Computational Materials Science 175 (2020) 109618

Contents lists available at ScienceDirect

Computational Materials Science

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

Using machine learning and feature engineering to characterize limited material datasets of high-entropy alloys

Dongbo Dai[a], Tao Xu[a], Xiao Wei[a,b], Guangtai Ding[a,b], Yan Xu[c], Jincang Zhang[b], Huiran Zhang[a,b,c,*]

[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

ARTICLE INFO

Keywords:
High-entropy alloy
Phase transformations
Machine learning
Feature engineering

ABSTRACT

The prediction of the phase formation of high entropy alloys (HEAs) has attracted great research interest recent years due to their superior structure and mechanical properties of single phase. However, the identification of these single phase solid solution alloys is still a challenge. Previous studies mainly focus on trial-and-error experiments or thermodynamic criteria, the previous is time consuming while the latter depends on the descriptors quality, both provide unreliable prediction. In this study, we attempted to predict the phase formation based on feature engineering and machine learning (ML) with a small dataset. The descriptor dimensionality is augmented from original small dimension to high dimension by non-linear combinations to characterize HEAs. The results showed that this method could achieve higher accuracy in predicting the phase formation of HEAs than traditional methods. Except the prediction of HEAs, this method also can be applied to other materials with limited dataset.

1. Introduction

In the past decade, HEAs have attracted great research interest for their superior properties, like excellent thermal and electrical conductivity [1], high corrosion resistance [2], significant refractory properties [3], great magnetic performance [4], etc. The concept of HEAs was proposed by Cantor et al. [5] and Yeh et al. [6] in 2004. Different from traditional alloys, this kind of alloys contains at least five principal elements whose concentration is between 5% and 35% by each [7]. Because of this, it is recognized that the HEAs have a huge search space in the materials researches [8]. On the other hand, the phase stability impacts the microstructure, which could lead HEAs to different physical and mechanical properties [9], for example, face-centered cubic (FCC) phase usually could improve an alloy's ductility and body-centered cubic (BCC) phase typical exhibits high strength [10]. Therefore, a critical question is posed latterly: can we predict which phase (solid solution phase, amorphous phase or intermetallic phase) will be formed for a given alloy system? [11].

Many methods have been used for acquiring the phase formation accurately. But at the same time, much computing time and resource have to be spent on them [12], just like first-principles calculations [13,14], calculation of phase diagram (CALPHAD) [15], or density functional theory (DFT) simulations with ML [16] and so on. Obviously, these methods are not practical for large search space in materials science. Based on experimental data, the parametric method proposed by Yeh et al. can solve these problems [6]. They used various empirical thermo-physical parameters, including enthalpy of mixing ΔH_{mix} [17], entropy of mixing ΔS_{mix} [18], atomic size difference δr [19], valence electron concentration (VEC) [20], dimensionless parameter Ω [21], atomic packing mismatch γ [22], etc. The parametric method is used for finding the rules that governing phase stability [23]. Usually, these parameters are used in pairs to construct a two-dimensional graph for separating different phases [11,19,24–29]. Different phase-regions may overlap with each other [18,30]. For example, FCC-phase-region and BCC-phase-region almost share the same area in $\delta r \sim \gamma$ diagram both on the dataset of Yeh et al. [23] and Gao et al. [9]. However, these parameters could only provide limited representation ability and thus it corresponds to the poor prediction performance. General speaking, the parametric method is not good enough to predict different phase formation with a two-dimensional plot.

Different from common computational and parametric methods mentioned above, ML is a simple yet efficient method that has been used in materials science [31]. It can recognize the inner data pattern and construct a model to make a quick prediction for unseen sample.

* Corresponding author.
E-mail address: hrzhangsh@shu.edu.cn (H. Zhang).

https://doi.org/10.1016/j.commatsci.2020.109618
Received 23 December 2019; Received in revised form 18 February 2020; Accepted 19 February 2020
Available online 25 February 2020
0927-0256/

Hence the process of materials discovery and design could be accelerated with ML. Huang et al. constructed different ML models based on five descriptors to predict the different phase formation of HEAs, with not so high accuracy [32,33]. Partha et al. utilized the data mining approach to predict new chalcopyrite compounds [34]. However, all these researches are implemented based on limited dataset, because we are often limited to hundreds or thousands in material science [31]. Researchers need to explore more potential of the dataset for making precise predictions.

In this paper, we proposed a methodology to construct a low-dimensional descriptors that allow predicting the stable phase of a HEA based on its composition. With feature engineering and ML strategy, over ten thousand descriptors could be constructed and through a selection strategy, 20 of these descriptors were chosen as the input features of the algorithm. The results show that ML with feature engineering could improve the ability of characterization and the performance of prediction on the phase formation of HEAs.

2. Method

2.1. Strategy

Feature engineering is a method of data processing. It could refactor the original dataset to new dataset in order to fit the learning algorithm [31]. Different from the exploration of the relationships between two sets of variables, such as canonical correlation analysis (CCA), we are aimed to characterize the material with suitable descriptors. The data processing includes data cleaning, feature standardization, dimensionality reduction, feature construction, feature selection, etc. Then, more features could be constructed from primary features for improving the size of the dataset [31,35–37]. The representation of descriptors is important for an ML algorithm towards a specific material property [36,38]. The properties of data and descriptors set limitations to certain ML methods. Therefore, it is important to create an efficient subset of descriptors to do the research.

In this work, we design a descriptor processing flow. The whole workflow involves several steps: (1) Coefficient analysis, which could remove the high relevant descriptors. (2) Dimensionality augmentation, which is based on fundamental structure parameters and prototypical functions (such as multiply). (3) Top important descriptors selection, which could find a proper descriptor subset for a better material characterization.

A systematic approach was applied to augment the dimensionality of descriptor to proceed with their representation [36]. The data in this paper refer to Refs. [9,18,23] which contains three collections of HEAs. After merging them together and removing the repeated data, 407 samples are left consisting of 48 HEAs with single FCC solid solution phase (SSP), 43 HEAs with single BCC SSP, 16 HEAs with single hexagonal close packed (HCP) SSP, 237 HEAs with multi-phase (MP), and 63 HEAs with amorphous phase (AM). Obviously, this is a five-classification problem with five class labels, namely FCC, BCC, HCP, MP and AM. It is worth mentioning that the imbalance of dataset is not the point we focus on, in contrast, the overall prediction performance is our main concern. The 14 original descriptors are shown in Table 1. The first step in ML area is to analyze the correlation among these descriptors. We choose the Pearson correlation coefficient which is defined as below to describe the correlations:

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

where x_i and y_i are the sample value of descriptor x and descriptor y, $\bar{x}$ and $\bar{y}$ are the mean value of descriptor x and descriptor y, s_x and s_y are the standard deviation of corresponding descriptor.

The original descriptors are then subjected to Pearson correlation filter to remove the highly correlated descriptors. The visualization of the Pearson correlation coefficient between these 14 descriptors is shown in Fig. 1. Value 1 means the two descriptors are almost completely relevant, like parameter d1 (δ: atomic size difference) and d5 (RMS: elastic residual strain root mean square) [9], one of them is redundant and should be removed. Correlated features can hinder the accuracy and efficiency of the model [39]. The coefficient value is adjusted so that 9 descriptors are left, and tests show that more descriptors could not benefit the ML model.

2.2. Dimensionality augmentation

To characterize the material from high dimensionality, different kind descriptors should be constructed. For example, Yan [40] constructed a descriptor set that including 70 descriptors and used genetic algorithm to select the best subset to study the phase formation of HEAs. We utilized the following steps to generate new non-linear descriptors (unmeaningful combinations are not considered such as size + energy [37]). Four fundamental functions: $|x|^{1/2}$, x^2, x^3, $log(1+|x|)$ with x being one of the 9 original descriptors, are used to generate 36 new descriptors. Additional 15,180 descriptors could be obtained by multiplying any 2 or 3 descriptors respectively. Moreover, a simple function $1/x$ is applied to generate new descriptors with all constructed descriptors, especially it would utilize a large number 100,000 to avoid an unnecessary calculating error if $x = 0$. This feature engineering approach finally provides us a data set with 30,450 descriptors.

2.3. Descriptors selection

An important aspect is to choose a descriptor subset from high-dimension descriptors to assist data-driven insight into the problems in ML process [41]. Dimensionality reduction and descriptor selection are popular methods to do this kind of work. However, dimensionality reduction method will change the original representation of descriptors, which would result in non-explanation. So we turn to feature selection strategy. We tested two different selection methods (Least absolute shrinkage and selection operator (LASSO), Recursive feature elimination (RFE)), and found RFE could result in better accuracy, with the detailed results evaluated by LASSO show in Fig. S1 in Supplementary material. RFE, a wrapper method of descriptor selection strategy, is designed to select descriptors by recursively minimizing the descriptor set with an external estimator, which should be replaced by kinds of classification ML methods. One aspect that RFE outperforms the other methods is that it implements backward feature elimination and removes irrelevant features, thus limiting the negative effect on cross-validation score [42]. Since the new descriptors are constructed by non-linear combinations, to accelerate the process of feature selection and simplify the complexity of the model, the simple and linear model, logistic regression (LR), is selected for RFE and ML, the performance of different external estimator is shown in Fig. S2 in Supplementary material.

The key for the whole multi-dimensional characterization of materials is descriptor selection. As mentioned above, the descriptor selection is the process of selecting a subset of original variables. The model is built on the dataset which only contains these descriptors. It avoids overfitting and improves the performance by getting rid of redundancies [41]. Top 20 descriptors were then gained with RFE.

3. Results and discussions

To determine the order of these descriptors input the model, we evaluated their relative importance to the model. Each descriptor was evaluated by removing it from 9 descriptors and testing the accuracy of the model with the left descriptor set. The model was trained on 70% randomly selected HEAs from the data set and then tested on the remaining 30% to predict the phase formation. The evaluating result was

Table 1
A snapshot of the first 5 instances of the dataset in this work. The columns are the 14 original descriptors of each instance. The explanation of each descriptor shows in Table S1 in Supplementary material.

Alloy	δ	ΔHmix	ΔSmix	φ	RMS	VEC	r
$Al_{0.5}CoCrCuFeNiTi_{0.2}$	4.93	−4.15	15.45	18.43	4.87	8.12	4.93
$Al_{0.3}CoCrFeNi$	3.49	−7.27	12.83	21.23	3.45	7.88	3.76
$Al_{0.5}CrCuFeNi_2$	4.2	−2.51	12.6	21.64	4.14	8.45	4.2
CoCrFeNi	1.03	−3.75	11.53	3789.72	1.03	8.25	0.3
CoFeMnNi	0.66	−4	11.53	25.96	0.66	8.5	3.55
Alloy	S_C	ΔH_{mix}^{ijmax}	ΔH_{mix}^{ijmin}	$\sqrt{\delta H_{mix}}$	$\sqrt{\delta H_{mix}^{0}}$	$\sqrt{\delta H_{mix}^{0+}}$	$\sqrt{\delta H_{mix}^{0-}}$
$Al_{0.5}CoCrCuFeNiTi_{0.2}$	1.86	13	−35	0.7	0.7	0.49	0.65
$Al_{0.3}CoCrFeNi$	1.54	0	−22	0.51	0.56	0	0.56
Al0.5CrCuFeNi2	1.54	0	−22	0.51	0.56	0	0.56
CoCrFeNi	1.39	0	−7	0.33	0.37	0	0.37
CoFeMnNi	1.39	0	−8	0.37	0.41	0	0.41

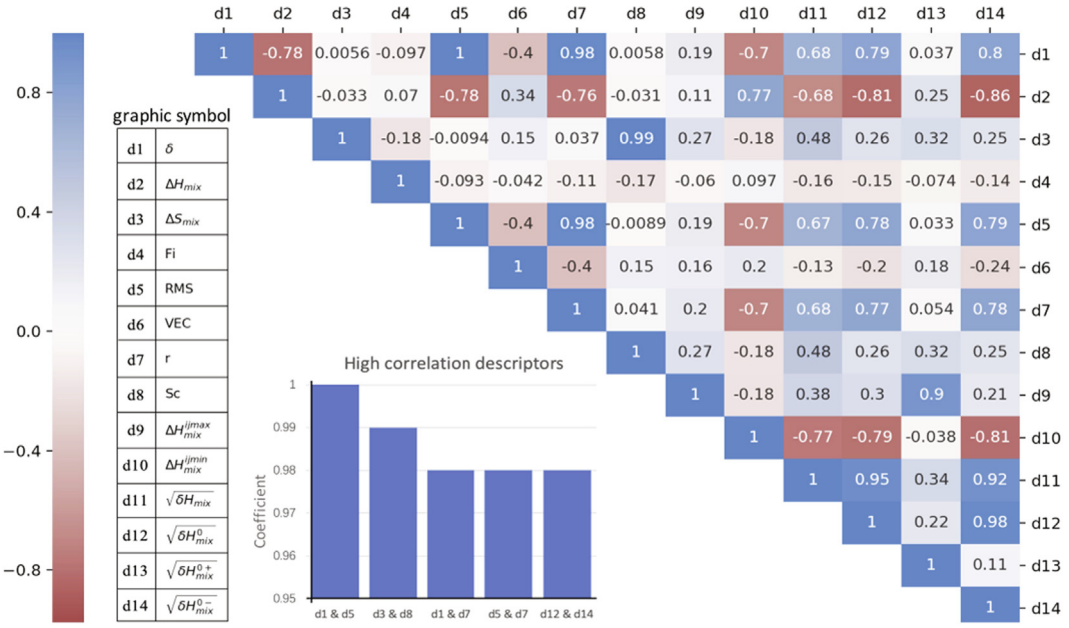

Fig. 1. The heatmap of Pearson correlation coefficient matrix between the original 14 descriptors. The value in the grid means the correlation coefficient between the corresponding two descriptors. Color intensity is proportional to the correlation coefficients. In the left side of the correlogram, the legend color shows the correlation coefficients and the corresponding color. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

shown in Fig. 2. It can be seen that VEC plays the most important role in this classification work, because the accuracy decreases the most when removing VEC from the descriptor set, while ΔS_{mix} shows the least effect. It should be pointed out that this evaluating result varies little when changes the sampling. The result is also consistent with the previous research [32–43]. The final sequence of these descriptors input the model in this order: VEC, $\sqrt{\delta H_{mix}^{0+}}$, ΔH_{mix}, $\sqrt{\delta H_{mix}}$, ΔH_{mix}^{ijmax}, ΔH_{mix}^{ijmin}, Fi, r, ΔS_{mix}. The order of the selected 20 descriptors follows the same way.

We used n-fold cross-validation to avoid overfitting in ML [44,45]. Overfitting, which means the model is too complicated, learns too much detail and performs well on the training set but has a bad generalization ability on unseen data. We adopted five-fold cross-validation which could evaluate the generalization ability of the model by dividing the data set into five subsets. The model was trained on four subsets and tested on the left one subset five times. The accuracy of cross-validation is the mean value of the five accuracy results.

The performance of the LR model with original descriptors and selected descriptors are presented both in Fig. 3. As can be seen, when the descriptors number input increased, both prediction accuracy of cross-validation initially increased and then decreased. It demonstrates not all descriptors could benefit the model. The initial increase indicates that the model is under-fitting and the later decrease is possibly because of over-fitting. Thus, the turning point of the curve could represent the best generalization performance of the model. For the original descriptor set, the cross-validation accuracy is 0.75 with 6 descriptors. For the selected descriptor set, the accuracy reaches 0.86 with 9 descriptors. It indicates that the model could explore more relationships between descriptors and properties with the selected non-linear descriptors. It means that it is useful to construct non-linear descriptors for linear machine learning methods. In other words, a proper descriptor set would be in favour of characterizing materials better to improve the performance of ML.

It is worth mentioning that in order to accelerate the processing speed of our LR model, standardization was used to preprocess the dataset, which could neutralize the effect of different scales across descriptors with the following equation:

$$z_{ij} = \frac{x_{ij} - \bar{x_i}}{s_i} \tag{2}$$

where $\bar{x_i}$ and s_i is the mean value and variance of all samples, x_{ij} is the true value and z_{ij} is the normalized value.

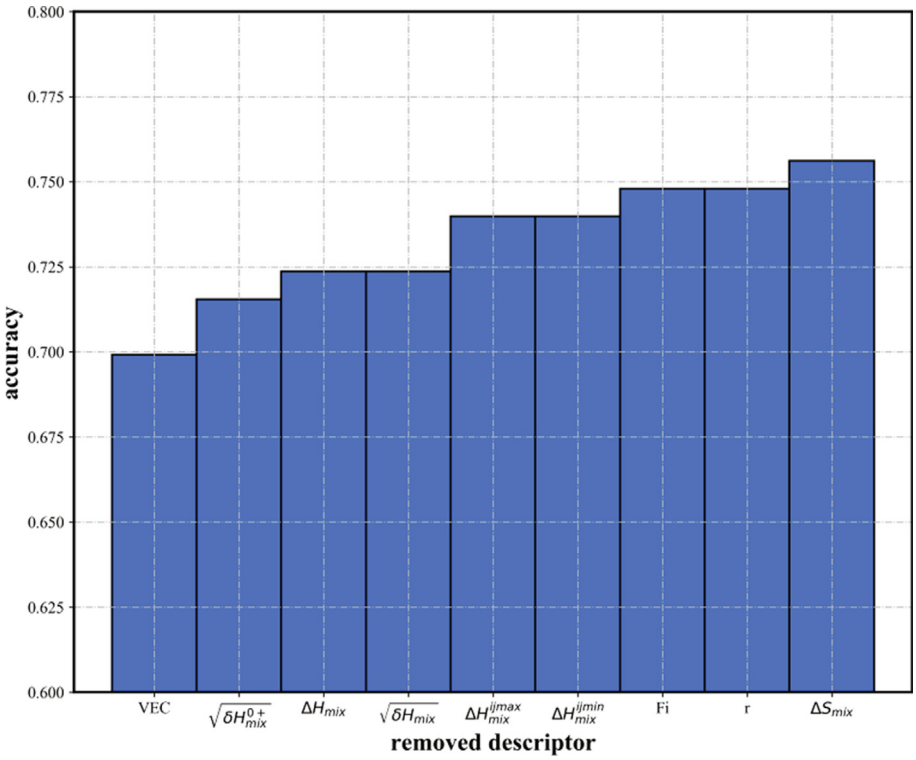

Fig. 2. The relative importance of nine descriptors. Descriptors on X-axis represents the removed one at each iterations. Value on Y-axis represents the accuracy when remove the corresponding descriptor from the descriptor set. The lower the accuracy, the greater effect of the descriptor on the result.

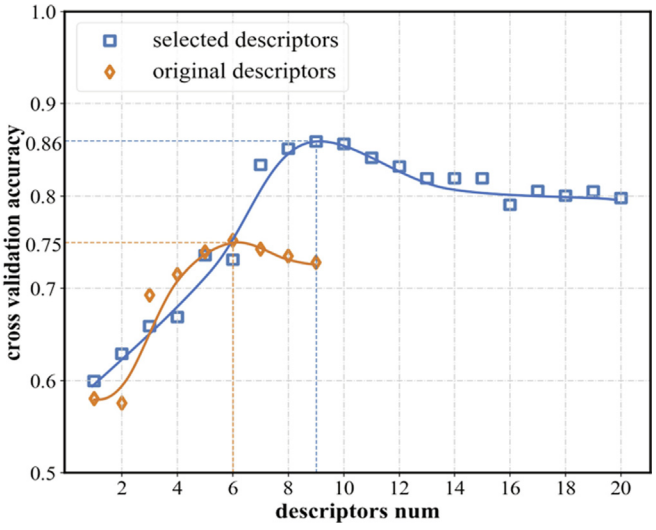

Fig. 3. The prediction accuracy curve with different descriptor number and descriptor set. Orange and blue dots represent the cross-validation accuracy of the original descriptor set and selected descriptor set, respectively. The curve is the fitting line of all the corresponding points. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Detailed results about the descriptor number and performance are then presented in Table 2. The LR model was also trained on 70% randomly selected HEAs and tested on the remaining 30%. Four indicators are adopted here to describe the performance: accuracy could evaluate the power of the model predicting correctly, the recall ratio reflects the ability of the classifier to find all positive samples, the precision is the capability of the classifier to find positive samples in all positive samples, and f1-score combines the results of precision and recall to reflect the comprehensive performance. It can be seen in

Table 2
Prediction performance of LR models with different descriptors number. The precision、recall and f1-score are calculated on X-Dimension descriptor set, with X ∈ [1,20].

Descriptor dimension	Accuracy	Precision	Recall	f1-score
1-D	0.58	0.43	0.58	0.44
2-D	0.62	0.51	0.62	0.53
3-D	0.65	0.6	0.65	0.58
4-D	0.69	0.65	0.69	0.65
5-D	0.78	0.78	0.78	0.76
6-D	0.81	0.82	0.81	0.8
7-D	0.87	0.88	0.88	0.87
8-D	0.88	0.9	0.89	0.88
9-D	0.89	0.9	0.89	0.89
10-D	0.88	0.9	0.89	0.87
11-D	0.9	0.91	0.9	0.89
12-D	0.9	0.91	0.9	0.89
13-D	0.9	0.91	0.9	0.89
14-D	0.9	0.91	0.9	0.89
15-D	0.9	0.91	0.9	0.89
16-D	0.9	0.91	0.9	0.89
17-D	0.91	0.92	0.91	0.9
18-D	0.91	0.92	0.91	0.9
19-D	0.88	0.9	0.89	0.88
20-D	0.87	0.89	0.88	0.86

Table 2, all these parameters grow when the descriptor dimension increases with the descriptor number ranges from 1 to 9. With more descriptors input the model, the performance tends to be stable. However, according to the result of cross-validation, the generalization ability of the model tends to decrease. Therefore, the first 9 selected descriptors could best represent the dataset in this work.

To show the efficiency of our model, we calculated the five-fold cross-validation accuracy of other five classic ML algorithms on different descriptors set. The comparison is shown in Fig. 4. After removed the relevant descriptors, there is no obvious difference between the performance of original 14 descriptors and the left 9 descriptors for all

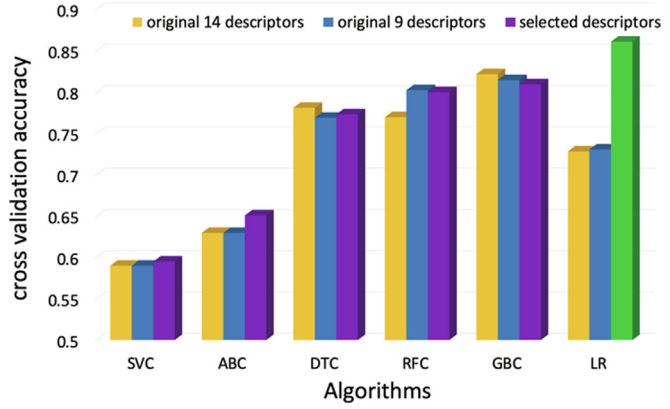

Fig. 4. The comparison between different descriptors sets and ML methods. SVC: Support Vector Classifier, ABC: Ada Boost Classifier, DTC: Decision Tree Classifier, RFC: Random Forest Classifier, GBC: Gradient Boosting Classifier.

algorithms, the result shows that it is rational to implement Pearson correlation analysis. And it is interesting to notice that all algorithms except LR indicate the similar trends which is slight fluctuation between the original descriptors set and the selected descriptors set. Nevertheless, it is about 11% improvement for LR model. The accuracy of our model with selected descriptor set (the green pillar) is higher than the other methods with the original descriptors set, which is clearly displayed in Fig. 4. That is, the prediction performance of our linear model is better than the other non-linear models. It indicates that, the non-linear combination of the original descriptors combined with the linear algorithm could boost the prediction performance for materials research. One aspect about this improvement possibly because the non-linear combination of primary features creates more complex descriptors that could better correlate with the target output. On the other hand, the nonlinear combination of descriptors could provide us some non-exist descriptors that express the relationship between the input and output variables, which can be used to discovery new physical laws.

Therefore, it is important to choose the proper algorithm when dealing with different material issues. ML method depends on the structure and quality of data sets. The appropriate selection of descriptors set and algorithms can improve the ML performance significantly [46,47].

4. Conclusions

In conclusion, a new method, which is based on feature engineering and ML, is proposed for predicting the phase formation of HEAs in this paper. We first analyzed the relationship between the original descriptors and removed the redundant descriptors. Then the low-dimension descriptors subset was selected from high dimensional descriptors space that constructed by several fundamental functions. The accuracy of cross-validation with original descriptors and selected descriptors shows that the constructed non-linear descriptors incorporating with linear algorithms outperform the original descriptors. The relative importance evaluation of the original descriptors indicates that VEC plays the most important role in predicting the phase formation of HEAs. Although this method achieves a satisfactory accuracy, there is still a lot of room to improve. For better characterization of materials, more suitable descriptors should be constructed in the future work. And larger dataset will also be helpful to improve the prediction accuracy. Furthermore, it is efficient to construct new descriptors by feature engineering to discover structure-property relationship and improve the prediction performance. However, though the interpretability of the selected descriptors and ML algorithms still remains outside our understanding even they could provide a superior result on average, it is believed that this efficient statistical methods will speed up material research [39]. In fact, this method can not only provide some help for the design and prediction of HEAs, but might also for the similar material issue with limited descriptors set.

CRediT authorship contribution statement

Dongbo Dai: Conceptualization, Writing - review & editing, Supervision. **Tao Xu:** Data curation, Formal analysis, Validation, Writing - original draft, Visualization. **Xiao Wei:** Writing - review & editing. **Guangtai Ding:** Writing - review & editing. **Yan Xu:** Writing - review & editing. **Jincang Zhang:** Writing - review & editing. **Huiran Zhang:** Conceptualization, Formal analysis, Methodology, Project administration, Writing - review & editing.

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.

Acknowledgement

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

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.commatsci.2020.109618.

References

[1] H.P. Chou, Y.S. Chang, S.K. Chen, J.W. Yeh, Microstructure, thermophysical and electrical properties in AlxCoCrFeNi ($0 \leq x \leq 2$) high-entropy alloys, Mater. Sci. Eng. B: Solid-State Mater. Adv. Technol. 163 (2009) 184–189, https://doi.org/10.1016/j.mseb.2009.05.024.

[2] Y.Y. Chen, U.T. Hong, H.C. Shih, J.W. Yeh, T. Duval, Electrochemical kinetics of the high entropy alloys in aqueous environments – a comparison with type 304 stainless steel, Corros. Sci. 47 (2005) 2679–2699, https://doi.org/10.1016/j.corsci.2004.09.026.

[3] O.N. Senkov, G.B. Wilks, J.M. Scott, D.B. Miracle, Mechanical properties of $Nb_{25}Mo_{25}Ta_{25}W_{25}$ and $V_{20}Nb_{20}Mo_{20}Ta_{20}W_{20}$ refractory high entropy alloys, Intermetallics 19 (2011) 698–706, https://doi.org/10.1016/j.intermet.2011.01.004.

[4] Y. Zhang, T. Zuo, Y. Cheng, P.K. Liaw, High-entropy alloys with high saturation magnetization, electrical resistivity, and malleability, Sci. Rep. 3 (2013) 1–7, https://doi.org/10.1038/srep01455.

[5] B. Cantor, I.T.H. Chang, P. Knight, A.J.B. Vincent, Microstructural development in equiatomic multicomponent alloys, Mater. Sci. Eng. A 375–377 (2004) 213–218, https://doi.org/10.1016/j.msea.2003.10.257.

[6] J.W. Yeh, S.K. Chen, S.J. Lin, J.Y. Gan, T.S. Chin, T.T. Shun, C.H. Tsau, S.Y. Chang, Nanostructured high-entropy alloys with multiple principal elements: novel alloy design concepts and outcomes, Adv. Eng. Mater. 6 (2004), https://doi.org/10.1002/adem.200300567 299–303 + 274.

[7] N.P. Tazuddin, K. Gurao, Biswas, In the quest of single phase multi-component multiprincipal high entropy alloys, J. Alloys Compd. 697 (2017) 434–442, https://doi.org/10.1016/j.jallcom.2016.11.383.

[8] F.G. Coury, P. Wilson, K.D. Clarke, M.J. Kaufman, A.J. Clarke, High-throughput solid solution strengthening characterization in high entropy alloys, Acta Mater. 167 (2019) 1–11, https://doi.org/10.1016/j.actamat.2019.01.029.

[9] M.C. Gao, C. Zhang, P. Gao, F. Zhang, L.Z. Ouyang, M. Widom, J.A. Hawk, Thermodynamics of concentrated solid solution alloys, Curr. Opin. Solid State Mater. Sci. 21 (2017) 238–251, https://doi.org/10.1016/j.cossms.2017.08.001.

[10] R. Chen, G. Qin, H. Zheng, L. Wang, Y. Su, Y.L. Chiu, H. Ding, J. Guo, H. Fu, Composition design of high entropy alloys using the valence electron concentration to balance strength and ductility, Acta Mater. 144 (2018) 129–137, https://doi.org/10.1016/j.actamat.2017.10.058.

[11] S. Guo, C.T. Liu, Phase stability in high entropy alloys: formation of solid-solution phase or amorphous phase, Prog. Nat. Sci. Mater. Int. 21 (2011) 433–446, https://doi.org/10.1016/S1002-0071(12)60080-X.

[12] V. Botu, R. Ramprasad, Adaptive machine learning framework to accelerate ab initio molecular dynamics, Int. J. Quantum Chem. 115 (2015) 1074–1083, https://doi.org/10.1002/qua.24836.

[13] C. Jiang, B.P. Uberuaga, Efficient ab initio modeling of random multicomponent alloys, Phys. Rev. Lett. 116 (2016) 1–5, https://doi.org/10.1103/PhysRevLett.116.105501.

[14] W.P. Huhn, M. Widom, Prediction of A2 to B2 phase transition in the high-entropy alloy Mo-Nb-Ta-W, JOM 65 (2013) 1772–1779, https://doi.org/10.1007/s11837-013-0772-3.
[15] R. Raghavan, K.C. Hari Kumar, B.S. Murty, Analysis of phase formation in multi-component alloys, J. Alloys Compd. 544 (2012) 152–158, https://doi.org/10.1016/j.jallcom.2012.07.105.
[16] T. Kostiuchenko, F. Körmann, J. Neugebauer, A. Shapeev, Impact of lattice relaxations on phase transitions in a high-entropy alloy studied by machine-learning potentials, NPJ Comput. Mater. 5 (2019) 55, https://doi.org/10.1038/s41524-019-0195-y.
[17] A. Takeuchi, A. Inoue, Quantitative evaluation of critical cooling rate for metallic glasses, Mater. Sci. Eng., A 304–306 (2001) 446–451, https://doi.org/10.1016/S0921-5093(00)01446-5.
[18] Y. Tan, J. Li, Z. Tang, J. Wang, H. Kou, Design of high-entropy alloys with a single solid-solution phase: average properties vs. their variances, J. Alloys Compd. 742 (2018) 430–441, https://doi.org/10.1016/j.jallcom.2018.01.252.
[19] Y. Zhang, Y.J. Zhou, J.P. Lin, G.L. Chen, P.K. Liaw, Solid-solution phase formation rules for multi-component alloys, Adv. Eng. Mater. 10 (2008) 534–538, https://doi.org/10.1002/adem.200700240.
[20] Z. Wang, S. Guo, C.T. Liu, Phase selection in high-entropy alloys: from nonequilibrium to equilibrium, JOM 66 (2014) 1966–1972, https://doi.org/10.1007/s11837-014-0953-8.
[21] X. Yang, Y. Zhang, Prediction of high-entropy stabilized solid-solution in multi-component alloys, Mater. Chem. Phys. 132 (2012) 233–238, https://doi.org/10.1016/j.matchemphys.2011.11.021.
[22] Z. Wang, Y. Huang, Y. Yang, J. Wang, C.T. Liu, Atomic-size effect and solid solubility of multicomponent alloys, Scr. Mater. 94 (2015) 28–31, https://doi.org/10.1016/j.scriptamat.2014.09.010.
[23] Y.F. Ye, Q. Wang, J. Lu, C.T. Liu, Y. Yang, High-entropy alloy: challenges and prospects, Mater. Today 19 (2016) 349–362, https://doi.org/10.1016/j.mattod.2015.11.026.
[24] Y. Zhang, W.J. Peng, Microstructural control and properties optimization of high-entrop alloys, Proc. Eng. 27 (2012) 1169–1178, https://doi.org/10.1016/j.proeng.2011.12.568.
[25] S. Guo, Phase selection rules for cast high entropy alloys: an overview, Mater. Sci. Technol. 31 (2015) 1223–1230, https://doi.org/10.1179/1743284715y.0000000018.
[26] M. Calvo-Dahlborg, S.G.R. Brown, Hume-Rothery for HEA classification and self-organizing map for phases and properties prediction, J. Alloys Compd. 724 (2017) 353–364, https://doi.org/10.1016/j.jallcom.2017.07.074.
[27] Q.W. Xing, Y. Zhang, Amorphous phase formation rules in high-entropy alloys, Chin. Phys. B. 26 (2017), https://doi.org/10.1088/1674-1056/26/1/018104.
[28] S. Guo, Q. Hu, C. Ng, C.T. Liu, More than entropy in high-entropy alloys: forming solid solutions or amorphous phase, Intermetallics 41 (2013) 96–103, https://doi.org/10.1016/j.intermet.2013.05.002.
[29] I. Toda-Caraballo, P.E.J. Rivera-Díaz-Del-Castillo, A criterion for the formation of high entropy alloys based on lattice distortion, Intermetallics 71 (2016) 76–87, https://doi.org/10.1016/j.intermet.2015.12.011.
[30] D.B. Miracle, O.N. Senkov, A critical review of high entropy alloys and related concepts, Acta Mater. 122 (2017) 448–511, https://doi.org/10.1016/j.actamat.2016.08.081.
[31] K.T. Butler, D.W. Davies, H. Cartwright, O. Isayev, A. Walsh, Machine learning for molecular and materials science, Nature 559 (2018) 547–555, https://doi.org/10.1038/s41586-018-0337-2.
[32] W. Huang, P. Martin, H.L. Zhuang, Machine-learning phase prediction of high-entropy alloys, Acta Mater. 169 (2019) 225–236, https://doi.org/10.1016/j.actamat.2019.03.012.
[33] N. Islam, W. Huang, H.L. Zhuang, Machine learning for phase selection in multi-principal element alloys, Comput. Mater. Sci. 150 (2018) 230–235, https://doi.org/10.1016/j.commatsci.2018.04.003.
[34] P. Dey, J. Bible, S. Datta, S. Broderick, J. Jasinski, M. Sunkara, M. Menon, K. Rajan, Informatics-aided bandgap engineering for solar materials, Comput. Mater. Sci. 83 (2014) 185–195, https://doi.org/10.1016/j.commatsci.2013.10.016.
[35] F. Nargesian, H. Samulowitz, U. Khurana, E.B. Khalil, D. Turaga, Learning feature engineering for classification, IJCAI Int. Joint Conf Artif. Intell. (2017) 2529–2535.
[36] G. Pilania, A. Mannodi-Kanakkithodi, B.P. Uberuaga, R. Ramprasad, J.E. Gubernatis, T. Lookman, Machine learning bandgaps of double perovskites, Sci. Rep. 6 (2016) 1–10, https://doi.org/10.1038/srep19375.
[37] 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. Mater. 2 (2018), https://doi.org/10.1103/PhysRevMaterials.2.083802.
[38] 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) 1–8, https://doi.org/10.1038/s41467-018-05761-w.
[39] J. Schmidt, Recent advances and applications of machine learning in solid- state materials science, NPJ Comput. Mater. (2019), https://doi.org/10.1038/s41524-019-0221-0.
[40] Y. Zhang, C. Wen, C. Wang, S. Antonov, D. Xue, Y. Bai, Y. Su, Phase prediction in high entropy alloys with a rational selection of materials descriptors and machine learning models, Acta Mater. (2019), https://doi.org/10.1016/j.actamat.2019.11.067.
[41] A. Mangal, E.A. Holm, A comparative study of feature selection methods for stress hotspot classification in materials, Integr. Mater. Manuf. Innov. 7 (2018) 87–95, https://doi.org/10.1007/s40192-018-0109-8.
[42] W. Li, R. Jacobs, D. Morgan, Predicting the thermodynamic stability of perovskite oxides using machine learning models, Comput. Mater. Sci. 150 (2018) 454–463, https://doi.org/10.1016/j.commatsci.2018.04.033.
[43] S. Guo, C. Ng, J. Lu, C.T. Liu, Effect of valence electron concentration on stability of fcc or bcc phase in high entropy alloys, J. Appl. Phys. 109 (2011), https://doi.org/10.1063/1.3587228.
[44] R. Kohavi, D. Sommerfield, Feature subset selection using the wrapper method: overfitting and dynamic search space topology, First Int Conf. Knowl. Discov. Data Min. 1995, pp. 192–197.
[45] D.M. Hawkins, The problem of overfitting, J. Chem. Inf. Comput. Sci. 44 (2004) 1–12, https://doi.org/10.1021/ci0342472.
[46] Y. Iwasaki, I. Takeuchi, V. Stanev, A.G. Kusne, M. Ishida, A. Kirihara, K. Ihara, R. Sawada, K. Terashima, H. Someya, K. Ichi Uchida, E. Saitoh, S. Yorozu, Machine-learning guided discovery of a new thermoelectric material, Sci. Rep. (2019) 1–7, https://doi.org/10.1038/s41598-019-39278-z.
[47] R. Liu, Y.C. Yabansu, A. Agrawal, S.R. Kalidindi, A.N. Choudhary, Machine learning approaches for elastic localization linkages in high-contrast composite materials, Integr. Mater. Manuf. Innov. 4 (2015), https://doi.org/10.1186/s40192-015-0042-z.