Journal of Alloys and Compounds 782 (2019) 110–118

Contents lists available at ScienceDirect

Journal of Alloys and Compounds

journal homepage: http://www.elsevier.com/locate/jalcom

Study on the factors affecting solid solubility in binary alloys: An exploration by Machine Learning

Shengzhou Li [a], Huiran Zhang [a, b, d, *], Dongbo Dai [a], Guangtai Ding [a, b], Xiao Wei [a, d], Yike Guo [a, c]

[a] School of Computer Engineering and Science, Shanghai University, Shanghai, 200444, China
[b] Materials Genome Institute of Shanghai University, Shanghai, 200444, China
[c] Data Science Institute, Imperial College, London, United Kingdom
[d] Shanghai Institute for Advanced Communication and Data Science, Shanghai, 200444, China

ARTICLE INFO

Article history:
Received 27 August 2018
Received in revised form
6 December 2018
Accepted 10 December 2018
Available online 13 December 2018

Keywords:
Support vector machine
Hume-Rothery rules
Solid solubility
Influence factors

ABSTRACT

The formation of solid solution alloy systems happens to two kinds of atoms with similar radii to comply with Hume-Rothery rules as a common feature. In recent years, as a useful tool, Machine Learning (ML) has been widely used in material science research to obtain useful information, including material preparation and process and so on. In this work, we use the method of Support Vector Machine (SVM) to predict solid solubility with a small dataset in order to provide evidence for the correctness of the Hume-Rothery rules and find factors of solid solubility. The results indicate that the main factors of solid solubility include three traditional Hume-Rothery factors. The solid solubility can also be evaluated by Support Vector Regression (SVR) with Radial basis function (RBF) kernel.

1. Introduction

The Hume-Rothery rules had taken a great reputation in the field of material science for a long time, as a series of basic rules describing the conditions which determine the solid solution of metal materials. In the Hume-Rothery rules, there are five factors affecting the stability of alloy phases without explicitly including: (1) A difference between the electronegativities (χ) of the elements involved; (2) A tendency for atoms of elements near the ends of the short periods and B subgroups to complete their octets of electrons; (3) Size factor effects, that is related to the relative atomic size difference between the solute and solvent elements which should be less than 15%; (4) A tendency for definite crystal structures to occur at characteristic numbers of electrons per unit cell, which, if all atomic sites are occupied, is equivalent to saying that similar structures occur at characteristic electrons per atom ratio *e/a* or the electron concentration; (5) Orbital-type restrictions in structures with certain types of hybrid bonding [1–8]. Here, the Hume-Rothery rules are summarized from experiments, which nevertheless suffered from the difficulty of reaching equilibrium at low temperatures in reasonable timescales [9]. Then, the Hume-Rothery rules could be validated by the help of ML methods (Artificial Neural Network (ANN) ML methods) and material datasets (408 silver and copper alloy systems, an extension of the 60-alloy systems first mentioned by Hume Rothery in 1934) [10,11].

In recent years, the ML methods, as an important aspect of artificial intelligence, have been widely applied to help multiple domains finding more information from big data in many fields, including the material science and technology. Typically, researchers rely on vast experiments to find a pattern of reactant properties and composition that governs material synthesis and characteristics. To speed up the process, they use ML techniques with a reasonable sized database of successful or failed experiments. One of interesting famous examples in Raccuglia et al. [12] suggested to use failed experiments in the ML-assisted materials discovery. They utilized a full database of reactions to train an SVM model; then some recommended reactions can be added to a historical reactions database. It is worth mentioning that an SVM-derived decision tree was given to uncover the schematic representation of the feedback mechanism in the dark reactions project finally. The decision tree, which is a human-interpretable hypothesis, is understood by chemists easily. Though the data is not always

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

https://doi.org/10.1016/j.jallcom.2018.12.136
0925-8388/

enough, the ML methods can be valid with a small dataset and beneficial to increase the volume of data. Johwi Lee et al. [13] proposed a prediction model of band gap for inorganic compounds by a combination of density functional theory (DFT) calculations and ML techniques. Owing to the hardness to synthesize high-quality single crystals, there is a limited experimental data of the band gap. Hence the first-principles calculation based on DFT is necessary to compute the band gap of inorganic compounds, which spends plenty of time and calculation cost. With the help of ordinary least squares regression (OLSR), least absolute shrinkage and selection operator (LASSO) and nonlinear SVR model, the relationship between band gap and predictors (such as valence, atomic, number, melting point, electronegativity, and pseudopotential radii of each element) was obtained. And it has reduced the number of k-points to the half in DFT calculation, which is of great use for researchers.

As it has been discussed in Zhang's research, determining the relative importance of each factor involved in his experiments is not easy [10]. Actually, it's mostly due that crystal structure factor is not independent from other influence factors, like electron concentration. When it comes to Cu-M (M = Zn, Al, Ga and Ge) alloy system [1] or $Al_xCo_yCr_zCu_{0.5}Fe_vNi_w$ systems [14], electron concentration has an obvious effect on crystal structure, where the crystal structures of fcc and bcc with complex unit cell are decided by *e/a*. Therefore, the relative importance of each parameter can be still evaluated by the ML methods if not including crystal structure factor; and more factors which are contribute to the solid solubility also could be found. In this work, we focus our attention on two aspects. One is whether the three main factors Hume-Rothery put forward are sufficient to determine the solubility in alloy solid systems and the relative importance of each factor. The other one is whether there exist other factors influencing the solid solubility of alloy systems and what're they. Support Vector Classification (SVC) and SVR, the two ML methods of SVM, are used to investigate these two questions. The results show that some main factors with solid solubility are found including three classic Hume-Rothery factors. The solid solubility of a small dataset can also be evaluated by SVR with RBF kernel. These will be very meaningful to predict the factors of effect on solid solubility by ML based on SVM and data mining.

2. Support Vector Machine

SVMs are learning machines implementing the structural risk minimization inductive principle to obtain good generalization on a limited number of learning patterns [15]. There are two main categories for SVM: SVC and SVR. The SVC is usually used to solve the classification problems, such as to distinguish alloy systems into three groups, *insoluble, partially soluble* and *absolutely soluble*. The SVR method was proposed by Vapnik et al. [16] in 1997. It is suitable for getting the exact value, which attempts to minimize the generalized error bound so as to achieve generalized performance rather than the observed training error. And the model produced by SVR only depends on a subset of the training data, because the cost function for building the model ignores any training data that is close to the model prediction. Hence, SVR can be used to predict solid solubility with a small metal alloy dataset and evaluate the relative importance of all factors by dividing them into several control group subsets. There are three common kernel functions of SVR algorithms as Table 1 shows, the effect of which is closely dependent on the dataset.

3. Data collection

In our work, the 408 silver and copper alloy systems are regarded as the original dataset, which are collected by Zhang et al. [10]. The physical parameters include solid solubility, the radii, valence, electronegativity and lattice parameters of solvent and solute. As stated above, the lattice parameters of solvent and solute will be ignored.

Table 1
Common kernel functions of SVR algorithms.

Kernel Type	Kernel Function
Linear	$K(x_i, x_j) = x_i^T x_j$
Polynomial	$K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0$
Radial basis function (RBF)	$K(x_i, x_j) = exp(-\gamma(\parallel x_i - x_j \parallel^2)), \gamma > 0$

According to the Hume-Rothery rules, the size factor, the electronegativity difference and the electron concentration effects are usually regarded as the most important factors that affect the stability of metallic phases. These factors will be transformed into other expressions in the input parameters of SVM. The size factor has an effect on the atomic arrangement in the alloy solid. When a solute differs in its atomic size by more than about 14–15% from the host atomic size, it is likely to have a low solubility in the host. So, the difference of the atomic size between the solvent and the solute will be one of the input parameters. Considering the electronic interactions among the constituent elements in FeNiP substitutional-solid-solution and $MoS_{2(1-x)}P_x$ solid solution, electrochemical effects play a key role in the description of the electronic interactions pictured in term of covalent bonding [17,18]. However, for binary alloys, a difference in the Pauling electronegativity between the two constituent elements determined the degree of charge transfer between neighboring unlike atoms or the degree of iconicity but not that of covalency. It's more likely to favor compound formation when a solute has a large difference in electronegativity relative to the host. It is also why the difference of the electronegativity of the two constituent elements should be one input parameter.

The situation for Hume-Rothery electron concentration rule is more complex. It has been established in the 1920–1930s in noble metals alloyed with polyvalent elements located to their right in the Periodic Table by Hume-Rothery and co-workers [19,20]. After the great success by Mott and Jones [21], the Hume-Rothery electron concentration rule can be obeyed in alloys or compounds, in which the electronic structure can be described from the view of the nearly free electron (NFE) model. Indeed, an extension to transition metal (TM) bearing compounds was still an unresolved question because the free electron model definitely fails. Until 1990s, Tsai et al. [22–24] made a great surprise and confusion about a series of thermally stable quasicrystals in Al-Cu-TM (TM = Fe, Ru, and Os) and Al-Pd-TM (TM = Mn and Re) alloy systems. A different set of *e/a* values for 3d-TM elements is available from this. A breakthrough has been brought about in the past decade by Mizutani and co-workers, who developed a method called the first-principles full-potential linearized augmented plane wave (FLAPW) formalism to self-consistently determine *e/a* value of 54 elements in the Periodic Table including 3d-, 4d- and 5d-TM elements. In addition, the interference condition is the very key word in the physics behind the Hume-Rothery electron concentration rule for compounds and alloys, no matter what constituent elements are involved. The three electronic parameters $(2k_F)^2$, *e/a* and $|G|_c^2$ were determined for each complex metallic alloy (CMA) and were employed to test whether validity of the interference condition is satisfied in Mizutani's researches, which led to the conclusion that the e/a dependent or Hume-Rothery stabilization mechanism holds even in systems where the departure from free-electron behavior is substantial because of strong orbital

hybridization in Al-TM-based CMAs including MI-type approximants [25–29].

As will be discussed in later works, in order to consider the Hume-Rothery rules as rigorously as possible, we should adopt complete uncontroversial parameters in numerical values. As mentioned before, there are two different parameters that may be used to define the electrons concentration: one is *e/a*, and the other VEC. With the help of FLAPW formalism, most the *e/a* value of 408 alloy systems could be found except some remaining elements (Si, Hg and Tl) and lanthanide and actinide elements [26]. Aside from that, it's necessary to know the specific complete constitute of alloy systems if calculating VEC, though free electron number of all the elements could be gained from Periodic Table. Based on the above points, any of these two electron concentrations couldn't be considered as a reliable input parameter for the later works. Here it should be kept in mind that a higher-valent metal is more soluble in a lower-valent metal than vice versa, proposed in the early Hume-Rothery rules [30]. Moreover, the valence of each element in 408 alloy systems seems to be more reliable due to its literature sources. Within these constraints, a difference among constitute elements in binary alloy systems could be regarded as an input parameter.

Here are the other expressions of the input parameters:

- For the size factor. The difference of the atomic diameters between solvent and solute atoms divided by the diameter of the solvent atoms is used.
- For the valence factor. These are integers, and the max value of them divided by the min value of them are used. The data is referenced from the original dataset.
- For the electronegativity factor. The difference of the electronegativity between the solvent and solute atoms is used.

4. Find more factors

Along with the three classic Hume-Rothery factors, we try to find more factors to build a more accurate representation of the solid solubility. Actually, Hume-Rothery, Mabbott, and Channel-Evans argued that atomic diameter should be defined as the closest distances of approach of atoms in the crystals of the elements. Atomic diameters are more closely correlated with solid solubility than with the Goldschmidt atomic diameters and with the Pauling single bond metallic radii according to Refs. [2,31,32]. In some studies, it is considered that maybe atomic diameters are replaced by atomic volumes [33,34]. It's proved by Hume-Rothery with the sequences Cu→Zn→Ga, Ag→Cd→In and Au→Hg→Tl that the atomic volumes do not correlate well with the solid solubility in alloy systems. There was no research indicated that the atomic weight is a possible factor. But if the atomic size factor is valid as Hume-Rothery rules say, the difference of the atomic weight of solvent and solute would also influence the solid solubility. Besides this, the atomic weight of solute is varied which could be an input parameter as well.

Furthermore, the change of energy in the metal system will also bring effect on the solid solubility [14,35]. The cohesive energy of a solid refers to the energy required to separate constituent atoms from each other and bring them to an assembly of neural free atoms. According to the range of the cohesive energies, it could be typically distributed into three bonding types for a solid: ionic, covalent and metallic bondings. In an ideally ionic alloy, cohesive energy can arise even without overlap of wave functions between the neighboring atoms, since ionic bonding originates from electrostatic interaction [36–40]. However, any realistic ionic alloys are certainly not in such an ideal state but form the valence band through overlap of wave functions of the outermost electrons on neighboring atoms [41]. The cohesive energy in both metallic and covalent bonding types is gained by lowering the total-energy relative to that of the assembly of free atoms [42,43]. In these cases, the change of bond energy would reflect the change of cohesive energy. Hence, a difference between bond energies of solvent and solute elements will be on behalf of the energy factor.

Above all, here are the added input parameters and their expressions:

- For the atomic weight factor. The difference between the atomic weight of solvent and solute atoms divided by the atomic weight of the solvent element is used. The data is referenced from NIST.
- For the atomic weight of solute factor. The atomic weight of solute atoms is used. The data is the same as the prior.
- For the energy factor. The difference between the bond energy of solvent and solute atoms divided by the bond energy of the solvent atoms is used. The bond energy data is referenced from Ref. [44].

5. Evaluate the relative importance

SVR method provides a relative importance evaluation based on performing an exhaustive search through all variable subsets. But it's usually not feasible because of the computational costs. Therefore, we start with the set of all variables and recursively delete one predictor in every step. It should be noticed that at this stage the SVR hyperparameters are the same and then the proper kernel function and hyperparameter are used for building every model with reduced number of predictors. So, we can only observe the influence of deleting variables in the training set instead of the changes associated with other aspects of the learning process [45].

As stated above, the whole parameters will be selected into 11 control group subsets. All the three classic Hume-Rothery rules parameters fall into a group named SVR-1. It is the universal set and its corresponding SVR model is also regarded as the best pattern in the Hume-Rothery rules. When the SVR-1 reduced one classic parameter, a control group subset is generated. The control group subsets are all listed in Table 2. Similarly, the complete input parameter control group with the added three factors is the universal set S in Fig. 1, and it could be selected into 6 control group subsets like in Table 3.

6. Main results

We use the SVC and SVR algorithm from the scikit-learn library with Python. To get more reliable results with a small dataset, the scale of the training set and test set (6:4) and the 4-fold cross validation will be used in the specific experiments.

In general, all alloy systems are divided into three groups based on the solid solubility: insoluble, partially soluble and absolutely soluble. If the solvent element or the solute element is the same and the other element belongs to the isotope, the same solid solubility

Table 2
The control group subsets and their input parameters. χ_a and χ_b are the electronegativities of the solvent and the solute individually. d is the difference between the atomic diameters of solvent and solute atoms divided by the diameter of the solvent atoms. P_{ab} is the ratio of the max valence and the others.

Control group subsets	Input parameters
SVR-1	χ_a χ_b d P_{ab}
SVR-2	χ_a χ_b d
SVR-3	χ_a χ_b P_{ab}
SVR-4	d P_{ab}

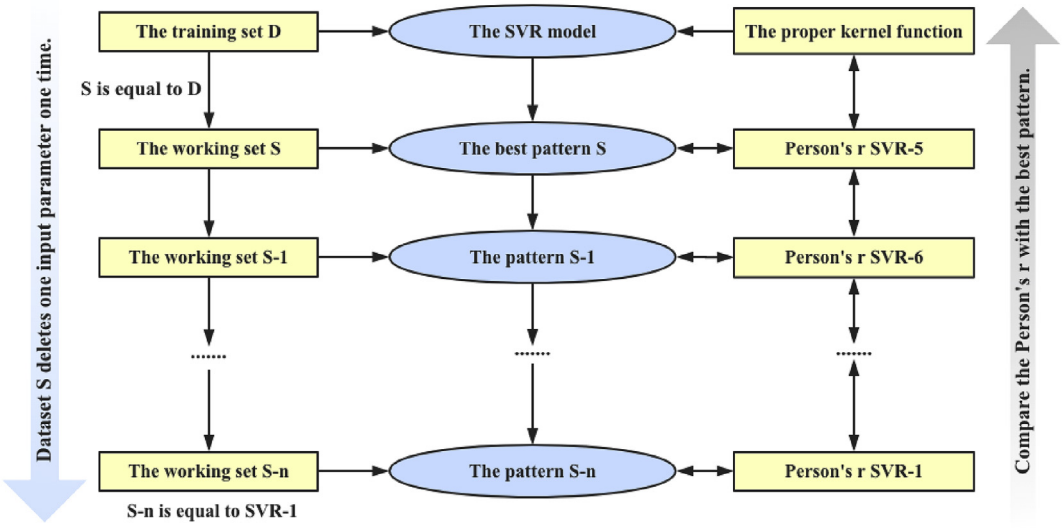

Fig. 1. The procedure of evaluating the relative importance. The training set D is the universal set, and the first working set S is equal to D. Generate different modifications of dataset S by excluding one of the input variables from dataset S at a time and build SVR models on these datasets using the same parameters until only the minimal input parameters group (SVR-4 in here) is left. When compare the correlation coefficient of the regression results with the best pattern S, the relative importance of the parameter will be obtained.

Table 3
The extended control group subsets and their input parameters. The prior four parameters are the same as Table 2. And *xE* is the difference of the atomic weight between solvent and solute elements, E_b is the atomic weight of the solute elements, and *xH* is the difference of the bond energy between solvent and solute elements.

Control group subsets	Input parameters	
	Original parameters	Extended parameters
SVR-5	χ_a χ_b d P_{ab}	xE E_b xH
SVR-6	χ_a χ_b d P_{ab}	E_b xH
SVR-7	χ_a χ_b d P_{ab}	xE xH
SVR-8	χ_a χ_b d P_{ab}	xE E_b
SVR-9	χ_a χ_b d P_{ab}	xE
SVR-10	χ_a χ_b d P_{ab}	E_b
SVR-11	χ_a χ_b d P_{ab}	xH

will be kept once or the different solid solubility will be all removed. Because of these, the dataset with 408 alloys will decrease to 231 alloys. These alloy systems will be used to testing the Hume-Rothery rules with the help of SVC method.

When two of the three classic Hume-Rothery factors and the solid solubility from original dataset (not prediction value) is set as the three-dimensional coordinates, all 231-alloy systems can be plotted as Fig. 2-a. Three distinct areas are separated due to the solid solubility with different colors, in which the green, red, blue hues signify insoluble, partially soluble and absolutely soluble. If the two axes are the solid solubility and the difference of atomic size, the distribution of all 231-alloy systems appears as Fig. 2-b. When looking into the difference of atomic size, the insoluble alloys lie in all the scope [0.0, 1.1], in contrast to partially soluble alloys and absolutely soluble alloys in the scope [0.0, 0.4]. All the dots focus on the scope of {1, 2, 3, 4, 5} in term of the valence factor, no matter whether the alloy system is soluble shown in Fig. 2-c.

Here it's a key point that only 62 partially soluble alloys of 231 silver and copper alloys (the red dots in Fig. 2) would be adopted to train and predict the solid solubility. With the representation discussed above, the final input control group subsets are presented in Tables 2 and 3. Besides these, SVR ML method would be the way to discover the quantitative relationships of solid solubility values and the influence factors.

6.1. Testing the proper kernel function

In order to compare the different kernel function's effect for alloy systems, we have to choose some measures. There are a lot of ways to evaluate the performance of SVM. The most straightforward way is using the value of the Pearson correlation coefficient (r), which is a measure of the linear correlation between predicted and experimental output. A problem occurs when r can be misleading if outliers are present and r is sensitive to the data distribution. In addition, the coefficient of determination (R^2) is the proportion of the parameter in the dependent variable that is predictable from the independent parameters, and it's robust so that all data will be preferably come out. This has the advantage of providing some balanced measurement indexes that can be used as a criterion for parameter selection in a looped optimization program when r and R^2 are combined to use. Besides this, another alternative method is to consider the mean error of the predicted value from the experimental value. There exist three ways in which this error can be calculated: (1) the mean absolute error (MAE) is a measure of difference between two continuous parameters; (2) the mean squared error (MSE) assesses the quality of a target parameter from a predictor; (3) the root mean squared error (RMSE) is a frequently used measure of the differences between predicted value and the target parameter value. The problem of (2) is that when many experimental values are zero or close to zero, the MSE value is very high, and the problem with (3) is that, the dataset range is likely to be affected by the size of dataset especially for a smaller dataset, respectively. Thus (1) provides the best criterion and, furthermore, it's an objective reflection of the direct source of errors between the target and the predictor. Hence, three measures are used to evaluate the performance: the Pearson correlation coefficient (r), the coefficient of determination (R^2) and the mean absolute error (MAE). The three computational formula shows in Table 4.

Obviously, the most important thing before carrying out experiments is to find the proper kernel function, so that the later work can be carried out in a series of suitable hyperparameters. Just as Tables 1 and 2 show, the proper kernel function with 62-alloy systems can be found by the comparison of the results from the

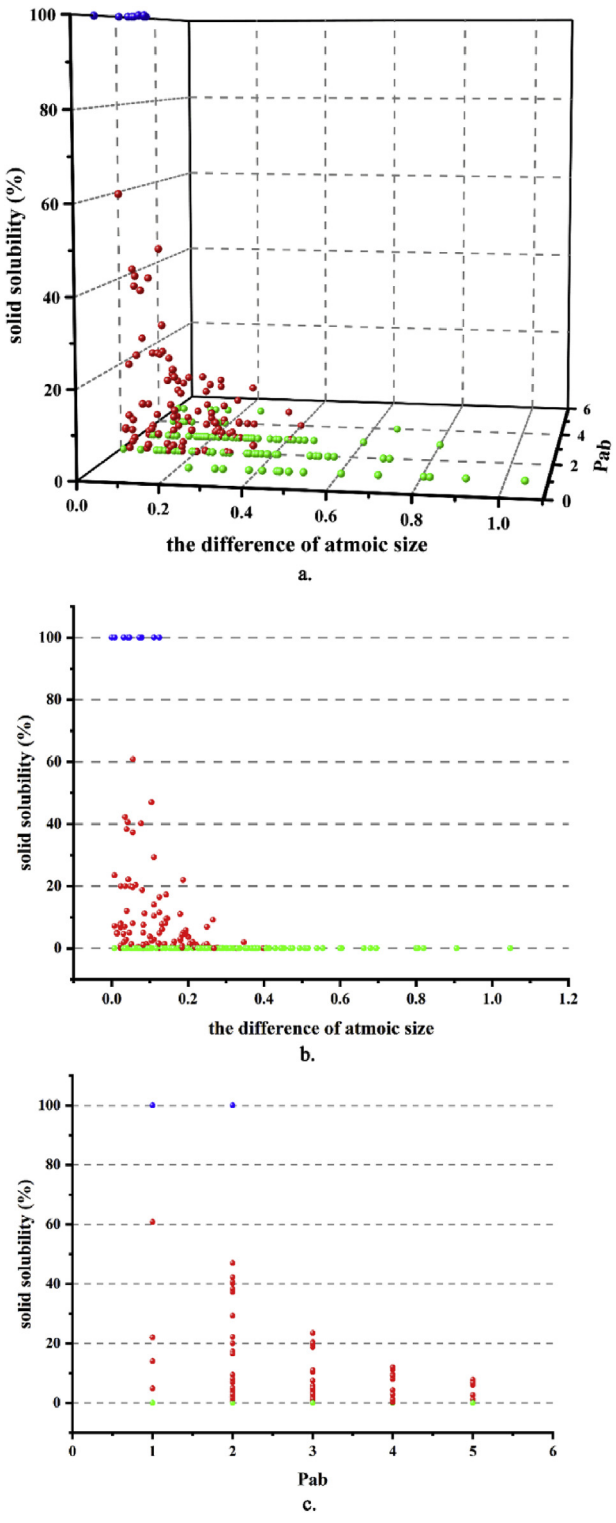

Fig. 2. The sample three-dimensional distribution and planar mapping of the 231 alloys. On the above, the blue, red and green hues signify absolutely soluble, partially soluble, and insoluble, respectively. When the difference of atomic size is X, the valence factor is Y and solid solubility is Z, the sample three-dimensional distribution is just like Fig. 2-a. After doing the planar mapping, the 231 alloys are clearly distinguished into three types like in Fig. 2-b and Fig. 2-c. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

linear kernel function to the RBF kernel function with SVR-1 input group. The results are shown in Table 5. The larger the value of

Table 4
Three measure variables.

Measure variables	How to compute
Pearson correlation coefficient	$r(X,Y) = Cov(X,Y)/\sqrt{Var[X]Var[Y]}$
Correlation coefficient	$R^2 = \sum_{i=1}^{n}\left(\widehat{y_i} - \overline{y_i}\right)^2 / \sum_{i=1}^{n}\left(y_i - \overline{y_i}\right)^2$
Mean Absolute Error	$MAE = \frac{1}{N}\sum_{i=1}^{N}\lvert f_i - y_i\rvert$

Table 5
Comparison of multiple kernel functions with SVR-1 input parameters group.

Kernel function	Person's r	R^2	MAE
RBF	0.665083	0.434101	3.050945
Ploy	0.550464	0.246252	3.681896
Linear	0.437999	0.117422	4.044618

Table 6
The results of SVC multi-classification.

Solid Solubility	True (Percent)	False (Percent)	Summary
Insoluble	113 (0.965812)	4 (0.034188)	117
Partially soluble	91 (0.892157)	11 (0.107843)	102
Absolutely soluble	10 (0.833333)	2 (0.166667)	12
Summary	214 (0.926407)	17 (0.073593)	231

Person's r and R^2, the better the kernel function. And the smaller MAE means the better kernel function. Based on these, RBF kernel function seems to work better and be a little higher accuracy. So, there is no doubt that the RBF kernel function is the proper kernel function with SVR for the next step.

6.2. Testing Hume-Rothery rules with 231-alloy systems

In cases such as these, the problem to be solved is a multi-classification problem, so the SVC method would like to be a viable option. Scikit-learn provides three ways to solve multi-classification with SVC method: SVC, NuSVC, and LinearSVC. Among these methods, SVC and NuSVC adopt the idea of “one-against-one” in the classification of multiple cases, while LinearSVC chooses the idea of “one-vs-the-rest”, so the SVC and NuSVC will lead to a nonlinear and more accurate multi-classification. When comparing with NuSVC, SVC has no need to set more hyper-parameters and it is a better choice for this problem.

Using the RBF kernel function in the SVC method, a multi-classification result is obtained in term of whether the solid alloy system is soluble. The *ovo* decision function shape is also chosen due to the multi-classification task. A most important hyper-parameter of SVC is the penalty parameter C, which is equivalent to punishing the slack variable. If the slack variable is close to zero, the penalty for misclassification will increase to the situation where the training set is fully paired. That is, the accuracy of the training set is high, and the generalization ability is relatively weak. When C is set to too small, the punishment effect will be weakened as well as more generalization ability. After several tries, a better penalty parameter C (10^3) is found. Last but not least, the gamma value of SVC is on behalf of the kernel coefficient for RBF kernel function and it renders at a great performance when the gamma value is 10. So, here are all the more proper hyperparameters for SVC method: (1) RBF kernel function; (2) *ovo* decision function shape; (3) Penalty parameter C is set to 10^3; (4) gamma value is set to 10.

From the result of the 231-alloy systems as Table 6 shows, the accuracy of the classification prediction is about 92.6% (214/231)

Table 7
SVR methods based on RBF kernel with multiple input parameters groups.

Control group	Person's r	R^2	MAE
SVR-1	0.665083	0.434101	3.050945
SVR-2	0.396181	0.097624	4.061656
SVR-3	0.623069	0.386363	3.310034
SVR-4	0.431203	0.103105	4.130320
SVR-5	0.969676	0.938405	0.294001
SVR-6	0.965078	0.929689	0.353936
SVR-7	0.828338	0.675126	2.121783
SVR-8	0.970069	0.939747	0.312838
SVR-9	0.701138	0.479264	2.635421
SVR-10	0.958863	0.919143	0.537298
SVR-11	0.802785	0.635369	2.469677

and SVC method works well. Even in the different solid solubility group, a lot of true positives are obtained according to the confusion table (Table 6), and the absolutely soluble alloy systems perform well with a small group dataset (12).

6.3. Testing 62-alloy systems with SVR

The input parameters groups in Table 2 and the chosen SVR kernel function will be used to test and validate the Hume-Rothery rules. After the dataset is reduced to 62-alloy systems, the problem becomes to predict the solid solubility, which is obviously a regression problem. Both results are listed in Table 7 in terms of the Person's r, R^2 and MAE of all the 62-alloy systems and Fig. 3. Comparing the results with the first control group, a slight difference is found: The Person's r and R^2 of the SVR-2 are both the smallest. There is no doubt that the absence of valence factor in the input parameters group has caused a great change in the prediction result. It also means that the valence factor plays a significant role in the prediction of solid solubility. Similarly, the results of other two groups SVR-3 and SVR-4 also have an obvious difference with the SVR-1 group. To summarize, all atomic size factor, valence factor and electronegativity factor are both contribute to the prediction of the solid solubility in the metal alloy systems. In other words, the three classic Hume-Rothery factors are indeed effective for judging the solid solubility of the binary alloy.

While more factors are extended into the input parameters group, we can get more accurate solid solubility, and the results show in Table 7 and Fig. 4. Comparing the SVR-1 with SVR-5, the predicting accuracy increases remarkably after extending the more three factors. Especially when the SVR-1 and SVR-5 in Fig. 4 are put together, there are more dots which congregate in tiny hot spots closing the zero in SVR-5. If SVR-1 is compared with SVR-9, SVR-10, and SVR-11 respectively, it can be inferred that each of these extended factors is contribute to predicting the solid solubility more accurately. This result also proves that the extended factors are effective for solid solubility.

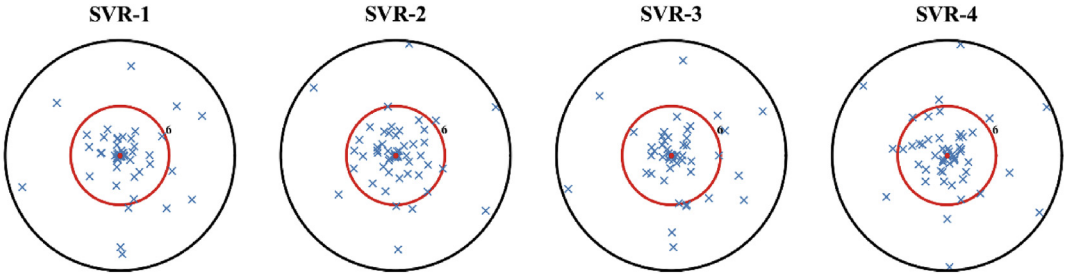

Fig. 3. The prediction error scatter diagram of the first four input groups. The dot which is more closed to the center of the circle (a red dot) indicates that the prediction error is smaller. And the degree of the dot in the circle is nothing and just distribute all dots from another. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

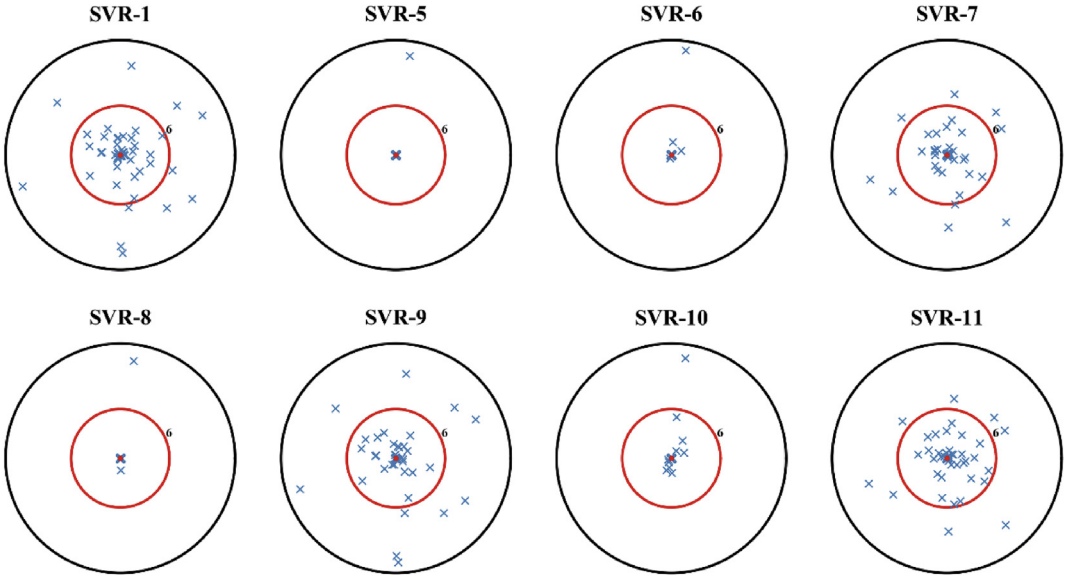

Fig. 4. The prediction of solubility using SVR based on RBF kernel with multiple input parameters subset SVR-1 and from SVR-5 to SVR-11 for the 62-alloy systems.

6.4. Testing relative importance of factors

The relative importance of each factor can be determined from Table 7 and Fig. 3. The input parameters groups from SVR-1 to SVR-4 are generated by excluding one of the input variables (atomic size factor, electronegativity factor and valence factor) from the universal set SVR-1. Hence, when we compare all the results in Fig. 3, we can calculate the dot beyond the border (including those points on the boundary) and regard 6 as the threshold value. In order of the input group, the number of the dots beyond the border in the row of Fig. 3 goes: 11, 9, 13, and 13. Due to a considerable gap between the SVR-1 and SVR-2 group, the valence factor should be treated as the most important factor relatively. In addition, the SVR-3 group and SVR-4 group in the dot value are very close. Not only there is a gap between the SVR-3 group and SVR-4 group, but also the prediction error of the SVR-4 group is higher than SVR-3 group (4.130320:3.310034). Therefore, the electronegativity factor is more relatively important than the atomic factor in predicting the solid solubility of the partially soluble alloy systems. In summary, the relative importance order of three classic Hume-Rothery factors by Hume-Rothery rules in predicting the solid solubility of the partially soluble alloy systems is: the valence factor > the electronegativity factor > the atomic size factor.

The extended factors can be also evaluated the relative importance for predicting the solid solubility in the alloy systems. In order of the input group, the number of the dots beyond the border goes in Fig. 4: 11, 1, 1, 9, 1, 10, 1 and 8. When the control group subsets are sorted descending by the Pearson's *r*, it's easy to reach a conclusion that, the relative importance of the extended factors is: the atomic weight of the solute factor > the bond energy factor > the atomic weight difference factor. On the one hand, if comparing the Pearson's *r* and R^2 of SVR-5 with SVR-6, SVR-7 and SVR-8 respectively, the difference between SVR-5 and SVR-6 are the maximal and the atomic weight of solute factor is the most important among the three extended factors certainly. On the other hand, the atomic weight difference factor seems to be less important when SVR-1 is compared with SVR-9, SVR-10, and SVR-11. And if SVR-7 and SVR-8 are put together, it states that the atomic weight of the solute factor takes on more importance than the bond energy factor.

7. Discussion

In our works, there are something worth discussing: (1) the selection of hyperparameters in SVC and SVR method, which affect the prediction performance from the algorithm itself; (2) the reliability of using the SVM method to confirm the Hume-Rothery rules; (3) the reliability of the extended factors and the relative importance of factors.

7.1. The selection of hyperparameters

The distribution of the dataset and the hyperparameters are the two most important parts in most of ML methods [46–48], while the dataset is settled in our case and the hyperparameters are the only way to implementing a better prediction performance. A better SVM method often includes a proper kernel function, a fit decision function shape and an appropriate penalty parameter *C*. It also contains a gamma value if the kernel function is RBF. By testing the three different kernel functions with SVM and gaining the result as Table 5, RBF kernel function can describe the alloy systems well. A fit decision function is mainly decided by the problem itself. The test of 231-alloy systems is a multi-classification problem, so that the fit decision function shape is “one-against-one”. Penalty parameter *C* and the gamma value all depend on the scale of the dataset, and both 231-alloy systems and 62-alloy systems are a small scale [45]. Therefore, a cubic of ten and ten are a better *C* and gamma value, respectively.

7.2. The reliability of Hume-Rothery rules with SVM

To validate the reliability of Hume-Rothery rules with SVM, the 231-alloy systems with SVC method and the 62-alloy systems with SVM method have been tested, and it comes out that SVM method can validate the Hume-Rothery rules. Through the distribution of 231-alloy systems in Fig. 2, it's easy to know that, the difference of atomic size in most of soluble alloy systems lies in [0.0, 0.2] and the threshold value of the difference of atomic size in Hume-Rothery rules (15%) is a credible guide [1,2]. Otherwise, the SVC method can distinguish different solid solubility between 231-alloy systems, which benefits from the three classic Hume-Rothery factors and a series of proper hyperparameters. Even though in the absolutely soluble alloy systems, the SVC method's true positive percentage is 10/12, and it also suggests that the SVC method can play well in a small scale.

7.3. The reliability of extended factors and the relative importance

Despite Hume-Rothery and his co-workers suggested that the three classic Hume-Rothery factors can be discovered from the 60-alloy systems mentioned by Hume-Rothery in 1934, other researches in later work [28,49–51] attempted to predict the solid solubility of the 408-alloy systems, and the result is that the three classic Hume-Rothery factors cannot be general because it just works in a part of 408-alloy systems. Hence, the performance of the prediction based on Hume-Rothery rules is limited. It's a beneficial try to extend the factors, which depend on the solid solubility. Some properties which describe the energy change or the difference of atomic weight, are the first choices on account that variety of solute atoms make a different degree of lattice distortion in variety of solvent atoms due to the variety of atomic weight [52–54] and the free electrons in whole alloy system would be limited along with the change of the energy of the alloy system [28,55,56].

According to Hume-Rothery rules, the atomic size factor should be the most important factor, followed by the electron concentration factor and the electronegativity factor. In our work, the order of three factors in Hume-Rothery rules is a judgement standard to determine if the solid alloy system is soluble, but the order is not always in accord with that in predicting the solid solubility of the soluble alloy system. However, based on the above experiment results, it can be inferred that, the electron concentration factor is the most important parameter for predicting the solid solubility of the soluble alloy systems. As a data-driven research, we need to still regard the result as the objectivity of the solid solubility dataset, and more information will be explained with domain knowledge.

8. Conclusions

This paper validated Hume-Rothery rules from a data standpoint and adopted SVR for predicting the solid solubility limit of alloy systems based on Hume-Rothery rules. Some main factors with solid solubility are detected including three classic Hume-Rothery factors. The solid solubility of a limit dataset can also be evaluated by SVR based on RBF kernel. The research results indicate that: (1) SVM is helpful for exploring Hume-Rothery rules and getting more information in material science; (2) Hume-Rothery rules work well in the 231-alloy systems and 62-alloy systems, such that the SVC and SVR can be utilized to predict solid solubility. Furthermore, when the 62-alloy systems are tested by SVR, the relative importance of the factors is given: the valence factor > the electronegativity factor > the atomic size factor > the atomic

weight of the solute factor > the bond energy factor > the atomic weight difference factor. Hence, not only the ML methods can validate the Hume-Rothery rules, but also it can dig more information inside the material data. In fact, the formation of solid solution is very complex and many factors all will affect it such as mixing enthalpy, mixing entropy and so on [57]. They all should be considered deeply in future research. As a preliminary exploration for the simple binary alloys, the present methods will be very meaningful to predicate the factors of effect on solid solubility by ML methods based on SVM and data mining.

Acknowledgements

This work is supported by the National Key Research and Development Program of China (No. 2016YFB0700502).

References

[1] U. Mizutani, Hume-rothery Rules for Structurally Complex Alloy Phases, 2012, https://doi.org/10.1557/mrs.2012.45.
[2] Z. Wang, Y. Huang, Y. Yang, J. Wang, C.T. Liu, Atomic-size effect and solid solubility of multicomponent alloys, Scr. Mater. 94 (2015) 28–31, https://doi.org/10.1016/j.scriptamat.2014.09.010.
[3] G.P. Tiwari, R.V. Ramanujan, Review the relation between the electron to atom ratio and some properties of metallic systems, J. Mater. Sci. 36 (2001) 271–283, https://doi.org/10.1023/A:1004853304704.
[4] A.R. Denton, N.W. Ashcroft, Vegard's law, Phys. Rev. A 43 (1991) 3161–3164, https://doi.org/10.1103/PhysRevA.43.3161.
[5] H. Huang, Y. Wu, J. He, H. Wang, X. Liu, K. An, W. Wu, Z. Lu, Phase-transformation ductilization of brittle high-entropy alloys via metastability engineering, Adv. Mater. 29 (2017) 1–7, https://doi.org/10.1002/adma.201701678.
[6] O.N. Senkov, D.B. Miracle, A new thermodynamic parameter to predict formation of solid solution or intermetallic phases in high entropy alloys, J. Alloys Compd. 658 (2016) 603–607, https://doi.org/10.1016/j.jallcom.2015.10.279.
[7] S.H. Park, S.H. Kim, Y.M. Kim, B.S. You, Improving mechanical properties of extruded Mg-Al alloy with a bimodal grain structure through alloying addition, J. Alloys Compd. 646 (2015) 932–936, https://doi.org/10.1016/j.jallcom.2015.06.034.
[8] J. Zhang, C. Ke, H. Wu, J. Yu, J. Wang, Y. Wang, Solubility limits, crystal structure and lattice thermal expansion of Ln_2O_3(Ln=Sm, Eu, Gd) doped CeO_2, J. Alloys Compd. 718 (2017) 85–91, https://doi.org/10.1016/j.jallcom.2017.05.073.
[9] H. Ohtani, K. Ishida, Application of the CALPHAD method to material design, Thermochim. Acta 314 (1998) 69–77, https://doi.org/10.1016/S0040-6031(97)00457-7.
[10] Y.M. Zhang, S. Yang, J.R.G. Evans, Revisiting Hume-Rothery's Rules with artificial neural networks, Acta Mater. 56 (2008) 1094–1105, https://doi.org/10.1016/j.actamat.2007.10.059.
[11] Y.M. Zhang, J.R.G. Evans, S. Yang, The prediction of solid solubility of alloys: developments and applications of Hume-Rothery's rules, J. Cryst. Phys. Chem. 1 (2010) 81–97.
[12] P. Raccuglia, K.C. Elbert, P.D.F. Adler, C. Falk, M.B. Wenny, A. Mollo, M. Zeller, S.A. Friedler, J. Schrier, A.J. Norquist, Machine-learning-assisted materials discovery using failed experiments, Nature 533 (2016) 73–76, https://doi.org/10.1038/nature17439.
[13] J. Lee, A. Seko, K. Shitara, I. Tanaka, Prediction model of band-gap for AX binary compounds by combination of density functional theory calculations and machine learning techniques, Phys. Rev. B 93 (2016), 115104, https://doi.org/10.1103/PhysRevB.93.115104.
[14] S. Guo, C. Ng, J. Lu, C.T. Liu, Effect of valence electron concentration on stability of fcc or bcc phase in high entropy alloys, J. Appl. Phys. 109 (2011), 103505, https://doi.org/10.1063/1.3587228.
[15] D. Basak, S. Pal, D.C. Patranabis, Support vector regression, Neuronal Inf. Process. Lett. Rev. 11 (2007) 203–224.
[16] H. Drucker, C.J.C. Burges, L. Kaufman, A. Smola, V. Vapnik, Support vector regression machines, Adv. Neural Inf. Process. Syst. 1 (1997) 155–161.
[17] M. Qian, S. Cui, D. Jiang, L. Zhang, P. Du, Highly efficient and stable water-oxidation electrocatalysis with a very low overpotential using FeNiP substitutional-solid-solution nanoplate arrays, Adv. Mater. 29 (2017) 1–6, https://doi.org/10.1002/adma.201704075.
[18] R. Ye, P. Del Angel-Vicente, Y. Liu, M.J. Arellano-Jimenez, Z. Peng, T. Wang, Y. Li, B.I. Yakobson, S.H. Wei, M.J. Yacaman, J.M. Tour, High-performance hydrogen evolution from $MoS_{2(1-x)}P_x$ solid solution, Adv. Mater. 28 (2016) 1427–1432, https://doi.org/10.1002/adma.201504866.
[19] W. Hume-Rothery, G.W. Mabbott, K.M.C. Evans, The freezing points, melting points, and solid solubility limits of the alloys of silver, and copper with the elements of the B sub-groups, Phil. Trans. Roy. Soc. Lond. 233 (1934) 1–97, https://doi.org/10.1098/rsta.1934.0014.
[20] W. Hume-Rothery, The Structure of Metals and Alloys, The Institute of Metals, 1937.
[21] N.F. Mott, H. Jones, The Theory of the Properties of Metals and Alloys, Courier Corporation, 1958.
[22] A.P. Tsai, A. Inoue, T. Masumoto, New stable icosahedral Al-Cu-Ru and Al-Cu-Os alloys, Jpn. J. Appl. Phys. 27 (1988) L1587, https://doi.org/10.1143/JJAP.27.L1587.
[23] A.P. Tsai, A. Inoue, Y. Yokoyama, T. Masumoto, Stable icosahedral Al–Pd–Mn and Al–Pd–Re alloys, Mater. Trans. JIM 31 (1990) 98–103, https://doi.org/10.2320/matertrans1989.31.98.
[24] Y. Yokoyama, A.P. Tsai, A. Inoue, T. Masumoto, H.S. Chen, Formation criteria and growth morphology of quasicrystals in Al–Pd–TM (TM= transition metal) alloys, Mater. Trans. JIM 32 (1991) 421–428, https://doi.org/10.2320/matertrans1989.32.421.
[25] U. Mizutani, T. Noritake, T. Ohsuna, T. Takeuchi, Hume-Rothery electron concentration rule across a whole solid solution range in a series of gamma-brasses in Cu-Zn, Cu-Cd, Cu-Al, Cu-Ga, Ni-Zn and Co-Zn alloy systems, Philos. Mag. 90 (2010) 1985–2008, https://doi.org/10.1080/14786430903246320.
[26] U. Mizutani, H. Sato, M. Inukai, Y. Nishino, E.S. Zijlstra, Electrons per atom ratio determination and hume-rothery electron concentration rule for P-based polar compounds studied by FLAPW–Fourier calculations, Inorg. Chem. 54 (2014) 930–946, https://doi.org/10.1021/ic502286q.
[27] U. Mizutani, H. Sato, Determination of electrons per atom ratio for transition metal compounds studied by FLAPW-Fourier calculations, Philos. Mag. 96 (2016) 3075–3096, https://doi.org/10.1080/14786435.2016.1224946.
[28] U. Mizutani, H. Sato, The physics of the Hume-Rothery electron concentration rule, Crystals 7 (2017) 9, https://doi.org/10.3390/cryst7010009.
[29] U. Mizutani, H. Sato, Energy gap formation mechanism through the interference phenomena of electrons in face-centered cubic elements and compounds with the emphasis on half-Heusler and Heusler compounds, Philos. Mag. 98 (2018) 1307–1336, https://doi.org/10.3390/cryst7010009.
[30] J.A. Alonso, S. Simozar, Prediction of solid solubility in alloys, Phys. Rev. B 22 (1980) 5583–5589, https://doi.org/10.1103/PhysRevB.22.5583.
[31] W. Hume-Rothery, Atomic diameters, atomic volumes and solid solubility relations in alloys, Acta Metall. 14 (1966) 17–20, https://doi.org/10.1016/0001-6160(66)90267-7.
[32] M.C. Troparevsky, J.R. Morris, M. Daene, Y. Wang, A.R. Lupini, G.M. Stocks, Beyond atomic sizes and hume-rothery rules: understanding and predicting high-entropy alloys, Jom 67 (2015) 2350–2363, https://doi.org/10.1007/s11837-015-1594-2.
[33] N.F. Mott, The cohesive forces in metals and alloys, Rep. Prog. Phys. 25 (1962) 218.
[34] T.B. Massalski, H.W. King, The lattice spacing relationships in H.C.P. ∈ and η phases in the systems Cu-Zn, Ag-Zn; Au-Zn and Ag-Cd, Acta Metall. 10 (1962) 1171–1181, https://doi.org/10.1016/0001-6160(62)90170-0.
[35] C.-J. Tong, Y.-L. Chen, J.-W. Yeh, S.-J. Lin, S.-K. Chen, T.-T. Shun, C.-H. Tsau, S.-Y. Chang, Microstructure characterization of $Al_xCoCrCuFeNi$ high-entropy alloy system with multiprincipal elements, Metall. Mater. Trans. A 36 (2005) 881–893, https://doi.org/10.1007/s11661-005-0283-0.
[36] Q.F. He, Y.F. Ye, Y. Yang, formation of random solid solution in multicomponent alloys: from hume-rothery rules to entropic stabilization, J. Phase Equilib. Diffus. 38 (2017) 416–425, https://doi.org/10.1007/s11669-017-0560-9.
[37] Y. Zhang, Y.J. Zhou, J.P. Lin, G.L. Chen, P.K. Liaw, Solid-solution phase formation rules for multi-component alloys, Adv. Eng. Mater. 10 (2008) 534–538, https://doi.org/10.1002/adem.200700240.
[38] F. Otto, Y. Yang, H. Bei, E.P. George, Relative effects of enthalpy and entropy on the phase stability of equiatomic high-entropy alloys, Acta Mater. 61 (2013) 2628–2638, https://doi.org/10.1016/j.actamat.2013.01.042.
[39] S. Guo, Q. Hu, C. Ng, C.T. Liu, More than entropy in high-entropy alloys: forming solid solutions or amorphous phase, Intermetallics 41 (2013) 96–103, https://doi.org/10.1016/j.intermet.2013.05.002.
[40] M. Stiehler, J. Rauchhaupt, U. Giegengack, P. Häussler, On modifications of the well-known Hume-Rothery rules: amorphous alloys as model systems, J. Non Cryst. Solids 353 (2007) 1886–1891, https://doi.org/10.1016/j.jnoncrysol.2007.01.052.
[41] I.D. Brown, Bond valence theory, Bond Val. 158 (2013) 11–58, https://doi.org/10.1007/430_2012_89.
[42] J. Yuhara, M. Yokoyama, T. Matsui, Two-dimensional solid solution alloy of Bi-Pb binary films on Rh(111), J. Appl. Phys. 110 (2011) 1–5, https://doi.org/10.1063/1.3650883.
[43] U. Häussermann, P. Viklund, M. Boström, R. Norrestam, S.I. Simak, Bonding and physical properties of Hume-Rothery compounds with the $PtHg_4$ structure, Phys. Rev. B 63 (2001) 1–10, https://doi.org/10.1103/PhysRevB.63.125118.
[44] Y.R. Luo, Bond Dissociation Energies, CRC Handbook of Chemistry and Physics, vol. 9, 2009, pp. 65–98. https://notendur.hi.is/~agust/rannsoknir/papers/2010-91-CRC-BDEs-Tables.pdf.
[45] M. Trzęsiok, The Importance of Predictor Variables for Individual Classes in SVM, vol. 235, Acta Universitatis Lodziensis, Folia Oeconomica, 2010, pp. 185–193.
[46] J. Wainer, G.C. Cawley, Empirical evaluation of resampling procedures for optimising SVM hyperparameters, J. Mach. Learn. Res. 18 (2017) 475–509.
[47] P. Probst, B. Bischl, A.-L. Boulesteix, Tunability: importance of hyperparameters of machine learning algorithms, 2018, pp. 1–27. http://arxiv.org/abs/1802.09596.

[48] H. Mayfield, C. Smith, M. Gallagher, M. Hockings, Use of freely available datasets and machine learning methods in predicting deforestation, Environ. Model. Softw. 87 (2017) 134–145, https://doi.org/10.1016/j.envsoft.2016.10.006.
[49] B.W. Zhang, Miedema theory for formation heat of alloy system, Shanghai Met. 15 (1993) 23–30.
[50] A. Cottrell, Concepts in the Electron Theory of Alloys, IOM Communications, London, 1998, pp. 56–92.
[51] K. Li, D. Xue, Estimation of electronegativity values of elements in different valence states, J. Phys. Chem. A 110 (2006) 11332–11337, https://doi.org/10.1021/jp062886k.
[52] J.S. Faulkner, The modern theory of alloys, Prog. Mater. Sci. 27 (1982) 1–187, https://doi.org/10.1016/0079-6425(82)90005-6.
[53] I. Basu, K.G. Pradeep, C. Mießen, L.A. Barrales-Mora, T. Al-Samman, The role of atomic scale segregation in designing highly ductile magnesium alloys, Acta Mater. 116 (2016) 77–94, https://doi.org/10.1016/j.actamat.2016.06.024.
[54] H.E. Ives, O. Stuhlmann, The result of plotting the separation of homologous pairs against atomic numbers instead of atomic weights, Phys. Rev. 5 (1915) 368–372, https://doi.org/10.1103/PhysRev.5.368.
[55] M. Calvo-Dahlborg, S.G.R. Brown, Hume-Rothery for HEA classification and self-organizing map for phases and properties prediction, J. Alloys Compd. 724 (2017) 353–364, https://doi.org/10.1016/j.jallcom.2017.07.074.
[56] I. Martin, S. Gopalakrishnan, E.A. Demler, Weak crystallization theory of metallic alloys, Phys. Rev. B 93 (2016) 1–9, https://doi.org/10.1103/PhysRevB.93.235140.
[57] X. Yang, Y. Zhang, Prediction of high-entropy stabilized solid-solution in multi-component alloys, Mater. Chem. Phys. 132 (2012) 233–238, https://doi.org/10.1016/j.matchemphys.2011.11.021.