Acta Materialia 185 (2020) 528–539

Contents lists available at ScienceDirect

Acta Materialia

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

Phase prediction in high entropy alloys with a rational selection of materials descriptors and machine learning models

Yan Zhang[a,b], Cheng Wen[a,b], Changxin Wang[a,b], Stoichko Antonov[a,c], Dezhen Xue[d,*], Yang Bai[a,b], Yanjing Su[a,b,*]

[a] Beijing Advanced Innovation Center for Materials Genome Engineering, University of Science and Technology Beijing, Beijing, 100083, China
[b] Corrosion and Protection Center, University of Science and Technology Beijing, Beijing, 100083, China
[c] State Key Laboratory for Advanced Metals and Materials, University of Science and Technology Beijing, Beijing, 100083, China
[d] State Key Laboratory for Mechanical Behavior of Materials, Xi'an Jiaotong University, Xi'an 710049, China

ARTICLE INFO

Article history:
Received 7 July 2019
Revised 19 November 2019
Accepted 28 November 2019
Available online 4 December 2019

Keywords:
High entropy alloys
Machine learning
Genetic algorithm
Active learning
Materials informatics

ABSTRACT

Materials informatics employs machine learning (ML) models to map the relationship between a targeted property and various materials descriptors, providing new avenues to accelerate the discovery of new materials. However, the possible ML models and materials descriptors are numerous, and a rational recipe to rapidly choose the best combination of the two is needed. In the present study, we propose a systematic framework that utilizes a genetic algorithm (GA) to efficiently select the ML model and materials descriptors from a huge number of alternatives and demonstrated its efficiency on two phase formation problems in high entropy alloys (HEAs). The optimized classification model allows an accuracy for identifying solid-solution and non-solid-solution HEAs to be up to 88.7% and further for recognizing body-centered-cubic (BCC), face-centered-cubic (FCC), and dual-phase HEAs to reach 91.3%. Furthermore, by employing an active learning approach, several HEAs with largest classification uncertainties were selected, experimentally synthesized and phase-identified, and augmented to the initial dataset to iteratively improve the ML model. The method serves as a general algorithm to select materials descriptors and ML models for various material problems including classification and optimization of targeted properties.

1. Introduction

Developing new materials in a faster and low-cost fashion is the scope of the materials genome engineering (MGE) and hence many reports have been devoted to exploring high-throughput experiments, high-throughput calculations and material database development [1]. Materials informatics manages and learns from the materials data, extracts the underlying physics, predicts properties of new materials and guides the next experiments or calculations [2,3]. This field has developed from the need for accelerating materials discovery. Examples include the predictions of crystal structures [4], physical properties [5], and thermodynamic stabilities using machine learning (ML) models [6], and the design of new experiments and calculations to efficiently search for global optima in vast search spaces using optimization algorithms [7,8].

Usually, the ML approach builds a surrogate model (f) to infer the interrelation between the targeted property (Y) and a set of materials descriptors (X), such that $Y = f(X)$. The materials descriptors (X) together with the model (f) determine the robustness of the ML prediction, which is crucial to the subsequent design of new materials. Therefore, for a given problem, an optimized combination of f and X should be identified.

The number of potential combinations is determined by the number of ML models (p), and the number of materials descriptors (n), which is given by $p \times (2^n - 1)$. Many ML models have been developed in various fields [9], including kernel-based ones, tree-based ones, neural networks and so on. These models can be utilized to solve the problems in materials science, giving rise to a large number of p readily available. Meanwhile, material properties are affected by various factors such as composition, microstructure, processing, and/or other thermodynamic and kinetic parameters. Thus different types of materials descriptors should be introduced, offering a large pool of n [10,11]. The large number of available p and n gives an enormously large combination space. For example, 5369 million different combinations can be reached for $p = 5$ and $n = 30$, which are typical values in a materials problem.

* Corresponding author.
E-mail addresses: xuedezhen@xjtu.edu.cn (D. Xue), yjsu@ustb.edu.cn (Y. Su).

https://doi.org/10.1016/j.actamat.2019.11.067
1359-6454/

A central question is then how to select the best combination of materials descriptors (X) and ML model (f) from this large space.

The performance of a ML model is usaually estimated by the test error *via* resampling methods such as cross-validation and bootstrapping [12]. An exhaustive search for all possibilities ensures a global optimization, but is always impossible due to an explosion of the required computational resource. Many reports have therefore been devoted to narrow down the search space of f and X.

Domain knowledge comes naturally to down-select the materials descriptors [13,14]. For example, it is known that the voltage and current measurements during a charge cycle are indicative of the capacity of a Li-ion battery cell [15]. Thus features from such a charge curve can be extracted, which include charge voltages (the initial and final voltages before and after charging), charge current (the current after charging) and charge capacities (the charge stored in the battery during the constant current charge step and the constant voltage charge step, respectively). These features can then be employed in accurately estimating the capacity of the battery throughout its life-time [15]. In many cases, domain knowledge is insufficient, and various algorithms have been adopted to rank materials descriptors according to their relevance and importance to the material property, such as the least absolute shrinkage and selection operator (LASSO), gradient tree boosting (GTB), random forest (RF), *etc*. For example, Ghiringhelli et al. [10] utilized LASSO to rank possible 10 000 materials descriptors and identified 3 materials descriptors to predict the crystal structure of octet binaries semiconductors. To work on highly correlated materials descriptors spaces, Ouyang et al. [11] proposed the sure independence screening and sparsifying operator (SISSO) to efficiently extract materials descriptors, which could overcome the limitation of LASSO. But those materials descriptors were valid only for a particular ML model (f), and if the f varies, its suitable materials descriptors can be different. The so-called wrapper methods have been employed to determine the optimal feature subsets, including the sequential forward selection (SFS), sequential backward selection (SBS), and so on [16]. However, they are greedy and can possibly ignore certain combinations of materials descriptors. Dimensionality reduction algorithms such as principal component analysis (PCA) can compress the feature space by certain transformation laws into a small set but lack of interpretability [17].

In the present study, we propose a rational framework for choosing the most appropriate combination of ML models and descriptors based on the Genetic algorithm (GA). GA is a highly parallel and global search algorithm that simulates the biological evolution mechanism of "natural selection, survival of the fittest", and is one of the most effective methods for the optimization problem [18]. We utilized GA to evaluate the ML models and descriptors simultaneously and identify the most suitable model with best materials descriptors in an accelerated manner. We validated the performance of our approach by improving the classification accuracy in predicting the phase formation of high entropy alloys (HEAs). Higher than 10% improvement for identifying solid-solution (SS) *vs.* non-solid-solution (NSS) HEAs, and for body-centered-cubic (BCC), face-centered-cubic (FCC), and dual-phase (DP) HEAs was achieved. To further improve the classification model efficiently, several HEAs with the largest classification uncertainties were recommended, experimentally synthesized and measured, and iteratively augmented to the initial dataset by employing an active learning approach. The method can serve as a general algorithm to select materials descriptors and ML models for various materials problems including classification and optimization of targeted properties.

2. Methodology

2.1. Materials descriptor space construction

One of the key factors in controlling the performance of a ML model is the set of materials descriptors used, thus it is important to create a general set of materials descriptors [10]. Since *prior* knowledge of which descriptor outperforms the others is usually lacking, as many as possible potentially "good" descriptors should be accumulated. It ensures that the subsequent selection of descriptors can well explain the targeted property. A materials descriptor space could be constructed by considering elemental parameters, thermodynamic and kinetic parameters, structural information as well as those from density functional theory (DFT) calculations.

Elemental parameters include the intrinsic quantities of elements such as atomic mass, period and group in the periodic table; the heuristic quantities of elements such as Pauling electronegativity and atomic radius; and the physical properties of elements such as melting temperature and specific heat [19]. These parameters represent a coarse-grained analogue of the electronic and bonding properties of the materials. Several physical and chemical properties of alloys, perovskites, and superconductors have been successfully predicted based on such elemental descriptors [7,20,21]. As the structure of a material determines its property, structural parameters such as the coordination number, Voronoi polyhedron of a central atom, angular and radial distribution function, and angular Fourier series can be used to represent different structures [22]. However, these structural descriptors are not yet widely used in the machine-learning assisted prediction of properties. The thermodynamic and kinetic parameters such as the free energy, the mixing entropy and enthalpy, and the diffusion coefficients are used to determine the stability of a particular phase, or the occurrence of a particular reaction [23]. The outputs of DFT calculations including unit cell volume, band gap, cohesive energy, elastic constants, dielectric constants, electronic structure and phonon properties are more abundant and are usually implanted in the ML framework as descriptors to improve the performance of the ML models [24].

Despite the above mentioned general aspects to construct a materials descriptor space, specific descriptors should also be added to particular problems. For example, temperature, time of wetness, exposure time, sulphur dioxide concentration, and chloride concentration are used as input to predict the annual corrosion depth of zinc and steel [25].

2.2. ML models pool construction

ML or data mining techniques have been widely used in web search algorithm[26], finance[27], medicine[28], and marketing[29] for decades. Various ML models have been developed to solve supervised (classification and regression) and unsupervised (clustering) problems. To utilize the most suitable algorithm for a particular problem is crucial for accelerating the discovery of new materials, as a good ML model allows for accurate prediction of unexplored materials.

From the "no free lunch" theorem [30], a best ML model for all materials problems does not exist, and a particular problem requires its own ML model. For example, linear algorithms lead to simple, easy-to-interpret models but always result in poor performance. Support Vector Machine (SVM) is a nonlinear analysis that makes use of the so-called kernel trick to map complex relationships and shows a capability advantage in dealing with scattered samples, nonlinearity and high number of dimensions[3]. Random forest leads to strong predictive accuracy but does not provide a

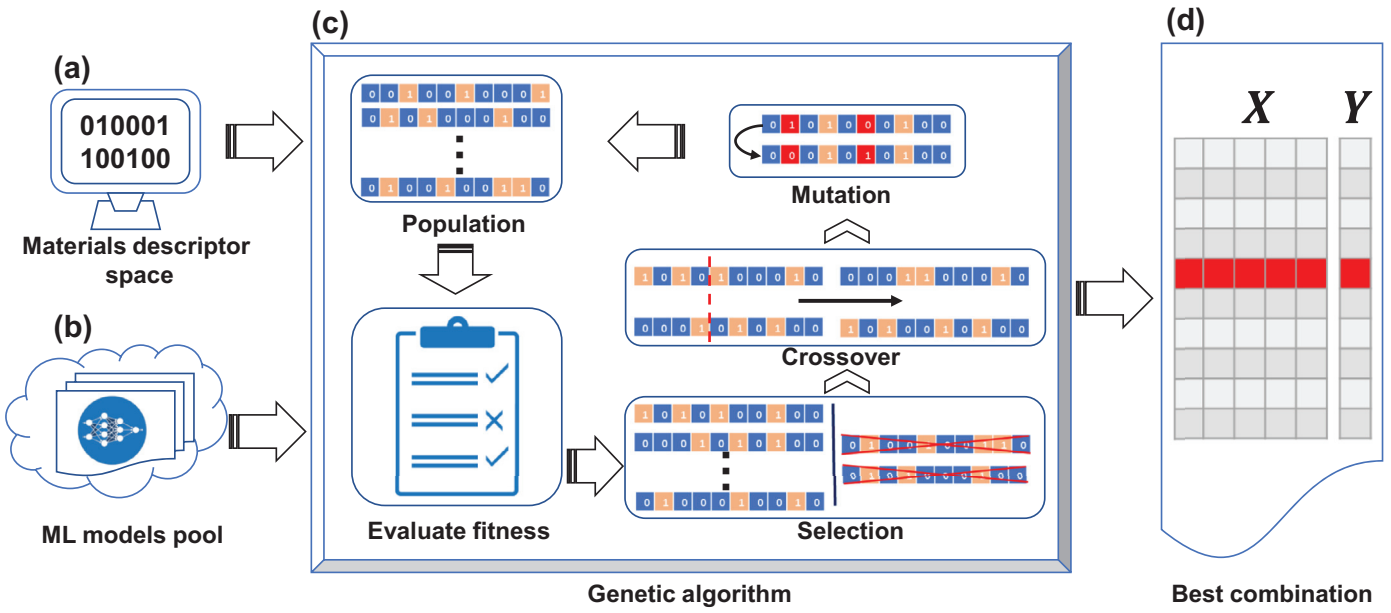

Fig. 1. Flowchart of our strategy to search for the best combination of materials descriptors and machine learning model for a given materials problem. (a) A materials descriptor space and (b) a machine learning models pool are input into (c) a GA iterative loop to search for the global optima by maximizing/minimizing a fitness function (e.g., the accuracy for classification or the root mean square error for regression) iteratively. Based on a stoping criterion on the performance of the fitness function, the GA outputs (d) the most appropriate combination of materials descriptors and machine learning model.

human-understandable model and easily results in overfitting [21]. Consequently, the ML models pool should be constructed according to the particular material problem and include more models which can be compared and selected from.

2.3. Genetic algorithm

After the construction of the ML models pool and materials descriptor space, the next step is to select the most appropriate combination which offers a more accurate prediction of the targeted property. The number of combinations is enormous, and GA is a means to rapidly find an acceptable solution or an approximation to the global optima in such a huge space. Each materials descriptor is encoded with an integer of 1 or 0, meaning it is selected or not. Thus, the whole materials descriptors set is described by a binary string of 0s and 1s. As shown in Fig. 1, GA randomly generates certain number of initial subsets of materials descriptors (various binary strings). A fitness function is defined as the classification accuracy, the forecast error or the particular properties based on specific problems. Based on the value of the fitness function of these initial subsets, three GA operators of selection, crossover and mutation on these binary digits generate new materials descriptor subsets as a new generation. The new generation will serve as the initial subsets and the loop is conducted repeatedly until a stopping criterion for the loop is satisfied. In this way, an optimized combination of ML model and materials descriptors can be achieved.

2.4. Active learning

The preselected ML model might not be as good as expected, and therefore, new samples with labels can be added to refine the ML model by intentionally choosing certain new experiments. Active learning defines a utility function, mostly a balance trade-off between exploitation and exploration, to dictate the next experiment or calculation to be performed [31]. This helps to efficiently improve the performance of the ML model and to find new materials with targeted properties in a faster manner [32–34].

3. Case study: Phase formation of HEAs

HEAs generally have five or more principle elements, and each principal element has a concentration between 5 and 35 at.% [35]. These alloys have attracted much attention and can serve as potential candidates for many industrial applications due to their excellent properties. These include superior strength and hardness [36], good resistances to wear [37], excellent corrosion resistance [38], high thermal stability [39], and good hydrogen storage properties [40].

The unique structure of the solid solution is considered as the dominant contribution to their excellent properties. For example, single-phase FCC HEAs are usually ductile but have relatively low strength, meanwhile single-phase BCC HEAs typically exhibit high strength but are often brittle [41]. Dual-phase HEAs can possess a balance of ductility and strength [42], and therefore, designing HEAs with the desired crystal structure is critical in developing new HEAs with better properties.

Due to their complexity in composition, the number of possible HEAs is enormous. The capability to accurately predict the stable structure or phase formation of unexplored multi-component HEAs is thus highly desired. Recently, several ML methods have been used to solve this phase formation problem in HEAs. For example, Islam et. al. [43], utilized five empirical materials descriptors, the valence electron concentration ($\overline{VEC}$), difference in the Pauling negativities ($\Delta\chi_P$), atomic size difference (δ), mixing enthalpy (ΔH_m), and mixing entropy (ΔS_m) to build a backpropagation neural network (NN) model to distinguish single-phase solid solution, amorphous structure and intermetallic compounds. On average, a predictive accuracy higher than 80% could be reached. Tancret et. al. [44] trained a Gaussian process ML model based on nine physical parameters to guide the design of single solid solutions HEAs, and the reliability of the predictions were increased by combining computational thermodynamics. K-nearest neighbors (KNN), support vector machine (SVM), and artificial neural network (ANN) using the same five materials descriptors in the literature [43] were employed for predicting the phase formation of new HEAs [45]. However, none of those works

Table 1
The materials descriptor space consisting of empirical and self-defined descriptors.

	Description	Abbreviation	Formula
Empirical Descriptors	Atomic Size mismatch	R	$\bar{R} = \sum_{i=1}^{n} c_i R_i$ $\delta R = \sqrt{\sum_{i=1}^{n} c_i (1 - R_i/\bar{R})^2}$
		Λ	$\Lambda = \Delta S_m / \delta^2$
		γ	$\gamma = (1 - \sqrt{\frac{(r_s+r)^2 - r^2}{(r_s+r)^2}}) / (1 - \sqrt{\frac{(r_l+r)^2 - r^2}{(r_l+r)^2}})$
	Electronegativity Mismatch (Pauling)	χ_P	$\overline{\chi_P} = \sum_{i=1}^{n} c_i \chi_{Pi}$ $\Delta\chi_P = \sqrt{\sum_{i=1}^{n} c_i * (1 - \chi_{Pi}/\overline{\chi_P})^2}$ $\delta\chi_P = \sqrt{\sum_{i=1}^{n} c_i * (\chi_{Pi} - \overline{\chi_P})^2}$
	Electronegativity Mismatch (Allen)	χ_A	$\overline{\chi_A} = \sum_{i=1}^{n} c_i \chi_{Ai}$ $\delta\chi_A = \sum_{i=1}^{n} c_i \lvert 1 - \chi_{Ai}/\overline{\chi_A} \rvert$
	Valence Electron Concentration	VEC	$\overline{VEC} = \sum_{i=1}^{n} c_i VEC_i$
	Mixing Entropy	ΔH_m	$\Delta H_m = 4 \sum_{i=1, i \neq j}^{n} c_i c_j H_{ij}^m$
	Mixing Enthalpy	ΔS_m	$\Delta S_m = -R \sum_{i=1}^{n} c_i \ln c_i$
	Combining effects of Mixing Entropy and Mixing Enthalpy	Ω	$\Omega = \frac{T_m \Delta S_m}{\Delta H_m}$
	Average Number of Itinerant Electrons Per Electrons	C_v	$C_v = \frac{\sum c_i e_{vi}}{\sum c_i Z_i}$
Self-defined Descriptors	Melting Point	MT	$\bar{X} = \sum_{i=1}^{n} c_i X_i$ $\delta X = \sqrt{\sum_{i=1}^{n} c_i (1 - X_i/\bar{X})^2}$
	Cohesive Energy of Solid	CE	
	Compression Modulus	MC	
	First Ionization Energy	FIE	
	Second Ionization Energy	SIE	
	Third Ionization Energy	TIE	
	Work Function	WF	
	Atomic Number	AN	
	Quantum Number	QN	
	Column in the Periodic Table	C	
	Relative Atomic Mass	RAM	
	Atom Volume	VA	
	Atomic Environment Number	AEN	
	Electronegativity (Martynov&Batsanov)	χ_{MB}	
	Electronegativity (Alfred-Rochow)	χ_{AR}	
	Absolute Electronegativity	χ_{ABS}	
	Chemical Potential (Miedema)	CPM	
	Effective Nuclear Charge (Slater)	NCE	
	Effective Charge Nuclear (Clementi)	CNE	
	Boiling Temperature	TB	
	Vaporization Enthalpy	EV	
	Melting Enthalpy	EM	
	Atomization Enthalpy	EA	
	Ionic Radii (Yagoda)	RI	
	Covalent Radii	RC	
	Valence Electron Distance(Schubert)	DVE	
	Core Electron Distance (Schubert)	DCE	
	Density	D	

systematically compare the performance of different ML models with different materials descriptors. Thus, we apply our systematic framework to the phase formation prediction problem of HEAs.

We collected 550 HEAs with available phase formation information from different literature sources. All the data was for as-cast samples, and the experimental data have shown that the phases formed in the as-cast state are quite stable [46]. Thus we neglect the possible interruption from processing conditions and thermal histories [47]. We considered the single body-centered cubic (BCC) phase, single face-centered cubic (FCC) phase, and dual FCC and BCC phase (DP) as the solid solution alloys (SS), while the HEAs with intermetallic compounds and amorphous phases are considered as the non-solid solution alloys (NSS). Thus each alloy in our dataset has two labels: one is SS or NSS and the other is FCC, BCC or DP. We will follow the steps shown in Section 2 to rapidly optimize the best ML model for distinguishing (1) the SS and NSS HEAs, and (2) FCC, BCC and DP HEAs.

3.1. Construction of descriptor space for the phase formation of HEAs

We collected materials descriptors that potentially affect the phase formation of HEAs. Fourteen empirical materials descriptors have been proposed in the literature. The mixing enthalpy (ΔH_m), the mixing entropy (ΔS_m), and a descriptor (Ω) that quantifies the predominance of the entropy with respect to the enthalpy are basic thermodynamic descriptors [48]. The elemental descriptors including the valence electron concentration ($\overline{VEC}$), the difference in the Pauling electronegativities ($\Delta\chi_P$), the Allen electronegativities ($\Delta\chi_A$), the atomic size difference (δ), and other geometrical descriptors (Λ) and (γ), are calculated from the extended classic Hume-Rothery rules [48–51].

Additionally, we include twenty-nine elemental parameters that approximately represent bond strength and electro-chemical properties, such as melting temperature (*MT*), ionization energy (*IE*), effective nuclear charge (*ENC*), and so on. Each elemental parameter of a HEA in our composition space can be represented by the molar average value ($\bar{X}$) of an elemental parameter (X_i) through,

$$\bar{X} = \sum_{i=1}^{n} c_i X_i \tag{1}$$

where c_i is the mole fraction of ith element. Another way to describe each HEA is to calculate the mismatch value (δX) between elemental parameters (X_i) of its components. The δX is given by

$$\delta X = \sqrt{\sum_{i=1}^{n} c_i (1 - X_i/\bar{X})^2} \tag{2}$$

where c_i is the mole fraction of ith element and $\bar{X}$ is the molar average value obtained by Eq. 1. Thus fifty-six self-defined materials descriptors are introduced and may contribute to the phase formation of HEAs as well. In total, the materials descriptor space contains 70 materials descriptors, which are listed in Table 1.

The initial training data includes the phase information and corresponding chemical compositions. The former is our targeted response. The latter compositions are used to calculate the 70 descriptors according to the formulas listed in Table 1 with the input of the properties of elements. As we are able to readily get access to these elemental properties, all the descriptors are available for both our training data and virtual data.

Our materials descriptors vary in magnitudes, units and range. As most of the machine learning algorithms use Eucledian distance between two data points, materials descriptors with high magnitudes will weigh more compared with those of low magnitudes. To bring all materials descriptors to the same level of magnitudes, each materials descriptor is normalized through,

$$X_i^{norm} = \frac{X_i - X_{min}}{X_{max} - X_{min}} \tag{3}$$

where X_{max} and X_{min} are the maximum and minimum values of descriptor X of our training data, respectively.

3.2. Construction of classifiers pool

Our problem requires us to build a binary classifier to distinguish NSS and SS HEAs and a multi-class classifier to separate FCC, BCC and DP HEAs. Our ML models pool consists of nine widely used classification models. These include a linear discriminant analysis (Lda), a decision tree model (Dtree), a Naive Bayes classifier (Nbayes), a Neural Network classifier (Nnet), a random forest model (RF) and support vector machine with a linear kernel (SVM.lin), a polynomial kernel (SVM.poly), a radial basis function (SVM.rbf) kernel or a sigmoid kernel (SVM.sigm). The machine learning calculations are performed using R language in the RSTUDIO environment. These classifiers are implemented in *e1017* package for Nbayse and SVM, in *nnet* package for Nnet, in *rpart* package for Dtree, in *randomForest* package for RF and in *MASS* package for Lda.

3.3. Model and materials descriptor selection based on GA

Nine ML models with seventy materials descriptors lead to a huge number of combinations, from which we need to identify the best performing combination. As the space is too large to be fully explored, we implement the GA to achieve an efficient search. Algorithm 1 illustrates the pseudo-code of GA, where the seventy materials descriptors are coded as a string with length of 70 bits. Each bit is either 1 or 0, indicating the presence or absence of a particular materials descriptor. Thus a materials descriptor subset is represented by a specific string. Our aim is to obtained an optimal subset, *i.e.*, an optimal string, for a particular ML model.

Algorithm 1: A pseudo code of genetic algorithm.

1. Randomly choose an initial population of n materials descriptor subsets (here n=100)
2. Evaluate the classification accuracy of each materials descriptor subset

repeat

3. Select the materials descriptor subsets with higher classification accuracy by the selection operator (*the stochastic tournament*)
4. Generate a new generation of n materials descriptor subsets by the crossover operator (*uniform crossover*) and then the mutation operator (*bit flip mutation*)
5. Evaluate the classification accuracy of each materials descriptor subset in the new generation

until *the stopping criterion (the number of generations)*;

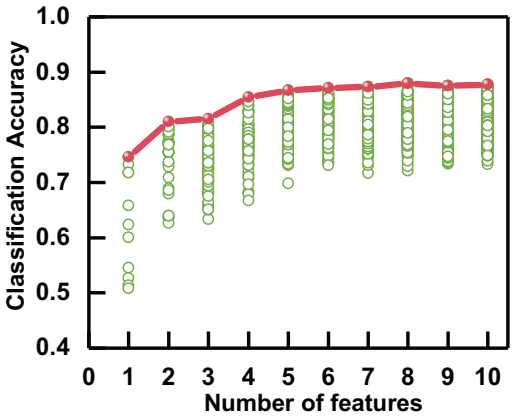

Fig. 2. The classification accuracy of Dtree model as a function of the number of descriptors. All the data are from various descriptor subsets that are selected and evaluated in the optimization process of GA. The red frontier tracks the best classifier for a given number of descriptors. The improvement of the model performance is negligible when the number of descriptors is more than four. The classification accuracy is the average value of 100 leave-10%-out testing sets. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The number of descriptors is usually determined by the best performance of the classifier. However, in order to reduce the complexity of the model, the number of descriptors can be limited to a certain value without sacrificing the model performance too much. In the optimization process of GA, various descriptor subsets with different size will be selected and evaluated according the model performance. Thus it is possible to track the model performance as a function of the number of descriptors. We plot the classification accuracy of Dtree model to classify the FCC, BCC and DP HEAs in Fig. 2. The red frontier tracks the best classifier for each number of descriptors. The classification accuracy increases with the number. However, the improvement of the model performance is negligible when the number of descriptors is more than four. Hence, we limit the number of selected descriptors to four in the subsequent analysis.

Initially, a population of n=100 materials descriptor subsets are randomly chosen. These 100 different strings thus serve as a parent generation. For each materials descriptor subset, we build a ML classifier and utilize the leave-out validation to evaluate its classification accuracy. We split the initial dataset randomly into a training set with 90% samples and a testing set with 10% samples.

Table 2
Parameters of GA process used in this work.

Parameter	Population Size	Number of generation	Crossing Probability	Mutation Probability	Number of Materials Descriptors
value	100	50	80%	1%	<=4

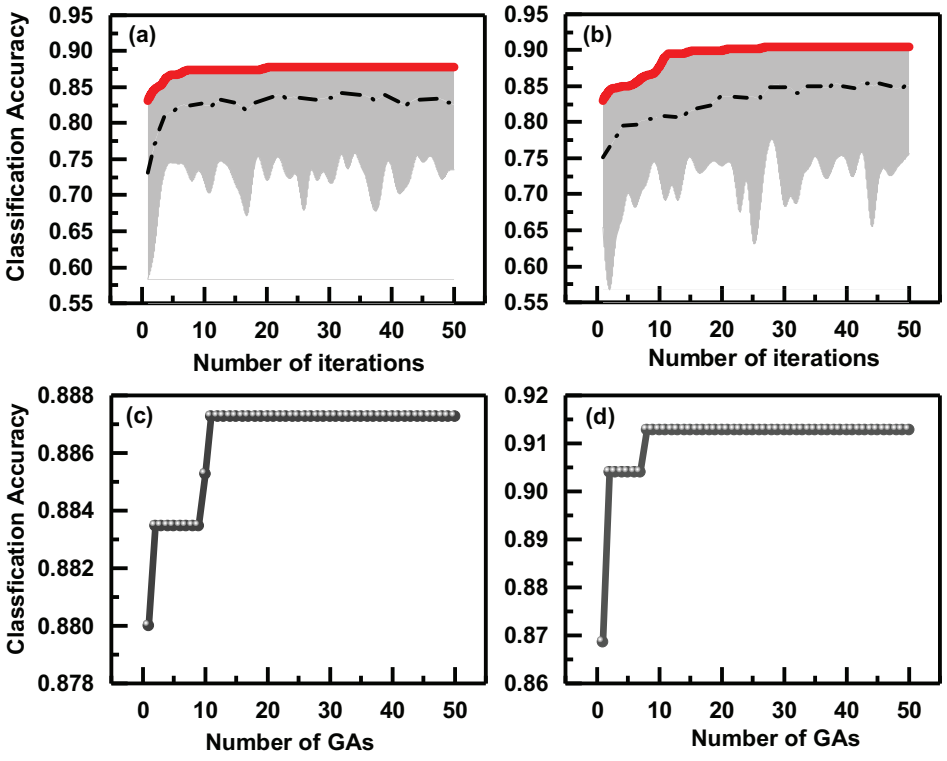

Fig. 3. Two typical examples of GA optimization process. (a) and (c) are the results of SVM.rbf model for classifying the SS and NSS HEAs. (b) and (d) are the results of Nnet model for classifying FCC, BCC and DP HEAs. (a) and (b) plot the classification accuracy *versus* the number of iterations during one GA run. The red solid line is the best performer of each iteration and the black dashed line represents the mean value of the classification accuracy for each iteration. Both the average and max accuracy increase rapidly in the early iterations and converge to an optimal solution. (c) and (d) show the classification accuracy of the "best" classifier after a particular number of GA runs. The "best" classifier is determined by comparing the current classification accuracy with the previous one and the larger one is retained as the "best" classifier. Stable solutions are obtained as the number of independent GA runs increases. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The ML classifier is built on the training set, and a classification accuracy is obtained by comparing the predicted and actual labels of the testing set. The process is repeated 100 times and the classification accuracy we used in the following steps is obtained by the average value of the 100 classification accuracies. We keep the materials descriptor subset with higher classification accuracy by the so-called stochastic tournament selection method. Specifically, we run "tournaments" for 100 times among two materials descriptor subsets chosen by a roulette wheel selection from the population, and within each tournament, the one with higher classification accuracy is selected. Then the crossover and mutation operators are employed to generate new materials descriptor subsets to diversify the search process. Specifically, we randomly choose two materials descriptor subsets from a preselected 80% of the population as parents. Each bit of the new materials descriptor subset is chosen from either parent with equal probability, known as the uniform crossover. Thus a large number of child subsets can be generated. We then setup a filter to examine the number of descriptors present in the child subset. Only the one with descriptors less than or equal to 4 is retained until enough subsets for the next mutation step are obtained. We then utilized the bit flip mutation which randomly selects bits from all 70×100 bits of the population with the probability of 1% and flip them. The same filter is utilized here again to limit the number of descriptors. The newly generated 100 materials descriptor subsets are appended to the next generation. The classification accuracies of the new generation are evaluated, and the selection, crossover and mutation are repeated. Such an iteration can be stopped either by controlling the number of generations or by setting a threshold on the classification accuracy. The parameters utilized in the present study are listed in Table 2. GA is a stochastic global optimization method and its results depend on its initial population, which is usually randomly selected. To obtain a stable solution, 50 independent GA runs are conducted for each ML classifier.

4. Results and discussion

4.1. Materials descriptor selection results for different ML models by GA method

In this section, the efficiency of our method to recognize the best materials descriptors for different ML models is validated. Two classification problems are studied here: classification I for the NSS and SS HEAs and classification II for the FCC, BCC and DP HEAs. Our method based on the GA rapidly selects the optimal materials descriptors for different ML models for the two problems.

We utilized the SVM.rbf model for classification I and Nnet model for classification II as typical examples to show the opti-

Table 3
Materials descriptors selected by the GA and SFS method for different ML models together with their classification accuracies. The materials descriptors selected by LASSO, RF and GTB are also included.

Machine Learning Models	Feature-selection Methods	Classification I		Classification II	
		Materials Descriptors	Accuracy	Materials Descriptors	Accuracy
Lda	GA	$\overline{EA}, \delta\chi_P, \delta AN, \delta RC$	81.3%	$\overline{\chi_{AR}}, \delta MC, \delta DCE, \delta D$	83.9%
	SFS	$\delta\chi_{AR}, \overline{RI}, \delta TIE, \overline{SIE}$	78.8%	$\overline{FIE}, \overline{MC}, \overline{AEN}, \delta D$	80.7%
Bayes	GA	$\overline{EV}, \delta AN, \delta VA, \delta TB$	80.4%	$\overline{D}, \overline{VEC}, \delta\chi_P, \delta SIE$	83.9%
	SFS	$\overline{RI}, \delta VA, \delta QN, \delta MT$	77.5%	$\delta SIE, \overline{VEC}, \delta CNE, \delta MC$	81.1%
Dtree	GA	$\overline{QN}, \overline{D}, \Delta\chi_P, \delta CE$	85.9%	$\overline{MT}, \overline{D}, \delta\chi_{ABS}, \delta DCE$	88.0%
	SFS	$\overline{TB}, \overline{QN}, \overline{RI}, \overline{DCE}$	84.8%	$\overline{AN}, \overline{TIE}, \overline{D}, \overline{EM}$	85.1%
Nnet	GA	$\overline{RI}, \delta AN, \delta VA, \delta TB$	87.4%	$\overline{VEC}, \overline{DCE}, \delta MC, \delta\chi_P$	91.3%
	SFS	$\delta\chi_{AR}, \delta MC, \overline{\chi_{MB}}, \delta CNE$	85.4%	$\overline{AN}, \overline{RM}, \delta\chi_P, \delta MC$	88.0%
RF	GA	$\overline{AN}, \delta QN, \delta TB, \delta CPM$	87.5%	$\overline{\chi_P}, \overline{\chi_{MB}}, \overline{D}, \delta RAM$	90.0%
	SFS	$\delta TB, \delta VA, \overline{TIE}, \delta AEN$	85.9%	$\gamma, \overline{RM}, \chi_P, \delta RAM$	87.5%
SVM.rbf	GA	$\overline{AN}, \delta RC, \delta\chi_{MB}, \delta TB$	88.7%	$\overline{\chi_P}, \overline{RI}, \overline{D}, \delta MC$	87.2%
	SFS	$\overline{EA}, \overline{DCE}, \delta TB, \delta VA$	86.9%	$\overline{VEC}, \delta CNE, \delta MC, \delta WF$	84.1%
SVM.lin	GA	$\delta\chi_P, \overline{CE}, \overline{RM}, \delta RC$	81.8%	$\overline{RM}, \overline{VEC}, \overline{\chi_A}, \delta MC$	85.6%
	SFS	$\delta\chi_{AR}, \overline{RI}, \overline{AEN}, \delta CNE$	78.6%	$\delta CNE, \delta MC, \overline{VEC}, C_v$	84.3%
SVM.poly	GA	$\delta\chi_P, \delta TB, \delta VA, \delta DVE$	81.5%	$\delta\chi_{ABS}, \delta\chi_{MB}, \overline{C}, \delta FIE$	75.6%
	SFS	$\overline{C}, \delta CNE, \delta\chi_P, \delta VA$	80.4%	$\overline{RI}, C_v, \delta DVE, \delta CE$	65.2%
SVM.sigm	GA	$\overline{VA}, \overline{CPM}, \overline{CE}, \delta CE$	64.3%	$\overline{VEC}, \delta C, \delta\chi_{ABS}, \delta MC$	81.9%
	SFS	$\overline{FIE}, \delta EV, \delta WF, \delta MC$	63.4%	$\overline{VEC}, C_v, \delta\chi_{MB}, \delta D$	80.7%
-	GTB	$\Delta H_m, \delta\chi_{AR}, \delta VA, \delta RC$	-	$\overline{R}, \overline{FIE}, \Delta\chi_{ABS}, \overline{TIE}$	-
-	RF	$\delta RC, \delta, \delta RAM, \delta VA$	-	$\overline{R}, \overline{VA}, \overline{VEC}, \overline{D}$	-
-	LASSO	$\delta RAM, \overline{EA}, \gamma, \delta VA$	-	$\overline{VEC}, \delta\chi_{ABS}, \delta MT, \delta D$	-

mization process of GA. Fig. 3 (a) and (b) plot the classification accuracy versus the number of iterations within one GA run for SVM.rbf and Nnet, respectively. The classification accuracies of 100 ML models within each generation are included. The red solid line is the best performer of each iteration and the black dashed line is the mean value of the classification accuracy for each iteration. The initial materials descriptor subsets contain stochastically selected materials descriptors, and their corresponding ML models have a low classification accuracy. Within each iteration, the materials descriptor subsets with higher classification accuracy are retained by the selection operator for reproduction. The crossover and mutation operators rapidly explore offsprings presumably with better prediction accuracy. Both the average and max accuracy increase rapidly for the initial few iterations and converge to an optimal solution. Therefore, we perform one GA run and obtain a "best" classifier on particular descriptors with its classification accuracy known.

We repeat the above GA run 50 times to obtain a stable solution as the result of GA is usually dependent on its initial population. Our aim is to keep the best performer within this 50 GA runs. Specifically, the independent GA runs with different initial populations are performed sequentially after each other. The "best" classifier is determined by comparing the current classification accuracy with the previous one and the larger one is retained as the "best" classifier. Fig. 3 (c) and (d) show the classification accuracy of the "best" classifier after a particular number of GA runs for the classification I and classification II problems, respectively. Stable solutions are obtained after 20 independent GA runs for both cases. After 50 GA runs, we can achieve a classifier with highest classification accuracy.

Besides the two typical models, GA selects the suitable materials descriptor subsets for different ML models, which are shown in Table 3. All the subsets contain four materials descriptors, since the number is limited to be equal to or less than 4 in the GA process. For classification I, the best performer is the SVM.rbf model with classification accuracy of 88.7%, and includes the atom number ($\overline{AN}$), the mismatches in Martynov and Batsanov electronegativities ($\delta\chi_{MB}$), in covalent radius (δRC) and in boiling temperature (δTB) as the optimal materials descriptors. The electronegativity describes the ability of an atom to attract a shared pair of electrons. According to the Hume-Rothery rules [52], elements with comparable electronegativities prefer to form a solid solution. Hence, a larger mismatch in electronegativity tends to form non solid solutions. A large radius mismatch would lead to substantial lattice distortions, which consequently destabilize the solid-solution phase. Our GA method converges to the $\Delta\chi_{MB}$ and δRC, which are related with the above two aspects and in agreement with empirical descriptors [48,49]. The other two selected descriptors are $\overline{AN}$ and δTB. The $\overline{AN}$ can be considered as a coarse estimation of the average number of electrons. Solid solution always have an electron concentration limit and excessive electrons make the structure unstable.

The boiling temperature reflects the bond strength between atoms. A low δTB results in homogeneous bond strength, and thus favors the solid solution. The four descriptors represent the mismatches in the elemental properties of the components, and we can reasonably infer that the uniformity between components, *i.e.,* a low mismatch, is an important factor to forming solid solutions.

For classification II, the Nnet model outperformed all others and had a classification accuracy as high as 91.3%. The best materials descriptor subset for this model comprised of the valence electron concentration ($\overline{VEC}$), the core electron distance ($\overline{DCE}$), the mismatch in Pauling electronegativities ($\delta\chi_P$), and the mismatch in compression modulus (δMC). Guo et al. [53] found that low $\overline{VEC}$ values correspond to BCC HEAs, while FCC structures are more likely to possess a high value of $\overline{VEC}$. Hence this property is closely related to the phase formation in HEAs, FCC or BCC. However, the underlying physics are not still understood. As FCC is more closely

packed than BCC, HEAs with larger $\overline{DCE}$ may have a better chance to possess BCC structure. The mismatch in electronegativity, $\delta\chi_P$, will influence the segregation behavior between components and consequently the structure [54]. Finally, the δMC can be thought of as a reflection of the bond strength, where a large δMC may lead to severe lattice distortions. BCC structures with a packing fraction of 68% could provide more space to reduce the overall lattice distortion compared with FCC structures, which have a packing fraction of 74% [55]. Thus BCC structures are likely have large δMC.

4.2. Performance of the ML models based on materials descriptors by other feature-selection methods

Wrapper methods such as the GA here and the SFS method can optimize the combination of ML model and its descriptors, while the embedded methods such as LASSO, RF or GTB extract materials descriptors that may be only valid for a particular ML model. GA and SFS can select the best descriptors for 9 different ML models by themselves, respectively. LASSO, RF and GTB selected the most relevant materials descriptors in advance and then build ML models based on them. Table 3 shows the materials descriptor subsets for different ML models selected from GA and SFS. The three materials descriptor subsets selected by LASSO, RF and GTB are listed as well. Note that we select from the 70 materials descriptors and do not enlarge the descriptors pool. In the following, we compare the efficiency and classification accuracy of the GA result with that from other methods.

As the embedded methods are always based on one ML model, a very high computation efficiency can be achieved. The SFS method adds one material descriptor at a time by sequentially selecting them and chooses the descriptor based on the classification accuracy. Thus considering the SFS method to deal with 70 descriptors in our case, the search space size of one ML model would be $\sum_{i=1}^{70} = 2485$. While the search space size for the GA method with an initial population of 100 and 50 generations is $50 \times 100 = 5000$. That is to say that these search space are comparable with each other.

We can then build in total 45 ML models, *i.e.*, nine ML models from our ML model pool using materials descriptor subsets from the GA, SFS, LASSO, RF and GTB methods as listed in Table 3. We then compare the performance of these ML models according to the classification accuracy. It is calculated by using a leave-10%-out as testing set and 90% as training set. The value of accuracy is determined by comparing the predicted and actual labels. We repeat this procedure 100 times and obtained the average accuracy of 100 classification accuracies for one ML model with one subset of materials descriptors.

The average classification accuracies of different models for classification I and classification II are shown by the histogram in Fig. 4 (a) and (b), respectively. It can be seen that SVM.rbf and Nnet with materials descriptors from the GA methods are the best performers for classification I and classification II, respectively. Within each type of model, the one with materials descriptors from the GA also outperforms the others with materials descriptors from LASSO, RF, GTB and SFS. This is true for all the models in our ML model pool. Thus the GA is a more robust method to choose materials descriptors for a particular type of ML model.

4.3. Traditional parametric approach

Besides the outlined recent ML studies, researchers have adopted the so-called parametric approach in the past decades to visualize different phases of HEAs in a two-dimensional plane of two empirical descriptors listed in Table 1. Typical examples of such two-dimensional scatters are shown in the Supplementary Materials. To quantitatively compare such a method with the ML

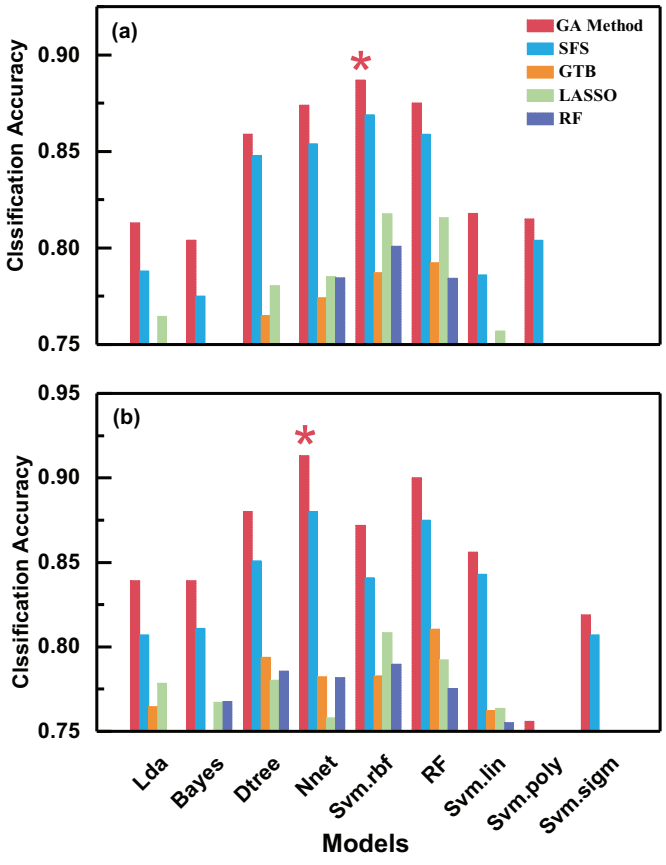

Fig. 4. Comparison of the performance of nine ML models using each materials descriptor subset selected by the LASSO, RF, GTB, SFS and GA methods. (a) are classifiers for classifying the SS and NSS HEAs and the best performer is SVM.rbf with materials descriptors from the GA method. (b) are classifiers for classifying FCC, BCC and DP HEAs and the best performer is Nnet with materials descriptors from the GA method. The histogram displays the average accuracy of 100 classification accuracies by 100 times leave-10%-out validation for each ML model.

ones, we calculate the classification accuracy by identifying regions in the two-dimensional plane. The details of the calculations are shown in the Supplementary Materials. Fig. 5 plots the classification accuracies of the parametric approaches as well as the GA method. Our method possesses an accuracy of more than ten percent higher for both classifications I and II.

It should be noted that although the prediction accuracy reaches 88.7% and 91.3%, the ML models here are complex and hard to interpret. A more simple formula such as a polynomial or exponential could be a more attractive choice to provide more fundamental insight for understanding the problem, even though the prediction accuracy may be compromised. In the future, more materials descriptors can be constructed based on the knowledge from domain experts, which would be helpful for further improving the accuracy and understanding.

4.4. The influence of the initial population on GA results

In the above sections, we utilize GA to select the best subset of materials descriptors by choosing the best ML model with highest classification accuracy. Another way to get the best subset is to do statistics on the subsets of 50 GA runs. However, the statistical analysis on the results of these 50 GA runs is rather scattered. One possible reason might be that several descriptors are linearly dependent on each other so that GA hardly distinguishes them based on our fairly small dataset. We proposed two possible solution for this: one is to increase the number of GA runs and the other is to

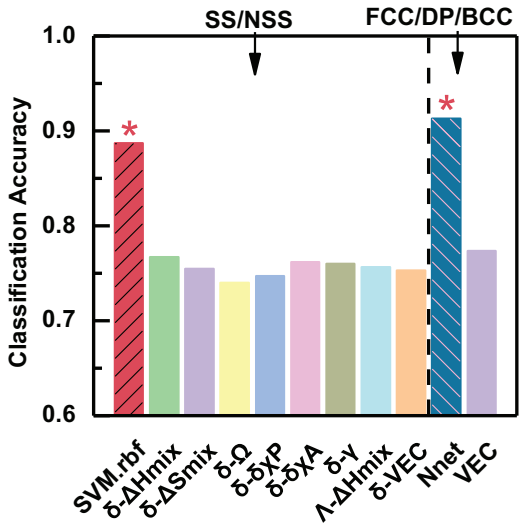

Fig. 5. Classification accuracies of the parametric approach as well as the SVM.rbf and Nnet with materials descriptors from GA method. Our models possess accuracies of more than ten percent higher.

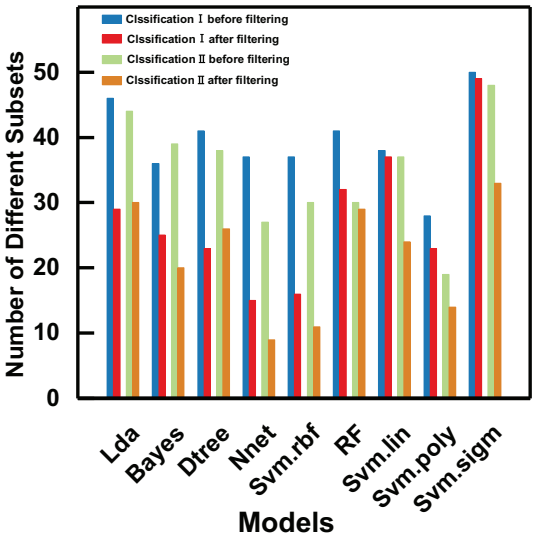

Fig. 6. Comparison of the number of different materials descriptor subsets within the results of 50 GA runs for each ML model before and after the Pearson correlation filter. The number of different subsets decreases after omiting the highly correlated descriptors..

setup a filter to remove the correlated descriptors beforehand. The former will increase the computation cost, so we show the effect of the latter solution.

For each ML classifier 50 independent GA runs would give a pool of 50 materials descriptor subsets, which can contain different descriptors. For example, for the SVM.rbf model of classifications I, δTB is selected 42 times and δRC is selected 15 times within the 50 GA runs, but the other two descriptors vary significantly, resulting in as much as 37 different subsets. A similar situation occurs in the GA runs for the other ML models. We calculate the Pearson correlation coefficients between different features to remove the highly correlated features with a correlation coefficient greater than 0.9. Accordingly, there are 46 descriptors for classifications I and 37 for classifications II after the filter. We again performed 50 independent GA runs on the new descriptor space. The number of different subsets for each ML model decreases after omiting the highly correlated descriptors, as shown in Fig. 6, indicating that removing correlated descriptors could help reduce the influence of initial population on the GA results.

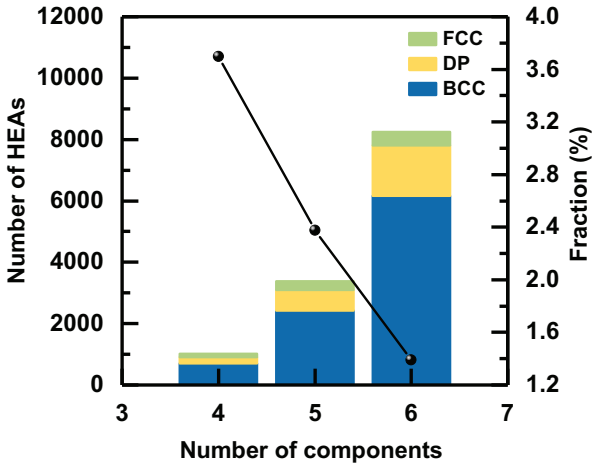

Fig. 7. The number of FCC, BCC and DP HEAs with four to six equimolar elements and the corresponding fraction of SS HEAs in the total candidates.

4.5. Prediction and experimental design

Given the better classification performance of our ML model with the associated descriptors, the trained classifiers then can be used to search for new HEAs with targeted structure. We consider the HEAs with four to six elements from the dataset including 30 elements in equimolar ratio as the unexplored candidates. This gives an unexplored search space with 763,686 possible alloys in total.

We employed the SVM.rbf model for classification I and Nnet for classification II. For each classification problem, we construct 1000 bootstrapped datasets from the original dataset and build 1000 classifiers to predict the structures of the whole search space. The classifier I (SVM.rbf) predicts 12,647 solid solution HEAs from the whole search space. Among them, classifier II (Nnet) predicts 845 HEAs with FCC structure, 9302 HEAs with BCC structure and 2500 HEAs with DP structure. We found that with increasing the number of components, the number of HEAs with SS phase increases, however, the proportion of HEAs with SS phase in the whole compositional space decreases, as shown in Fig. 7. The total proportion HEAs with SS phase in the space with 4 to 6 components is about 7.46%. This is due to the increased possibility to form intermetallic compounds with increasing number of components.

In order to offer a more robust prediction, the accuracy of the classifiers should be further enhanced. We utilize the so-called active learning framework to intentionally augment more labeled alloys to refine our classifier. The flowchart of our active learning is illustrated in Fig. 8. In our case, the synthesis of new alloys and the characterization of the structure are time-consuming and costly. Therefore, selecting the most informative alloys to do experiments on, so as to minimize the number of possible experiments, is very beneficial. Accordingly, we have the experimental design panel in Fig. 8, which allows us to choose HEAs with greatest uncertainty in the classifier's prediction.

We employ the "bootstrap method" to obtain the uncertainty associated with the classifier's prediction. The dataset is resampled for 1000 times with replacement to generate 1000 bootstrapped datasets. Based on the 1000 datasets, 1000 classifiers are built. Each HEA in the search space has 1000 predictions from the 1000 classifiers, and the probability of this HEA belonging to a particular class can be calculated. Then, the candidates for the next

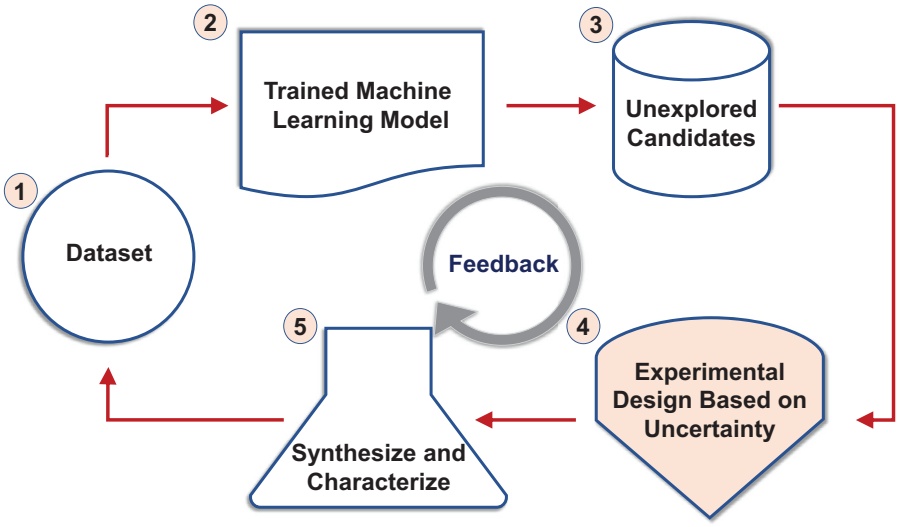

Fig. 8. Flowchart of our active learning. An optimal ML model with materials descriptors selected by the GA method has been trained for each classification problem. The ML model is used to predict the phase of the unexplored candidates and the associated uncertainties are determined by the "bootstrap method". The HEAs with greatest uncertainty are chosen to synthesize and characterize, and augmented to the initial dataset to refine the ML model.

Table 4
The phase and labels of the 10 newly selected, synthesized and phase-identified HEAs.

Compositions	The phase by XRD	Labels in classification I	Labels in classification I
TiMnCoCu	IM	NSS	-
TiCoZrNbMo	IM	NSS	-
AlTiVCoNiCu	SS	SS	DP
AlTiCrMnFeCu	IM	NSS	-
AlTiCrMnCoCu	IM	NSS	-
SiTiVCrNiCu	IM	NSS	-
TiVNiZrNbMo	IM	NSS	-
NbMoHfTa	SS	SS	BCC
TiFeCoNiAg	IM	NSS	-
VFeCoNiCuSn	IM	NSS	-

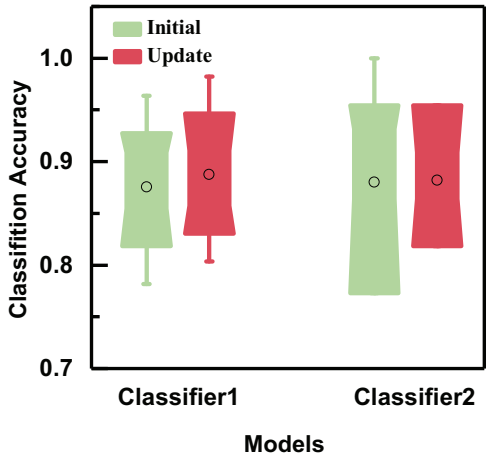

Fig. 9. Comparison of the accuracy of classifier I (SVM.rbf model) and classifier II (Nnet) on initial dataset and updated dataset with feedback, respectively. The classification accuracies of the two classifiers all improved.

experiment are recommended through the following two steps. First, we chose the HEAs with a 45% to 55% probability of belonging to a particular class. This means that the classifier is not sure about the label of the alloy. Second, the maximum melting point of the components should be at least 100 K lower than the lowest boiling point of components to make sure a successful synthesis of the HEAs.

Following these steps, we select 10 HEAs to experimentally synthesize and characterize. After characterization, eight of them are found to be NSS and two are SS with DP (AlTiVCoNiCu) and BCC structure (NbMoHfTa). The results are also shown in Table 4. The labels (NSS and SS) for the 10 samples are augmented to the SVM.rbf model of classification I and the AlTiVCoNiCu alloy with label of DP and NbMoHfTa alloy with label of BCC are augmented to the Nnet model of classification II.

The classification accuracies of the two classifiers all improved as shown by the box plot in Fig. 9. Therefore, the phase formation prediction of HEAs can be more accurate by augmenting new HEAs with greater uncertainty in ML model prediction. More iterations can be conducted to further improve the accuracy, which will be performed in the future.

5. Summary

In summary, we propose a framework to select the best combination of materials descriptor subset and ML model based on a genetic algorithm. We validate our method by applying it to the phase formation problem of HEAs. The classification accuracy is improved more than ten percent over the traditional two-dimensional empirical approach. With this framework we also identify several new materials descriptors which are related to the phase formation. Furthermore, an active learning approach is employed to improve the accuracy of the classifiers iteratively, by considering the prediction uncertainties. We expect that our method can serve as a recipe for selecting the best combination

of ML model and materials descriptor subset for various materials problems.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was financially supported by the National Key Research and Development Program of China(Grant No. 2016YFB0700505), National Natural Science Foundation of China (Grant No. 51671157) and Higher Education Discipline Innovation Project.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.actamat.2019.11.067.

References

[1] J. Rickman, T. Lookman, S. Kalinin, Materials informatics: from the atomic-level to the continuum, Acta Materialia. 168 (2019) 473–510.
[2] K.T. Butler, D.W. Davies, H. Cartwright, O. Isayev, A. Walsh, Machine learning for molecular and materials science, Nature. 559 (2018) 547–555.
[3] P. Raccuglia, K. Elbert, P.D.F. Adler, C. Falk, M. Wenny, A. Mollo, M. Zeller, S.A. Friedler, J. Schrier, A. Norquist, Machine-learning-assisted materials discovery using failed experiments, Nature. 533 (2016) 73–76.
[4] C. Stefano, M. Dane, P. Kristin, R. John, C. Gerbrand, Predicting crystal structures with data mining of quantum calculations, Phys. Rev. Lett. 91 (2003) 135503.
[5] E. Blisle, Z. Huang, S.L. Digabel, A.E. Gheribi, Evaluation of machine learning interpolation techniques for prediction of physical properties, Comput. Mater. Sci. 98 (2015) 170–177.
[6] B. Meredig, A. Agrawal, S. Kirklin, J.E. Saal, J.W. Doak, A. Thompson, K. Zhang, A. Choudhary, C. Wolverton, Combinatorial screening for new materials in unconstrained composition space with machine learning, Phys. Rev. B. 89 (2014) 094104.
[7] D. Xue, D. Xue, R. Yuan, Y. Zhou, P.V. Balachandran, X. Ding, J. Sun, T. Lookman, An informatics approach to transformation temperatures of niti-based shape memory alloys, Acta Materialia. 125 (2017) 532–541.
[8] C. Wen, Y. Zhang, C. Wang, D. Xue, Y. Bai, S. Antonov, L. Dai, T. Lookman, Y. Su, Machine learning assisted design of high entropy alloys with desired property, Acta Materialia. 170 (2019) 109–117.
[9] J. Fürnkranz, D. Gamberger, N. Lavrač, Machine Learning and Data Mining, Springer, Berlin Heidelberg, 2012. 1–17
[10] L.M. Ghiringhelli, Jan, big data of materials science: critical role of the descriptor, Phys. Rev. Lett. 114 (2014) 105503.
[11] R. Ouyang, S. Curtarolo, E. Ahmetcik, M. Scheffler, L.M. Ghiringhelli, Sisso: a compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates, Phys. Rev. Mater. 2 (2018) 083802.
[12] R. KOHAVI, A study of cross-validation and bootstrap for accuracy estimation and model selection,, Proceedings of the international joint Conference on Artificial intelligence (1995) 1137–1143.
[13] D. Xue, P.V. Balachandran, H. Wu, R. Yuan, Y. Zhou, X. Ding, J. Sun, T. Lookman, Material descriptors for morphotropic phase boundary curvature in lead-free piezoelectrics, Appl. Phys. Lett. 111 (3) (2017) 032907.
[14] Q. Xu, Z. Li, M. Liu, W.J. Yin, Rationalizing perovskite data for machine learning and materials design, J. Phys. Chem. Lett. 9 (24) (2018) 6948–6954.
[15] C. Hu, G. Jain, C. Schmidt, C. Strief, M. Sullivan, Online estimation of lithium-ion battery capacity using sparse Bayesian learning, J. Power Sources. 289 (2015) 105–113.
[16] J. Prez-Benitez, L. Padovese, Feature selection and neural network for analysis of microstructural changes in magnetic materials, Expert Syst. Appl. 38 (2011) 10547–10553.
[17] S.R. Broderick, J.R. Nowers, B. Narasimhan, K. Rajan, Tracking chemical processing pathways in combinatorial polymer libraries via data mining, J. Combinator. Chem. 12 (2010) 270–277.
[18] D.E. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Wesley, Boston, MA,
[19] A. Seko, A. Togo, I. Tanaka, Descriptors for Machine Learning of Materials Data, Springer, Singapore, 2018. 3–23
[20] S.G. Javed, A. Khan, A. Majid, A.M. Mirza, J. Bashir, Lattice constant prediction of orthorhombic Abo3 perovskites using support vector machines, Comput. Mater. Sci. 39 (2007) 627–634.
[21] V. Stanev, C. Oses, A.G. Kusne, E. Rodriguez, J. Paglione, S. Curtarolo, I. Takeuchi, Machine learning modeling of superconducting critical temperature, NPJ Comput. Mater. 4 (2018) 29.
[22] A. Bartok, R. Kondor, G. Csnyi, On representing chemical environments, Phys. Rev. B. 87 (18) (2013) 184115.
[23] Y. Ye, Q. Wang, J. Lu, C. Liu, Y. Yang, Design of high entropy alloys: a single–parameter thermodynamic rule, Scripta Materialia 104 (2015) 53–55.
[24] A. Seko, T. Maekawa, K. Tsuda, I. Tanaka, Machine learning with systematic density-functional theory calculations: application to melting temperatures of single- and binary-component solids, Phys. Rev. B. 89 (2014) 054303.
[25] S. Fang, M. Wang, W. Qi, F. Zheng, Hybrid genetic algorithms and support vector regression in forecasting atmospheric corrosion of metallic materials, Comput. Mater. Sci. 44 (2008) 647–655.
[26] C.-F. Tsai, Y.-F. Hsu, C.-Y. Lin, W.Y. Lin, Intrusion detection by machine learning: a review, Expert Syst. Appl. 36 (2009) 11994–12000.
[27] R. Dash, P.K. Dash, A hybrid stock trading framework integrating technical analysis with machine learning techniques, J. Finance Data Sci. 2 (2016) 42–57.
[28] Y. Liu, Active learning with support vector machine applied to gene expression data for cancer classification, ChemInform. 36 (2005) 1936–1941.
[29] J. Patel, S. Shah, P. Thakkar, K. Kotecha, Predicting stock market index using fusion of machine learning techniques, Expert Syst. Appl. 42 (2015) 2162–2172.
[30] D.H. Wolpert, W.G. Macready, No free lunch theorems for optimization, IEEE Trans. Evolut. Comput. 1 (1997) 67–82.
[31] T. Lookman, P. Balachandran, D. Xue, R. Yuan, Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design, NPJ Comput. Mater. 5 (2019) 21.
[32] R. Yuan, Z. Liu, P.V. Balachandran, D. Xue, Y. Zhou, X. Ding, J. Sun, D. Xue, T. Lookman, Accelerated discovery of large electrostrains in batio3-based piezoelectrics using active learning, Adv. Mater. 30 (7) (2018) 1702884.
[33] P.V. Balachandran, D. Xue, J. Theiler, J. Hogden, T. Lookman, Adaptive strategies for materials design using uncertainties, Sci. Rep. 6 (2016) 19660.
[34] R. Dehghannasiri, D. Xue, P.V. Balachandran, M.R. Yousefi, L.A. Dalton, T. Lookman, E.R. Dougherty, Optimal experimental design for materials discovery, Comput. Mater. Sci. 129 (2017) 311–322.
[35] J.W. Yeh, Recent progress in high-entropy alloys, Eur. J. Control 31 (2006) 633–648.
[36] Y. Deng, C. Tasan, K. Pradeep, H. Springer, A. Kostka, D. Raabe, Design of a twinning-induced plasticity high entropy alloy, Acta Materialia. 94 (2015) 124–133.
[37] C.-Y. Hsu, J.-W. Yeh, S.-K. Chen, T.T. Shun, Wear resistance and high-temperature compression strength of FCC cuconicral0.5fe alloy with boron addition, Metallurg. Mater. Trans. A 35 (5) (2004) 1465–1469.
[38] Y.-J. Hsu, W.-C. Chiang, J.K. Wu, Corrosion behavior of feconicrcux high-entropy alloys in 3.5% sodium chloride solution, Mater. Chem. Phys. 92 (1) (2005) 112–117.
[39] B. Schuh, F. Mendez-Martin, B. Vlker, E. George, H. Clemens, R. Pippan, A. Hohenwarter, Mechanical properties, microstructure and thermal stability of a nanocrystalline cocrfemnni high-entropy alloy after severe plastic deformation, Acta Materialia. 96 (2015) 258–268.
[40] Y.-F. Kao, S.-K. Chen, J.-H. Sheu, J.-T. Lin, W.-E. Lin, J.-W. Yeh, S.-J. Lin, T.-H. Liou, C.W. Wang, Hydrogen storage properties of multi-principal-component cofemntixvyzrz alloys, Int. J. Hydrogen Energy 35 (17) (2010) 9046–9059.
[41] O. Senkov, S. Senkova, C. Woodward, Effect of aluminum on the microstructure and properties of two refractory high-entropy alloys, Acta Materialia 68 (2014) 214–228.
[42] S. Guo, Phase selection rules for cast high entropy alloys: an overview, Mater. Sci. Technol. 31 (10) (2015) 1223–1230.
[43] N. Islam, W. Huang, H.L. Zhuang, Machine learning for phase selection in multi-principal element alloys, Comput. Mater. Sci. 150 (2018) 230–235.
[44] F. Tancret, I. Toda-Caraballo, E. Menou, P.E.J.R. Daz-Del-Castillo, Designing high entropy alloys employing thermodynamics and gaussian process statistical analysis, Materials Design 115 (2017) 486–497.
[45] W. Huang, P. Martin, H.L. Zhuang, Machine-learning phase prediction of high--entropy alloys, Acta Materialia. 169 (2019) 225–236.
[46] S. GUO, C. LIU, Phase stability in high entropy alloys: formation of solid-solution phase or amorphous phase, Progr. Natural Sci. Mater. Int. 21 (6) (2011) 433–446.
[47] L. Wen, H. Kou, J. Li, H. Chang, X. Xue, L. Zhou, Effect of aging temperature on microstructure and properties of alcocrcufeni high-entropy alloy, Intermetallics. 17 (4) (2009) 266–269.
[48] Y. Zhang, Y. Zhou, J. Lin, G. Chen, P. Liaw, Solid-solution phase formation rules for multi-component alloys, Adv. Eng. Mater. 10 (6) (2008) 534–538.
[49] M. Poletti, L. Battezzati, Electronic and thermodynamic criteria for the occurrence of high entropy alloys in metallic systems, Acta Materialia 75 (2014) 297–306.
[50] Z. Wang, Y. Huang, Y. Yang, J. Wang, C. Liu, Atomic-size effect and solid solubility of multicomponent alloys, Scripta Materialia. 94 (2015) 28–31.

[51] A.K. Singh, N. Kumar, A. Dwivedi, A. Subramaniam, A geometrical parameter for the formation of disordered solid solutions in multi-component alloys, Intermetallics. 53 (2014) 112–119.
[52] U. Mizutani, Hume-Rothery Rules for Structurally Complex Alloy Phases, 37, MRS Bulletin, 2012, p. 169.
[53] 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) 213.
[54] Z. Nong, J.-C. Zhu, Y. Cao, X. Yang, Z.-H. Lai, Y. Liu, Stability and structure prediction of cubic phase in as cast high entropy alloys, Mater. Sci. Technol. 30 (2014) 363–369.
[55] S.A. Kube, S. Sohn, D. Uhl, A. Datye, A. Mehta, J. Schroers, Phase selection motifs in high entropy alloys revealed through combinatorial methods: large atomic size difference favors bcc over fcc, Acta Materialia. 166 (2019) 677–686.