Journal Name

ARTICLE TYPE

Cite this: DOI: 00.0000/xxxxxxxxxx

Supplemental Material: Descriptors for phase prediction of high entropy alloys using interpretable machine learning†

Shang Zhao,[a] Ruihao Yuan,*[ab] Weijie Liao,[a] Yatong Zhao,[a] Jun Wang,[a] Jinshan Li*[a] and Turab Lookman*[c]

1 The data and descriptors used in the present study

To illustrate the classification process, Figure S1 is provided, making the predictions in order from Category-1 to Category-4.

Figure S2 shows the 33 elements in the data and their frequencies of occurrence in the 541 samples. These elements spread over a large range, including refractory metals, light metals, transition metals, nonmetals and rare earth element etc. As there are 541 alloys in total and some alloy systems only comprise one or two alloys, thus we only list a part of the alloy systems, as shown in Table S1. There are 37 alloy systems in the table and different systems contain different amount of alloys. Moreover, to further confirm the diversity of our alloys, we have also evaluated the change of chemical compositions in several alloy systems (Figure S3), that is, the Al-Co-Cr-Fe-Ni, Mo-Nb-Ti-V-Zr and Cr-Cu-Fe-Mn-Ni with a relatively large amount of alloys and the constitutive elements are very different. Figure S4 shows the distribution of the 541 alloys, the majority of the alloys distribute closely and only several points deviate from them, they are AlLiMgZnSn, $Al_{80}Li_5Mg_5Zn_5Sn_5$, $Al_{80}Li_5Mg_5Zn_5Cu_5$, and AlCoCuNiTiZn, respectively.

There are three kinds of empirical descriptors, as listed in Table S2. Specifically, 26 atomic descriptors, 5 thermodynamic descriptors and 2 theoretical descriptors. For each descriptor, both the corresponding description and abbreviation are given.

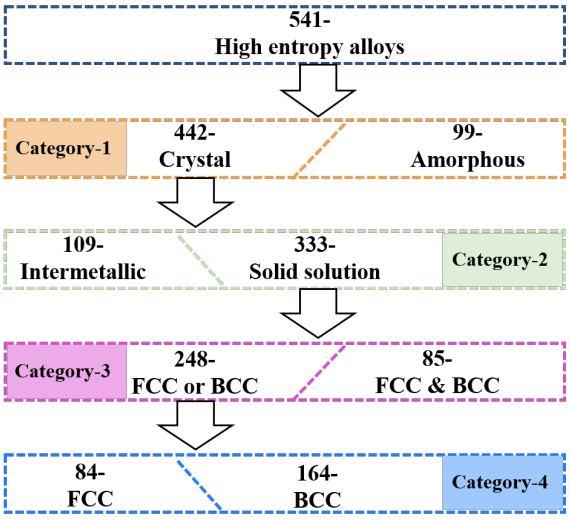


Fig. S1 The process of high entropy alloys phase structure prediction.

[a] State Key Laboratory of Solidification Processing, Northwestern Polytechnical University, Xi'an 710072, China. E-mail: rhyuan@nwpu.edu.cn, ljsh@nwpu.edu.cn
[b] Chongqing Innovation Center, Northwestern Polytechnical University, Chongqing 401120, China
[c] AiMaterials Research LLC, Santa Fe, New Mexico 87501, USA. E-mail: turablookman@gmail.com
† Electronic Supplementary Information (ESI) available: [details of any supplementary information available should be included here]. See DOI: 00.0000/00000000.

Table S1 Typical alloy systems in the raw data. The number in '( )' represents the amount of alloys in related systems.

Element	System	Element	System
Light metal + 3d transition metal	AlCoCrFeNi(24)	Light metal/nonmetal + refractory metal	HfMoNbTiZrSi(5)
	AlCoCrFeMnNi(11)		AlMoNbTiV(5)
	AlCoCrCuFeNiTi(11)		AlMoTaTiV(2)
	AlCoCrFeNiTi(17)		AlHfNbTaTiZr(9)
	AlFeMnNiTi(3)	Light metal/nonmetal + 3d transition metal	AlCuMgMnZnSi(1)
	AlCrCuFeMnNi(7)		AlCuMgSiZn(10)
	AlCoCrFeNiV(4)		AlCoCrFeNiSi(5)
	AlCoCrCuFeNiV(12)		CoCrFeMnNiC(4)
	AlCoFeNi(4)		AlCCoCrFeNi(8)
	AlCrCuFeNi(8)		FeMnMoCrWCB(3)
3d transition metal	CoCrFeMnNi(6)	Others	CoCrHfFeNi(6)
	CrCuFeMnNi(13)		CoCrFeNbNi(9)
	CoCrFeNiV(6)		CoCrCuFeNiNb(4)
	CrCuFeMnV(4)		ZrTiCuNiBeAl(4)
Refractory metal	MoNbTaTiW(4)		LaCeAlNiCuCo(4)
	NbTaTiZr(7)		FeCoCrMoCBY(4)
	MoNbTiVZr(20)		ZrTiCuNiBeFe(4)
	HfMoNbTaTiZr(4)		NdAlNiCuFe(4)
	HfMoNbTiZr(12)		...

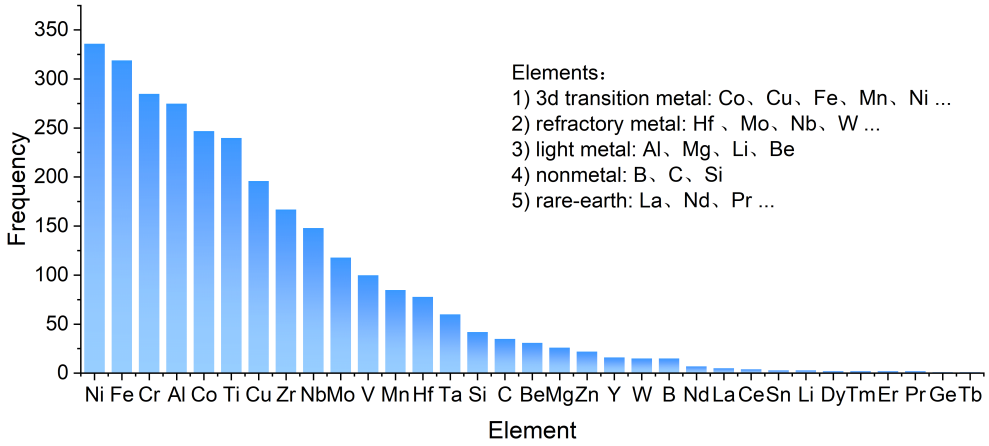


Fig. S2 Statistics of occurrence frequency of 33 elements in 541 high entropy alloys.

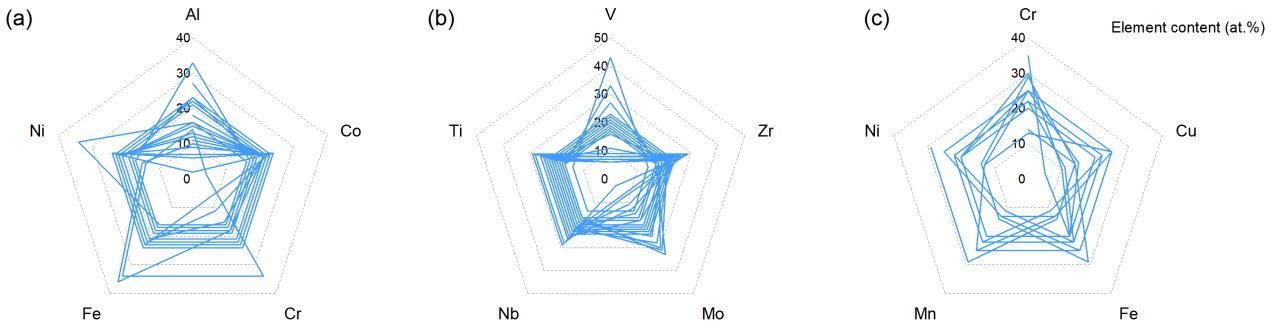


Fig. S3 The change of chemical compositions in several typical systems. (a)Al-Co-Cr-Fe-Ni, (b)Mo-Nb-Ti-V-Zr and (c)Cr-Cu-Fe-Mn-Ni.

Table S2 List of descriptors used in this article.

Category	Description	Abbreviation	Formulation
Atomic Descriptors	Atomic Number	AN	
	Atomic Radius	r	
	Covalent Radius	r_{cov}	
	Van Der Waals Radius	r_{van}	
	Atomic Volume	AV	
	Molar Volume	MV	
	Pauling Electronegativity	χ_P	
	Valence Electron Concentration	VEC	
	Allred Rochow Electronegativity	χ_A	
	Heat of Vaporization	H_V	
	First Ionization Energy	FIE	
	Second Ionization Energy	SIE	$P = \sum_{i=1}^{n} C_i * P_i$
	Electron Affinity	EA	$\delta P = \sqrt{\sum_{i=1}^{n} C_i * (1 - \frac{P_i}{P})^2}$
	Relative Atomic Mass	RAM	
	Melting Point Temperature	T_m	$\Delta P = max[C_i * (1 - \frac{P_i}{P})^2] - min[C_i * (1 - \frac{P_i}{P})^2]$
	Boiling Point Temperature	B_m	
	Density	ρ	
	Cross Section	CS	
	Specific Heat Capacity	SHC	
	Enthalpy of Fusion	E_F	
	Enthalpy of Vaporization	E_V	
	Enthalpy of Atomization	E_A	
	Thermal Conductivity	TC	
	Resistivity	R_Ω	
	Bulk Modulus	BM	
	Thermal Expansion	TE	
Theoretical Descriptors	X_e parameter	X_e	$X_e = 4.12 * \delta r \sqrt{\frac{BM * AV}{k_B * T_m}}$
	X_c parameter	X_c	$X_c = 2 * \sqrt{\frac{\delta_{\Delta H_m}}{k_B * T_m}}$
Thermodynamic Descriptors	Mixing enthalpy	ΔH_m	$\Delta H_m = 4 \sum_{i=1, i<j}^{n} H_{ij} C_i C_j$
		$\delta_{\Delta H_m}$	$\delta_{\Delta H_m} = \sqrt{\sum_{i=1.i<j}^{n} C_i C_j (H_{ij} - \Delta H_m)^2}$
	Mixing entropy	ΔS_m	$\Delta S_m = -R \sum_{i=1}^{n} C_i ln C_i$
	Λ parameter	Λ	$\Lambda = \frac{\Delta S_m}{\delta r^2}$
	Ω parameter	Ω	$\Omega = \frac{T_m * \Delta S_m}{\lvert \Delta H_m \rvert}$

Footnotes: (i) In the above formulas, R represents the molar gas constant with a value of 8.314 J/(mol·K), k_B represents the Boltzman constant with a value of $1.38*10^{-23}$(J/K), C_i and P_i represent the atomic fraction and the value of atomic descriptor of the ith element. (ii) X_e and X_c stand for the extended atomic size difference and chemical bond misfit, respectively. These two attributes can quantify the potential energy landscape brought about by the chemical and mechanical interaction among constituent elements.

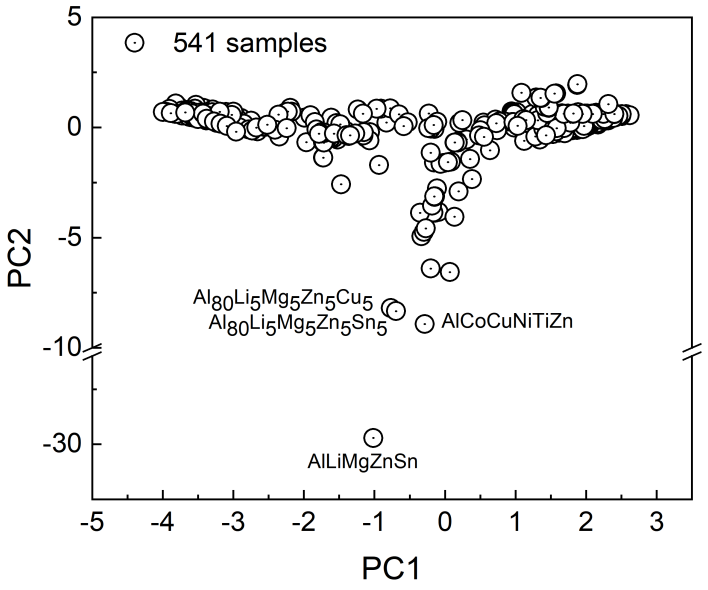


Fig. S4 Principal component analysis of 541 high-entropy alloy samples using chemical composition as input.

2 The effect of mathematic operators and descriptor space size on the phase prediction accuracy

The final descriptor space size is determined by the number of original descriptors and mathematic operators, together with the times the operators will be used. The size of the descriptor space after expansion is roughly given by $\#\Phi_n = (\#\Phi_0)^{2^n} \times (\#\hat{H}_2)^{2^n-1}$, where Φ_0 is the initial descriptor space and $\#\Phi_0$ is the corresponding number of descriptors in Φ_0, and $\hat{H}_2$ and $\#\hat{H}_2$ represent the operator space and the number of binary operators used, respectively. In this study, all the 85 descriptors were used instead of ranking and selecting the descriptors with feature selection methods such as the Pearson correlation analysis and gradient boosting tree. Besides, we used three unary operators and five binary operators mentioned in *subsection 2.3*. Therefore, $\#\Phi_0$=85 and $\#\hat{H}_2$=5. In addition, n represents the operator iteration, sepcifically, n=1 means that the operator is used only once and the resultant descriptor is given by, *e.g.* $x_1 * x_2$. Similarly, when the operator is used twice (n=2), the descriptor form is analogous to $(x_1+x_2)*(x_1-x_2)$. By setting the operator iteration n with a value of 2, the expanded descriptor space size $\#\Phi_n$ can reach 10^9, which is quite huge for the next computation by SIS and SO. With n further increases, the descriptor space size expands exponentially and the new descriptors obtained will be too complicated to understand.

Suppose the initial descriptors, mathematic operators and their iterations are fixed, the most key factor affecting the prediction accuracy and computational efficiency is the size of subspace that contains optimal descriptors ranked and selected by SIS. In general, the larger the subspace, the higher the possibility to find optimal descriptors for phase prediction. However, the whole descriptor space is rather huge ($\sim 10^5$ for n=1, and $\sim 10^9$ for n=2) and it is not realistic to enumerate every descriptor. Thus, to balance the computational resource required and the prediction accuracy, we first examine the dependence of prediction accuracy on the descriptor subspace size.

Figure S5(a) and (b) show the change of prediction accuracy with the descriptor subspace size. For simplicity, n=1 is used as an example. In each of the four categories, the accuracy slightly increases and tends to a stable value with increasing subspace size. This suggests that an optimal subspace with a critical size is adequate to capture the information of the whole space. Thus, the use of such a much smaller subspace can save computational resource without sacrificing the prediction accuracy. To be specific, from Category-1 to Category-4 the stable subspace size are 5000, 1000, 1000 and 5000, as listed in Table S3. The prediction accuracy as a function of subspace size is also investigated for n=2 and the optimal size is listed in Table S3, there is a similar tendency to Figure S4(a).

In addition to the descriptor subspace size, another important factor is the mathematic operators iterations, *i.e.*, the times operators are used. The operator iterations not only affect the form of new descriptors and the resultant prediction accuracy, but also the size of the whole descriptor space. With the optimized subspace size in Table S3, we examine how the prediction accuracy changes with n. Figure S4(c) plots the best prediction accuracy for Category-1 to Category-4, except for n=1 and 2, n=0 is also included as comparison, *i.e.*, no operators are used. For each category, the dataset is randomly divided into a training data and a test data with a ratio of 4:1, to check the generalizability of the learned descriptors. Overall, the accuracies for all the four categories increase with increasing n. For Category-1 and Category-4, the accuracies can reach a value $\sim$90% when n equal to 0 and increase slightly with n. However, the accuracies are improved adequately for Category-2 and Category-3 with n. Specifically, the accuracy for Category-2 is optimized from $\sim$50% to $\sim$75% with n changes from 0 to 2. For Category-3, the accuracy for n=0 and 2 is $\sim$60% and $\sim$80%, respectively. Furthermore, the accuracies in the test data agree well with that in the training data, indicating the good generalizability of newly constructed descriptors for phase prediction of HEAs.

Table S3 The optimized subspace size for the 2D descriptors. The 10^5 and 10^9 mean the size of total descriptor space with n=1 and n=2, respectively

Category	n=0	n=1	n=2
Category-1	85/85	5000/10^5	20000/10^9
Category-2	85/85	1000/10^5	20000/10^9
Category-3	85/85	1000/10^5	1000/10^9
Category-4	85/85	5000/10^5	20000/10^9

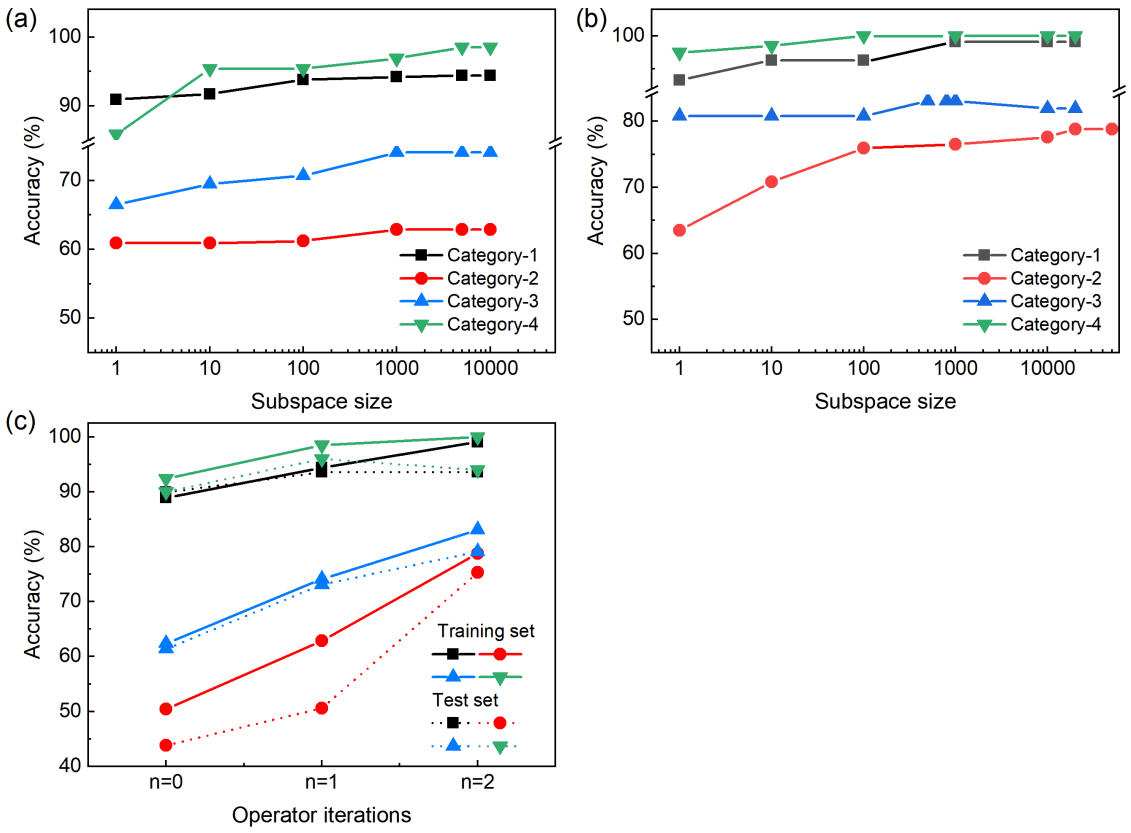


Fig. S5 Determination of important parameters of SISSO algorithm. (a) and (b) are the prediction accuracy varies with the descriptor subspace size for the 2D descriptors when n=1 (mathematic operators are used once) and n=2 (mathematic operators are used twice). (c) The prediction accuracy as a function of operator iterations (the times the operators are used).

3 The best 2D descriptors for phase prediction on test set

Figure S6 shows the phase prediction performance of the best 2D descriptors on the test set, the results agree with that in the training data in Figure 2 in the main text. Several samples are outside the convex hulls, this can be ascribed to the division of training data and test data. The random division cannot guarantee that the upper/lower limits of the descriptors in the test data are covered by that in the training data. In order to obtain prediction accuracy, the outside samples are labelled by calculating the distance to two convex hulls.

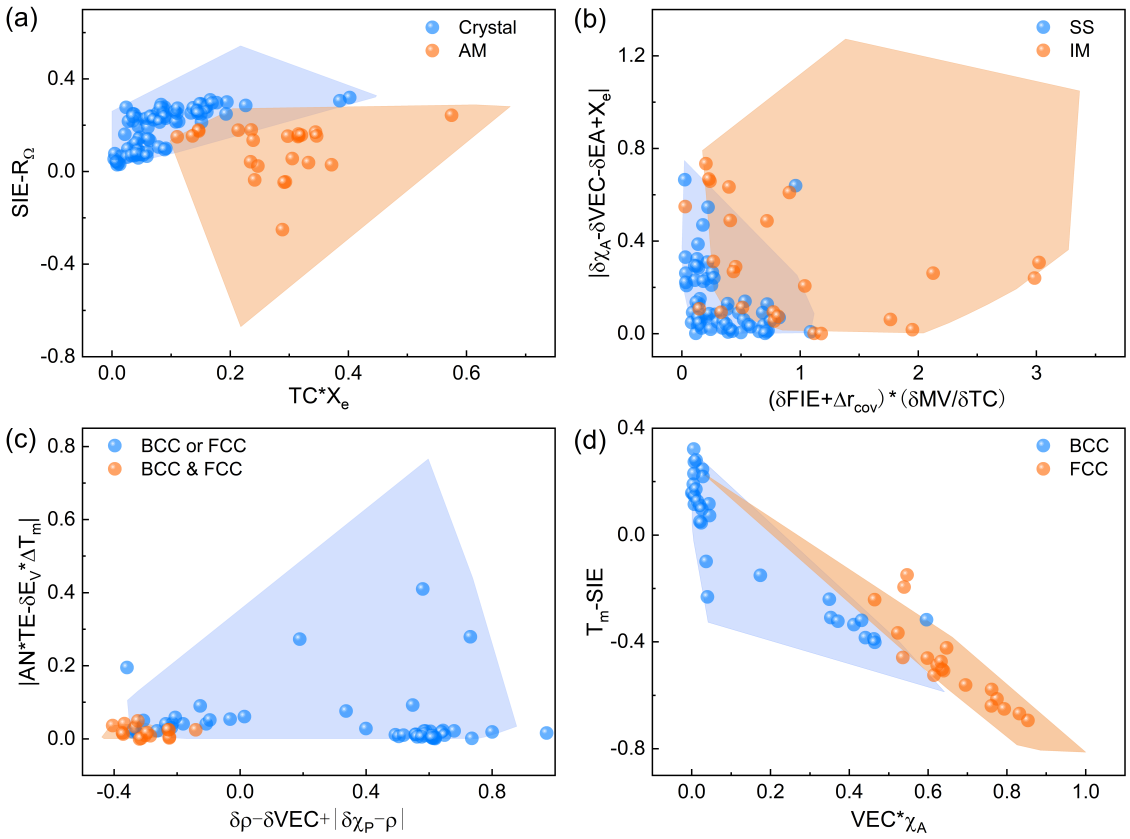


Fig. S6 The best 2D descriptors for phase prediction of test set. (a) to (d) represent the result for Category-1 to Category-4, respectively.

4 Data point statistics for the overlapping part of the convex hulls

The points in the overlapping region in Category-1 and Category-4 in the training set are listed in Table S4. In addition, due to the excessive data in the overlapping regions of Category-2 and Category-3, they are shown in data2.csv and data3.csv, respectively.

Table S4 The training data located in the Category-1 and Category-4 overlapping regions

Category-1	Phase structure	Category-4	Phase structure
AlCoCrCuNiTi	Crystal	$Al_{0.3}CrFe_{1.5}MnNi$	BCC
$Mo_{0.5}NbHf_{0.5}ZrTiSi_{0.9}$	Crystal	AlNbTiV	FCC
AlCoCrCuFeNiTi	Crystal	$Al_{0.59}CoCrFeNi$	FCC
ZrTiHfCuNiFe	Crystal		
AlCoFeNiTiVZr	AM		
FeMoNiTiVZr	AM		
AlFeNiTiVZr	AM		
$Ni_{42}Ti_{20}Zr_{20.5}Al_8Cu_5Si_{4.5}$	AM		
$AlCoCrCu_{0.5}FeNiSi$	AM		
CoCuFeNiTiVZr	AM		
CoCrCuFeNiTiVZr	AM		
AlCrTaTiZr	AM		
$Al_8Cu_7Ni_{19}Zr_{66}$	AM		
$Zr_{17}Ta_{16}Ti_{19}Nb_{22}Si_{26}$	AM		
AlMoNbSiTaTiVZr	AM		
$Be_{18}Cu_9Ni_8Ti_{65}$	AM		
$AlCoCrCu_{0.5}FeNiTi$	AM		
$Al_{10}Ni_{40}Cu_5Ti_{17}Zr_{28}$	AM		
CoCrFeMoNiTiVZr	AM		
$Al_{10}Ni_{40}Cu_6Ti_{16}Zr_{28}$	AM		
$Ti_{50}Zr_{10}Cu_{20}Ni_{20}$	AM		

5 Cluster analysis of the overlapped samples in Category-2 and Category-3

We conduct principal component analysis and unsupervised cluster analysis on the overlapped samples in Category-2 and Category-3, as shown in Figure2 (b) and (c) in the main text. In Figure S7, the purple and green ellipses represent the two clusters learned by K-means based on the first two principal components. In each cluster, the samples do not belong to a certain phase, indicating that there is no clear tendency underlying the overlapped samples.

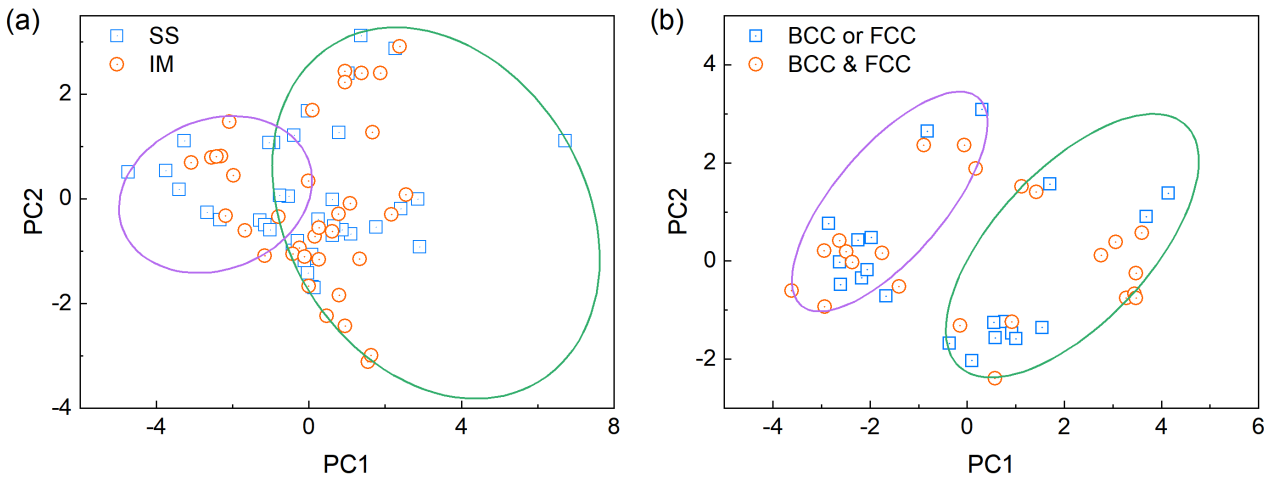


Fig. S7 Unsupervised cluster analysis of the overlapped samples in Category-2 (a) and Category-3 (b). The x and y axes represent the first two principal components calculated by the linear combination of empirical descriptors, respectively.

6 The prediction accuracies of empirical descriptors based 2D maps for Category-2

The most commonly used six empirical descriptors (δr, ΔH_m, Ω, ΔS_m, VEC and $\delta\chi_P$) are employed for phase prediction of Category-2. The random combination of these six descriptors gives up to 15 descriptors pairs. For each descriptor pair, the prediction accuracy is calculated using the same convex hull method. Table S5 lists the accuracies for the 15 descriptor pairs. Figure S8 shows the phase prediction performance of the descriptor pairs, for simplicity, the results for 4 descriptor pairs are merely presented.

Table S5 The 15 commonly used descriptor pairs and their accuracies for phase prediction of Category-2

Empirical descriptor	Accuracy (%)	Empirical descriptor	Accuracy (%)
δr & ΔH_m	26.3	δr & Ω	29.5
δr & ΔS_m	13.0	δr & $\delta\chi_P$	8.5
δr & VEC	16.4	ΔS_m & ΔH_m	32.9
ΔS_m & Ω	33.1	ΔS_m & $\delta\chi_P$	11.3
ΔS_m & VEC	20.4	ΔH_m & Ω	30.9
ΔH_m & $\delta\chi_P$	17.8	ΔH_m & VEC	24.9
Ω & $\delta\chi_P$	31.4	Ω & VEC	33.4
$\delta\chi_P$ & VEC	12.7		

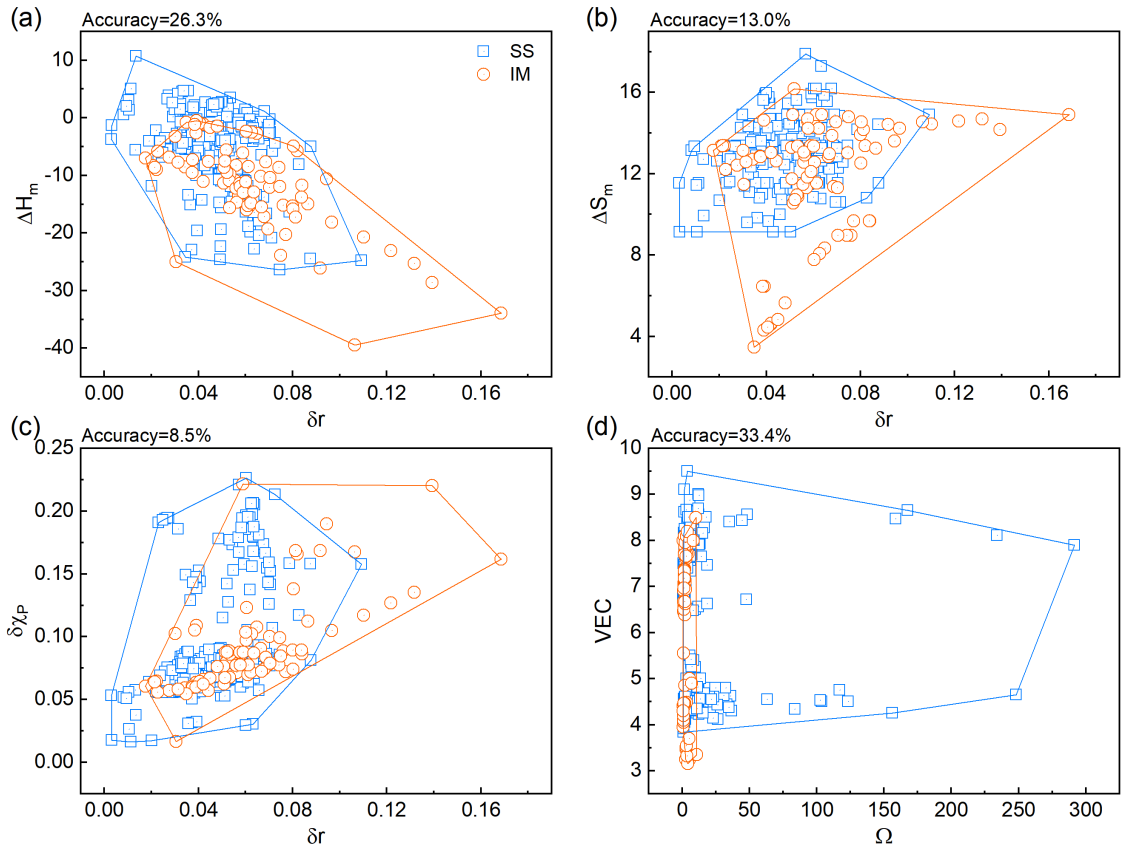


Fig. S8 Four of fifteen descriptor pairs for phase prediction of Category-2 are presented as examples, (a) δr & ΔH_m, (b) δr & ΔS_m, (c) δr & $\delta \chi_P$, (d) Ω & *VEC*.

7 The prediction accuracies of empirical descriptors based 2D maps for Category-3

The 15 descriptor pairs are also generally used for Category-3. Thus, the corresponding accuracies are calculated with the same convex hull method and listed in Table S6. In addition, Figure S9 shows the phase prediction performance of the descriptor pairs, for simplicity, the results for 4 descriptor pairs are merely presented.

Table S6 The 15 commonly used descriptor pairs and their accuracies for phase prediction of Category-3

Empirical descriptor	Accuracy (%)	Empirical descriptor	Accuracy (%)
δr & ΔH_m	21.4	δr & Ω	15.8
δr & ΔS_m	19.9	δr & $\delta\chi_P$	40.2
δr & VEC	47.7	ΔS_m & ΔH_m	15.4
ΔS_m & Ω	7.1	ΔS_m & $\delta\chi_P$	39.5
ΔS_m & VEC	46.2	ΔH_m & Ω	4.9
ΔH_m & $\delta\chi_P$	38.3	ΔH_m & VEC	47.0
Ω & $\delta\chi_P$	37.2	Ω & VEC	47.7
$\delta\chi_P$ & VEC	52.6		

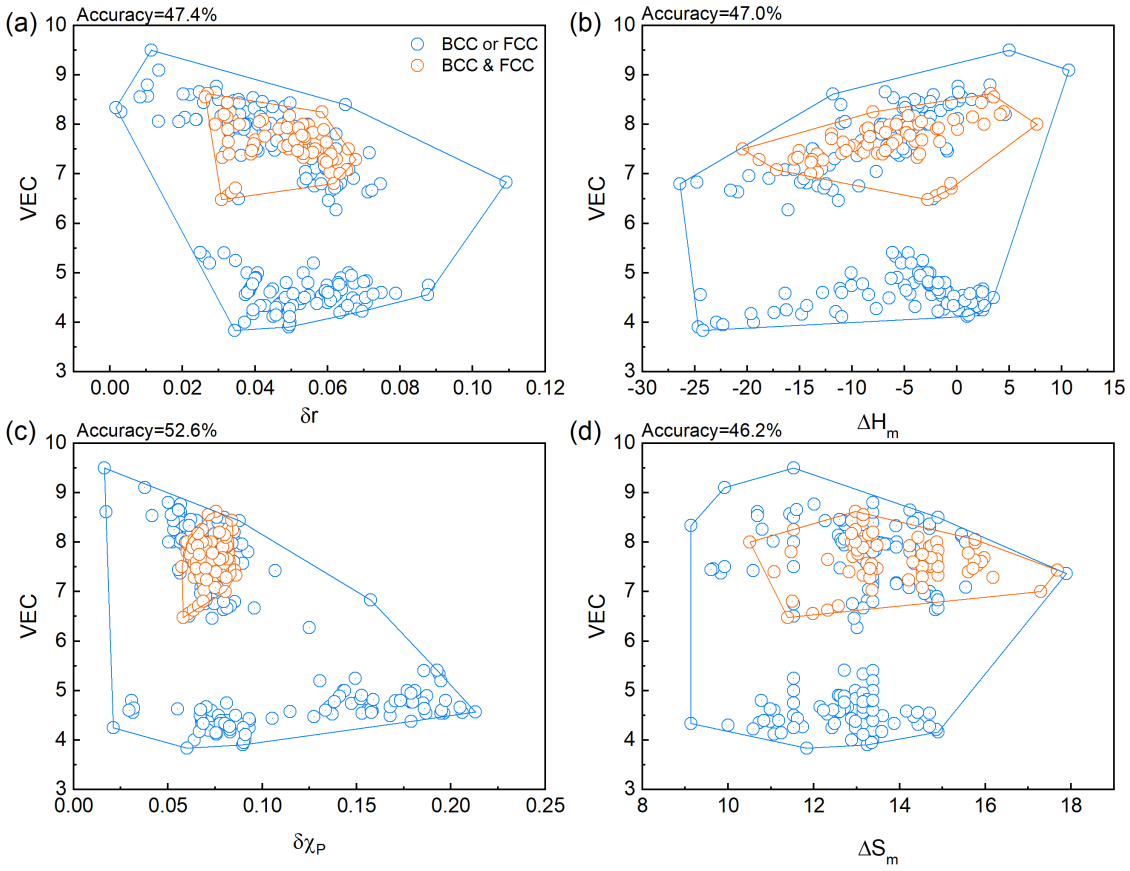


Fig. S9 Four of fifteen descriptor pairs for phase prediction of Category-3 are presented as examples. (a) δr & *VEC*, (b) ΔH_m & *VEC*, (c) $\delta \chi_P$ & *VEC*, (d) ΔS_m & *VEC*.

8 The calibration of prediction accuracies for 1D descriptors

Figure S10 shows the 4 descriptors with calibration accuracy larger than 90% for Category-1, and Figure S11 shows the 3 descriptors with calibration accuracy larger than 90% for Category-4.

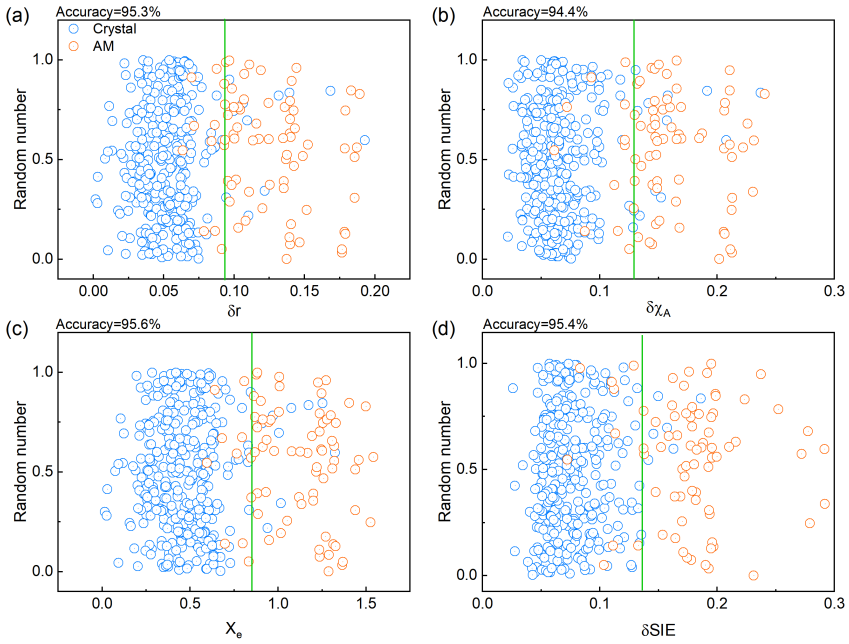


Fig. S10 The calibration of prediction accuracy for the 1D descriptors (n=0), for Category-1. (a) δr, (b) $\delta\chi_A$, (c) X_e, (d) δSIE. The green vertical line is used for the calibration and the y-axis is set as random numbers to avoid the overlap of samples.

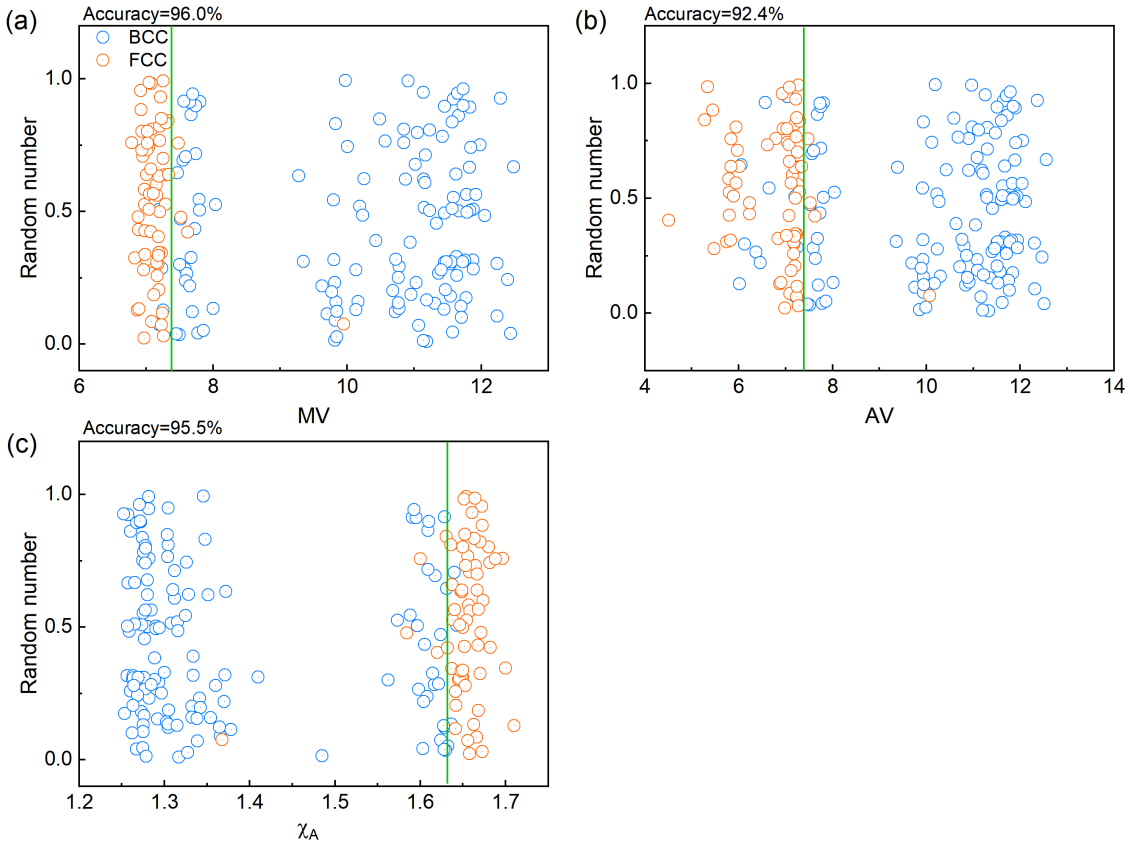


Fig. S11 The calibration of prediction accuracy for the 1D descriptors (n=0), for Category-4, (a) *MV*, (b) *AV*, (c) χ_A. The green vertical line is used for the calibration and the y-axis is set as random numbers to avoid the overlap of samples.

9 Analysis and experimental test of new alloy compositions.

Although Al-Co-Cr-Cu-Fe-Ni and V-Zr-Mo-Nb-Ti are alloy systems present in the raw data, which is different from the raw data, as shown in Figure S12. This paper respectively focuses on analyzing the effect of changes in the element content of Al, Cu and V, Zr respectively. For each alloy, the prediction is made from Category-1 to Category-4, sequentially, as shown in Figure S1.

We have measured the backscattering electron images of Al-Cu-Co-Cr-Fe-Ni alloys using SEM. As show in Figure S13(a)-(d), there is no minor phases in all the four alloys. In addition, we also synthesized four alloys consisting of Hf, Mo, Nb, Ta and W, to again validate the performance of the new descriptors. To our best knowledge, this system has not been reported or synthesized before. Figure S14 shows the XRD measured phase structures of these alloys, which agree well with the predictions.

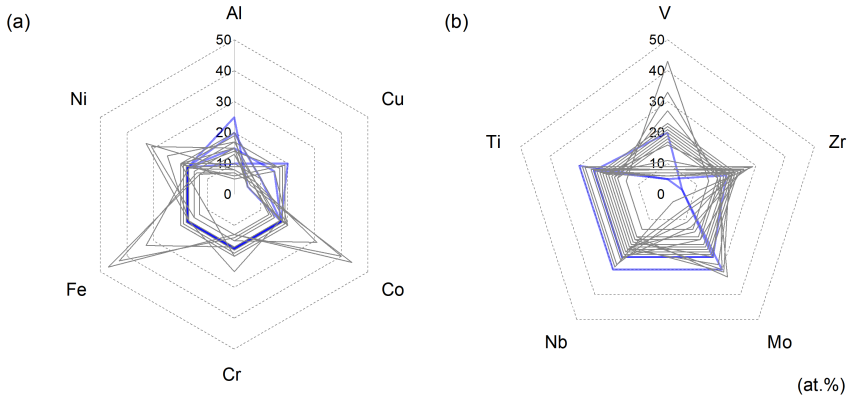


Fig. S12 Comparison of the alloy compositions in (a) $Al_xCu_y(CoCrFeNi)_{(100-x-y)}$ and (b) $V_xZr_y(MoNbTi)_{(100-x-y)}$ systems. The gray lines represent the initial compositions, and the blue lines are experimental data.

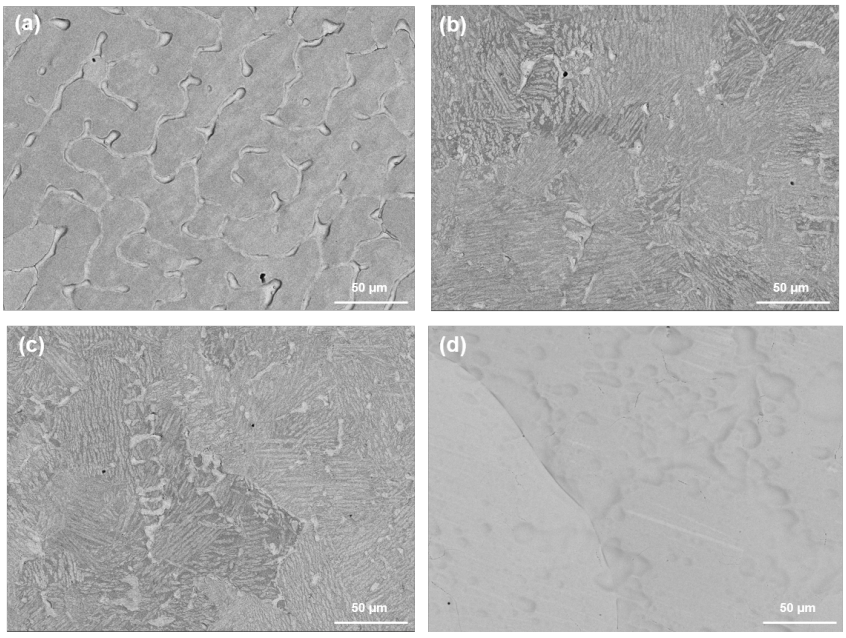
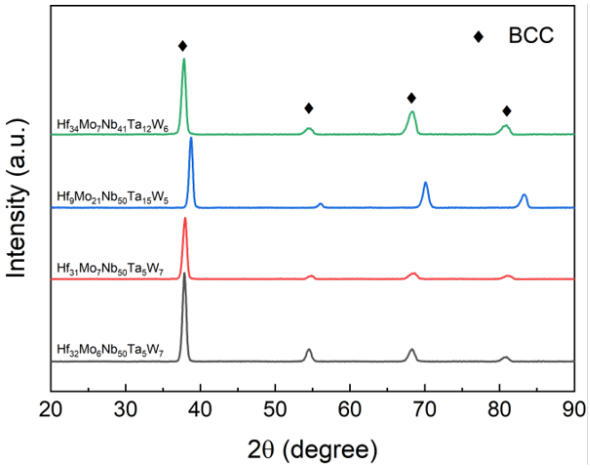


Fig. S13 The SEM backscattering electron images of $Al_xCu_y(CoCrFeNi)_{(100-x-y)}$ alloys, (a) $Al_{10}Cu_{20}(CoCrFeNi)_{70}$, (b) $Al_{15}Cu_{15}(CoCrFeNi)_{70}$, (c) $Al_{20}Cu_{10}(CoCrFeNi)_{70}$, (d) $Al_{25}Cu_5(CoCrFeNi)_{70}$.



Composition	Prediction	XRD
$Hf_{32}Mo_6Nb_{50}Ta_5W_7$	BCC	BCC
$Hf_{31}Mo_7Nb_{50}Ta_5W_7$	BCC	BCC
$Hf_9Mo_{21}Nb_{50}Ta_{15}W_5$	BCC	BCC
$Hf_{34}Mo_7Nb_{41}Ta_{12}W_6$	BCC	BCC

Fig. S14 The experimental and predicted results of Hf-Mo-Nb-Ta-W HEAs system.

10 The frequency of empirical descriptors for Category-2 in different training and test data

We have randomly divided the initial data for another four times in Category-2. Figure S15 shows the first 10 empirical descriptors that have higher frequency of occurrence in the top 100 descriptors.

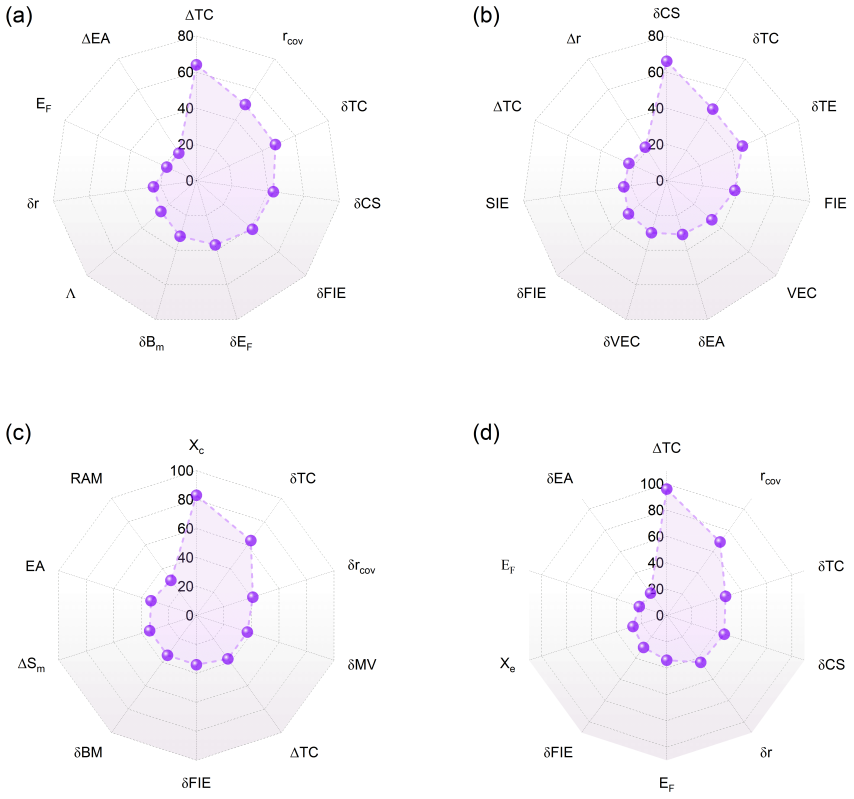


Fig. S15 The appearance frequency of empirical descriptors in Category-2 with different data division.