Applied Materials Today 36 (2024) 102007

Contents lists available at ScienceDirect

Applied Materials Today

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

Solid-liquid phase transition temperature prediction of alloys based on machine learning key feature screening

Jiheng Fang [a,b], Shangrong Yang [b], Ming Xie [b,*], Jieqiong Hu [b], Hongsheng Sun [c], Guohua Liu [b], Shangqiang Zhao [b], Yongtai Chen [b], Youcai Yang [b], Dekui Ning [b], Xingqun He [d,*], Qinglin Jin [a,*]

[a] Faculty of Materials Science and Engineering, Kunming University of Science and Technology, Kunming 650093, China
[b] Kunming Institute of Precious Metals, Kunming 650106, China
[c] Kunming Institute of Physics, Kunming 650106, China
[d] Xi'an Noble Rare Metal Materials Co. Ltd., Xi'an 710201, China

ARTICLE INFO

Keywords:
Machine learning
Multi-component alloy
Feature screening
Solid-liquid phase transition temperature
Symbolic regression

ABSTRACT

A machine learning strategy is proposed based on the demand for prediction of solid-liquid phase transition temperature properties of multi-component precious metal alloys. Firstly, the candidate feature set are constructed by mathematical operation of the physical and chemical parameters of the material according to the chemical ratio of the alloy chemical formula. Then, a novel feature selection framework of correlation screening → genetic algorithm screening → feature weight ranking → exhaustive screening is proposed to effectively identify key feature combinations affecting solid-liquid phase transition temperature. Finally, the support vector regression algorithm is used to establish the "key feature combination-solid phase transition temperature" model with an error less than 9.83% and the "key feature combination-liquid phase transition temperature" model with an error less than 9.35%. The proposed new feature framework overcomes the inability of conventional feature selection techniques to simultaneously meet the requirements of interpretability, low computational complexity, and strong feature generality. Moreover, the comprehensive equilibrium complexity and accuracy (R^2) constructed "solid temperature-key feature combination" with an error of less than 6.50% and "liquid phase transition temperature-key feature combination" with an error of less than 6.71 % by symbolic regression algorithm, which strengthened the understanding of the influence of key independent variable feature parameters on solid-liquid phase transition temperature prediction. This established machine learning strategy has excellent reliability in the prediction of solid-liquid phase transition temperature of multi-component alloys, and the proposed feature selection framework is expected to develop into a new feature selection technique.

1. Introduction

Precious metal alloy solders occupies an irreplaceable position in a wide range of fields including electronics industry, microelectronic packaging, vacuum multi-stage brazing, high temperature technology, jewelry manufacturing and aerospace [1–3], and solid-liquid phase transition temperature is one of the key indicators to determine the specific service scenarios of precious metal alloy solder. The solid-liquid phase transition temperature of binary and ternary precious metal alloy solders can be obtained from conventional phase diagrams or from some thermodynamic calculation software (Factsage, Pandat, JmaPro, Thermo-Calc, etc.). However, the lack of data on the phase equilibria and thermodynamic properties of multi-component precious metal alloy solders (e.g., quaternary, quintuple and above) and the complexity of multi-element interactions lead to the fact that relevant solid-liquid phase transition temperature predictions are not yet possible, ultimately leading to a lack of guidance in designing multi-component precious metal alloy solders with ideal solid-liquid phase transition temperatures. In addition to the solder field, solid-liquid temperature is also used as a key parameter in some multi-component precious metal alloy application scenarios, for example, liquid phase transition temperature is an important parameter in the design of superalloys, electrical contact materials and refractory alloys [4–6]; Solid phase transition temperature plays a critical role in alloy solidification, semi-solid or completely liquid metal alloy molding design [7,8]; While the solid-liquid phase transition temperature difference is generally

* Corresponding authors.
E-mail addresses: 921539423@qq.com (M. Xie), jdhxq@qq.com (X. He), 15559870920@163.com (Q. Jin).

https://doi.org/10.1016/j.apmt.2023.102007
Received 30 May 2023; Received in revised form 8 November 2023; Accepted 22 November 2023
Available online 5 December 2023

applied in alloy design (eutectic alloy), solidification crack judgment and solidification feature analysis [9,10]. However, the solid-liquid phase transition temperature of multi-component precious metal alloys in other fields mentioned above is the same as that of multi-component precious metal solder alloys, which indicates that there is an urgent need to propose an effective method to predict the solid-liquid phase transition temperature of multi-component precious metal alloys.

In the past few years, machine learning has become the frontier and hot field of material research [11–16]. For example, Hart et al. [17] reviewed the present state of machine-learning-driven alloys research, including metallic glasses, high-entropy alloys, shape-memory alloys, magnets, superalloys, catalysts and structural materials, and discussed the approaches and applications in this field. The applications of machine learning in rechargeable battery materials design and discovery are reviewed by Liu et al. [18], including the property prediction for liquid electrolytes, solid electrolytes, electrode materials, and the discovery of novel rechargeable battery materials through component prediction and structure prediction. A summary of recent advances in the intersection of the fields of 2D materials and ML was provided by Yin et al. [19], who discussed atomistic structure analysis, property prediction, ML-assisted preparation of 2D material and revealed their connections. Machine learning allows us to discover new materials in a new way that is different from the "trial-and-error" approach: based on material databases, ML makes new predictions by understanding the hidden patterns in the data, providing guidance for material design, which is expected to accelerate the design of new materials and shorten the material development cycle [20]. It is particularly suitable for the prediction of alloy material properties because it can overcome the problems of traditional experimental methods (high cost and low efficiency) as well as cross-scale computational simulations (heavy workload). Currently, some examples of successful application of machine learning technology in the prediction of alloy material properties include electrical conductivity and hardness of electrical contact alloy materials [21], critical casting size of amorphous alloy [22], creep fracture stress of Ni-base superalloy [23], solid solution formation ability of high entropy alloy [24], mechanical properties of medical magnesium alloy (ultimate tensile strength, yield strength and elongation) [25], catalytic properties of multicomponent precious metal alloy catalytic materials [26], etc. This implies that it is feasible to apply machine learning technology to predict the solid-liquid phase transition temperature properties of multi-component precious metal alloys, and there is almost no research on machine learning-assisted solid-liquid phase transition temperature prediction.

Data and features determine the upper limit of machine learning, and models and algorithms approximate this upper limit [17], which suggests the importance of feature selection. In materials research, each feature set is generally only for the application of specific conditions, and there is no uniform feature that is valid for all applications, which leads to the selection of the most appropriate feature for each machine learning process as one of the other challenges. Currently, a lot of work is being done on feature selection. It is worth noting that the feature selection method can be divided into four types according to the form of feature selection: (i) feature selection based on domain knowledge. This technique is excellent in explainability, but it encounters difficulties with insufficient domain knowledge in many cases; (ii) Filter selection features [27], Commonly used filtering methods include correlation coefficient [28], variance screening method [29], mutual information [30]. Its advantage is that it is efficient in calculation time and has high robustness to overfitting problems, but its disadvantage is that it does not take into account the association between features, resulting in useful association features may be mistakenly eliminated; (iii) Wrapper selection feature [31,32], commonly used packaging methods include branch and bound search [33], breadth-first traversal [34], bi-directional search (BDS) [35], sequence forward selection (SFS) [36], randomly generated sequence selection algorithm [37], genetic algorithm [38]. Compared with the Filter method, the feature subset classification performance found by this method is usually better. However, its disadvantage is that the screened features are not universal, and it needs to re-select the features for the learning algorithm when it changes the learning algorithm. Since the classifier is trained and tested for each evaluation of a subset, the computational complexity of the algorithm is high. In addition, this method has a large number of key features screened when the candidate feature set is large, resulting in its poor interpretability; and (iv) Embedded feature selection [39], including: feature selection method based on tree model and feature selection method based on penalty term [40]. It is fast and effective, but parameter setting requires deep background knowledge. According to the above analysis, the above feature selection methods can realize feature selection in the machine learning process, but the above feature selection methods can not meet the requirements of feature selection such as less domain knowledge, low computational complexity, strong generality and high interpretability, especially when there are a large number of candidate features in small sample data or complex application scenarios. Therefore, it is necessary to propose a reasonable framework for selecting the best feature sets to optimize such problems, which has always been a difficult problem in the field of using machine learning technology to predict material properties.

The black-box model of machine learning is often criticized for failing to provide new "laws of physics" or to obtain explicit expressions, which limits its potential in some cases. Symbolic regression is an interpretable machine learning method, which combines mathematical operators (such as +, -, *, /, sin, cos, log, etc.) and material features by intelligent algorithms, so as to provide specific mathematical expressions between objective functions and independent variable feature parameters, which can be used to represent structure-activity relationships [41]. Therefore, symbolic regression can be applied to construct the mathematical expression of solid-liquid phase transition temperature of multi-component precious metal solder alloy, and the influence of key independent variable feature parameters on solid-liquid phase transition temperature prediction can be enhanced based on the expression.

In conclusion, based on the demand of solid-liquid phase transition temperature prediction of multi-component precious metal alloys, machine learning strategy is proposed to realize solid-liquid phase transition temperature performance prediction. When screening a large number of machine learning candidate features, conventional feature selection techniques can not meet the requirements of interpretability, low computational complexity and strong generality of features at the same time. In this work, a reasonable feature selection framework of correlation screening→ genetic algorithm screening→ feature weight ranking → exhaustive screening is proposed to identify key feature combinations affecting solid-liquid phase transition temperature. In addition, in order to enhance the understanding of the influence of the key independent variable feature parameters on the solid-liquid phase transition temperature prediction, the specific mathematical expressions between the target properties (solid-liquid phase transition temperature) and independent feature parameters are provided by symbolic regression method, which provide a reference for the study of solid-liquid phase transition temperature of alloy materials.

2. Methodology

2.1. Overall machine learning strategy for predicting solid-liquid phase transition temperature

Based on the designed machine learning framework, the key features of the solid-liquid phase transition temperature prediction model are screened and the solid-liquid phase transition temperature performance is predicted, and the mathematical expression for predicting the solid-liquid phase transition temperature is constructed based on the symbolic regression algorithm. The overall idea is shown in Fig. 1. Firstly,

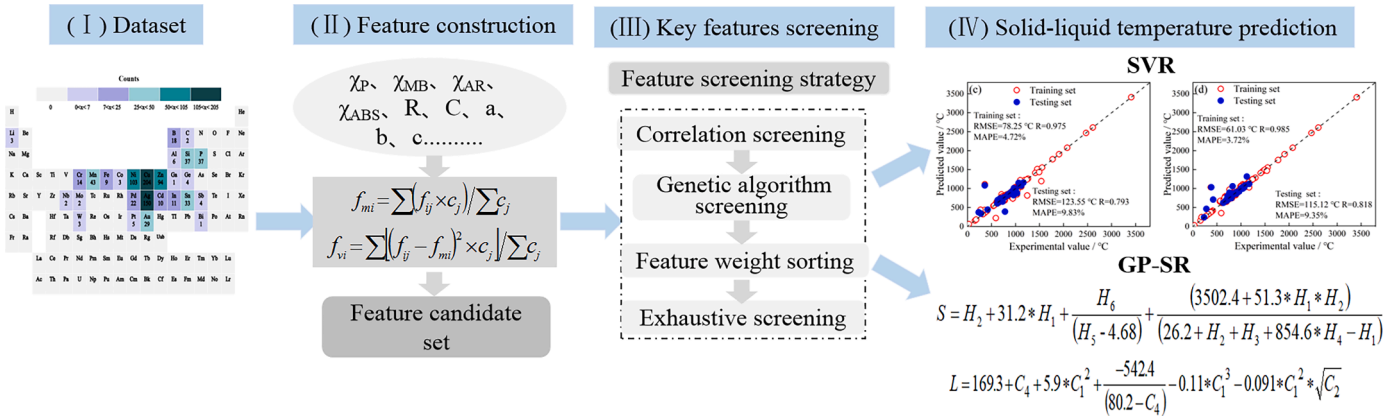

Fig. 1. Overall machine learning strategy for predicting solid-liquid phase transition temperature.

the composition-performance data of the solder alloy are collected, and then the set of physical and chemical parameters is constructed, and a feature set is constructed to evaluate the influence of each physico-chemical parameter on the target properties according to the chemical ratio of the chemical formula of the collected solder alloy. Then, the key features are screened based on the proposed feature screening combination strategy, based on the screened key features, a machine learning prediction model for solid-liquid transition phase temperature was established using support vector regression algorithm, and a mathematical expression for predicting solid-liquid phase transition temperature was constructed using symbolic regression algorithm.

2.2. Data collection and alloy feature construction

2.2.1. Data collection

Alloy solder mainly includes precious metal alloy solder and non-precious metal alloy solder. Therefore, the original intention of our design is to establish a solid-liquid phase transition temperature properties prediction model with good prediction effect based on the data of precious metal alloy solder, and then extend it to the dataset of non-precious metal alloy solder to verify the reliability of the model, and finally realize the prediction of solid-liquid phase transition temperature properties of all alloy solders.

(1) Selected precious metal alloy solder dataset: The original data of machine learning focuses on the precious metal alloy solder system, which mainly includes silver-based, gold-based, palladium-based and a small amount of platinum-based solder, as shown in Fig. 2. Ag, Au, Pd and Pt are located in the 10th and 11th columns of the periodic table. The data of precious metal alloy solder is relatively small. Considering that the chemical properties of the elements of the same main set are similar, the collected data include nickel-based and copper-based alloy solder data. The above data of Ag, Au, Pd, Pt, Ni and Cu are mainly from

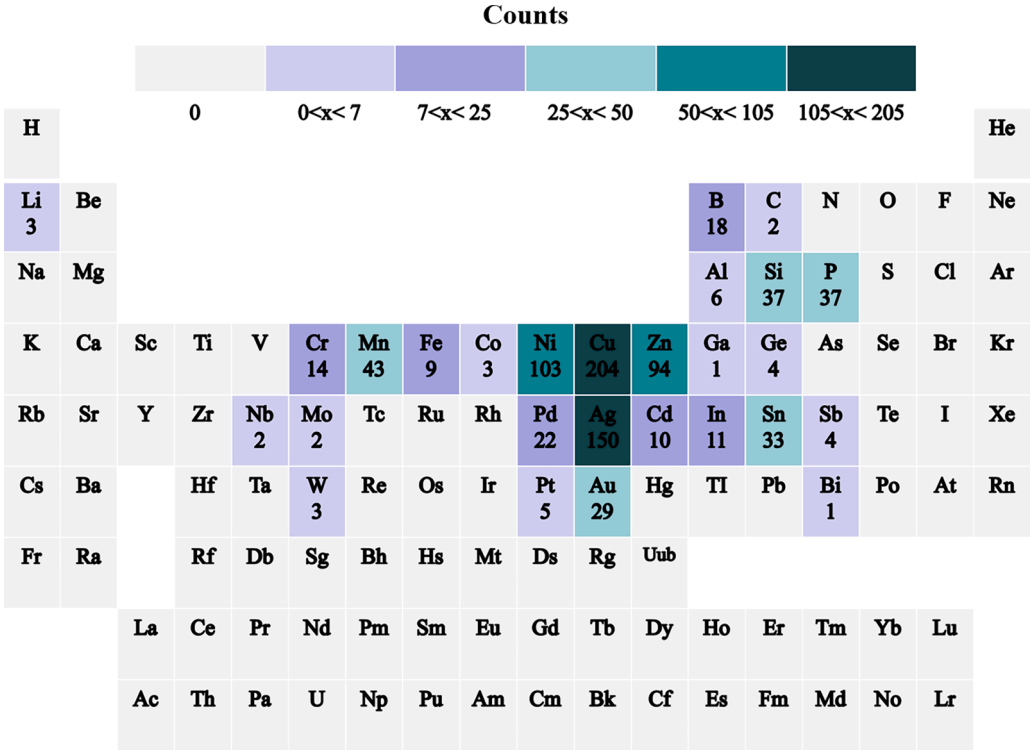

Fig. 2. Distribution of elements in the collected data. (The number under each element name represents the number of times the element occurs in the study system, and each element is also coloured, with grey indicating that the collected dataset does not contain this element.).

our laboratory, Chinese national standards [42–45], books [46, 47] and references [48–53]. A total of 267 sets of valid data were obtained. The process and state are basically the same, and the composition space is shown in Fig. 2. Among them, the proportion of one-element alloy, binary alloy, ternary alloy, quaternary alloy, five-element alloy and six-element alloy in the total data set is 11.24 %, 16.48 %, 34.21 %, 29.58 %, 6.74 %, 2.62 %. For specific data, see **supplementary material S1**.

(2) Selected non-precious metal alloy solder data set: in order to verify the general applicability of the key feature combinations used in other non-precious metal alloy solder, 60 sets of non-precious metal alloy solder data under the same process conditions were collected from tin-lead solder, manganese-based solder, lead-free tin solder and aluminum-based solder [54–57], which can be found in **supplementary material S2**.

2.2.2. Feature set construction

It is almost impossible to establish a machine learning model with strong generalization ability for the features that determine the material properties or service behavior, such as material composition, structure / structure, preparation process, etc. This, coupled with the small amount of data collected, leads to the construction of new features of the elemental physical and chemical parameters in the data set through mathematical operations according to material theory or prior knowledge, so as to establish a high-quality regression model based on small sample data. 50 parameters such as nuclear electron distance, metal radius, atomic enthalpy and bulk modulus are extracted to construct the set of physical and chemical parameters. For details of the specific parameters, see the **supplementary material S3**. Then, a feature set used to evaluate the influence of each parameter on the target quantity is constructed according to the chemical ratio of the collected alloy chemical formula, which replaces the direct input of the chemical formula. In order to investigate the effects of the average and difference levels of the basic physical and chemical parameters in the alloy on the properties of the alloy, two kinds of alloy features were constructed by Eq. (1) and (2) (see **supplementary material S4**). The specific construction process is as follows: Eq. (1) is used to calculate 50 feature mean factors f_{mi} of each alloy, Eq. (2) is used to calculate 50 feature variance factors f_{vi} of each alloy, and f_{mi} and f_{vi} are used as inputs to the machine learning performance prediction model.

$$f_{mi} = \sum\left(f_{ij} \times c_j\right) \Big/ \sum c_j \tag{1}$$

$$f_{vi} = \sum\left[\left(f_{ij} - f_{mi}\right)^2 \times c_j\right] \Big/ \sum c_j \tag{2}$$

where f_{ij} represents the physical/chemical parameters of the element (i represents the individual physical/chemical parameters of the element, $i = 1, 2, \ldots, 50$; j represents the individual elements of the alloy, $j = 1, 2, 3, 4, 5, 6, 7, \ldots, 27$) and c_j represents the elemental content.

2.3. Key feature screening strategy and machine learning model construction

2.3.1. Key feature screening strategy

In order to realize the prediction of solid-liquid phase transition temperature properties of precious metal alloy solder and non-precious metal alloy solder, a solid-liquid phase transition temperature properties prediction model with good prediction effect was established by using the dataset of precious metal alloy solder as training and testing sets. The dataset of precious metal alloy solder is randomly divided into two parts: training set (80 %) and testing set (20 %). Then, the non-precious metal alloy solder dataset is used as the verification dataset to verify the reliability of the model. The fact that the precious metal alloy solder dataset is not divided into a separate verification set is helpful to solve the problem of small amount of precious metal alloy solder dataset. At the same time, the non-precious metal alloy solder dataset as a verification set is more helpful to reveal whether the solid-liquid phase transition temperature properties prediction model has general applicability in all alloy data. For the dataset, new features constructed by calculating the physical and chemical parameters of the chemical formula are used as the input, and the solid phase transition temperature and liquid phase transition temperature are the predicted target values of the machine learning model. The training set is used to complete the screening of key features, the establishment of machine learning model and the construction of mathematical expressions; the test set is used to verify the effect of feature screening, machine learning model and mathematical expression evaluation. Correspondingly, a feature selection framework of correlation screening→ genetic algorithm screening→ feature weight ranking →exhaustive screening is constructed by fully combining the advantages of feature selection techniques such as "Filter method" and "Wrapper method", which will be used for machine learning key feature screening of solid-liquid phase transition temperature. The basic process and principles of the feature selection framework used in this study are as follows:

(a) The remaining n features after linear correlation screening form the feature set:

$$F = \{X_1, X_2, X_3, X_4 \cdots X_i, X_{i+1} \cdots X_n\} \tag{3}$$

Where, X_i denotes the ith feature screened, and n is the number of features remaining after linear correlation screening.

(a) The genetic algorithm is used to perform k screenings, as shown in Eq. (4) below, and m features are screened out each time ($m = 10$ in this study, the basis for the determination of m is detailed in Section 3.1.2), and regression modelling carried out based on m features, with a model prediction accuracy of p_k: (i.e. p_k=1-*MAPE*)

$$1 = \{V_1, V_2, V_3, V_4 \cdots V_i, V_{i+1} \cdots V_m\} \rightarrow P_1$$

$$2 = \{X_1, X_2, X_3, X_4 \cdots X_i, X_{i+1} \cdots X_m\} \rightarrow P_2$$

$$3 = \{Y_1, Y_2, Y_3, Y_4 \cdots Y_i, Y_{i+1} \cdots Y_m\} \rightarrow P_3$$

$$K = \{Z_1, Z_2, Z_3, Z_4 \cdots Z_i, Z_{i+1} \cdots Z_m\} \rightarrow P_k \tag{4}$$

where, V_i, X_i, Y_i and Z_i all denote the ith feature screened.

(a) The feature weights are shown in Eq. (5), and the feature weights are equal to the sum of the products of the same features and the prediction accuracy of the model after each genetic algorithm screening respectively, and the sum of the products of the features and the prediction accuracy of different classes of features will be ranked later.

$$W_A = \sum_{i=1}^{k} p_k D_n^{A,k} \tag{5–1}$$

$D_n^{A,k}$ is whether feature A is screened at the kth screening, and $D_n^{A,k} = 1$ if the screened feature is A, otherwise $D_n^{A,k} = 0$;

$$W_B = \sum_{i=1}^{k} p_k D_n^{B,k} \tag{5–2}$$

$D_n^{B,k}$ is whether feature B is screened at the kth screening, and $D_n^{B,k} = 1$ if the screened feature is B, otherwise $D_n^{B,k} = 0$; and so on..

$$W_N = \sum_{i=1}^{k} p_k D_n^{N,k} \tag{5–3}$$

$D_n^{N,k}$ is whether feature N is screened at the kth screening, and $D_n^{N,k} = 1$ if the screened feature is N, otherwise $D_n^{N,k} = 0$.

(a) Feature ranking: the sum of the product of the features and prediction accuracy of the different classes of features processed by the feature weighting formula is ranked and the plot drawn is the ranking for calculating the cumulative sum of the accuracy generated by these features:

$$I_n = rank(W_A, W_B, W_C, ..., W_M) \tag{6}$$

According to the above principle of feature weight ranking, the sum of the product of different types of features and their prediction accuracy W_N provides a score that indicates the usefulness or value of each feature when constructing the target performance value in the model. When it is used to predict the target performance value, the higher the prediction accuracy of a feature contribution and the higher the frequency of using this feature, the higher the relative importance of this feature. The results of W_N are obtained on the basis of K times screening and K times target performance prediction by genetic algorithm. Compared with the single or less screening process, the feature weight ranking method proposed in this study can significantly reduce the chance caused by the influence of genetic algorithms in the ranking of the importance of key features.

(a) Exhaustive screening:

Exhaustive method, also known as enumeration method, is the concrete embodiment of brute force strategy, is a simple and direct method to solve the problem [58]. Its basic idea is to enumerate all the situations involved in the problem one by one, and test which are solutions and which should be excluded according to the conditions put forward by the problem. The candidate features of the top M most important features constitute exhaustive screening are screened out separately by feature weight ranking (Considering the computational complexity of exhaustive analysis, the number of M should not be too large. Based on the results of feature importance ranking, it is recommended to select features within 15 that have the highest importance ranking). Then, the feature combination with the best prediction accuracy of the model is exhaustively screened, which avoids the huge amount of computation when exhaustive screening of all candidate feature sets, and it overcomes the problem of not considering the correlation effect between features during correlation screening→ genetic algorithm screening→ feature weight ranking filtering.

2.3.2. Machine learning model construction

The properties prediction of this study belongs to the regression problem, and the commonly used regression algorithms include: linear regression, polynomial regression, support vector regression, tree regression, multi-layer perceptron regression, Gaussian regression, K-nearest neighbors regression, stochastic gradient descent regression, ensemble tree regression, naive bayes regression, random forest regression, gradient boosting regression, XGBoost regression, LightGBM regression, neural network regression, etc. [59,60]. Based on the characteristics of each algorithm, the number of datasets, and the linear or nonlinear relationship of the data, the above regression algorithms were screened, and the support vector regression (SVR), Gaussian regression (GPR), tree regression (TR) and ensemble tree regression (ETR) algorithms were randomly selected as candidate algorithms for this study. Subsequently, after comparing the predictive effectiveness of several typical machine learning regression algorithms, namely SVR, GPR, TR, and ETR, the SVR algorithm has higher accuracy. Therefore, in this paper, the support vector regression algorithm, which is suitable for small data samples, is selected as the machine learning modeling algorithm in the screening of key feature combinations and the model building process. the SVR algorithm has the characteristics of minimizing the risk of structuralization, and has strong generalization ability, etc. [61]. In the process of feature screening and final prediction model verification, root mean square error (RMSE), average absolute percentage error (MAPE) and linear regression correlation coefficient (R) are employed to reflect the overall deviation degree of the prediction model and the effect of comprehensive analysis and modeling. The calculation equations of RMSE, MAPE and R are shown in Eqs. (7), (8) and (9), respectively. In addition, for the screened key features of solid-liquid phase transition temperature, Gaussian regression (GPR), tree regression (TR) and ensemble tree regression (ETR) algorithms are tried to be incorporated into modeling and prediction to verify the applicability of the key feature combinations screened by the feature screening framework (using support vector regression algorithm) to other models.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{7}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\% \tag{8}$$

$$R = \frac{\sum_{i=1}^{N}(x_i - \bar{x}_i)(y_i - \bar{y}_i)}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x}_i)^2 \sum_{i=1}^{N}(y_i - \bar{y}_i)^2}} \tag{9}$$

where N is the number of samples, y_i is the experimental value, $\hat{y}_i$ is the predicted value, and $\bar{y}_i$ is the mean of the experimental values. x_i is the experimental value of another observation and $\bar{x}_i$ is its average value. When equation (5) is used for feature correlation screening, x_i and y_i represents the two different features of the ith alloy ($i = 1, 2, ..., N$), and $\bar{x}_i$ and $\bar{y}_i$ represents the average of these two different features across the N alloys. The correlation coefficient R ranges from −1 to +1, with positive values indicating a positive correlation, negative values indicating a negative correlation, and higher absolute values indicating a stronger correlation.

2.4. Construction of mathematical expression of solid-liquid phase transition temperature

In the process of constructing the mathematical expression of solid-liquid transition temperature, the symbolic regression algorithm (GP-SR) based on genetic programming is selected as the machine learning algorithm [62]. Symbolic Regression (SR) is a supervised machine learning method used to find some hidden mathematical expression or function to best fit a given dataset. The most commonly used algorithm for solving SR problems is Genetic Programming (GP). GP is the core algorithm of SR. By introducing self-defined functions and dynamic program service methods, it has achieved remarkable results in the fields of machine learning, artificial intelligence, combinatorial optimization, adaptive systems and control technology. Compared with the genetic algorithm which uses fixed-length string coding, GP overcomes the shortcomings such as the inability to describe hierarchy and the lack of dynamic variability by using functional expressions [63]. For a detailed discussion of GP-SR formalism, please refer to the work of Koza et al. [64], Goel et al. [65], Vyas et al. [66] and Langdon et al. [67].

Symbolic regression calculation is realized by Eureqa Formulize software (the specific usage of the software can be referred to [68,69]). In order to establish the regression formula by using the symbolic

regression method, it is necessary to determine the input independent and dependent variables, mathematical operation symbol set and fitness function, in which the selection of mathematical operation symbol set and fitness function is the key to the optimization of regression algorithm. The screened key feature combination of solid-liquid phase transition temperature (X_1, X_2……X_n) is taken as input variabl/e, solid-liquid phase transition temperature Y as output variable, and the mathematical expression of target is Y = f (X_1, X_2……X_n), where n represents the number of key features screened. The selected function symbols include +, -, ×, /, exp (x), log (x) and sqrt (x), and the complexity coefficient of each symbol adopts the default complexity value of the software. The expression is evaluated by the fitness function of R^2. After the Eureqa Formulize software runs automatically, the resulting regression model will be automatically updated and evaluated until the resulting model is stable and meets the measurement criteria. Root mean square error (RMSE), mean absolute error (MAE) and coefficient of determination R^2 were used to evaluate the fitting effect of the mathematical expression. The specific calculation formulas are shown in Eqs. (7) (10) and (11):

$$MAE = \frac{1}{N}\sum_{1}^{N}|y_i - \hat{y}_i| \tag{10}$$

$$R^2 = 1 - \frac{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{\frac{1}{N}\sum_{i=1}^{N}(\bar{y}_i - y_i)^2} \tag{11}$$

Where N is the number of samples, y_i is the experimental value, $\hat{y}_i$ is the predicted value, and $\bar{y}_i$ is the average of the experimental values. the closer the value of R^2 is to 1, the better the fit, and the smaller the values of RMSE and MAE, the smaller the dispersion and the better the fit.

2.5. Experimental methods

For the verification test of Cu-Ag-Zn-Mn-Ni-Si-B-P multi-component alloy system, Ag, Cu, Zn, Mn, Ni and Si with purity ≥ 99.99 %, as well as Ni-B and Cu-P master alloys were selected as experimental materials. The alloy ingots with diameter of Φ 20 mm were prepared by vacuum melting and steel mold casting, and the alloy was homogenized at 700C for 4 h. Then, the composition analysis and solid-liquid phase transition temperature test were carried out from the three positions of the head, middle and tail of the homogenized ingot. In the composition analysis test, 2 g samples were taken from each sample, and the alloy composition was determined by inductively coupled plasma atomic emission spectrometry (ICP-AES). The solid-liquid phase transition temperature of each sample was 0.5 g. The solid-liquid phase transition temperature of the alloy was measured by NETZSCH STA 409PG/PC synchronous thermal analyzer. High purity nitrogen was added to the sample chamber as a protective atmosphere at a heating rate of 10 °C min^{-1}.

3. Results and discussion

3.1. Selection of key features of solid-liquid phase transition temperature of alloy and prediction of solid-liquid phase transition temperature

In order to clarify the key machine learning features affecting the solid-liquid phase transition temperature of the alloy, three feature screening strategies were tried: (i) linear correlation screening, that is, genetic algorithm screening based on the unlimited number of features; (ii) linear correlation screening, that is, screening → feature weight ranking based on genetic algorithm limiting the number of features; and (iii) Linear correlation screening, that is, genetic algorithm screening based on limiting the number of features→ feature weight ranking→ exhaustive screening. The following analysis of the feature results screened by the three strategies: The feature results screened by the three strategies are analyzed below:

3.1.1. Filtering features based on strategy I (linear correlation filtering → genetic algorithm filtering based on unlimited number of features)

Results of linear correlation screening. In the linear correlation screening, the linear correlation degree of each alloy feature is analyzed, and the alloy features with strong linear correlation are classified into the same set according to the strong linear correlation of Pearson correlation coefficient greater than or equal to 0.95. In each set, the alloy feature with the lowest modeling error with a single feature is screened to represent the alloy feature. The screened alloy features continued to calculate their Pearson correlation coefficients for grouping. The alloy features that model the best single eigenquantities are repeatedly screened until the Pearson correlation coefficient between the screened alloy features < 0.95, which is considered to exclude strong linear correlations between alloy features. As shown in Fig. 3, after linear correlation screening, the remaining alloy features of both the solid phase transition temperature model and the liquid phase transition temperature model are 55 (see **supplementary material S5**). After grouping screening, there was a strong linear correlation among the alloy factors in each set (|R|≥0.95), but there was no strong linear correlation among the alloy factors (|R|<0.95).

Results of feature screening for strategy I. Based on the unrestricted number of features genetic algorithm screening, the model parameters are: 100 generations, 200 populations, model parameter optimization, 50 random sampling, each sampling using 50 % discount cross-validation. Taking the candidate feature set of solid-liquid phase transition temperature screened by linear correlation as an example, the process of screening features by unrestricted feature number genetic algorithm are shown in Fig. 4(a) (b) respectively. Finally, 37 features were screened from solid phase transition temperature candidate features, including features 5, 35, 38, 39, 45, 48, 51, 53, 55, 58, 61, 62, 63, 64, 66, 67, 68, 70, 71, 72, 73, 75, 77, 78, 79, 80, 81, 82, 85, 86, 88, 89, 95, 96, 97, 99 and 100; and 16 features were screened from liquid phase transition temperature candidate features, including features 11, 13, 58, 62, 65, 68, 70, 73, 75, 79, 88, 95, 97, 98, 99 and 100. The specific feature types can be checked according to the screened feature number and compared with the **supplementary material S4**. However, the number of features screened out by the unrestricted feature genetic algorithm is large, which is undoubtedly not conducive to the interpretability of the model, and it is difficult to find the key features affecting the solid-liquid phase transition temperature.

3.1.2. Filtering features based on strategy ii (linear correlation filtering → genetic algorithm filtering based on limiting the number of features → feature weight ranking)

According to the feature screening carried out by the above-mentioned genetic algorithm based on the number of unrestricted features, the number of features screened is larger, which is obviously not conducive to the interpretability of the model, and it is difficult to determine the key features that affect the solid-liquid phase transition temperature. Therefore, the second strategy is adopted, that is, linear correlation filtering → genetic algorithm filtering based on limiting the number of features → feature weight ranking. The linear correlation screening process and results of strategy II are the same as those of strategy I, which will not be repeated here. In strategy II, limiting the number of features in genetic algorithm screening helps to improve the interpretability of the model. In addition, limiting the number of features also helps to reduce the number of information redundant features into the screening of subsequent key features. Based on the interpretable and small computational requirements of this study, and subsequent exhaustive screening, only the features with the highest importance (less than the first 15 features) are selected as the candidate feature set,

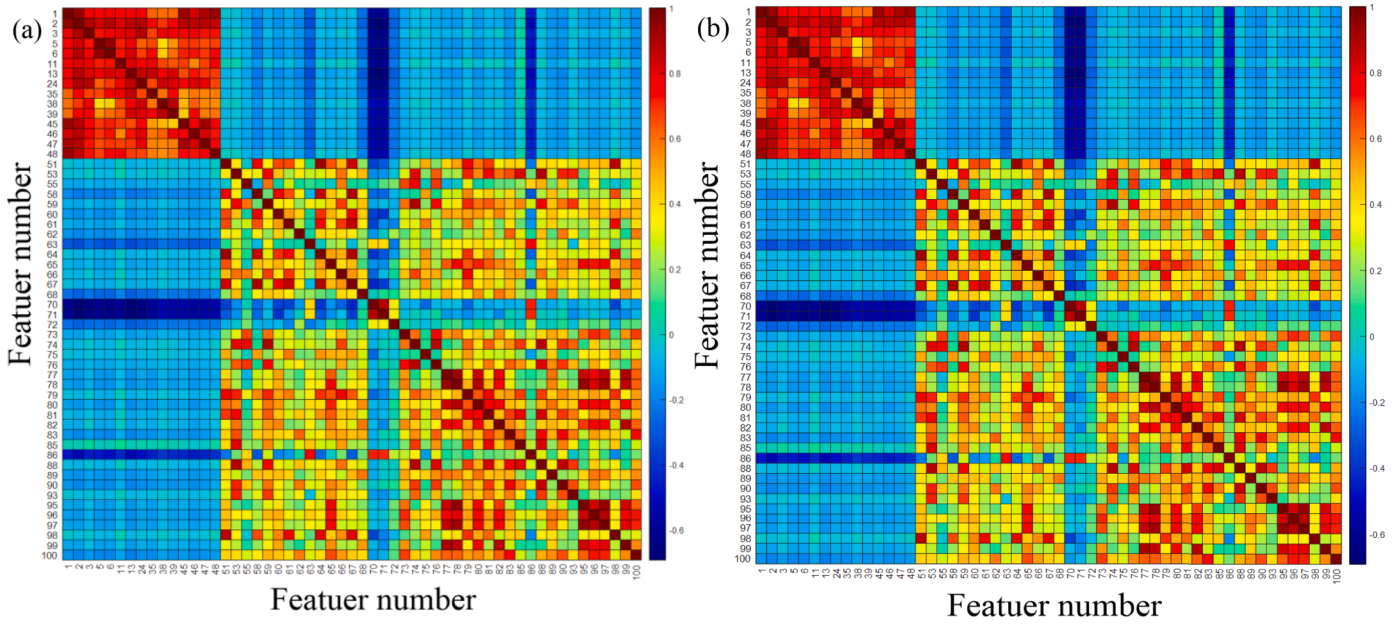

Fig. 3. The correlation performance of alloy features after linear correlation screening: (a) solid phase transition temperature, (b) liquid phase transition temperature.

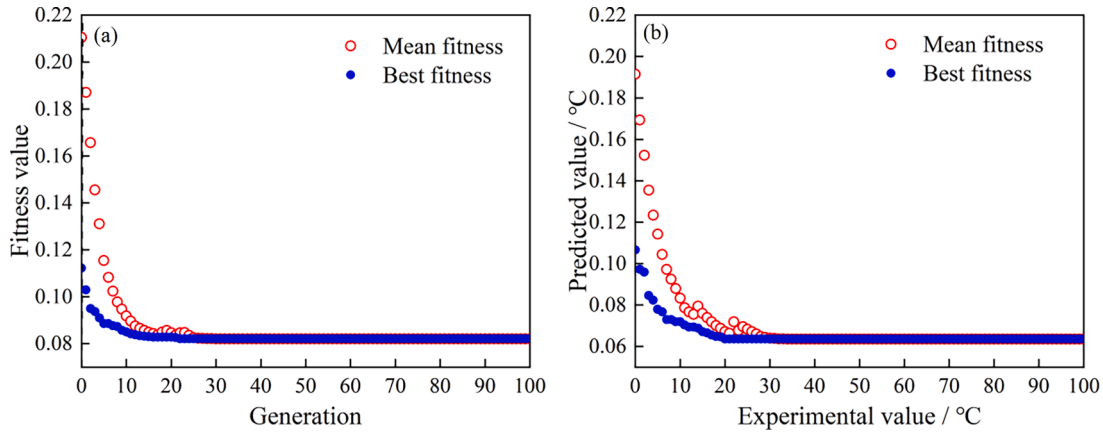

Fig. 4. Screening of solid / liquid phase transition temperature features carried out based on genetic algorithm screening of unrestricted number of features (a) model error of genetic algorithm in screening solid phase transition temperature feature processes; (b) model error of genetic algorithm in screening liquid phase transition temperature feature processes.

therefore, in strategy II, the number of features selected by genetic algorithm is set to 10. **supplementary material S6** shows that limiting the number of different feature screenings (5, 10, and 15) within a certain range has a relatively small impact on the final feature screening results. Therefore, this study limits the number of screened features to 10. By combining the genetic algorithm based on limiting the number of features (10) with the feature weight ranking method, the fast screening and ranking of key features can be realized, and the ranked key features exclude chance and are characterized by feature universality.

Feature screening results and corresponding modeling effects for strategy ii. The screening results of the key features of solid phase transition temperature after adopting strategy 2 are shown in Fig. 5(a). According to the sum order of feature prediction accuracy, the first five key features affecting solid phase transition temperature are numbered 35, 38, 81, 74 and 85, respectively, corresponding to the average melting enthalpy, the average bulk modulus, atomic radius 2 (coordination number 12) variance, environmental atomic number variance and melting enthalpy variance features, especially 35 and 38 have higher fitness values, indicating that 35 and 38 are the most important machine learning features that affect the prediction of solid phase transition temperature. The screened 35 and 38 features are directly modeled by the SVR algorithm, and the prediction effect of the model on the solid phase transition temperature is shown in Fig. 5(c). The average absolute percentage error MAPE=11.86 % on the model of the test set is 11.86 %, that is, the model has a good prediction effect on the solid phase transition temperature, which verifies the effectiveness of strategy 2 (linear correlation filtering→genetic algorithm screening based on limiting the number of features (10) → feature weight ranking).

After the same strategy is applied to screen the features of liquid phase transition temperature, the results are shown in Fig. 5(b). The first five key features affecting liquid phase transition temperature are numbered 35, 38, 81, 85 and 95, corresponding to the average melting enthalpy, liquid phase transition temperature bulk modulus, atomic radius 2 (coordination number 12) variance, melting enthalpy variance and mass decay coefficient CrKalpha variance, respectively. Obviously, the fitness values of 35 and 38 are also the highest. Compared with the first five features of solid phase transition temperature, four features of

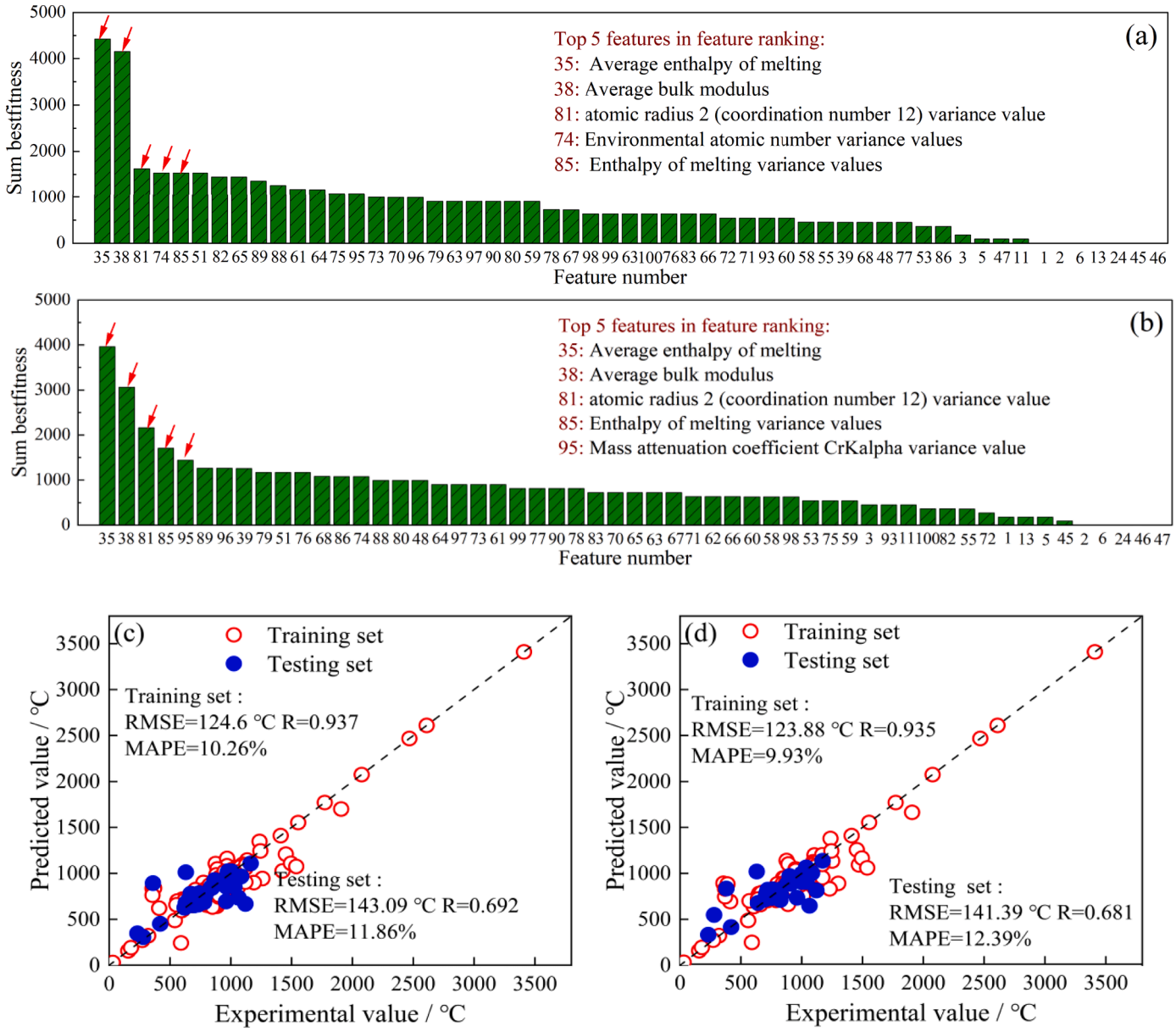

Fig. 5. Feature screening results and their modelling prediction results carried out using strategy 2 (linear correlation filtering → genetic algorithm screening by limiting the number of features (10) → feature weight ranking). (a) prediction accuracy summation and ranking results of solid phase transition temperature features; (b) prediction accuracy summation and ranking results for liquid phase transition temperature features; (c) prediction results for solid phase transition temperature after modelling based on both 35 and 38 features; (d) prediction results for liquid phase transition temperature after modelling based on both 35 and 38 features.

liquid phase transition temperature are the same, and the order of the first three features is the same. Most fundamentally, 35 and 38 are also the most important machine learning features that affect the prediction of liquid phase transition temperature. The screened 35 and 38 features are modeled directly by the SVR algorithm, and the prediction effect of the model on liquid phase transition temperature is shown in Fig. 5(d). The average absolute percentage error MAPE=12.39 %, of the test set on the model, that is, the prediction effect of the model on the liquid phase transition temperature is good, which verifies the effectiveness of strategy 2 again.

Performance of prediction models based on strategy ii with different number of features. More importantly, based on the feature prediction accuracy addition and ranking results of the solid phase transition temperature in Fig. 5(a), the model is established by the first two key features (35, 38), the first three key features (35, 38, 81) and the first five key features (35, 38, 81, 74, 85) combined with SVR algorithm. The prediction effect of the model for solid phase transition temperature is shown in Figs. 6(a) (b) (c). The average absolute percentage error MAPE of the test set on the model is 11.86 %, 10.99 % and 9.97 %, respectively. It is obvious that when the number of modeling features screened out by strategy 2 $\leq$ 5, with the increase of the number of features, the prediction accuracy of the solid phase transition temperature model increases, and the accuracy of the model is high, which ensures that the model has good explanatory property. For liquid phase transition temperature, similarly, the model is established by the previous two key features (35, 38), the first three key features (35, 38, 81), and the first five key features (35, 38, 81, 85, 95) combined with SVR algorithm. The prediction effect of the model on the liquid phase transition temperature is shown in Figs. 6 (d) (e) (f). The average absolute percentage error MAPE of the test set on the model is 12.39 %, 11.21 % and 10.62 %, respectively.

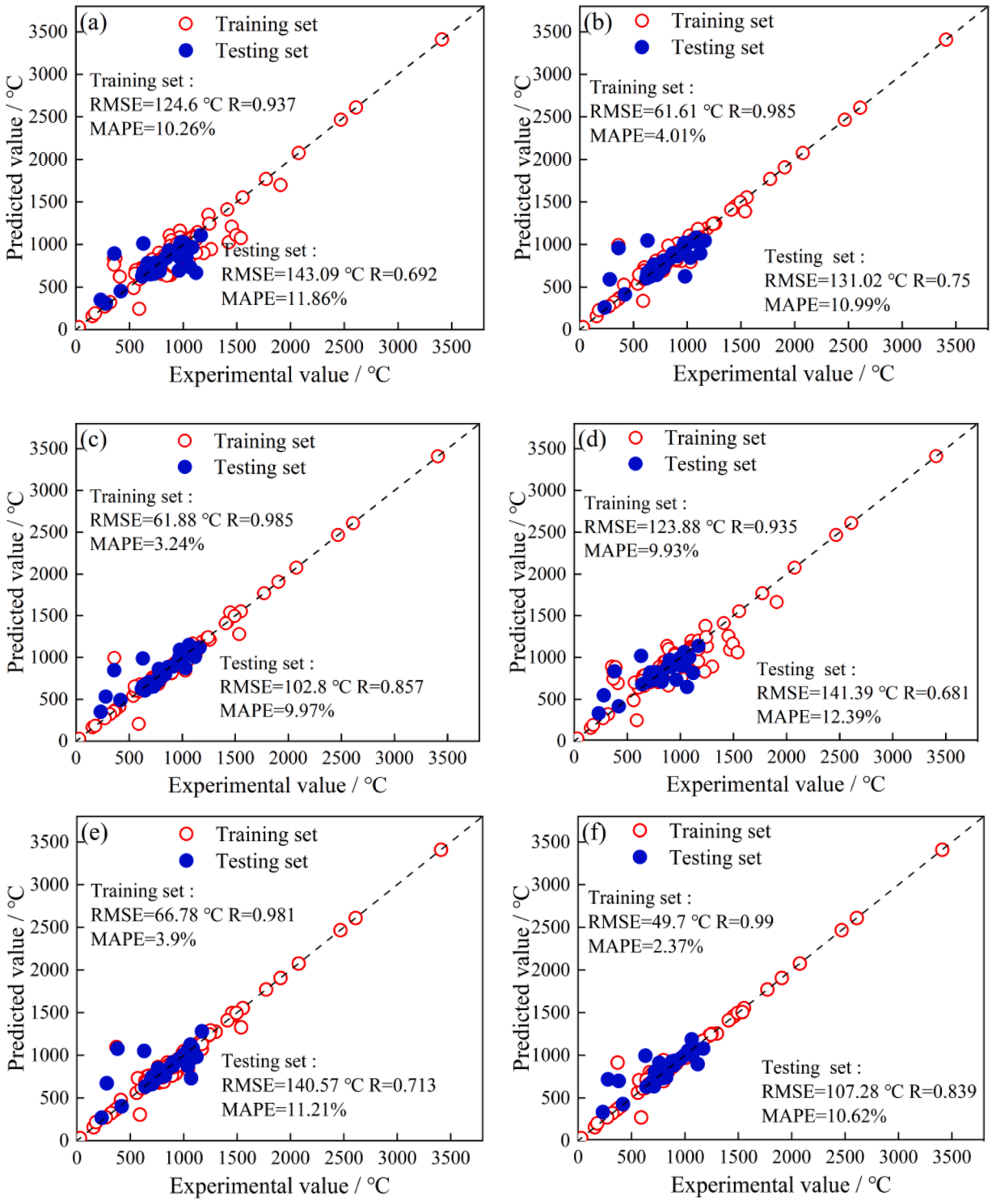

Fig. 6. Predicted performance of solid-liquid phase transition temperature models based on different numbers of modelling features. Solid phase transition temperature models: (a) modelled based on 35 and 38 features; (b) modelled based on 35, 38 and 81 features; (c) modelled based on 35, 38, 81, 74 and 85 features; liquid phase transition temperature models: (d) modelled based on 35 and 38 features; (e) modelled based on 35, 38 and 81 features; (f) modelled based on 35, 38, 81, 85 and 95 features.

3.1.3. Filtering features based on strategy III (linear correlation filtering → genetic algorithm filtering based on limiting the number of features → feature weight ranking → exhaustive screening)

In general, the prediction accuracy and explanation of feature screening using strategy 2 are good, but it does not take into account the correlation between features, resulting in useful correlation feature errors may be removed. The most critical features screened to form the feature set are modeled directly through the SVR algorithm based on the feature set, but it can not guarantee that all the most critical feature sets can achieve the best fit degree of the model. Basically, the best candidate feature combination of the model is not composed of the first few features screened by feature weight ranking. In view of the fact that strategy 2 does not take into account the correlation between features and in order to further improve the prediction accuracy of the model,

strategy 3 is adopted to carry out feature screening. In view of the shortcomings of strategy 2, employment strategy 3 carries out feature screening. The method of strategy 3 is to add exhaustive filtering on the basis of strategy 2. Specifically, the first 12 and 14 most important key features are screened to form exhaustive candidate features through the feature weight ranking(strategy 2). It should be noted that the prediction accuracy summation and ranking results of Fig. 5(a) and 5(b) show higher fitness values for the first 12 and 14 features of the solid / liquid phase transition temperature feature, respectively. Therefore, the first 12 and 14 features are selected as candidate feature sets respectively for the most important key features of solid-liquid phase transition temperature in exhaustive screening. Then, the feature combination with the best prediction accuracy of the model is screened through exhaustive screening, which constitutes strategy 3. Strategy 3 avoids the huge amount of computation when exhaustive screening is used for all candidate feature sets, and overcomes the fact that the correlation between features is not considered in strategy 2. According to the analysis of the above strategy 2, the features with universal features can be screened by the combination of genetic algorithm based on limiting the number of features (10) and feature weight ranking method. The features of exhaustive screening enumerating the problems one by one determine that the features screened by strategy 3 are remain universal, and this conclusion can be verified by subsequent different algorithms and different data.

Feature screening results and corresponding modeling effects for strategy III. Strategy 3 is applied to screen key feature combinations of solid phase transition temperature and liquid phase transition temperature. By exhaustively listing all the top 12/14 feature combinations screened by feature weights, the feature combinations with the lowest relative error are used for modeling training to improve the accuracy and generalization ability of the model. The exhaustive results of solid phase transition temperature and liquid phase transition temperature are shown in Fig. 7(a,b). For the solid phase transition temperature model, when the key alloy features are 6, the model average absolute percentage error (MAPE) is the smallest, which is 9.83 %; for the liquid phase transition temperature model, when the number of features is 5, the model average absolute percentage error (MAPE) is the smallest, which is 9.35 %. In the solid phase transition temperature and liquid phase transition temperature models, the specific alloy feature types screened out by exhaustive selection are shown in Tables 1 and 2, respectively.

According to the above alloy screening results, the support vector regression (SVR) algorithm, which is consistent with the alloy feature selection, is used for regression modeling. The modeling results of solid phase transition temperature and liquid phase transition temperature are shown in Fig. 7(c) (d). The results show that for the solid phase transition temperature prediction model, the root mean square error of the training set is 78.25 °C, the regression coefficient R is 0.975, and the percentage error is 4.72 %; the root mean square error of the test set is

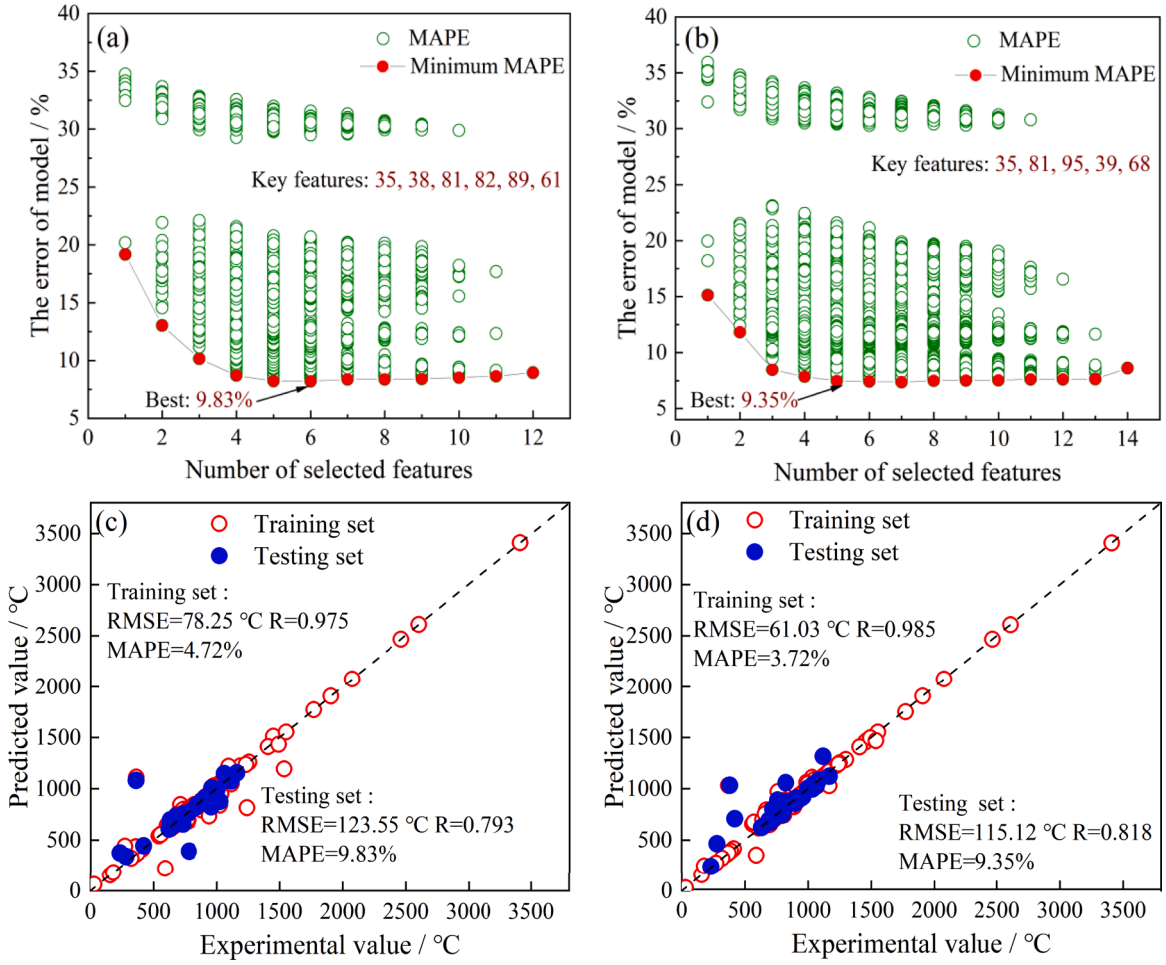

Fig. 7. The results of feature screening and their modelling predictions using strategy 3 (linear correlation filtering → genetic algorithm screening for limiting the number of features (10) → feature weight ranking → exhaustive screening). (a) results after exhaustive screening of solid phase transition temperature features; (b) results after exhaustive screening of liquid phase transition temperature features; (c) prediction results for solid phase transition temperature after modelling the best combination of features based on exhaustive screening; (d) prediction results for liquid phase transition temperature after modelling the best combination of features based on exhaustive screening.

Table 1
Screening results of alloy features for the solid phase transition temperature model.

Feature number	Feature name
35	Average melting enthalpy
38	Average bulk modulus
81	Atomic radius 2 (coordination number 12) variance
82	Atomic volume variance
89	Young's modulus variance
61	Electron affinity energy variance

Table 2
Screening results of alloy features for the liquid phase transition temperature model.

Feature number	Feature name
35	Average enthalpy of melting
81	Atomic radius 2 (coordination number 12) variance
95	Mass attenuation coefficient CrKalpha variance
39	Average Young's modulus
68	Melting point variance

123.55 °C, the regression coefficient R is 0.793, and the percentage error is 9.83 %. The errors are small, indicating that the training effect of the model is better and the generalization ability is better. For the liquid phase transition temperature prediction model, the root mean square error of the training set is 61.03 °C, the regression coefficient R is 0.985, and the percentage error is 3.72 %; the root mean square error of the test set is 115.12 °C, the regression coefficient R is 0.818, and the percentage error is 9.35 %. The error is small. According to the results, strategy 3 selects the key feature combinations that improve the generalization ability of the model again on the basis of strategy 2, and meets the requirements of low computational complexity, strong generality and interpretability of the model. In general, the training effect of the model is excellent and the prediction accuracy is high.

In summary, strategy 3 (correlation screening→ genetic algorithm screening→ feature weight ranking →exhaustive screening) can meet the requirements of interpretability, low computational complexity, strong feature versatility and good model prediction effect during screening of key features of solid-liquid phase transition temperature. Therefore, the framework of feature selection of correlation screening→ genetic algorithm screening→ feature weight ranking →exhaustive screening is expected to be developed into a new feature selection strategy. In addition, the machine learning strategy modeled by feature construction + key feature combination screening + support vector regression algorithm established a "solid phase transition temperature-key feature combination" model with an error of less than 9.83 % and a "liquid phase transition temperature-key feature combination" model with an error of less than 9.35 %, respectively, which realized the prediction of the solid-liquid phase transition temperature performance of noble metal alloys.

3.2. Construction of mathematical expression of solid-liquid phase transition temperature

Based on the key feature combinations screened by the above feature screening strategy, the mathematical expression of the target quantity is constructed by symbolic regression algorithm. The data of the key feature sets of solid-liquid phase transition temperature are used as the benchmark data set respectively. The detailed data can be found in the **supplementary material S7**. The "Treat all data points equally mode" of the software is screened according to the partition method of training data set and testing data set. According to the symbolic regression algorithm, the quantitative relationship between solid phase transition temperature S and key feature combination(35, 38, 81, 82, 89, 61, for the convenience of subsequent formulas, H_1, H_2, H_3, H_4, H_5 and H_6 are replaced respectively) and liquid phase transition temperature L and key feature combination(35, 81, 95, 39, 68, subsequently replaced by C_1, C_2, C3, C_4 and C_5) was established, and 25 sets of solid phase transition temperature and 24 sets of liquid phase transition temperature regression expressions were optimized respectively. For detailed data, see **supplementary material S8**. The effect of 25 sets of expressions on the solid phase transition temperature simulation is shown in Fig. 8(a). The range of model complexity coefficient is 3~43, and the range of RMSE, MAE and R^2 for simulated solid phase transition temperature and measured temperature is 89.9~287.7 °C, 50.7~167.3 °C, and 0.39~0.94, respectively. The effect of 24 sets of expressions on the liquid phase transition temperature simulation is shown in Fig. 8(b). The range of model complexity coefficient is 3~44, and the range of RMSE, MAE and R^2 for simulating solid phase transition temperature and measured temperature is 96.6~271.0 °C, 66.2~156.3 °C, and 0.43~0.93, respectively. Fig. 8 suggests that with the increase of the complex coefficient [70] of the regression model, RMSE and MAE tend to decrease and R^2 tends to increase, that is, the more accurate the regression formula is to simulate the temperature data. Sorted according to the size of the determination coefficient R^2, the first symbolic regression estimation formula of solid / liquid phase transition temperature is shown in Eqs. (12) and (13), respectively. The style of the formula is more complex, that is, it is almost impossible to determine the formula with larger coefficient R^2. Considering the equilibrium complexity and accuracy (R^2), the final selection of the symbolic regression estimation formula of solid / liquid phase transition temperature can be found in Eqs. (14) and (15), respectively.

$$S = H_2 + 30.2 * H_1 + \frac{26.0 * H_3}{(H_5^2 - 21.0)} + \frac{(3338.8 + 5.8 * H_5 + 52.3 * H_1 * H_2)}{(26.0 + H_1 + H_3 + 75.8 * H_1 * H_4 - H_1)} \tag{12}$$

$$L = 221.0 + C_4 + 5.8 * C_1^2 + 0.00038 * C_1 * C_2 * C_4 + \frac{-422.3}{(80.1 - C_4)} - 0.1 * C_1^3 - 0.17 * C_1^2 * \sqrt{C_2} \tag{13}$$

$$S = H_2 + 31.2 * H_1 + \frac{H_6}{(H_5 - 4.68)} + \frac{(3502.4 + 51.3 * H_1 * H_2)}{(26.2 + H_2 + H_3 + 854.6 * H_4 - H_1)} \tag{14}$$

$$L = 169.3 + C_4 + 5.9 * C_1^2 + \frac{-542.4}{(80.2 - C_4)} - 0.11 * C_1^3 - 0.091 * C_1^2 * \sqrt{C_2} \tag{15}$$

where S is the solid phase transition temperature and H1, H2, H3, H4, H5 and H6 represent the six key features of the solid phase transition temperature: mean enthalpy of melting, mean bulk modulus, variance of atomic radius 2 (coordination number 12), variance of atomic volume, variance of Young's modulus and variance of electron affinity energy. L is the liquid phase transition temperature and C1, C2, C3, C4 and C5 represent the five key features of the liquid phase transition temperature: mean enthalpy of melting, variance of atomic radius 2 (coordination number 12), variance of mass decay coefficient CrKalpha, mean Young's modulus and variance of melting point.

4. Discussion

4.1. Physical significance analysis of key alloy features for solid-liquid phase transition temperature

4.1.1. Physical significance analysis of key alloy features for solid phase transition temperature

The analysis of the final alloy features screened by the feature selection framework using correlation screening→ genetic algorithm

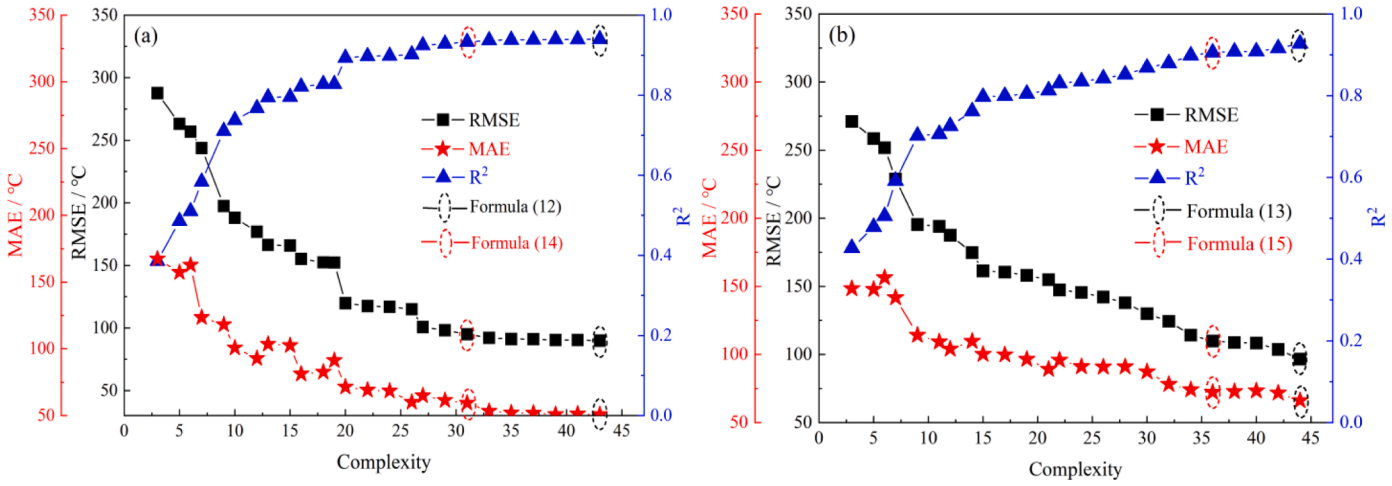

Fig. 8. The relationship between the complexity coefficient and simulation accuracy of the solid-liquid phase transition temperature regression formula. (a) solid phase transition temperature; (b) liquid phase transition temperature.

screening→ feature weight ranking →exhaustive screening supports the understanding of the main ways in which alloy features affect performance.

For the solid phase transition temperature model, the average melting enthalpy, the average bulk modulus, atomic radius 2 (coordination number 12) variance, atomic volume variance, Young's modulus variance and electron affinity variance play a key role in the model accuracy. The main factors affecting the solid state temperature are the melting heat of the alloy and the metal bond. Melting heat refers to the heat absorbed by a unit mass crystal when it becomes a liquid substance at the same temperature, and it is also equal to the heat released by a unit mass of the same substance when it changes from a liquid to a solid at the same pressure. The melting heat of a metal is approximately consistent with the melting point of the metal as follows:$\Delta H_m = aT_m^{b-1}$, where a and b represent the empirical constants related to the structure type of metal crystals, so the melting heat indirectly reflects the melting point, and the solid phase transition temperature is closely related to the liquid phase transition temperature, so the melting heat is the key factor affecting the solid phase transition temperature. The solid state temperature depends on the atomic binding force or metal bond in the substance [71]. Generally, the stronger the binding force or metal bond between atoms, the higher the solid state temperature of the metal. The main results are as follows: (1) Among the screened feature parameters, the melting enthalpy is the most important factor affecting the solid phase transition temperature. Under certain conditions, the melting enthalpy is equal to the melting heat. For example, at a certain temperature and pressure, the melting heat absorbed by the system before and after the melting process is equal to the increase of the melting enthalpy of the system before and after the process, and the melting enthalpy is closely related to the melting heat. Thus, the feature value of melting enthalpy is the key feature of the solid phase transition temperature. The higher the melting enthalpy, the higher the solid phase transition temperature. (2) The bulk modulus is the second key feature affecting the order of solid phase transition temperature, and the bulk modulus is included in the elasticity modulus. The elasticity modulus is a physical quantity that describes the elasticity of a substance, which is a general term and can be expressed by "Young's modulus", "shear modulus", "bulk modulus", etc. Elasticity modulus is an important performance parameter of engineering materials. From a microscopic point of view, it is a reflection of the bonding strength between atoms, ions or molecules. The elasticity modulus of the substance with high solid phase transition temperature is also large. In addition, the relationship between bulk modulus and elasticity modulus is: $K=E/(3 \times (1-2 \times \nu))$, where E is elasticity modulus and ν is Poisson's ratio. Therefore, there is a linear relationship between bulk modulus and elasticity modulus, that is, the bulk modulus of substances with high solid phase transition temperature is also large. (3) Atomic radius: the melting point and boiling point of general metals increase with the strength of metal bonds. The strength of the metal bond is usually inversely related to the radius of the metal ion and positively related to the free electron density in the metal (which can be roughly regarded as a positive correlation with the number of electrons outside the atom). Therefore, the atomic radius is the key factor affecting the solid phase transition temperature. The larger the atomic radius, the lower the solid phase transition temperature. (4) Atomic volume: the space occupied by each atom in the crystal is equal to the cell volume divided by the quotient of the number of atoms in the cell. The value of the atomic volume is more stable than the atomic radius, and the application is more convenient. There is a corresponding relationship between the atomic volume and the atomic radius. The effective nuclear charge of the main set elements in the same period increases significantly from left to right, but the electron layer (n) does not increase. The gravitation of the outer electrons is enhanced, which leads to the inward contraction of the outer electrons, resulting in the obvious decrease of the atomic radius r and the corresponding decrease of the atomic volume. For the main set elements, the atomic radius increases significantly from top to bottom, and the corresponding atomic volume increases. The atomic radius affects the solid phase transition temperature through metal bonds, that is, the atomic volume is the main factor affecting the solid phase transition temperature. The larger the atomic volume, the smaller the solid phase transition temperature. (5) Young's modulus: the Young's modulus is included in the elasticity modulus, so the material with high solid phase transition temperature has a large Young's modulus. In addition, when the temperature is lower than 300 K, $E = \frac{100kT_m}{V_a}$, where k represents constant, T_m represents the melting point, and V_a represents the volume of atoms or molecules. Therefore, Young's modulus is one of the key determinants of solid phase transition temperature. The higher the Young's modulus, the higher the solid phase transition temperature. (6) The magnitude of the electron affinity energy depends on factors such as the atomic radius of the atom. The smaller the atomic radius, the greater the attraction of the nucleus to electrons, the more energy released after combining electrons, and the greater the electron affinity energy, while smaller the atomic radius of the metal element, the stronger the metal bond, the higher the melting boiling point of the metal, that is, the greater the electron affinity energy, the higher the melting boiling point, so the electron affinity energy is closely related to the solid phase transition temperature. The higher the electron affinity energy, the higher the solid phase transition temperature.

In summary, the greater the melting enthalpy, volume modulus,

Young's modulus and electron affinity, the greater the liquid phase transition temperature; while the larger the atomic radius and atomic volume, the smaller the liquid phase transition temperature. The comprehensive effects of the six key features screened by the feature screening framework on the solid phase transition temperature properties of the alloy are finally determined by the solid phase transition temperature of the alloy.

4.1.2. Physical significance analysis of key alloy features for liquid phase transition temperature

For the liquid phase transition temperature model, the average melting enthalpy of each element, the variance of atomic radius 2 (coordination number 12), the variance of mass decay coefficient (CrKalpha), the average value of Young's modulus and the variance of melting point play a key role in the accuracy of the model. Similar to the solid phase transition temperature, the main factors affecting the liquid phase transition temperature are the melting heat of the alloy and the metal bond. The main results are as follows: (1) Among the screened feature parameters, the melting enthalpy is the most important factor affecting the liquid phase transition temperature. As mentioned above, the melting heat indirectly reflects the melting point, and under certain conditions, the melting enthalpy is equal to the melting heat, so the eigenvalue of the melting enthalpy is the key feature of the liquid phase transition temperature. The higher the enthalpy of melting, the higher the liquid phase transition temperature. (2) Atomic radius: one of the main influencing factors of metal bond is the atomic radius of metal elements. Generally speaking, the smaller the atomic radius is, the stronger the metal bond is, and the higher the melting point of the alloy is, so the atomic radius is the key factor affecting the liquid phase transition temperature. The larger the atomic radius, the lower the liquid phase transition temperature. (3) Mass decay coefficient CrKalpha; mass decay coefficient is approximately related to the atomic number of the absorbing material to the third power, while the atomic number is related to the melting point: with the increase of the atomic number of elements in the same period, the melting point of the metal element increases, while the melting point of the metal element of the same set of elements decreases from top to bottom, so the mass decay coefficient indirectly affects the liquid phase transition temperature through the atomic coefficient. The liquid phase transition temperature increases with increasing mass decay coefficients for the same period elements and decreases with increasing mass decay coefficients for elements of the same main group. (4) Young's modulus is included in the elasticity modulus, and the relationship between the elasticity modulus and the melting point is $E = kT_m^a c^b$, where E represents the elasticity modulus, T_m represents the melting point, c represents the specific heat capacity, and k, a, b represents the constant, $a \approx 1$, $b \approx 2$. Obviously, there is a linear relationship between the elasticity modulus and the melting point, and the elasticity modulus of the material with high melting point is also large. Therefore, there is a linear relationship between Young's modulus and melting point, and the material with high liquid phase transition temperature also has a large Young's modulus. (5) Melting point: the melting point of each element in the alloy plays a key role in the melting point of the alloy. When the melting point of each component is higher, the melting point of the alloy is generally higher. However, compared with the pure metal, the size of the atoms in the alloy is different and the arrangement is not as neat as that of the pure metal, which reduces the interaction force between the atoms, so the melting point of most alloys is generally lower than that of various constituent metals.

In summary, the larger the melting enthalpy, Young's modulus and melting point, the higher the liquid phase transition temperature; the larger the atomic radius, the smaller the liquid phase transition temperature; While the liquid phase transition temperature increases with increasing mass decay coefficients for the same period elements and decreases with increasing mass decay coefficients for elements of the same main group. The comprehensive effect of the above five key features selected by the feature screening framework on the liquid phase transition temperature properties of the alloy are finally determined by the liquid phase transition temperature of the alloy.

4.2. Reliability and universal applicability of the screened key feature combinations

4.2.1. Reliability analysis of screened key feature combinations

In order to further clarify how the test error depends on the potential chemical properties, the solid-liquid phase transition temperature predictions of all 267 sets of effective data are calculated based on the established solid-liquid phase transition temperature machine learning model. Then, combined with the measured values, the average absolute percentage error of the predicted value of each set of data is calculated. Tables 3 and 4 provide the alloy composition whose average absolute percentage error of solid-liquid phase transition temperature is more than 11.5 %, respectively. The number of data in the two tables is 11 and 10 respectively, accounting for only 4.12 % and 3.75 % of the 267 data, respectively, indicating once again that the established solid-liquid phase transition temperature machine learning model is successful. However, some predictions are not satisfactory, such as some alloys containing P, Al and Ge, eutectic alloys and metal elements. For alloys containing P, Al and Ge, there are more alloys of this type in 267 sets of effective data, but there are few data that produce similar obvious errors. P, Al and Ge are low melting point elements. The alloy composed of them and high melting point elements contains many complex intermediate phases, and the curve of melting point test contains more endothermic peaks, which hinders the calibration of the starting point and end point of solid-liquid phase transition temperature, and is vulnerable to measurement data errors. and more complex intermediate phases also make the prediction of solid-liquid phase transition temperature machine learning model face challenges. For eutectic alloys and metal elements, there is only a single phase in the alloy or element. Compared with a large number of multi-phase alloy training data, the number of similar single-phase training data is small, which leads to a large prediction error of machine learning, which will need to be further studied and improved in the future, such as adding similar data for iterative optimization or combining alloy phase recognition steps in the process of modeling.

4.2.2. Universal applicability of screened key feature combinations to other algorithms

In order to verify the universal applicability of the key feature combinations selected by the feature screening framework (using support vector regression algorithm) on other algorithms, Gaussian regression, tree regression and ensemble tree regression were used to establish a solid phase transition temperature prediction model using 35, 38, 81, 82, 89 and 61 as the modeling features, respectively. The

Table 3
Alloy composition with an average absolute percentage error of more than 11.5 % in predicting solid phase transition temperature.

Alloy composition / Wt%	Measured values / °C	Predicted value / °C	MAPE /%	Remarks
89.5Au-0.5Ag-10Ge	356	435.2	22.2	Contains Ge
94Ag-1Mn-5Al	780	1018.1	30.5	Contains Al
80Au-20Sn	280	337.4	20.5	Eutectic
Ag	961.9	823.6	14.4	Simple
Fe	1538	1195.1	22.3	element
Mn	1244	816.7	34.3	
Sn	231.9	374	61.3	
Bi	271.5	440.7	62.3	
P	590	219.3	62.8	
C	3500	1150.3	67.1	
Ga	29.8	62	108.0	

Table 4
Alloy composition with an average absolute percentage error of more than 11.5 % in predicting liquid phase transition temperature.

Alloy composition / Wt%	Measured values / °C	Predicted value / °C	MAPE /%	Remarks
89.03Cu-4.95Ag-6.02P	696.7	798.9	14.7	Contains P
80.36Cu-14.65Ag-4.99P	669.9	801.8	19.7	
94Ag-1Mn-5Al	825	1058	28.2	Contains Al
48Ni-32Mn-20Pd	1120	923.3	17.6	Eutectic
80Au-20Sn	280	661.2	136.1	
Fe	1538	1232.9	19.8	Simple substance
Sn	231.9	342.1	47.5	
Sb	630	1010.3	60.4	
P	590	193.5	67.2	
C	3500	1082.6	69.1	

performance of the model is evaluated by testing the modeling prediction error of the model on the data set. Similarly, with 35, 81, 95, 39 and 68 as modeling features, the prediction models of liquid phase transition temperature were established by Gaussian regression, tree regression and ensemble tree regression, respectively. As shown in Fig. 9, the prediction percentage errors of Gaussian regression, tree regression and ensemble tree regression for solid phase transition temperature test sets are 10.08 %, 9.81 % and 11.43 %, respectively. The percentage errors of the three algorithm models on the liquid phase transition temperature test set are 10.84 %, 9.87 % and 11.49 %, respectively, which directly indicates that the screened key feature combinations are generally applicable to other algorithms.

4.2.3. Universal applicability of screened key feature combinations to other data

Performance on non-precious metal solder data set. We validate the general applicability of the adopted combination of key features to other alloy brazing materials by means of a collected dataset of non-precious metal brazing materials. According to the type and percentage of elements in each alloy, the key solid phase transition temperature features (35, 38, 81, 82, 89 and 61) and key liquid phase transition temperature features (35, 81, 95, 39 and 68) of each alloy in the non-precious metal brazing dataset were calculated. Then, the solid phase transition temperature prediction model with 35, 38, 81, 82, 89 and 61 as input and the liquid phase transition temperature prediction model with 35, 81, 95, 39 and 68 as input were established by SVR. The performance of the model is evaluated by testing the modeling prediction error of the model on the data set. As shown in Fig. 10, the percentage errors of the solid phase transition temperature SVR prediction model and the solid phase transition temperature SVR prediction model based on the screened key feature combinations on the validation data set are 6.71 % and 3.19 %, respectively, which indicates that the key feature combinations used have good applicability to other filler alloy data.

The performance on Cu-Ag-Zn-Mn-Ni-Si-B-P multi-component alloy system. In addition, the accuracy of the model in predicting the solid-liquid phase transition temperature of the multi-component alloy system is verified again by selecting the Cu-Ag-Zn-Mn-Ni-Si-B-P alloy system, and the results are shown in Table 5. The predicted average absolute percentage errors of solid-liquid phase transition temperature are 7.18 % and 8.03 % respectively, indicating that the established machine learning model can achieve better prediction of Cu-Ag-Zn-Mn-Ni-Si-B-P multi-component alloy system. It is worth noting that the limitations of the basic thermodynamic database of commercial software lead to the failure to fully cover the above eight elements in the prediction of Cu-Ag-Zn-Mn-Ni-Si-B-P multi-component alloy system, so it is impossible to predict Cu-Ag-Zn-Mn-Ni-Si-B-P alloy system. This also indirectly confirms that the proposed machine learning strategy can make up for the defect that commercial software can not predict the solid-liquid phase transition temperature of some multi-component alloy elements.

4.3. Comparison with other feature selection techniques

4.3.1. Comparison of the proposed feature selection framework with other feature selection techniques

In order to verify the feasibility of the proposed feature selection framework, this study focuses on comparing two commonly used feature selection techniques, namely filter method (linear correlation screening) and wrapper method (GA intelligent algorithm screening), with the proposed feature selection framework. (1) filter method: based on the feature screening results of filter method (linear correlation screening), SVR is used for modeling and prediction. In "3.1.1.1 results of linear correlation screening", after linear correlation screening, the remaining alloy features of both solid temperature and liquid temperature model are 55. With the screened features modeled directly by the SVR algorithm, the predictions of the solid-liquid phase transition temperature model are shown in Fig. 11(a)(b), respectively. The mean absolute percentage error of the test set on the model were MAPE = 15.54% and MAPE = 15.68%, respectively, and the model had poor prediction effect on solid-liquid phase transition temperature. Compared with the feature selection framework proposed in this study, the SVR model established by filter method (linear correlation screening) has more input features, and the performance of the training set is very good, but the amount of information input is larger, resulting in poor generalization ability of the model, so the error of the test set is larger. (2) wrapper method: based on the feature screening results of wrapper method (GA intelligent algorithm screening), SVR is used for modeling and prediction. In "3.1.1.2 feature screening results of strategy I", 37 features were selected from 55 solid phase transition temperature candidates by genetic algorithm screening technique with unrestricted number of features, and 16 features were selected from 55 liquid phase transition temperature candidates. With the screened features modeled directly by the SVR algorithm, the predictions of the solid-liquid phase transition temperature model are shown in Fig. 11(c)(d), respectively, and the mean absolute percentage error of the test set on the model were MAPE=5.44 % and MAPE=7.83 %, respectively, and the model had a good prediction effect on solid-liquid phase transition temperature. Compared with the feature selection framework proposed in this study, a large number of features are screened by the wrapper method (GA intelligent algorithm), which is not conducive to the interpretability of the model, and it is difficult to find the key features that affect the solid-liquid temperature. In addition, Liu et al. [72,73] proposed a feature selection method with embedded material domain knowledge, which is used to select high quality features. Compared to a variety of other feature selection methods commonly used in materials informatics, the feature selection method with embedded material domain knowledge selects a subset of features with a more appropriate number of highly correlated features, and improves the prediction accuracy. This study attempts to explain the main reasons affecting the solid phase transition temperature and liquid phase transition temperature of alloys from the basic physical-chemical parameters, but the main physical and chemical parameters affecting the solid phase transition temperature and liquid phase transition temperature of alloys have not been reported so far, and there is no domain knowledge to refer to. In this paper, the method of increasing domain knowledge cannot be utilized for feature screening and modeling for the time being.

In summary, compared with many other feature selection methods commonly used in materials informatics, the proposed feature selection framework can simultaneously meet the requirements of interpretability, low computational complexity, strong generality of features and good prediction effect of the model in the screening of key features of solid-liquid phase transition temperature, and requires less domain knowledge.

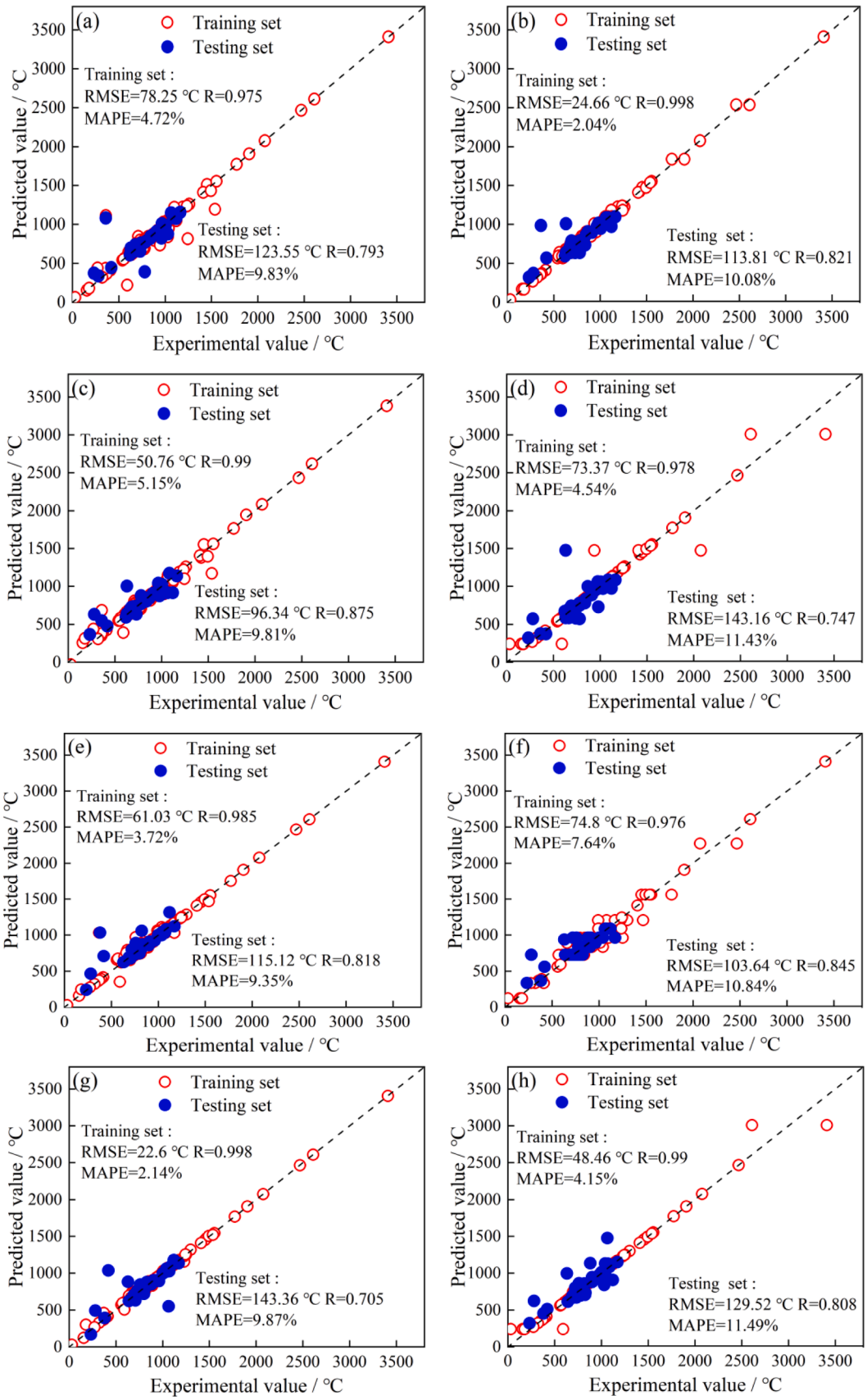

Fig. 9. The impact of different machine learning algorithms on prediction results based on screened key alloy features, solid phase transition temperature: (a) support vector regression (b) Gaussian regression (c) tree regression (d) ensemble tree regression; liquid phase transition temperature: (e) support vector regression (f) Gaussian regression (g) tree regression (h) ensemble tree regression.

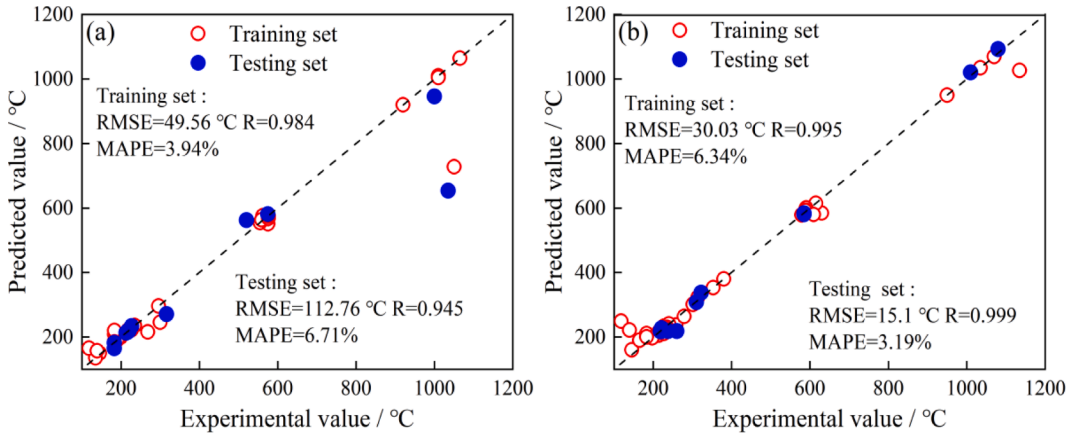

Fig. 10. Performance of the established machine learning strategy on non-precious metal brazing alloy data (a) prediction results of solid phase transition temperature; (b) prediction results of liquid phase transition temperature.

Table 5
Prediction of solid-liquid transition phase temperature in CuAgZnMnNiSiBP multi-component alloy system.

Alloy composition / Wt%								Measured values / °C		Predicted value / °C		MAPE /%	
Cu	Ag	Zn	Mn	Ni	Si	B	P	S	L	S	L	S	L
36	29.8	1	21	11.5	0.3	0.2	0.2	779.9	851.8	842.6	917.3	7.44	7.14
38	27.8	1	21	11.5	0.3	0.2	0.2	785.3	864.1	837.7	934.1	6.26	7.49
40	25.8	1	21	11.5	0.3	0.2	0.2	778.3	877.3	833.3	949.3	6.60	7.58
42.8	23	1	21	11.5	0.3	0.2	0.2	771.8	884.0	828	967	6.79	8.58
44	21.8	1	21	11.5	0.3	0.2	0.2	763.8	890.6	826	973.2	7.53	8.49
46	19.8	1	21	11.5	0.3	0.2	0.2	758.4	902.3	823.2	981.4	7.87	8.06
48	17.8	1	21	11.5	0.3	0.2	0.2	746.7	911.0	821	986.8	9.05	7.68
43.3	23	0.5	21	11.5	0.3	0.2	0.2	775.9	892.9	830.4	968.6	6.56	7.82
41.8	23	2	21	11.5	0.3	0.2	0.2	770.1	880.0	823.3	962.7	6.46	8.59
40.8	23	3	21	11.5	0.3	0.2	0.2	759.8	871.7	818.8	957.1	7.21	8.93
Average of MAPE												7.18	8.03

Note: S stands for solid phase transition temperature and L for liquid phase transition temperature.

4.3.2. Comparison of the proposed feature weight ranking method with other feature importance ranking methods

As shown in Fig. 5(a) (b) above, the first five key solid phase transition temperature features (35, 38, 81, 74, 85) and the first five key liquid phase transition temperature features (35, 38, 81, 85, 95) were screened by the proposed feature weight ranking method. Combined with the SVR algorithm for modeling, the effectiveness of the model in predicting the solid/liquid phase transition temperatures is shown in Fig. 6(c,f), respectively. The mean absolute percentage error of the test set on the solid / liquid phase transition temperature model are MAPE=9.97 %, and MAPE=10.62 %, respectively.

The feature importance of solid-liquid phase transition temperature can be obtained by the proposed method, but the decision tree and its variants can also obtain the feature importance and predict the solid-liquid phase transition temperature. In order to compare with the proposed feature weight ranking method, the decision tree method (Among the model parameters of the decision tree, the number of trees and the number of leaf nodes are the two key parameters, and in order to achieve the best prediction accuracy of the final properties model, the best model parameters were found by constant experimentation in the range of the number of trees from 100 to 400, as well as in the range of the number of leaf nodes from 5 to 500. Ultimately, the number of trees used in the solid phase transition temperature is 150 and the number of optimal leaf nodes is 5. The number of trees used in the liquid solid phase transition temperature is 100 and the number of optimal leaf nodes is 5.) is used to rank the feature importance of solid phase transition temperature and liquid phase transition temperature, and the ranking results are shown in Fig. 12(a,b). The first four key features affecting the solid temperature are numbered 35, 39, 13 and 38, corresponding to the average values of melting enthalpy, Young's modulus, binding energy and bulk modulus, respectively. The first four key features affecting liquid phase transition temperature are numbered 35, 13, 39 and 38, corresponding to the average values of melting enthalpy, binding energy, Young's modulus and bulk modulus, respectively. Among the key features screened by the decision tree, 35 and 38 also belong to the key features selected in the feature weight ranking method, which verifies the effectiveness of the proposed feature weight ranking method again. Four key solid phase transition temperature features (35, 39, 13 and 38) and four key solid phase transition temperature features (35, 13, 39 and 38) were modeled by SVR algorithm. The prediction effect of the model for solid / liquid phase transition temperature is shown in Fig. 12(c,d), respectively. The mean absolute percentage errors of the test set on the solid / liquid temperature model are MAPE=10.65 %, and MAPE=11.43 %, respectively. The solid / liquid phase transition temperature model also has high prediction accuracy, but the prediction accuracy of the model based on decision tree is lower than that of the model based on feature weight ranking method. Decision tree-based feature importance ranking relies entirely on feature relationships between data for classification. The data selected in this work are small data, the generalization ability of the selected feature modeling is slightly poor, and the prediction is limited. The feature weight ranking method proposed in this study adopts the principle of structural risk minimization, which is more suitable for small data, avoids chance, and is more conducive to obtaining important features with high generalization ability.

In summary, compared to the decision tree method, the feature weight ranking method proposed in this study is more likely to obtain important features with higher generalization ability in a usage

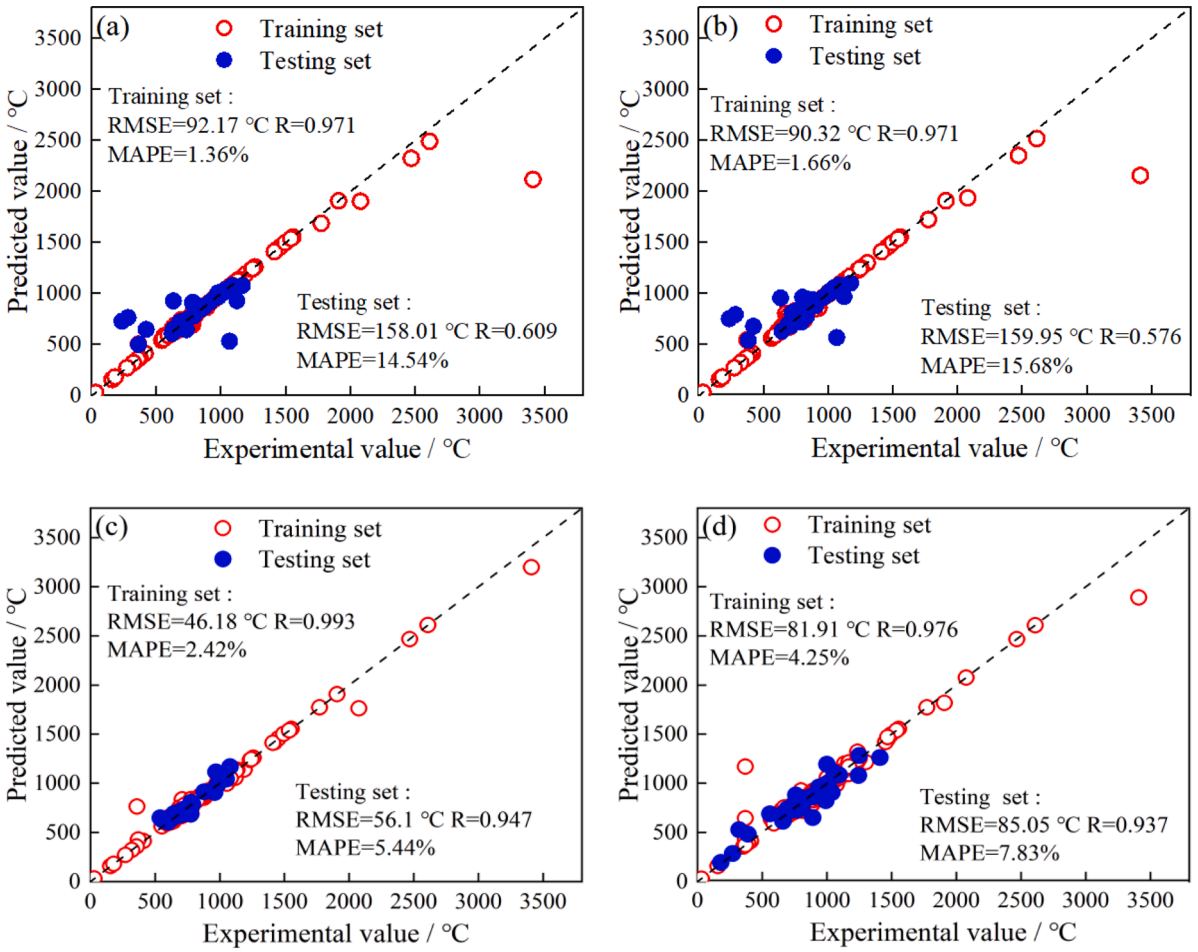

Fig. 11. Prediction effect of SVR modeling based on the results of different feature selection methods. (a) (b) the prediction effect of SVR modeling based on the feature results selected by filter method (linear correlation screening); (c) (d) the prediction effect of SVR modeling based on the feature results selected by wrapper method (GA intelligent algorithm screening). (a) (c) solid phase transition temperature; (b) (d) liquid phase transition temperature.

environment with less data.

4.4. *The expression of constructing formulas by symbolic regression*

In order to verify the accuracy of the final screened solid phase transition temperature Eq. (14) and liquid phase transition temperature Eq. (15), 47 sets of solid-liquid key feature sets were used as verification samples, see **supplementary material S9** for detailed data. Compared with the measured solid-liquid phase transition temperature, the accuracy of the symbolic regression expression of solid-liquid phase transition temperature is analyzed (see Fig. 13). According to Fig. 13(a,b), the solid-liquid phase transition temperature can be well calculated by two sets of symbolic regression formulas of solid-liquid phase transition temperature. During the verification period, the RMSE, MAE and MAPE of the two sets of formulas for the estimated and measured values of solid-liquid phase transition temperature were 0.80, 70.59 °C, 48.62 °C, 6.24 % and 0.67 84.04 °C, 50.88 °C, 6.05 %, respectively. In general, the established mathematical expression of "solid / liquid phase transition temperature-key feature combination" can meet the accuracy requirements of solid-liquid phase transition temperature prediction of precious metal alloys. the excellent performance of the mathematical expression of "solid / liquid phase transition temperature-key feature combination" confirms that the features screened by the proposed feature screening strategy are universal.

In addition, Fig. 14 counts the number of occurrences of input variables in the symbolic regression formula, and Figs. 14(a) (b) show the number of times that different input variables of solid phase transition temperature appear in each formula and the cumulative number of times they appear in all 26 equations. Obviously, the frequency of H_1 (feature 35) and H_2 (feature 38) is the highest, indicating that 35 and 38 have the greatest contribution in the process of constructing symbolic regression formula, which is consistent with the previous results of feature importance ranking. It is once confirmed that the average melting enthalpy (feature 35) and the average bulk modulus (feature 38) are the two key features that affect the solid phase transition temperature. Fig . 14(c,d) shows that C_1 (feature 35) and C_4 (feature 39) have the highest frequency, indicating that 35 and 39 have the greatest contribution in the process of constructing liquid phase symbolic regression formula. This is not consistent with the previous results of feature importance ranking, because the liquid phase feature combination screened by strategy 3 does not contain feature 38, so it does not reflect the contribution of feature 38 in the process of constructing liquid phase symbolic regression formula. However, it is seen from Section 4.1 "Physical significance analysis of key alloy features for solid-liquid phase transition temperature" that features 38 and 39 are the average bulk modulus and Young's modulus, respectively, and the two features are included in the category of elasticity modulus. The effects of the two features on liquid phase transition temperature are basically the same, so feature 39 plays the role of feature 38 in the process of constructing liquid phase symbolic regression formula. The reliability of the feature importance ranking method is verified again by the above statistical results about the occurrence times of variables in the symbolic regression formula.

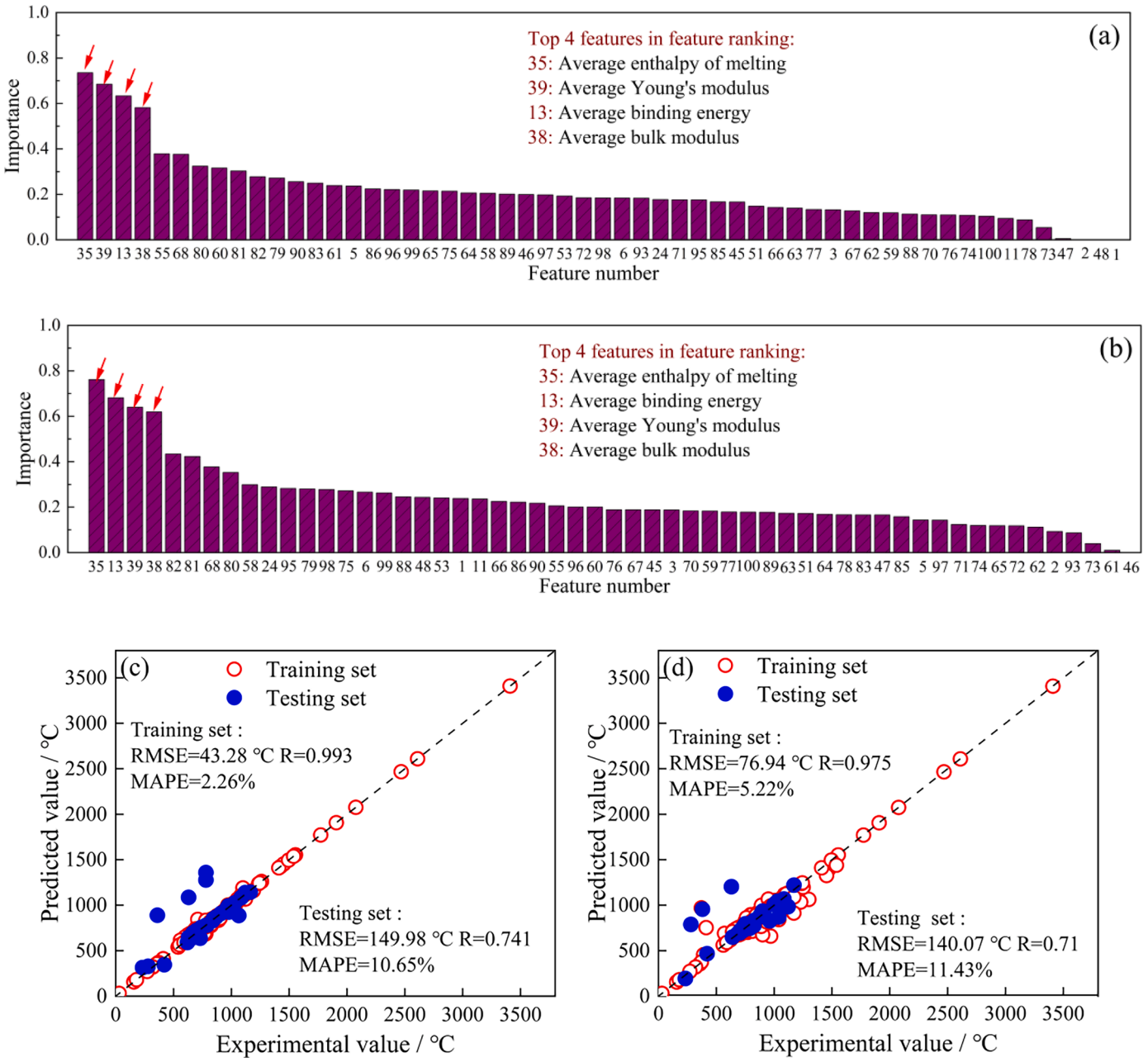

Fig. 12. Results of feature importance ranking and its modeling prediction using decision tree approach method. (a) ranking results of the importance of solid phase transition temperature features; (b) ranking results of the importance of liquid phase transition temperature features; (c) prediction of solid phase transition temperature after modeling the best features screened based on the ranking of the decision tree method; (d) prediction of liquid phase transition temperature after modeling the best features screened based on the ranking of the decision tree method.

5. Conclusion

Based on the machine learning strategy and symbolic regression algorithm of feature screening framework, the solid-liquid phase transition temperature prediction of multi-component precious metal alloy solder and the construction of the mathematical expression of "solid/liquid phase transition temperature-key feature combination" are realized. The main conclusions are as follows: (1) The first proposed feature selection framework of "correlation screening→ genetic algorithm screening→ feature weight ranking →exhaustive screening" is adopted to identify the key feature combinations affecting the solid phase transition temperature of the alloy (the average melting enthalpy, the average bulk modulus, atomic radius 2 (coordination number 12) variance, atomic volume variance, Young's modulus variance and electron affinity energy variance) and liquid phase transition temperature (melting enthalpy mean, atomic radius 2 (coordination number 12) variance, mass decay coefficient CrKalpha variance, Young's modulus mean and melting point variance). The selected key feature combinations of solid-liquid phase transition temperature have good universal applicability in other algorithms and non precious metal alloy solder data;

(2) The "key feature combination-solid phase transition temperature" prediction model with error less than 9.83 % and the "key feature combination-liquid phase transition temperature" prediction model with error less than 9.35 % are respectively established through feature construction + key feature combination screening + support vector regression algorithm modeling machine learning strategy.

(3) The mathematical expressions of "solid phase transition

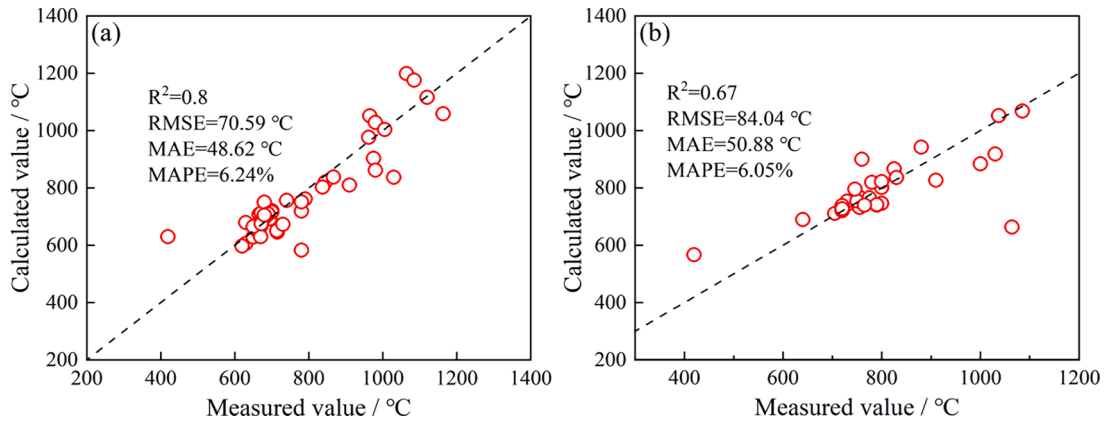

Fig. 13. The calculated and measured values of the regression formula for solid-liquid phase transition temperature symbols, (a) predicted results of Eq. (14) for the solid phase transition temperature and (b) predicted results of Eq. (15) for the liquid phase transition temperature.

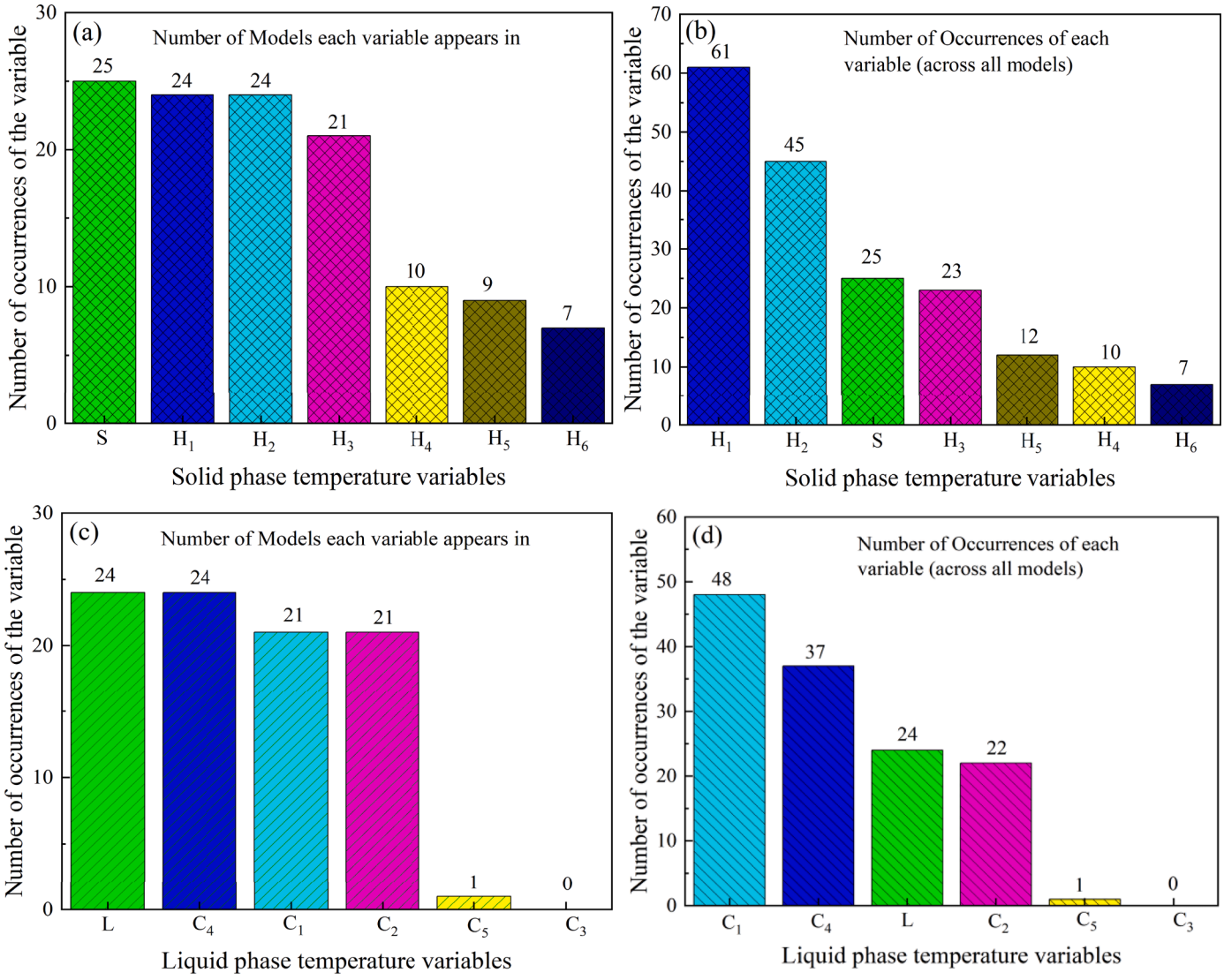

Fig. 14. Number of occurrences of input variables in the symbolic regression equations, (a) (c) depicts the number of occurrences of different input variables for solid/liquid phase transition temperature in each equation, where (a) is for the solid phase and (c) is for the liquid phase; (b) (d) depicts the cumulative number of occurrences of different input variables for solid/liquid phase transition temperature in all 26 equations, where (b) is for the solid phase and (d) is for the liquid phase.

temperature-key feature combination" with error less than 6.50 % and "liquid phase transition temperature-key feature combination" with error less than 6.71 % were established through symbolic regression algorithm. Based on the established mathematical expression, the influence of feature parameters of key independent variables on solid-liquid phase transition temperature prediction is better understood.

The machine learning strategy based on feature screening framework and symbolic regression algorithm provide a new and reliable method for predicting the solid-liquid phase transition temperature of multi-component complex alloys, and also provide a new idea for predicting the properties of other materials. However, its solid-liquid phase transition temperature prediction of eutectic alloy and metal element is still not satisfactory. The follow-up work may further improve the prediction accuracy of the solid-liquid phase transition temperature prediction model in eutectic alloys and metal elements by adding eutectic and elemental data for iterative optimization or adding alloy phase identification steps. In addition, compared with the traditional feature screening technology, the proposed feature screening method can meet the requirements of interpretability, low computational complexity, strong versatility of features and good model prediction, and is expected to become a new feature selection technology. Next, it is suggested that the proposed feature selection framework be built into a user-friendly software, so that the feature selection method can be easily extended to other application scenarios where a large number of machine learning candidate feature sets are selected for key feature combinations in other material informatics fields.

CRediT authorship contribution statement

Jiheng Fang: Writing – original draft, Conceptualization, Software, Methodology. **Shangrong Yang:** Data curation, Formal analysis. **Ming Xie:** Methodology, Writing – review & editing, Funding acquisition, Resources. **Jieqiong Hu:** Data curation, Formal analysis. **Hongsheng Sun:** Validation, Resources. **Guohua Liu:** Visualization, Resources. **Shangqiang Zhao:** Software. **Yongtai Chen:** Software. **Youcai Yang:** Investigation. **Dekui Ning:** Investigation. **Xingqun He:** Methodology, Writing – review & editing, Funding acquisition, Resources. **Qinglin Jin:** Methodology, Writing – review & editing, Funding acquisition, Resources.

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.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the National Natural Science Foundation of China [Project No. 52101040, 51961016], Science and technology talents and platform plan of Yunnan Province [Project No. 202105AC160002], Central guidance for local scientific and technological development funds [Project No. 202207AA110034], Yunnan Province Basic Research Special Project - Key Projects [Project No. 202201AS070339], Research Project of Yunnan Precious Metals Lab Co.,LTD [Project Nos. YPML-2022050226, YPML-2023050230, YPML-2022050228 and YPML-2022050229].

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.apmt.2023.102007.

References

[1] M. Way, J. Willingham, R. Goodall, Brazing filler metals[J], Int. Mater. Rev. 65 (5) (2020) 257–285.
[2] X. Yin, Q. Ma, B. Cui, et al., Current review on the research status of cemented carbide brazing: filler materials and mechanical properties[J], Met. Mater. Int. 27 (4) (2021) 571–583.
[3] B. Fan, J. Xu, H. Lei, et al., Microstructure and mechanical properties of Al2O3/Cu joints brazed with Ag-Cu-Ti+Zn composite fillers[J], Ceram. Int. 48 (13) (2022) 18551–18557.
[4] O.A. Idowu, N.L. Richards, M.C Chaturvedi, Effect of bonding temperature on isothermal solidification rate during transient liquid phase bonding of Inconel 738LC superalloy[J], Mater Sci Eng A Struct Mater 397 (1–2) (2005) 98–112.
[5] R.A. Coutu Jr, P.E. Kladitis, K.D. Leedy, et al., Selecting metal alloy electric contact materials for MEMS switches[J], J. Micromech. Microeng. 14 (8) (2004) 1157.
[6] M. Srikanth, A.R. Annamalai, A. Muthuchamy, et al., A review of the latest developments in the field of refractory high-entropy alloys[J], cryst. 11 (6) (2021) 612.
[7] L.R. Narayan, R Hebert, Rapid solidification of hypoeutectic aluminum copper alloys using fast-scanning calorimetry[J], J. Alloys Compd. 925 (2022), 166829.
[8] A. Dehghan-Manshadi, M. Bermingham, M.S. Dargusch, et al., Metal injection moulding of titanium and titanium alloys: challenges and recent development[J], Powder Technol. 319 (2017) 289–301.
[9] M. Sadeghi, B Niroumand, Design and characterization of a novel MgAlZnCuMn low melting point light weight high entropy alloy (LMLW-HEA)[J], Intermetallics 151 (2022), 107658.
[10] G.W. Park, S. Shin, J.Y. Kim, et al., Analysis of solidification microstructure and cracking mechanism of a matrix high-speed steel deposited using directed-energy deposition[J], J. Alloys Compd. 907 (2022), 164523.
[11] P. Rajendra, A. Girisha, T.G Naidu, Advancement of machine learning in materials science[J], Mater. Today: Proc. 62 (8) (2022) 5503–5507.
[12] D. Packwood, L.T.H. Nguyen, P. Cesana, et al., Machine learning in materials chemistry: an invitation[J], Mach Learn Appl 8 (2022), 100265.
[13] L.E. Vivanco-Benavides, C.L. Martínez-González, C. Mercado-Zúñiga, et al., Machine learning and materials informatics approaches in the analysis of physical properties of carbon nanotubes: a review[J], Comput. Mater. Sci. 201 (2022), 110939.
[14] T. Lookman, P.V. Balachandran, D. Xue, et al., Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design[J], npj Comput. Mater. 5 (1) (2019) 1–17.
[15] Y. Liu, O.C. Esan, Z. Pan, et al., Machine learning for advanced energy materials[J], Energy and AI 3 (2021), 100049.
[16] E.W. Huang, W.J. Lee, S.S. Singh, et al., Machine-learning and high-throughput studies for high-entropy materials[J], Materials Science and Engineering: R: Reports 147 (2022), 100645.
[17] G.L.W. Hart, T. Mueller, C. Toher, et al., Machine learning for alloys[J], Nat. Rev. Mater. 6 (8) (2021) 730–755.
[18] Y. Liu, B. Guo, X. Zou, et al., Machine learning assisted materials design and discovery for rechargeable batteries[J], Energy Storage Materials 31 (2020) 434–450.
[19] H. Yin, Z. Sun, Z. Wang, et al., The data-intensive scientific revolution occurring where two-dimensional materials meet machine learning[J], Cell Rep. Phys. Sci. 2 (7) (2021), 100482.
[20] M.I. Jordan, T.M Mitchell, Machine learning: trends, perspectives, and prospects [J], Science 349 (6245) (2015) 255–260.
[21] X.Q. He, H.D. Fu, H.T. Zhang, et al., Machine learning assisted rapid discovery of high-performance silver alloy electrical contact materials[J], Acta Metall. Sinica 58 (6) (2022) 816–826, in Chinese.
[22] Y.X. Zhang, G.C. Xing, Z.D. Sha, et al., A two-step fused machine learning approach for the prediction of glass-forming ability of metallic glasses[J], J. Alloys Compd. 875 (2021), 160040.
[23] E. Menou, J. Rame, C. Desgranges, et al., Computational design of a single crystal nickel-based superalloy with improved specific creep endurance at high temperature[J], Comput. Mater. Sci. 170 (2019), 109194.
[24] K. Kaufmann, K.S Vecchio, Searching for high entropy alloys: a machine learning approach[J], Acta Mater. 198 (2020) 178–222.
[25] H. Hou, J. Wang, L. Ye, et al., Prediction of mechanical properties of biomedical magnesium alloys based on ensemble machine learning[J], Mater. Lett. (2023), 134605.
[26] J.K. Pedersen, T.A.A. Batchelor, A. Bagger, et al., High-entropy alloys as catalysts for the CO2 and CO reduction reactions[J], ACS Catal 10 (3) (2020) 2169–2176.
[27] N. Sánchez-Marono, A. Alonso-Betanzos, M Tombilla-Sanromán, Filter methods for feature selection–a comparative study[C], in: Proceedings of the International Conference on Intelligent Data Engineering and Automated Learning, Berlin, Heidelberg, Springer, 2007, pp. 178–187.
[28] A.G. Asuero, A. Sayago, A.G Gonzalez, The correlation coefficient: an overview[J], Crit. Rev. Anal. Chem. 36 (1) (2006) 41–59.
[29] M. Cherrington, F. Thabtah, J. Lu, et al., Feature selection: filter methods performance challenges[C], in: Proceedings of the International Conference on Computer and Information Sciences (ICCIS), IEEE, 2019, pp. 1–4.
[30] B. Gierlichs, L. Batina, P. Tuyls, et al., in: Mutual information analysis[C] Proceedings of the International Workshop on Cryptographic Hardware and Embedded Systems, Berlin, Heidelberg, Springer, 2008, pp. 426–442.

[31] N. El Aboudi, L Benhlima, Review on wrapper feature selection approaches[C], in: Proceedings of the International Conference on Engineering & MIS (ICEMIS), IEEE, 2016, pp. 1–5.
[32] M.M. Kabir, M.M. Islam, K Murase, A new wrapper feature selection approach using neural network[J], Neurocomputing 73 (16–18) (2010) 3273–3283.
[33] E.L. Lawler, D.E Wood, Branch-and-bound methods: a survey[J], Oper. Res. 14 (4) (1966) 699–719.
[34] A. Bundy, L. Wallen, Breadth-first search[M], Springer, 1984.
[35] N. Sturtevant, A Felner, A brief history and recent achievements in bidirectional search[C], in: Proceedings of the AAAI Conference on Artificial Intelligence 32, 2018.
[36] K.N Berk, Forward and backward stepping in variable selection[J], J. Stat. Comput. Simul. 10 (3–4) (1980) 177–185.
[37] A.G. Karegowda, M.A. Jayaram, A.S Manjunath, Feature subset selection problem using wrapper approach in supervised learning[J], Int. J. Comput. Appl. 1 (7) (2010) 13–17.
[38] Kumar M., Husain M., Upreti N., et al. Genetic algorithm: review and application [J]. Available at SSRN 3529843, 2010.
[39] T.N. Lal, O. Chapelle, J. Weston, et al., Embedded methods[M], Springer, 2006.
[40] Wang S., Tang J., Liu H. Embedded unsupervised feature selection[C], Proceedings of the AAAI Conference on Artificial Intelligence. 2015, 29(1).
[41] B. Tan, Y.C. Liang, Q. Chen, et al., Discovery of a new criterion for predicting glass-forming ability based on symbolic regression and artificial neural network[J], J. Appl. Phys. 132 (12) (2022), 125104.
[42] GB/T 10859-2008, Nickel base brazing filler metals[S].
[43] GB/T 6418-2008, Copper base brazing filler metals[S].
[44] GB/T 18762-2017, Specification for filler brazing materials made of precious metals and their alloy[S].
[45] GB/T 10046-2018, Silver brazing filler metals[S].
[46] Z. Qiyun, Z. Hongshou, Brazing Manual [M], China Machine Press, 2008.
[47] M.M Schwartz, Brazing[M], ASM international, 2003.
[48] M. Hasanabadi, A. Shamsipur, H.N. Sani, et al., Interfacial microstructure and mechanical properties of tungsten carbide brazed joints using Ag-Cu-Zn+ Ni/Mn filler alloy[J], Trans. Nonferrous Met. Soc. China 27 (12) (2017) 2638–2646.
[49] Z. Yang, P. He, L. Zhang, et al., Microstructural evolution and mechanical properties of the joint of TiAl alloys and C/SiC composites vacuum brazed with Ag–Cu filler metal[J], Mater. Charact. 62 (9) (2011) 825–832.
[50] V.K. Beura, V. Xavier, T. Venkateswaran, et al., Interdiffusion and microstructure evolution during brazing of austenitic martensitic stainless steel and aluminum-bronze with Ag-Cu-Zn based brazing filler material[J], J. Alloys Compd. 740 (2018) 852–862.
[51] M. Lei, Y. Li, H Zhang, Interfacial microstructure and mechanical properties of the TiC-Ni cermet/Ag-Cu-Zn/Invar joint[J], Vacuum 168 (2019), 108830.
[52] W. Zhu, H. Zhang, C. Guo, et al., Wetting and brazing characteristic of high nitrogen austenitic stainless steel and 316L austenitic stainless steel by Ag–Cu filler [J], Vacuum 166 (2019) 97–106.
[53] T. Venkateswaran, V. Xavier, D. Sivakumar, et al., Brazing of stainless steels using Cu-Ag-Mn-Zn braze filler: studies on wettability, mechanical properties, and microstructural aspects[J], Mater. Des. 121 (2017) 213–228.
[54] GB/T 3131-2020, Tin-lead solder[S].
[55] GB/T 13679-2016, Manganese base brazing filler metal[S].
[56] GB/T 20422-2018, Lead-free solders[S].
[57] GB/T 13815-2008, Aluminium base brazing filler metals[S].
[58] Z.W. Ulissi, M.T. Tang, J. Xiao, et al., Machine-learning methods enable exhaustive searches for active bimetallic facets and reveal active site motifs for CO2 reduction [J], ACS Catal 7 (10) (2017) 6600–6608.
[59] V Nasteski, An overview of the supervised machine learning methods[J], Horizons 4 (2017) 51–62, b.
[60] R Muhamedyev, Machine learning methods: an overview[J], Comput. Modell. New Technol. 19 (6) (2015) 14–29.
[61] A.J. Smola, B Schölkopf, A tutorial on support vector regression[J], Stat. Comput. 14 (2004) 199–222.
[62] M. Kovačič, B Šarler, Application of the genetic programming for increasing the soft annealing productivity in steel industry[J], Mater. Manuf. Processes 24 (3) (2009) 369–374.
[63] R. Riolo, T. Soule, B Worzel, Genetic Programming Theory and Practice vi[M], Springer Science & Business Media, 2008.
[64] J.R Koza, Genetic programming as a means for programming computers by natural selection[J], Stat. Comput. 4 (2) (1994) 87–112.
[65] P. Goel, S. Bapat, R. Vyas, et al., Genetic programming based quantitative structure–retention relationships for the prediction of Kovats retention indices[J], J. Chromatogr. A 1420 (2015) 98–109.
[66] R. Vyas, P. Goel, S.S Tambe, Genetic Programming Applications in Chemical Sciences and Engineering[M]//Handbook of Genetic Programming Applications, Springer, Cham, 2015, pp. 99–140.
[67] W.B. Langdon, R Poli, Foundations of Genetic programming[M], Springer Science & Business Media, 2013.
[68] S. Sharma, S.S Tambe, Soft-sensor development for biochemical systems using genetic programming[J], Biochem. Eng. J. 85 (2014) 89–100.
[69] R Dubčáková, Eureqa: software review, Genet Program Evolvable Mach 12 (2011) 173–178.
[70] F. Yuan, T Mueller, Identifying models of dielectric breakdown strength from high-throughput data via genetic programming[J], Stat. Comput. 7 (1) (2017) 1–12.
[71] Y.M. Yang, X.L. Cui, K.J Zhang, Relationship between the heat of fusion and the melting point of metal[J], J. Inner Mongolia Univ. Technol. (Nat. Sci. Ed.) (02) (1993) 90–94.
[72] Y. Liu, X. Zou, S. Ma, et al., Feature selection method reducing correlations among features by embedding domain knowledge[J], Acta Mater. 238 (2022), 118195.
[73] Y. Liu, J.M. Wu, M. Avdeev, et al., Multi-layer feature selection incorporating weighted score-based expert knowledge toward modeling materials with targeted properties[J], Adv. Theory Simul. 3 (2) (2020), 1900215.