Research Article

Predicting single-phase solid solutions in as-sputtered high entropy alloys: High-throughput screening with machine-learning model

Ji-Chang Ren[a], Junjun Zhou[a], Christopher J. Butch[b,*], Zhigang Ding[a], Shuang Li[a], Yonghao Zhao[a], Wei Liu[a,*]

[a] Nano and Heterogeneous Materials Center, School of Materials Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China
[b] Department of Biomedical Engineering, Nanjing University, Nanjing 210023, China

A B S T R A C T

Searching for single-phase solid solutions (SPSSs) in high-entropy alloys (HEAs) is a prerequisite for the intentional design and manipulation of microstructures of alloys in vast composition space. However, to date, reported SPSS HEAs are still rare due to the lack of reliable guiding principles for the synthesis of new SPSS HEAs. Here, we demonstrate an ensemble machine-learning method capable of discovering SPSS HEAs by directly predicting quinary phase diagrams based only on atomic composition. A total of 2198 experimental structure data are extracted from as-sputtered quinary HEAs in the literature and used to train a random forest classifier (termed AS-RF) utilizing bagging, achieving a prediction accuracy of 94.6% compared with experimental results. The AS-RF model is then utilized to predict 224 quinary phase diagrams including ~32,000 SPSS HEAs in Cr-Co-Fe-Ni-Mn-Cu-Al composition space. The extrapolation capability of the AS-RF model is then validated by performing first-principle calculations using density functional theory as a benchmark for the predicted phase transition of newly predicted HEAs. Finally, interpretation of the AS-RF model weighting of the input parameters also sheds light on the driving forces behind HEA formation in sputtered systems with the main contributors being: valance electron concentration, work function, atomic radius difference and elementary symmetries.

© 2022 Published by Elsevier Ltd on behalf of The editorial office of Journal of Materials Science & Technology.

1. Introduction

High entropy alloys (HEAs) [1,2] have been emerging as a new research field in the materials community. The complex local chemical environment in HEAs significantly alters the mechanical and electronic properties of these materials, leading to favorable combinations of strength and ductility, high fracture toughness, and high thermal stability [3–18]. To date, only a small number of alloys have been demonstrated experimentally, primarily through a trial-and-error approach to the hyper-compositional space of HEAs. Transitioning this trial and error approach to one of intentional design and manipulation of the HEA microstructures requires tools to reliably explore single-phase solid solutions (SPSS) with simple FCC or BCC crystal structures [1,2,19]. However, the multiple principal elements in HEAs lead to complex phase diagrams making the design of desirable HEAs with SPSS challenging and frustrating the development of accurate phase selection rules for HEAs [20–22].

Presently, the design tools available for the selection of new SPSS-HEAs include empirical parametric analysis and density functional theory (DFT). The parameter method [23] provides phase estimates from the thermo-physical parameters of elementary constituents of potential HEAs [24–28] and has been demonstrated to be able to decrease the phase overlap between SPSS and intermetallic phase in HEAs [20,29–31]. However, with the number of thermos-physical factors going beyond three [22,23], the empirical methods can hardly provide reliable phase separations for HEAs. On the other hand, DFT can provide accurate descriptions of the complex chemical environment of HEAs, including lattice distortion [32], phase stabilities [33] and transitions [34]. However, the high computational cost of DFT significantly hinders the full exploration of phase formation in HEAs.

As an alternative to the traditional approaches discussed above, several groups have applied machine learning (ML) algorithms to the complex relationship between composition and phase forma-

---

* Corresponding authors.
E-mail addresses: chrisbutch@gmail.com (C.J. Butch), weiliu@njust.edu.cn (W. Liu).
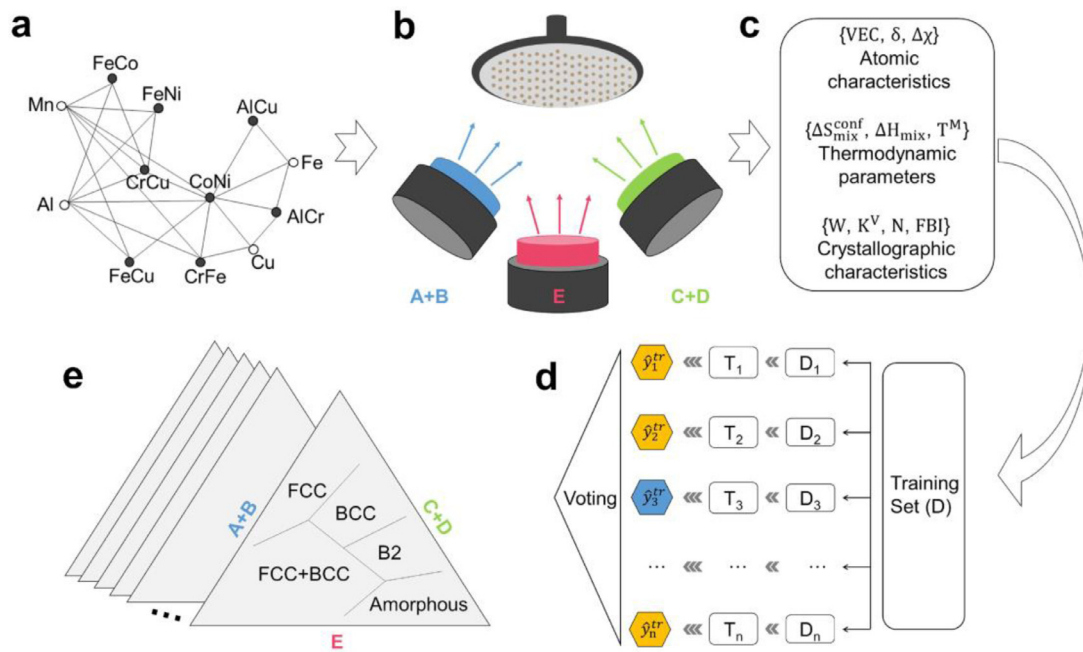
Fig. 1. A schematic diagram of the random forest model for the phase diagram predictions. (a) Elemental network for the formation of five-principal HEAs. (b) A schematic graph represents the fabrication processes of as-sputtered HEAs. (c) A list of thermo-physical parameters for the descriptions of the HEA samples. (d) The architecture of the AS-RF machine learning model. D_n is the subset of the training data. T_n is a single decision tree. $\hat{y}_n^{tr}$ is decision made by T_n. The filled colors in the figure denote different phases of HEAs. During the training process, to reduce the similarity and correlations between decision tresses, 60% of the training data have been randomly selected for each tree. To make the final predictions in the RF model, each tree first collects the results from the leaf nodes and votes for the one with largest probabilities, and then the RF collects the votes from each tree and provides the final prediction. (e) A list of schematic phase diagrams as the output of the AS-RF model.

tion in HEAs [35–42]. Artificial neural networks (ANN), have been applied to predict the formation of solid solution phases, intermetallic compound phases and mixed phases [37] based on a training set of 401 as-cast HEAs. This model yielded a prediction accuracy of 75% due in part to the small training dataset (the ANN model usually requires thousands of samples for training and testing [43]). Interestingly, by combining an artificial HEA dataset (created through synthetic minority over sampling technique) with a dataset of 322 as-cast samples, an alternative model using support vector machines (SVMs) was able to increase prediction accuracy beyond 90% for distinguishing SPSS with BCC and FCC structures [38].

Each of the above ML algorithms is limited in part by the available experimental data on which they were trained, and bias in that data due to the imbalance of BCC structures in reported experimental quinary HEAs. The underlying issue here is the difficulty in obtaining collect enough HEA samples for training ML models, since the conventional HEA casting process is time-consuming and ill-suited to high-throughput experiments. Moreover, the collected data are complex, influenced by variations in the fabrication process [19,44], and often incompletely characterized [45]. Fortunately, high-throughput HEA experimental methods have recently been developed by employing combinatorial magnetron co-sputtering and this method was used to produce 2198 high-quality quinary HEAs [46]. From a kinetic and thermodynamic point of view, sputtering processes undergo different thermodynamic conditions from those of casting processes, and the cooling rate in the sputtering processes is at least five orders of magnitude larger than that for casting processes. Compared with casting and electrochemical deposition methods, sputtering provides uniform experimental conditions with well-controlled compositions making it well suited for experimental exploration of the HEA space. A schematic overview of these 2198 HEAs (Fig. 1(a)) shows seven

elements (Fe, Co, Ni, Mn, Cr, Al, Cu) combined to form BCC type SPSSs, FCC type SPSSs, BCC+FCC type multi-phase solid solutions (MPSS), B2 (IM) and amorphous phases (See Ref. [46] and Fig. S1 in Supplemental Materials for more details). However, the complexity of this chemical space makes it a nontrivial task to construct an explicit relationship between the atomic compositions and phase formations in HEAs.

In this contribution, we overcome this challenge by constructing a random forest ML model capable of predicting the formation of SPSS HEAs trained on the database of high-throughput as-sputtered HEAs [46]. Random Forest ML model has the built-in ability to estimate feature importance contributions, a characteristic that allows the model to be interpreted with the order and effect size of the feature associated with the outcome. Compared to other ML models, this as-sputtered random forest (AS-RF) model has the advantage of the easy interpretation of the factors underlying the predictions, allowing us to quickly identify relationships between the features of alloys and phase formation. By screening and selecting the most relevant features, our AS-RF model can distinguish five phases, including BCC type SPSS, FCC type SPSS, BCC+FCC type multi-phase solid solution (MPSS), B2 (IM) and amorphous phases, with a high accuracy of 94.6% for the prediction of HEAs with SPSS phase. We then used the AS-RF model to predict 224 quinary phase diagrams including ~32,000 SPSS HEAs in the Cr-Co-Fe-Ni-Mn-Cu-Al composition space. Finally, we compared the AS-RF predicted HEAs to first-principles calculations of the same compositions using DFT and show the results to be strongly predictive. These results demonstrate that the AS-RF model accurately reveals the interplays between the driving forces (especially valance electron concentration, work function, atomic radius difference and elementary symmetries) on the phase formations of as-sputtered HEAs, providing a tool to guide the exploration of HEAs with SPSS structures.

2. Computational methods

2.1. Random forest model

As an ensemble learning method, RF is built upon decision trees [47], in which the number and depth of the tree control the complexity of the model. To make reliable decision making, we applied one of the popular algorithms, so-called Gini-index method [48], for the split of the nodes and the decision making in each node is achieved by minimizing the value of Gini-index. For the evaluation of the quality of the model, ten-fold cross-validation (CV) was applied to achieve the tradeoff between bias and variance in our RF model. We have repeated 300 independent training and testing processes to guarantee the stability of our RF model. We have also performed "learning curve" estimation, precision and recall calculations, F1-score comparison between different models, receiver operating characteristic curve and confusion matrix estimations of the AS-RF model, as seen in Figs. S2–S5 in the Supplemental Materials.

For the split of the nodes, there are two commonly applied algorithms, *i.e.*, Gini-index method and the entropy criterion. The Gini-index favors large distributions of the dataset and is simple to implement, whereas the entropy criterion favors small distributions of the dataset with a variety of diverse values and is good at classifications of a highly imbalanced dataset. In our work, the dataset structure of the HEAs is almost balanced and possesses big distributions. Therefore, here the Gini-index method was adopted for the split of the nodes, which is expressed as

$$\text{Gini} = 1 - \sum_{m=1}^{M} (p_m)^2 \tag{1}$$

where p_m denotes the probability of the training sample belonging to class m. M is the number of classifications. The criterion for splitting is to choose the feature with the least Gini index for the corresponding node.

In each decision tree, the relationship between the predicted class $\hat{y}^{\text{tr}}$ and features x can be determined by

$$\hat{y}^{\text{tr}} = \sum_{m=1}^{M} c_m I\{x \in R_m\} \tag{2}$$

where R_m is the subset date in the leaf node, and c_m is the training class in the leaf node R_m. $I\{x \in R_m\}$ is the identity function that returns 1 if x is in the subset R_m and otherwise returns 0.

In the AS-RF model, each tree is treated as one independent estimator and votes for one class. Then, RF collects the votes from trees and makes the final prediction with the majority vote. Here, we used the so-called soft majority voting algorithm [49] to estimate the phase predictions, with the following formula:

$$\hat{y} = \arg\max_j \sum_{i=1}^{M} \omega_i p_{ij}^{\text{tr}} \tag{3}$$

where p_{ij}^{tr} is the probability of predicted class label j from the i-th tree, and $\hat{y}$ is the final predicted class with the RF model.

Based on the scikit-learn library [50], we built the AS-RF model systematically by growing the depth and the number of the trees. After a comprehensive grid search for the parameters of the tree and depth (see Fig. S6 in Supplemental Materials for more details), 100 trees with 30 depths have been adopted for the final optimized model. The correlations between trees are minimized with a randomly selection of features.

2.2. Mann-Whitney-Wilcoxon U test

The Mann-Whitney-Wilcoxon U test is performed using the following equations:

$$U = \text{AUC} \cdot n_{\text{BCC}} \cdot n_{\text{FCC}} \tag{4}$$

$$z = \frac{U - \frac{n_{\text{BCC}} \cdot n_{\text{FCC}}}{2}}{\sqrt{\frac{n_{\text{BCC}} \cdot n_{\text{FCC}} \cdot (N+1)}{12}}} \tag{5}$$

where U is Mann-Whitney-Wilcoxon U statistic, AUC is the area under the ROC curve (in Fig. 5(c) and (d)), n_{BCC} is the number of BCC samples by DFT (19), n_{FCC} is the number of FCC samples by DFT (56), N is the total number of the samples (75) and z is z-score.

2.3. Density functional theory calculations

As an external validation step for the machine learning predictions, the phase stabilities of some randomly selected HEAs were estimated by performing density functional theory (DFT) calculations as implemented in Vienna Ab-initio Simulation Package (VASP) code [51]. For DFT validation, we have selected 41 HEAs from five phase diagrams, $(Al_x Cu_y[(FeCoNi)_{0.33}]_z$, $Co_x Cu_y[(AlCrFe)_{0.33}]_z$, $Fe_x Co_y[(AlNiCu)_{0.33}]_z)$, $(Al_x Ni_y[(FeCoCu)_{0.33}]_z$, $Co_x Cu_y[(CrFeNi)_{0.33}]_z)$, and for each HEA system, a supercell size of $3 \times 3 \times 3$ has been applied for the total energy calculations. The details can be found in Table S3 in Supplemental Materials. Special quasi-random structure method [52] was applied to construct chemical disordered HEAs, as implemented in Alloy Theoretic Automated Toolkit package [53]. Thee exchange correlation interactions between valance electrons were described within the generalized gradient approximation with Perdew-Burke-Ernzerhof types [54]. The Projector-Augmented Wave pseudopotentials were applied for the electron-ion interactions. To solve the Kohn-Sham equations, a $3 \times 3 \times 3$ k-point mesh was applied to sample the Brillouin Zone, along with an energy cutoff of 400 eV for the construction of the basis sets. For the convergence criteria, a residual force threshold of 0.001 eV/Å was used for the geometric optimizations and an error bar of 0.1 meV/atom was adopted for the final total energy estimations.

3. Results and discussion

3.1. The AS-RF machine learning model workflow

To appropriately characterize the as-sputtered HEAs with different phases, we applied several thermodynamic and intrinsic physical characteristics of HEAs as the basic features for the HEA samples (Fig. 1(c)) (*c.f.* Tables S1–S2 for the detailed expressions in Supplemental Materials). Among the features, mixing enthalpy (ΔH_{mix}), mixing configurational entropy $(\Delta S_{\text{mix}}^{\text{conf}})$ and averaged melting temperature (T^{M}) stand for thermodynamic effects on the phase formations of as-sputtered HEAs. And inspired by Home-Rothery rules that the elements possessing similar electronegativity and similar valance electron numbers would favor to form solid solution phases, here we introduce valence electron concentration (VEC) and electronegativity difference $(\Delta \chi)$ as another two basic features for HEAs. Further, we introduce atomic size difference (δ) and averaged space group (N) to represent strain and atomic structural properties of the samples. Different from as-cast samples, sputtering HEAs are deposited on the substrate and form thin films through polymorphic crystallization, and the formation of phases is largely determined by the crystalline orientations and surface energies. To capture these surface electronic and elementary structural
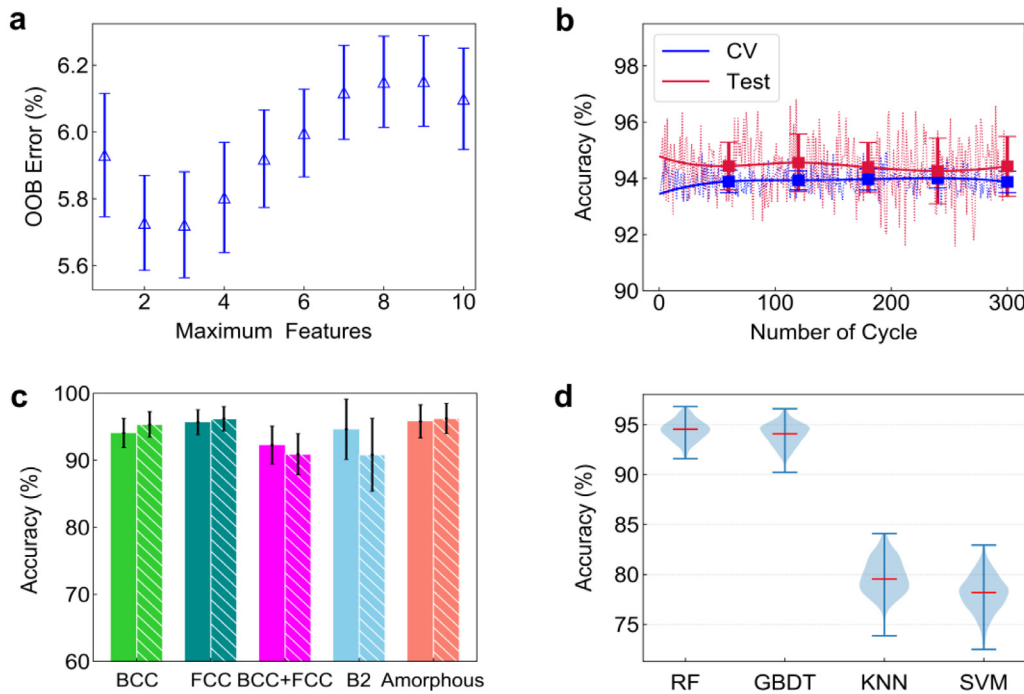
Fig. 2. The validation of the quality of our RF model. (a) Out-of-bag (OOB) error of our RF model as a function of maximum features that randomly selected in each individual tree. (b) The accuracies of prediction and cross-validation (CV) for 300 independent processes. The solid dot lines are their average values for every 60 times predictions, and their error bars are present at each dot point. (c) The prediction accuracies of precision (solid) and recall (slashed) for BCC, FCC, BCC+FCC, B2 and amorphous phases. The corresponding error bars are denoted in the histogram. (d) The Violin plots of the prediction accuracies of the four models over 300 independent experiments under ten-fold cross validation.

effects, we introduced averaged work function (W) and FCC-BCC index (FBI) as basic features to study their possible effects on the phase selections of as-sputtered HEAs.

The ten parameters descriptor-set was then calculated for the 2198 as-sputtered HEA samples along with the known phase (see Table S2 in Supplemental Materials for the detailed calculations of the descriptors), and these data were used to train and optimize the AS-RF model. The AS-RF model (shown schematically in Fig. 1(d) and (e)) is constructed by an ensemble of decision tress, and for each tree, the bootstrapping strategy is adopted by randomly selecting 63.2% of the training data with replacement. During the training processes, the out-of-bag (OOB) method [55] was adopted for the estimation of the variance of the model by applying the rest 36.8% of the training data. To reduce the variance, the correlation between trees was minimized by bagging different subsets of the training data and randomly selecting the considered features at each decision-making step in each tree.

3.2. The quality of the AS-RF model

By systematically increasing the number of the selected features for each tree, in Fig. 2(a) the smallest OOB error (around 5.7% at three features) was obtained. With the optimized parameters, we calculated the overall phase prediction accuracy for each of the five phases. As seen in Fig. 2(b), the averaged prediction accuracy is 94.6% with a low variance over 300 independent training processes. We additionally investigated the prediction balance for each phase by calculating their precisions and recalls, respectively [56]. A large difference between the values of precision and recall indicates either too tight or too loose phase selection rules. In Fig. 2(c), there is no obvious difference between precision and recall for in any phase, indicating high prediction quality in the optimized AS-RF model.

We then evaluated the performance of our AS-RF model by using the same training data with three alternate well-known ma-

chine learning models: SVM, *k*-nearest neighbor (KNN) [57], and gradient boosted decision trees (GBDT) [58]. Fig. 2(d) and Fig. S7 show that the AS-RF model exhibits the best performance among the four models, with the largest prediction accuracy and the lowest bias and variance. Unsurprisingly, SVM and KNN perform worse in the phase predictions. For both algorithms [59,60], the decision functions are based either on the local support vectors or *k* nearest points as opposed to those in AS-RF and GBDT models, where the prediction is dependent on all samples in the feature space. In contrast, the RF and GBDT models vote based on randomly selected data, leading to an improved balance of bias and variance. However, compared with the AS-RF model, the GBDT model is relatively sensitive to the collected dataset, especially for multi-classification problems [61]. This makes the convergence and generalization of the GBDT model less reliable than the AS-RF model for the phase predictions in HEAs.

3.3. The phase prediction of quinary HEAs

Using the AS-RF model, we first predicted 14 phase diagrams based on compositions consistent with high-throughput experiments and showed them to be in good agreement with the experimental data (Figs. 3(a), S8(1–14)). We then predicted a comprehensive set of five-principal element phase diagrams (with atomic ratio >5% and <35%) of 210 previously uncharacterized compositions including ∼67,000 unknown alloys (see Supplemental Materials for details). To provide a statistical analysis of HEAs with SPSS phase, it is essential to construct direct relationships between constituent elements and phase formations of as-sputtered HEAs. To do this, we constructed a full-connected network within the Cr-Co-Fe-Ni-Mn-Al-Cu chemical space, containing all possible five-principal alloys with component concentrations ranging from 5% to 35%. By mapping 67,000 alloys separately onto the network according to their phase structures, different correlations between elements can then be revealed, as seen in Fig. 4. In particular for HEAs with SPSS,
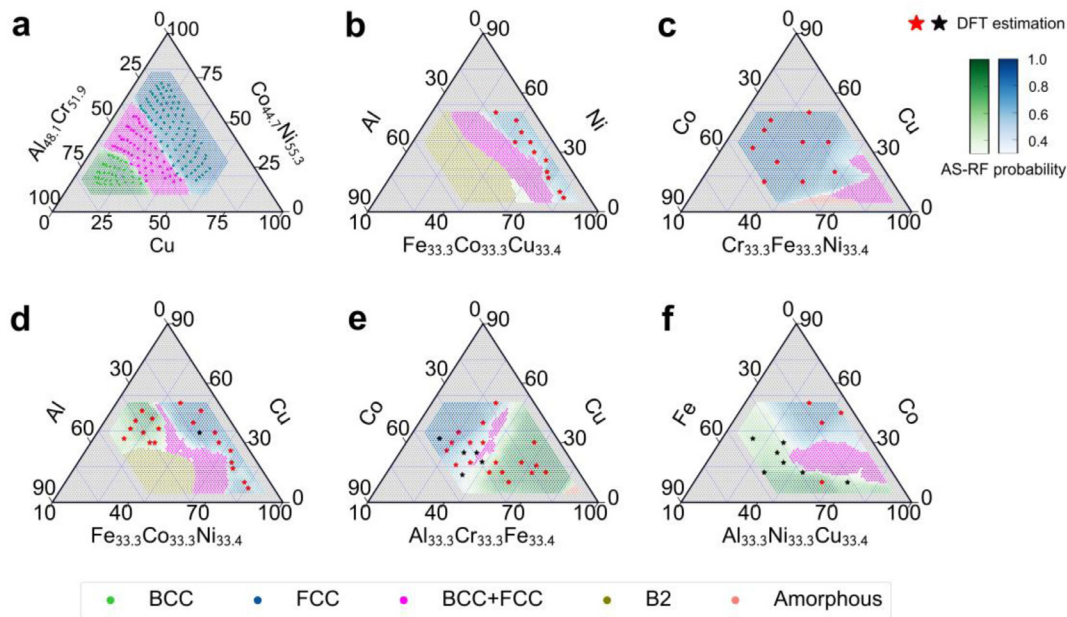
Fig. 3. Phase diagrams of quinary HEAs with Al, Cu, Fe, Mn, Cr, Co, and Ni constituent elements. The atomic percentage unit is adopted at the axes of phase diagrams. The colored dots in (a) represent the phases of HEAs that experimentally fabricated, and in (b–f) the pentagon represent DFT estimations of the HEA phase, with the red ones denoting the consistent results between AS-RF prediction and DFT estimation while the (single) black one representing the inconsistent observation. The colored background in (a–f) represents predicted phases, with green, dark-blue, pink, yellow, and salmon denoting BCC, FCC, BCC+FCC, B2, and amorphous phase, respectively. The prediction probabilities for BCC and FCC phases are represented with heatmaps in the diagrams. The whole 224 phase diagrams are in Fig. S8 in Supplemental Materials.

the combination between Al, Cr and Fe element plays the most important role in the formation of BCC structure, whereas the effect of Cu-Ni-Co correlation dominates the formation of FCC structure.

The different characteristics of the BCC and FCC networks indicate different roles of the participation elements. In SPSS phases, Cr-Fe (with BCC type lattice order) and Ni-Co (with FCC type lattice order) serve as the basic components by introducing the orders of the lattices into the base materials. Al and Cu play roles to stabilize the lattice symmetries (BCC or FCC type SPSSs), and the competition of the two results in the FCC+BCC type solid solution. Further, based on the first three networks, to explore more HEAs with solid solution phases, it is essential to replace Mn element with Al or Cu in Cantor-related as-sputtered alloys. This observation is consistent with previous experimental studies [62,63], especially since the replacement of Mn with Al has become the main branch for the exploration of SPSS HEAs. In this study, with the aid of ML model, we take a step forward towards the rational design of SPSS HEAs, by demonstrating the strong composition-dependent phase formations (Al-Cr-Fe for BCC type SPSS and Cu-Ni-Co for FCC type SPSS). In addition, Al element also plays a critical role in the formation of the intermetallic B2 phase and its pair correlations with Cu, Ni, Co, and Mn drive the as-sputter HEA to form the intermetallic phase with B2 structures. It should be noted that in this study the amorphous phase has been defined in a strict manner (SS + amorphous treated as amorphous phase), which may result in larger amorphous regions in the predicted phase diagrams compared with experimental results.

3.4. Validation of the AS-RF model

While the ability to derive experimental trends from the AS-RF model is a promising sign that the predictions have value in predicting new experimental alloys, further analysis is needed to be certain of the model utility in extrapolating new HEA compositions from the reported experimental data. Thus, to provide an independent test of the model performance, we selected five quinary HEA compositions from the 210 proposed phase diagrams

$(Al_xNi_y[(FeCoCu)_{0.33}]_z$, $Co_xCu_y[(CrFeNi)_{0.33}]_z$, $Al_xCu_y[(FeCoNi)_{0.33}]_z$, $Co_xCu_y[(AlCrFe)_{0.33}]_z$, and $Fe_xCo_y[(AlNiCu)_{0.33}]_z)$ and conducted comprehensive density functional theory (DFT) calculations to verify the accuracy of our machine learning predictions, as shown in Fig. 3(b–f).

For each of the selected HEA compositions, we compared the phase predictions made by the AS-RF classifier to the calculated DFT lattice energies for BCC and FCC SPSSs. Only BCC and FCC energies were considered as these are the phases of greatest experimental interest, and because the DFT methodology for such calculations are well established and can be conducted in relatively high throughput, as compared to amorphous, B2, and BCC+FCC calculations which, while they can be conducted, require significant manual intervention and refinement to achieve meaningful results. Of the five selected phase systems, three have clear FCC/BCC phase transitions $(Al_xCu_y[(FeCoNi)_{0.33}]_z$, $Co_xCu_y[(AlCrFe)_{0.33}]_z$, and $Fe_xCo_y[(AlNiCu)_{0.33}]_z)$, while two $(Al_xNi_y[(FeCoCu)_{0.33}]_z$, $Co_xCu_y[(CrFeNi)_{0.33}]_z,)$ have a single FCC phase. For each phase diagram, alloys were selected to span the composition space predicted by the AS-RF to be FCC or BCC, and DFT was used to calculate the energy difference between the BCC and FCC structures of each alloy.

Of the 75 alloys considered, the AS-RF and DFT methodologies were in agreement in 62 cases (82.6%), as shown in Fig. 5(a). Applying McNemar's test statistic on the confusion matrix generated from the two predictors gives a p-value of 0.57 indicating no statistically significant bias between the two models. Said differently, the deviations of the AS-RF model from the DFT appear to be randomly distributed, rather than systematically biased. We then constructed ROCs in Fig. 5(c) and (d) for both the prediction probability of BCC (P_{BCC}) and FCC (P_{FCC}) outputs of the AS-RF model based on the agreement with the DFT model. The AUCs of the two curves, respectively, are 0.886 and 0.848. These AUCs can be converted to Mann-Whitney-Wilcoxon U statistics and subsequently z-scores based on Eqs. (4) and (5), yielding z-scores of 5.0 and 4.5. These scores correspond to p-values of <0.001, indicating the rankings to be significantly different from random sampling. We also ap-
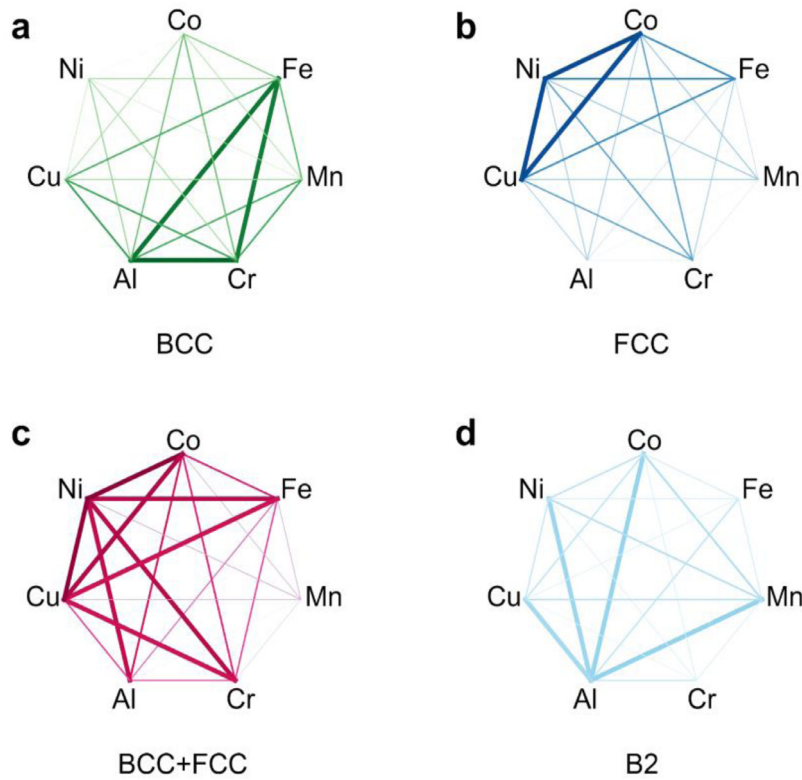
Fig. 4. The relationships between element correlations and phase formations of BCC type SPSS (a), FCC type SPSS (b), BCC+FCC type mixing solid solution phase (c) and B2 type intermetallic phase (d). The weights of the lines in the networks represents different correlation strengths between elements. The larger the weight of the line the stronger the correlations between elements would be.
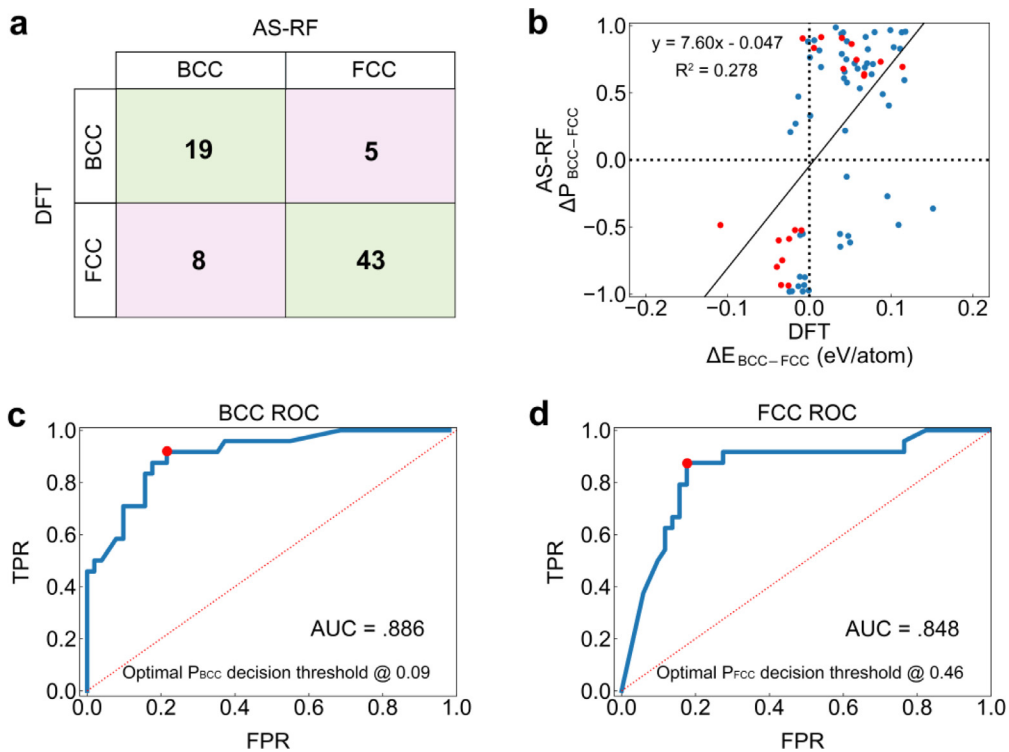

Fig. 5. (a) The confusion matrix between DFT and AS-RF model predictions. (b) The relationship between the energy difference $\Delta E_{BCC\text{-}FCC}$ between BCC and FCC phases as predicted by DFT and the relative certainty ($\Delta P_{BCC\text{-}FCC} = P_{BCC} - P_{FCC}$) of the AS-RF model, with $P_{FCC/BCC}$ the AS-RF prediction probability for FCC/BCC phase. The data points are all the systems within DFT validation, and the case of the Fe-Co-Ni-Cu-Al system are denoted with red. (c) and (d) are the receiver operating characteristics (ROCs) for BCC and FCC phases, respectively, with the x-axis the false positive ratio (FPR) and the y-axis the true positive ratio (TPR).

Table 1
The comparisons of phase structures between AS-RF model predictions and experimental measurements.

Composition	Fabrication method	Characterization	Experimental phase structure	Refs.	AS-RF prediction
AlCoCrFeNi	Radio frequency	XRD/TEM	FCC+BCC+B2	[67]	FCC+BCC
AlCoCrCu$_{0.5}$FeNi	Radio frequency	XRD	FCC+BCC	[68]	FCC+BCC
CoCrFeMnNiV$_{0.3}$	Radio frequency	XRD/TEM	FCC+Amorphous	[69]	Amorphous
CoCrFeMnNiV$_{0.7}$	Radio frequency	XRD/TEM	FCC+Amorphous	[69]	Amorphous
CoCrFeMnNiV$_{1.1}$	Radio frequency	XRD/TEM	FCC+Amorphous	[69]	Amorphous
CoCrFeNiCu	Direct current	XRD	FCC	[70]	FCC
Al$_{2.5}$CoCrFeNiCu	Direct current	XRD	BCC	[70]	B2

plied Youden's J statistic to each ROC, and identified optimal decision thresholds of 0.09 for BCC and 0.46 for FCC in this binary system. It should be noted that there are relatively large discrepancies between results from the AS-RF predictions and DFT calculations (Fig. 3(e) and (f)). To clearly understand the discrepancy in these regions, we compared the distributions of the training dataset with that of the "wrong" predicted samples in the descriptor-space, and found that most of the wrongly predicted HEAs are located in the "mixing" regions, where both BCC and FCC phase structures appear in these regions (see Fig. S9 in Supplemental Materials for more details). It indicates that the AS-RF model is not well-trained due to the complex phase distributions of the training dataset in these regions, for which activation learning may be performed by collaborating with an experimentalist.

We then considered the relationship between the energy difference of the two lattices as predicted by DFT and the relative certainty of the AS-RF model, as shown in Fig. 5(b). While some correlation exists between these two figures, the R^2 coefficient is small at 0.278. Considering specifically the case of the Fe$_x$Co$_y$[(AlNiCu)$_{0.33}$]$_z$ phase system (the red dots in Fig. 5(b)), it is obvious why the correlation is poor: the AS-RF model yields a sharp jump in the probability of FCC at very close to the point at which DFT predicts no energy difference between the two phases. In fact, as the $\Delta E_{BCC\text{-}FCC}$ increases in magnitude, the probabilities of either phase decreases, as the predicted probability of the increase of the additional phase.

Given the poor relationship between DFT $\Delta E_{BCC\text{-}FCC}$ and $\Delta P_{BCC\text{-}FCC}$, we sought an alternate evaluation of the statistical value of an increase in $P_{FCC/BCC}$ provided by the AS-RF model. To do this we considered the relationship between accuracy and $\Delta P_{BCC\text{-}FCC}$ by introducing a minimum probability for classification (see Fig. S10 in Supplemental Materials for more details). The trend is strongly positive, with a high R^2 value of 0.846, indicating a strong correlation between the magnitude of $P_{FCC/BCC}$ with correct identification of the DFT phase. This is further illustrated by consideration of the location within the phase diagrams of the alloys for which the AS-RF model and DFT energies are not in agreement. These points occur primarily near phase boundaries, suggesting that the best application for the AS-RF model in the search for new high entropy alloys for subsequent experimental analysis is the identification of high probability regions in the middle of a phase region in a given quinary diagram.

In addition to the comparison with DFT calculations, we also applied the AS-RF model to experimentally well-studied as-sputtered HEAs not included in the training dataset. As displayed in Table 1, despite different fabrication processes, our AS-RF model predictions are overall in excellent agreement with the measured phase structures, indicating excellent transferability of the model. The one inconsistent result is for Al$_{2.5}$CoCrFeNiCu, for which B2 type intermetallic phase has been predicted with the AS-RF model while it was experimentally determined to be BCC type SPSS phase by XRD. However, since BCC type SPSS and B2 phase share the same lattice structure, we argue that XRD cannot distinguish these two phases. Moreover, both experimental observations and theoretical calculations have demonstrated that higher Al con-

tent leads to AlNi [64] B2 type intermetallic phase precipitation, which is consistent with the predictions of the AS-RF model. It should be noted that the as-sputtered HEAs have different thermodynamic conditions from that of as-cast HEAs. Since the AS-RF model is trained with as-sputtered alloys, the AS-RF model is expected to provide reliable predictions of the phase formations of as-sputtered HEAs. Since there is a lack of information in the training dataset for the constituent elements far away from the seven elements in this study, the reliable prediction ability of the AS-RF model might be limited for the HEAs with components within or nearby these seven elements.

3.5. Interpreting rules for SPSS-HEA phase formation from the AS-RF model

To shed light on the nature of phase formation in as-sputtered HEAs, we analyzed the weighting of the various features within the AS-RF model. First, individual feature importance was estimated (Fig. 6) for the different phases showing VEC, FBI and W to have dominant roles in the formation of both FCC and BCC SPSS. Further, we found the feature importance remains invariable by applying GBDT machine learning model, indicating the reliable of our model on the estimation of the feature importance (see Fig. S11 in Supplemental Materials). The significant effect of VEC, *i.e.*, larger (smaller) VEC favoring FCC (BCC) lattice structure, is widely accepted by the community in the formations of solid solution HEAs [28]. The importance of elementary structural symmetry (FBI) can be simply understood as the extension of Hume-Rothery rules, though this effect is strongly coupled with other features, as we will further discuss below. The effect of W mainly differentiates the formation of solid solution phases from B2 and amorphous phases.

Interestingly, the formation of amorphous phase in as-sputtered HEAs is almost only determined by the $\Delta \chi$ factor (Fig. 6(e)). The experimental dataset (see Fig. S1 in Supplemental Materials) shows the formation of amorphous phases largely originates from Mn inclusion. In our high-throughput predictions, we find that 93% of the amorphous alloys contain Mn (see Fig. S12 in Supplemental Materials). The substitution of Mn with Al increases the value of δ but substantially reduces the effect of $\Delta \chi$. By replacing Mn with Al constituent, the amorphous phase can be largely suppressed (see Supplemental Materials for details).

Next, we provide quantitative estimates for the effects of the driving forces on the formation of HEA phases. Since direct interpretation of our AS-RF model is difficult due to the complexity of the ensemble model with hundreds of deep trees, dimensional reduction is used to project the high dimensional space of the AS-RF model by mapping from the subset of the feature space to the phase space, with the aid of the Forest Floor package [65]. To reveal the formations of HEA with solid solution phases, in Fig. 7 the interpretations of our AS-RF model are visualized by mapping physical features VEC, FBI and W to BCC SPSS, FCC SPSS and BCC+FCC mixing phase solid solution, respectively. In these geometric representations of the AS-RF model, the training nodes distributions are colored by the values of the projected features. The
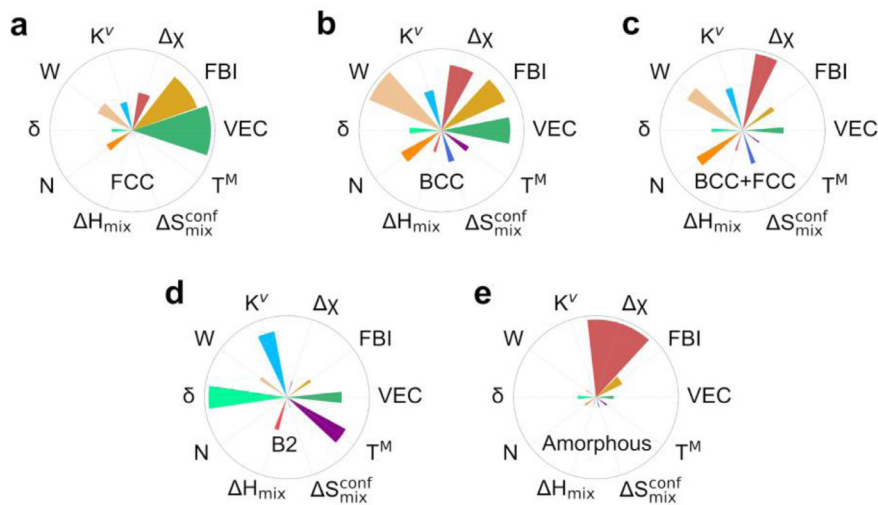
Fig. 6. Relative feature importance for the phase formation of BCC type SPSS (a), FCC type SPSS (b), BCC+FCC type mixing solid solution phase (c), B2 type intermetallic phase (d), amorphous phase (e). The sizes of the contributions are shown in each bar chart with their relevant values.
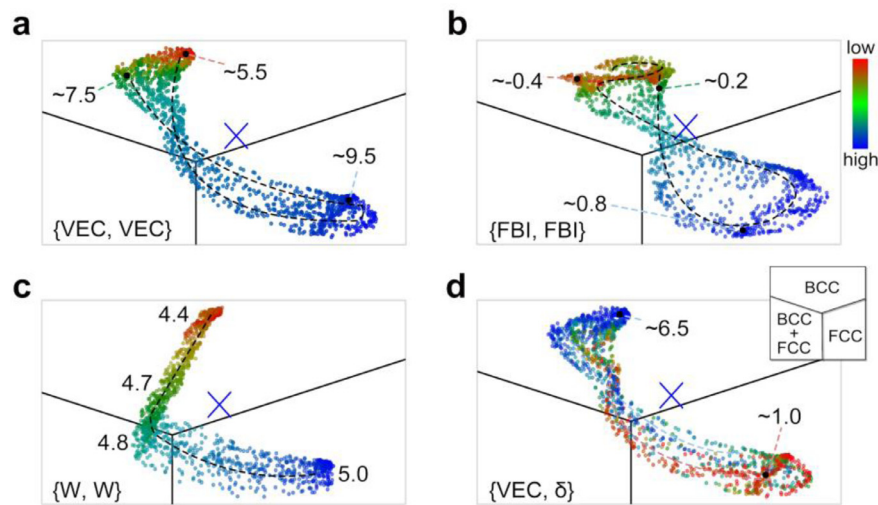

Fig. 7. The explicit representations of the mapping from physical features to FCC-BCC-FCC+BCC phases. The interpretations of the AS-RF model are represented by the node distributions of the ensemble trees in (a–d). The crossing points are "root node", obtained with bootstrap increment (representing random decision making of the forest), and the dots denote "training nodes". $\{X, Y\}$ represents the distribution of nodes projected on X attribute and colored by the value of Y attribute, for which $X=Y$ denotes the main effect of the feature X and $X \neq Y$ represents the second order effect between X and Y.

further the training node is from the crossing point (the point without training), the better the prediction ability.

Fig. 7(a–c) shows the main effects of VEC, FBI and W on the phase formations. As shown in Fig. 7(a), the feature distribution of VEC indicates that the alloys with VEC values of 5.5 or 7.5 possess the largest probability to form BCC type SPSS whereas the ones with VEC values of 9.5 favor FCC type HEAs. This is consistent with previous experimental observations [23] that FCC type HEAs form with VEC > 8.0 while the alloys with VEC < 7.0 favor to BCC type SPSS. On the other hand, Fig. 7(b) also displays a clear trend of the effect of the elementary symmetry (FBI): a large amount of FBI values locates at ∼0.8 (more FCC characteristics) towards FCC phase and in the range from -0.4 to 0.2 (more BCC characteristics) pointing to BCC phase. Fig. 7(c) provides more insight into the effect of work function: with gradually decreasing the W values from 4.8 to 4.4 eV, alloys tend to form BCC structure, but its effect on the formation of FCC structure is unclear, due to the much loosely distributed W values in the regime of FCC phase.

Beside the main effects of the individual features, the visualization of the second-order effects can provide a deeper understand-

ing of the phase formations by mapping the value of attribute Y on the feature contributions of attribute X. As displayed in Fig. 7(d), VEC and δ show a negative correlation on the formation of SPSS HEAs: alloys with smaller VEC (∼5.5) and larger δ (∼6.5%) favoring BCC structure. This can guide the design of the HEA phase structure. For example, the Al element possesses both small VEC value and large atomic radius. Controlling the concentration of Al composition in HEA can obtain BCC type or FCC+BCC type solid solutions. Indeed, experimentally, the Al element has been frequently introduced into Co-Cr-Fe-Ni base alloys (FCC structure) [62,66] to precipitate BCC phase. Other feature correlations, such as {FBI, δ} and {VEC, W}, are more complex (see Fig. S13 in Supplemental Materials for details), indicating higher orders of effects between these factors.

Finally, to construct a full contribution of the features on the phase formations, we applied principal component analysis (PCA) for classifying the predicted HEAs data. PCA disentangles the phases distributions by generating a new orthogonal basis set by diagonalizing the previous highly coupled feature vectors, with the first eigenvector PC1 describing the direction of the most vari-
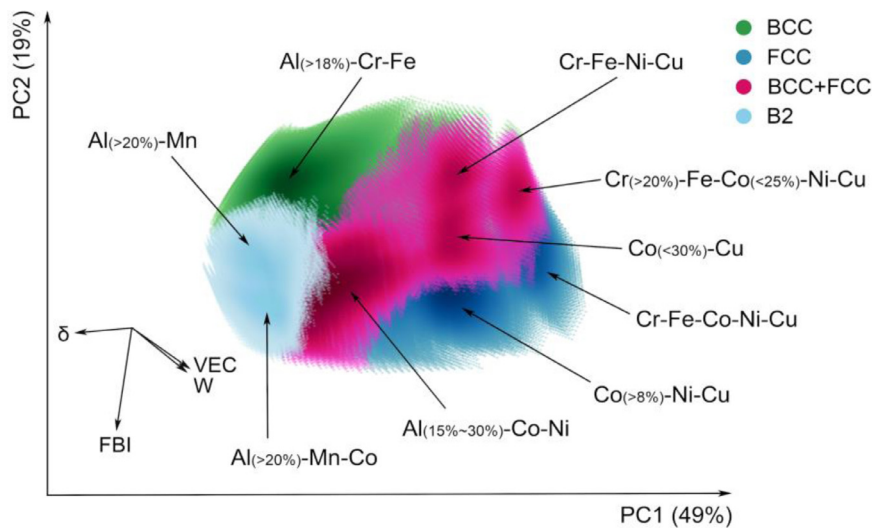
Fig. 8. The selection of SPSS phases in as-sputtered HEAs with the aid of principle component analysis. The colored dots represent the predicted HEAs with different phase structures. They are well-separated on the two-dimensional projected space formed by the two main principal vectors, PC1 and PC2. Several most common types of chemical constitutions for different phases are denoted in the figure. The insert vectors show four features that dominate the distributions of the as-sputtered HEA samples.

ance in the data set, and the second eigenvector PC2 for the second most variance, etc. As shown in Fig. 8, the phase distributions are well-disentangled on a two-dimensional subspace, formed by the first two PC basic vectors. Both PC vectors can be expressed as the linear combination of the features. In particular, PC1 $(49\%) = (0.38VEC + 0.36W + 0.36T^M) - (0.41\delta + 0.32\Delta\chi + 0.1FBI)$ and PC2 $(19\%) \approx -0.66FBI$. It can be found that to separate the phase formations between FCC and BCC types of SPSSs, the features VEC, W, FBI and δ play roles. Specifically, the effect of δ impacts on the formations of SPSSs is opposite to that of W and VEC. The elementary symmetry (FBI) weakly correlates with the other three factors. For the phase distributions in Fig. 8, some high-density regions can be found, where the most chemical compositions have been denoted. For BCC type SPSS, there is one main clustering region contributed from Al-Cr-Fe composition, with high Al concentrations ($> 18\%$), while for FCC type SPSS, two main distinct clustering regions corresponding to Co-Ni-Cu and Cr-Fe-Co-Ni-Cu appear. These may guide the design of quinary as-sputtered HEAs with SPSS phases in the Cr-Co-Fe-Mn-Ni-Cu-Al chemical space.

4. Conclusion

By applying the random forest machine learning model to high throughput experimental HEA phase data, we have developed a high accuracy model (94.6%) for the prediction of single-phase solid solutions (with FCC and BCC structures) in quinary HEAs. We further used first-principles DFT calculations to demonstrate the accuracy (82%) of the AS-RF model in previously unknown HEAs, and show a strong trend between the magnitude of AS-RF predicted probability and agreement with the DFT results. We also predicted a comprehensive set of 224 quinary HEAs phase diagrams (210 new) in the chemical space containing Cr, Co, Fe, Ni, Mn, Al, and Cu elements. It is found that the combination of Al-Cr-Fe favors to form BCC type SPSS, whereas the effect of Cu-Ni-Co correlation dominates the formation of FCC type SPSS. The interpretability of our machine learning model also sheds light on the driving forces behind FCC and BCC SPSSs in as-sputtered HEAs, with the main contributors being valance electron concentration, work function, atomic radius difference and elementary symmetries. This machine learning based HEA discovery framework may

provide a valuable tool to accelerate the rational design and exploration of HEAs.

Code availability

The source code of the Random forest model can be downloaded from https://github.com/NJUST-WEILIU-s-group/HEA.

Data availability

Datasets

We applied the dataset of as-sputtered quinary HEAs from the experimental work of Ref. [46], where the high-throughput experiments were performed by the combinatorial magnetron co-sputtering techniques. The public experimental data have been obtained from http://materialsatlasproject.org.

Declaration of Competing Interest

The authors declare no competing financial interests.

CRediT authorship contribution statement

Ji-Chang Ren: Visualization, Investigation, Software, Formal analysis, Writing – original draft. **Junjun Zhou:** Software, Formal analysis, Writing – original draft. **Christopher J. Butch:** Software, Formal analysis, Writing – original draft, Software, Formal analysis. **Zhigang Ding:** Software, Formal analysis, Writing – original draft. **Shuang Li:** Software, Formal analysis, Writing – original draft. **Yonghao Zhao:** Software, Formal analysis, Writing – original draft. **Wei Liu:** Visualization, Investigation, Software, Formal analysis, Writing – original draft.

Acknowledgements

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jmst.2022.07.059.

References

[1] B. Cantor, I.T.H. Chang, P. Knight, A.J.B. Vincent, Mater. Sci. Eng. A 375-377 (2004) 213–218.
[2] J.W. Yeh, S.K. Chen, S.J. Lin, J.Y. Gan, T.S. Chin, T.T. Shun, C.H. Tsau, S.Y. Chang, Adv. Eng. Mater. 6 (2004) 299–303.
[3] O. El-Atwani, N. Li, M. Li, A. Devaraj, J.K.S. Baldwin, M.M. Schneider, D. Sobieraj, J.S. Wróbel, D. Nguyen-Manh, S.A. Maloy, E. Martinez, Sci. Adv. 5 (2019) eaav2002.
[4] W.Y. Ching, S. San, J. Brechtl, R. Sakidja, M. Zhang, P.K. Liaw, NPJ Comput. Mater. 6 (2020) 45.
[5] T. Hori, T. Nagase, M. Todai, A. Matsugaki, T. Nakano, Scr. Mater. 172 (2019) 83–87.
[6] C. Chen, H. Zhang, Y. Fan, W. Zhang, R. Wei, T. Wang, T. Zhang, F. Li, J. Magn. Magn. Mater. 502 (2020) 166513.
[7] S. Ranganathan, Curr. Sci. 85 (2003) 1404–1406.
[8] B. Gludovatz, A. Hohenwarter, D. Catoor, E.H. Chang, E.P. George, R.O. Ritchie, Science 345 (2014) 1153–1158.
[9] A.J. Zaddach, R.O. Scattergood, C.C. Koch, Mater. Sci. Eng. A 636 (2015) 373–378.
[10] L. Liu, J.B. Zhu, C. Zhang, J.C. Li, Q. Jiang, Mater. Sci. Eng. A 548 (2012) 64–68.
[11] Y. Lu, X. Gao, L. Jiang, Z. Chen, T. Wang, J. Jie, H. Kang, Y. Zhang, S. Guo, H. Ruan, Y. Zhao, Z. Cao, T. Li, Acta Mater. 124 (2017) 143–150.
[12] X. Gao, Y. Lu, B. Zhang, N. Liang, G. Wu, G. Sha, J. Liu, Y. Zhao, Acta Mater. 141 (2017) 59–66.
[13] Q. Ding, Y. Zhang, X. Chen, X. Fu, D. Chen, S. Chen, L. Gu, F. Wei, H. Bei, Y. Gao, M. Wen, J. Li, Z. Zhang, T. Zhu, R.O. Ritchie, Q. Yu, Nature 574 (2019) 223–227.
[14] S.W. Wu, G. Wang, Q. Wang, Y.D. Jia, J. Yi, Q.J. Zhai, J.B. Liu, B.A. Sun, H.J. Chu, J. Shen, P.K. Liaw, C.T. Liu, T.Y. Zhang, Acta Mater. 165 (2019) 444–458.
[15] Z. Zhang, M.M. Mao, J. Wang, B. Gludovatz, Z. Zhang, S.X. Mao, E.P. George, Q. Yu, R.O. Ritchie, Nat. Commun. 6 (2015) 10143.
[16] H. Zhang, Y. He, Y. Pan, Scr. Mater. 69 (2013) 342–345.
[17] B. Schuh, F. Mendez-Martin, B. Völker, E.P. George, H. Clemens, R. Pippan, A. Hohenwarter, Acta Mater. 96 (2015) 258–268.
[18] O.N. Senkov, G.B. Wilks, J.M. Scott, D.B. Miracle, Intermetallics 19 (2011) 698–706.
[19] D.B. Miracle, O.N. Senkov, Acta Mater. 122 (2017) 448–511.
[20] S. Guo, Q. Hu, C. Ng, C.T. Liu, Intermetallics 41 (2013) 96–103.
[21] J.W. Yeh, JOM 65 (2013) 1759–1771.
[22] Y.F. Ye, Q. Wang, J. Lu, C.T. Liu, Y. Yang, Mater. Today 19 (2016) 349–362.
[23] S. Guo, Mater. Sci. Technol. 31 (2015) 1223–1230.
[24] A. Takeuchi, K. Amiya, T. Wada, K. Yubuta, W. Zhang, JOM 66 (2014) 1984–1992.
[25] B.S. Murty, J.W. Yeh, S. Ranganathan, P.P. Bhattacharjee, High-Entropy Alloys, Elsevier Science, 2019.
[26] S. Guo, C.T. Liu, Prog. Nat. Sci. Mater. Int. 21 (2011) 433–446.
[27] S. Fang, X. Xiao, L. Xia, W. Li, Y. Dong, J. Non-Cryst. Solids 321 (2003) 120–125.
[28] S. Guo, C. Ng, J. Lu, C.T. Liu, J. Appl. Phys. 109 (2011) 103505.
[29] X. Yang, Y. Zhang, Mater. Chem. Phys. 132 (2012) 233–238.
[30] Y. Zhang, Z.P. Lu, S.G. Ma, P.K. Liaw, Z. Tang, Y.Q. Cheng, M.C. Gao, MRS Commun. 4 (2014) 57–62.
[31] Z. Wang, Y. Huang, Y. Yang, J. Wang, C.T. Liu, Scr. Mater. 94 (2015) 28–31.
[32] H. Oh, D. Ma, G. Leyson, B. Grabowski, E. Park, F. Körmann, D. Raabe, Entropy 18 (2016) 321.
[33] Y. Wang, M. Yan, Q. Zhu, W.Y. Wang, Y. Wu, X. Hui, R. Otis, S.L. Shang, Z.K. Liu, L.Q. Chen, Acta Mater. 143 (2018) 88–101.
[34] D. Ma, B. Grabowski, F. Körmann, J. Neugebauer, D. Raabe, Acta Mater. 100 (2015) 90–97.
[35] N. Islam, W. Huang, H.L. Zhuang, Comput. Mater. Sci. 150 (2018) 230–235.
[36] C. Wen, Y. Zhang, C. Wang, D. Xue, Y. Bai, S. Antonov, L. Dai, T. Lookman, Y. Su, Acta Mater. 170 (2019) 109–117.
[37] W. Huang, P. Martin, H.L. Zhuang, Acta Mater. 169 (2019) 225–236.
[38] Y. Li, W. Guo, Phys. Rev. Mater. 3 (2019) 095005.
[39] Z. Zhou, Y. Zhou, Q. He, Z. Ding, F. Li, Y. Yang, npj Comput. Mater. 5 (2019) 128.
[40] K. Kaufmann, D. Maryanovsky, W.M. Mellor, C. Zhu, A.S. Rosengarten, T.J. Harrington, C. Oses, C. Toher, S. Curtarolo, K.S. Vecchio, NPJ Comput. Mater. 6 (2020) 42.
[41] S. Feng, H. Fu, H. Zhou, Y. Wu, Z. Lu, H. Dong, npj Comput. Mater. 7 (2021) 10.
[42] Y. Sun, Z. Lu, X. Liu, Q. Du, H. Xie, J. Lv, R. Song, Y. Wu, H. Wang, S. Jiang, Z. Lu, Appl. Phys. Lett. 119 (2021) 201905.
[43] A.R. Barron, Mach. Learn. 14 (1994) 115–133.
[44] O.N. Senkov, D.B. Miracle, K.J. Chaput, J.P. Couzinie, J. Mater. Res. 33 (2018) 3092–3128.
[45] J.H. Li, M.H. Tsai, Scr. Mater. 188 (2020) 80–87.
[46] S.A. Kube, S. Sohn, D. Uhl, A. Datye, A. Mehta, J. Schroers, Acta Mater. 166 (2019) 677–686.
[47] J.R. Quinlan, Mach. Learn. 1 (1986) 81–106.
[48] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and regression trees, 1st ed., Chapman and Hall/CRC, New York/Boca Raton, FL, 1984.
[49] M. Robnik-Šikonja, Improving random forests, Springer Berlin Heidelberg, 2004.
[50] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[51] G. Kresse, J. Furthmüller, Phys. Rev. B 54 (1996) 11169–11186.
[52] A. Zunger, S.H. Wei, L.G. Ferreira, J.E. Bernard, Phys. Rev. Lett. 65 (1990) 353–356.
[53] A. van de Walle, M. Asta, G. Ceder, Calphad 26 (2002) 539–553.
[54] J.P. Perdew, K. Burke, M. Ernzerhof, Phys. Rev. Lett. 77 (1996) 3865–3868.
[55] Breiman L. , Out-of-bag Estimation, https://www.stat.berkeley.edu/~breiman/OOBestimation.pdf. (10 August 2016), 1996b.
[56] K.A. Spackman, in: Proceedings of the sixth International Workshop on Machine Learning, 1989, pp. 160–163.
[57] D.J. Hand, P. Smyth, H. Mannila, Principles of Data Mining, MIT Press, 2001.
[58] J. Friedman, Ann. Stat. 29 (2001) 1189–1232.
[59] R. Short, K. Fukunaga, IEEE Trans. Inf. Theory 27 (1981) 622–627.
[60] M.A. Hearst, S.T. Dumais, E. Osuna, J. Platt, B. Scholkopf, IEEE Intell. Syst. App. 13 (1998) 18–28.
[61] Y. Freund, R.E. Schapire, European Conference on Computational Learning Theory, Springer, 1995.
[62] W.R. Wang, W.L. Wang, S.C. Wang, Y.C. Tsai, C.H. Lai, J.W. Yeh, Intermetallics 26 (2012) 44–51.
[63] N. Li, C.L. Jia, Z.W. Wang, L.H. Wu, D.R. Ni, Z.K. Li, H.M. Fu, P. Xue, B.L. Xiao, Z.Y. Ma, Y. Shao, Y.L. Chang, Acta Metall. Sin. (Engl. Lett.) 33 (2020) 947–956.
[64] Y.P. Wang, B.S. Li, H.Z. Fu, Adv. Eng. Mater. 11 (2009) 641–644.
[65] Welling S.H., Refsgaard H.H.F., Brockhoff P.B., Clemmensen L.H., arXiv [Internet] (2016). Available from: http://arxiv.org/abs/1605.09196.
[66] H.P. Chou, Y.S. Chang, S.K. Chen, J.W. Yeh, Mater. Sci. Eng. B 163 (2009) 184–189.
[67] C. Wang, X. Li, Z. Li, Q. Wang, Y. Zheng, Y. Ma, L. Bi, Y. Zhang, X. Yuan, X. Zhang, C. Dong, P.K. Liaw, Thin Solid Films 700 (2020) 137895.
[68] N.A. Khan, B. Akhavan, H. Zhou, L. Chang, Y. Wang, L. Sun, M.M. Bilek, Z. Liu, Appl. Surf. Sci. 495 (2019) 143560.
[69] S. Fang, C. Wang, C.-L. Li, J.H. Luan, Z.B. Jiao, C.T. Liu, C.H. Hsueh, J. Alloys Compd. 820 (2020) 153388.
[70] Y. Ma, Y.H. Feng, T.T. Debela, G.J. Peng, T.H. Zhang, Int. J. Refract. Met. Hard Mater. 54 (2016) 395–400.