and control synthesis of the layered double hydroxide with the desired basal spacing, Chemomet. Intell. Lab. Syst. 144 (2015) 11–16, http://dx.doi.org/10.1016/j.chemolab.2015.03.005.
[49] N.J. Browning, R. Ramakrishnan, O.A. von Lilienfeld, et al., Genetic optimization of training sets for improved machine learning models of molecular properties, J. Phys. Chem. Lett. 8 (7) (2017) 1351–1359, http://dx.doi.org/10.1021/acs.jpclett.7b00038.
[50] M. Nekoei, M. Mohammadhosseini, E. Pourbasheer, QSAR study of VEGFR-2 inhibitors by using genetic algorithm-multiple linear regressions (GA-MLR) and genetic algorithm-support vector machine (GA-SVM): a comparative approach, Med. Chem. Res. 24 (7) (2015) 3037–3046, http://dx.doi.org/10.1007/s00044-015-1354-4.
[51] E.L. Dyer, A.C. Sankaranarayanan, R.G. Baraniuk, Greedy feature selection for subspace clustering, J. Mach. Learn. Res. 14 (2013) 2487–2517.
[52] I. Guyon, A.e. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (2003) 1157–1182.
[53] K.C. Chou, H.B. Shen, REVIEW: Recent advances in developing web-servers for predicting protein attributes, Nat. Sci. 01 (02) (2009) 63–92, http://dx.doi.org/10.4236/ns.2009.12011.
[54] Z. Liu, X. Xiao, W.R. Qiu, et al., iDNA-Methyl: identifying DNA methylation sites via pseudo trinucleotide composition, Anal. Biochem. 474 (2015) 69–77, http://dx.doi.org/10.1016/j.ab.2014.12.009.
[55] B. Liu, L. Fang, R. Long, et al., iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition, Bioinformatics 32 (3) (2016) 362–369, http://dx.doi.org/10.1093/bioinformatics/btv604.
[56] X. Cheng, X. Xiao, K.C. Chou, pLoc-mPlant: predict subcellular localization of multi-location plant proteins by incorporating the optimal GO information into general PseAAC, Mol. BioSyst. 3 (2017), pp. 1722–1727. doi: 10.1039/C7MB00267J.
[57] X. Cheng, X. Xiao, K.C. Chou, pLoc-mVirus: Predict subcellular localization of multi-location virus proteins via incorporating the optimal GO information into general PseAAC, Gene 628 (2017) 315–321, http://dx.doi.org/10.1016/j.gene.2017.07.036.
[58] X. Cheng, X. Xiao, K.C. Chou, pLoc-mEuk: Predict subcellular localization of multi-label eukaryotic proteins by extracting the key GO information into general PseAAC, Genomics 110 (1) (2018) 50–58, http://dx.doi.org/10.1016/j.ygeno.2017.08.005.
[59] X. Cheng, X. Xiao, K.C. Chou, pLoc-mGneg: Predict subcellular localization of Gram-negative bacterial proteins by deep gene ontology learning via general PseAAC, Genomics (2017) 1–9, http://dx.doi.org/10.1016/j.ygeno.2017.10.002.
[60] X. Xiao, X. Cheng, S. Su, et al., pLoc-mGpos: incorporate key gene ontology information into general PseAAC for predicting subcellular localization of gram-positive bacterial proteins, Nat. Sci. 09 (09) (2017) 330–349, http://dx.doi.org/10.4236/ns.2017.99032.
[61] W.R. Qiu, B.Q. Sun, X. Xiao, et al., iKcr-PseEns: Identify lysine crotonylation sites in histone proteins with pseudo components and ensemble classifier, Genomics (2017), http://dx.doi.org/10.1016/j.ygeno.2017.10.008.
[62] X. Cheng, S.G. Zhao, W.Z. Lin, et al., pLoc-mAnimal: predict subcellular localization of animal proteins with both single and multiple sites, Bioinformatics 33 (22) (2017) 3524–3531, http://dx.doi.org/10.1093/bioinformatics/btx476.
[63] K.C. Chou, Impacts of bioinformatics to medicinal chemistry, Med. Chem. 11 (3) (2015) 218–234, http://dx.doi.org/10.2174/1573406411666141229162834.
[64] K.C. Chou, An unprecedented revolution in medicinal chemistry driven by the progress of biological science, Curr. Top. Med. Chem. 17 (21) (2017) 2337–2358, http://dx.doi.org/10.2174/1568026617666170414145508.