Contents lists available at ScienceDirect

Computational Materials Science

journal homepage: http://ees.elsevier.com

Predictive descriptors in machine learning and data-enabled explorations of high-entropy alloys

Ankit Roy, Ganesh Balasubramanian *

Department of Mechanical Engineering & Mechanics, Lehigh University, Bethlehem, PA, USA

ARTICLE INFO

Keywords
High-entropy alloys
Machine learning
Data-analytics
Multi-principal element alloys
Mechanical properties

ABSTRACT

Located at the intersection of intriguing material phases and potentially superior mechanical properties, high-entropy alloys (HEAs) have been gaining increasing interest across academia and industry, in particular for high temperature applications. The extremely vast compositional space (~10^{12} possibilities) for these complex metallic alloys require rigorous predictive strategies to scavenge the expansive realm of unexplored alloy composition-processing-structure–property landscape. Enabled by the advances in artificial intelligence and machine learning methods, data-driven exploration of HEAs are burgeoning, not only for the discovery of new materials but also for predicting properties that are challenging to measure using experiments or require resource and time-intensive computations. Nevertheless, success of such data-enabled models in delivering accurate estimates of microstructures and properties depend on the choice of appropriate descriptors that suitably represent the underlying structural and transport mechanisms. This review provides a synopsis of the contemporary and effective data-centric methods employed to examine HEAs, with special emphasis on the selection and role of feature descriptors. We highlight some of the current challenges with these approaches that the computational materials community is facing, and offer recommendations to address them.

1. The need for data-analytics in HEA research

A subset of multi-principal element alloys (MPEAs), HEAs typically contain 5 or more metals in equimolar concentrations [1]. According to Yeh et al. [1], the high configurational entropy arising from the significant fractions of multiple elements minimize the Gibbs free energy and stabilize a single-phase solid-solution. However, recent reports have indicated that formation enthalpy of MPEAs does exert a considerable contribution to the crystallographic phase stability, and hence the 'high-entropy' composition may not necessarily be the most stable [2–4]. While multicomponent alloys are known for their preference to form intermetallic compounds with complex microstructures [5,6], one of the notable exceptions and widely studied HEAs has been the equiatomic FeNiCoCrMn alloy that is found to assume a single-phase solid-solution [7]. The literature reports are predominantly for two major elemental groups, *viz.*, (a) 3d-transition metal (TM) HEAs [8] comprising of Co, Cr, Cu, Mn, Fe, Mn, Ni, Ti and V, and typically with yield strengths >1000 MPa for temperatures below 600 °C, and (b) refractory HEAs (RHEAs) [9] consisting metals from subgroups IV (Ti, Zr and Hf), V (V, Nb, Ta) and VI (Cr, Mo, W), with extremely high melting temperatures (>1600 °C) [10,11]. Considering a palette of 75 stable elements that could be used to synthesize a 3-to-6 element HEA, a 10% change in the fraction of each element would result in a staggering 592 billion new possible compositions [12]. Such an enormous design space defines a crucial research challenge: *how can we explore the entire landscape of elemental compositions, their microstructural phases and mechanical properties, circumventing the arduous and time-inefficient approach of characterizing each alloy*. Herein, emerges the need for data-centric approaches.

Artificial Intelligence (AI) refers to replicating human intelligence in machines that can be trained to make decisions based on past identical encounters or experiences [13]. If a new encounter is markedly different from prior knowledgebase, then there is a greater likelihood of an error in the decision-making. But this error may be recorded as a lesson learned and appended to the dataset already recorded in the system memory. The amalgamation of past encounters and new experiences improves the decision-making abilities of AI systems. In particular, over the last decade, material scientists have successfully quantified these concepts to a mathematical form [14], ushering an era of data-driven and machine learning (ML) approaches for the screening and discovery of new materials. Data from experimental characterization and computational predictions on material processing, structure, properties and performance, as

* Corresponding author.
E-mail address: bganesh@lehigh.edu (G. Balasubramanian)

https://doi.org/10.1016/j.commatsci.2021.110381
Received 31 October 2020; Received in revised form 25 January 2021; Accepted 8 February 2021
Available online xxx
0927-0256/

available in the literature serve as a resource to generate the required critical information for mining the data and estimating the targeted physical quantities. In the space of HEAs, it is worthwhile to mention some excellent reviews [8,15] that provide exhaustive data useable to construct previously unknown correlations between structure and properties of these complex materials. Nevertheless, challenges arise when such a predictive framework attempts to extrapolate beyond the realm of the available data, making uncertainty quantification very essential.

In this review, we attempt to provide an overview of successful data-intensive approaches accomplished in recent years for discovery and analyses of HEAs. Specifically, we discuss the choice and role of descriptors that facilitate constructing data-guided models for the predictive frameworks. Subsequently, technical challenges associated with such methods with regards to HEAs are presented, and we conclude with potential future directions to overcome the current challenges and robustify this field of research.

2. Descriptors for data-enabled examinations of HEAs

The major steps involved in a ML based interrogation of HEAs are illustrated in Fig. 1. The HEAs and their 'descriptors' are the input and the property of interest is the output. The descriptors for every HEA are its chemical signature or fingerprint. If large volumes of data are available, the ML algorithm leverages the pattern existing within the available data, and estimates the property for an alloy outside the original dataset, given that the HEA comprises of elements and microstructures prevalent amongst the materials represented in the original dataset. It is important to note that construction of a dataset is not a mundane copy-editing activity; the process requires critical scrutiny and expertise to translate the underlying physics associated with the microstructure, material properties and mechanisms, to quantitative features (i.e., descriptors). Likewise, selection of the ML algorithm depends on the size and nature of the dataset, and the desired objective. If the target objective of interest is a continuous physical quantity (e.g., Young's modulus or yield strength), a regression algorithm is typically considered, while for a discreet quantity such as lattice structure, a classification algorithm is employed. Thus, in a nutshell data-analytics is a sequential chain involving the creation of relevant dataset, formulation of descriptors, mapping of the descriptors to the target property using a suitable algorithm, validation of the predictive framework and subsequently augmenting the dataset to improve accuracy of the model output.

Formulation of HEA descriptors: The process of assigning numerical values to the physical or chemical features anticipated to influence the target property, creates feature vectors. Determining the features relevant for a specific material characteristic requires deep knowledge of the material science to formulate features based on the factors that govern the physical or chemical property. For instance, if the objective is to estimate the hardness of a HEA, then a suitable feature is the melting temperature of the alloy because it is an indirect representation of the metallic bond strength [16]. Likewise, Hume-Rothery rules [17] have been adopted as the classic set of features in most ML efforts where the objective is to determine the probability of a HEA composition to form a single-phase solid-solution.

Prior literature also reveals the adoption of thermodynamic and phenomenological routes to predict solid-solution formation. Troparevsky *et. al.* [18] proposed a model to predict elemental combinations that would form single-phase HEA. The underlying assumption of the model was that a given combination of elements will form a single-phase alloy if the enthalpy of formation of all possible binary combinations in this set of the metals assumed values within previously stipulated limits, such that within this range only single-phase HEAs will be formed. The minimum of this range is the entropy of mixing i.e. $-T_{ann}\,\Delta S_{mix}$, where T_{ann} is the annealing temperature used to homogenize the alloy. The upper limit is the highest possible enthalpy that signifies the threshold below which phases do not segregate due to insolubility of any two elements in the composition. The model also provides for customization of alloys by allowing the addition of new metals to attain the desired properties such as density and cost, while maintaining the enthalpies within the said range for single-phase HEAs.

It is widely accepted that with multicomponent alloys, there is always a tendency for the formation of intermetallic compounds [19–21]. Guo *et al.* [22] have delineated a narrow range for mixing enthalpy $\Delta H_{mix} = 4\sum_{i=1,i\neq j}^{n} c_i c_j H_{ij}^{m}$ values, where c_i is the concentration of individual species and H_{ij} is the binary mixing enthalpy for all possible pairs. Certain conditions were required to be simultaneously satisfied for the formation of solid-solution phases in equiatomic multicomponent alloys [23]: $-22\text{kJ/mol} \le \Delta H_m \le 7\text{kJ/mol}$, $0 \le \delta r \le 8.5$, $11\text{J/Kmol} \le \Delta S_{mix} \le 19.5\text{J/Kmol}$, where $\delta r = \sqrt{\sum_{i=1}^{n} c_i\left(1-\frac{r_i}{\bar{r}}\right)^2}$, r_i being the radius of an atom of individual element species, $\bar{r}$ the mean radius of all species, mixing entropy $\Delta S_{mix} = -R\sum_{i=1}^{n} c_i ln c_i$, with R being the gas constant (=8.314 J/K.mol). Yang *et al.* [24] improved upon a thermodynamic model for the prediction of stabilized solid-solution phases as proposed by Takeuchi and Inoue [25], and defined a descriptor $\Omega = \frac{T_m \Delta S_{mix}}{|\Delta H_{mix}|}$, where T_m is the melting temperature of a *n-element* alloy. It was suggested that for a solid-solution to form, $\Omega>1$, while the occurrence of intermetallic compounds would be more likely if $\Omega<1$. Both Ω and δr were posited to play a role in determining the formation of solid-solution but the role of processing variables, such as rapid cooling rate, could also influence the formation of the crystal structures [26]. Ω and δr were calculated for over 130 alloys and it was concluded that a solid-solution was formed when $\Omega \ge 1.1$ and $\delta r \le 6.6\%$. Singh *et al.* [27] coined a new purely geometric descriptor $\lambda = \frac{\Delta S_{mix}}{\delta r^2}$, where δr^2 is analogous to the measure of strain energy. It was suggested that a large value of λ was favorable for the formation of disordered solid-solutions, with empirical reports for single-phase disordered solid-solutions ($\lambda > 0.96$), two phase mixtures ($0.24 < \lambda < 0.96$) and for compounds ($\lambda < 0.24$). Upon analyzing over seventy-six multicomponent alloys, it was shown that λ was strongly correlated to the type of the crystallographic phases relative to ΔH_{mix}, δ, and Ω. King et al. [28] adopted the Miedema's macroscopic atom model [29] to extend Yang's [23] model for multicomponent materials. A new descriptor Φ was conceived, which is the ratio of the Gibbs free energy of a totally disordered solid-solution to that of the most likely intermetallic or segregated binary system and defined as $\Phi = \frac{\Delta G_{SS}}{-|\Delta G_{\max}|}$ where ΔG_{SS} is the change in Gibbs free energy for the formation of a fully disordered solid-solution from constituent elements, ΔG_{max} is maximum magnitude of Gibbs free energy change (lowest for intermetallics and highest for segregation of elements) obtainable from the formation of binary system. A value of $\Phi \ge 1$ indicates the formation of a complete solid-solution whereas a negative value would imply that a solid-solution would not be formed due to a positive value formation enthalpy (endothermic nature). This method was validated against a testing set of 185 alloys with only 16 exceptions encountered.

Tables 1 and 2 summarize a list of ML efforts on HEAs, while Table 3 provides a comprehensive list of widely used descriptors. >90 descriptors have been proposed based on atomic, mechanical and chemical properties of HEAs and environmental factors like temperature and humidity for investigations related to corrosion of alloys. Additionally, Fig. 2 provides a general classification of these features based on their type and length scales, *viz.,* thermodynamic descriptors (T), atomic descriptors (A), physical property descriptors

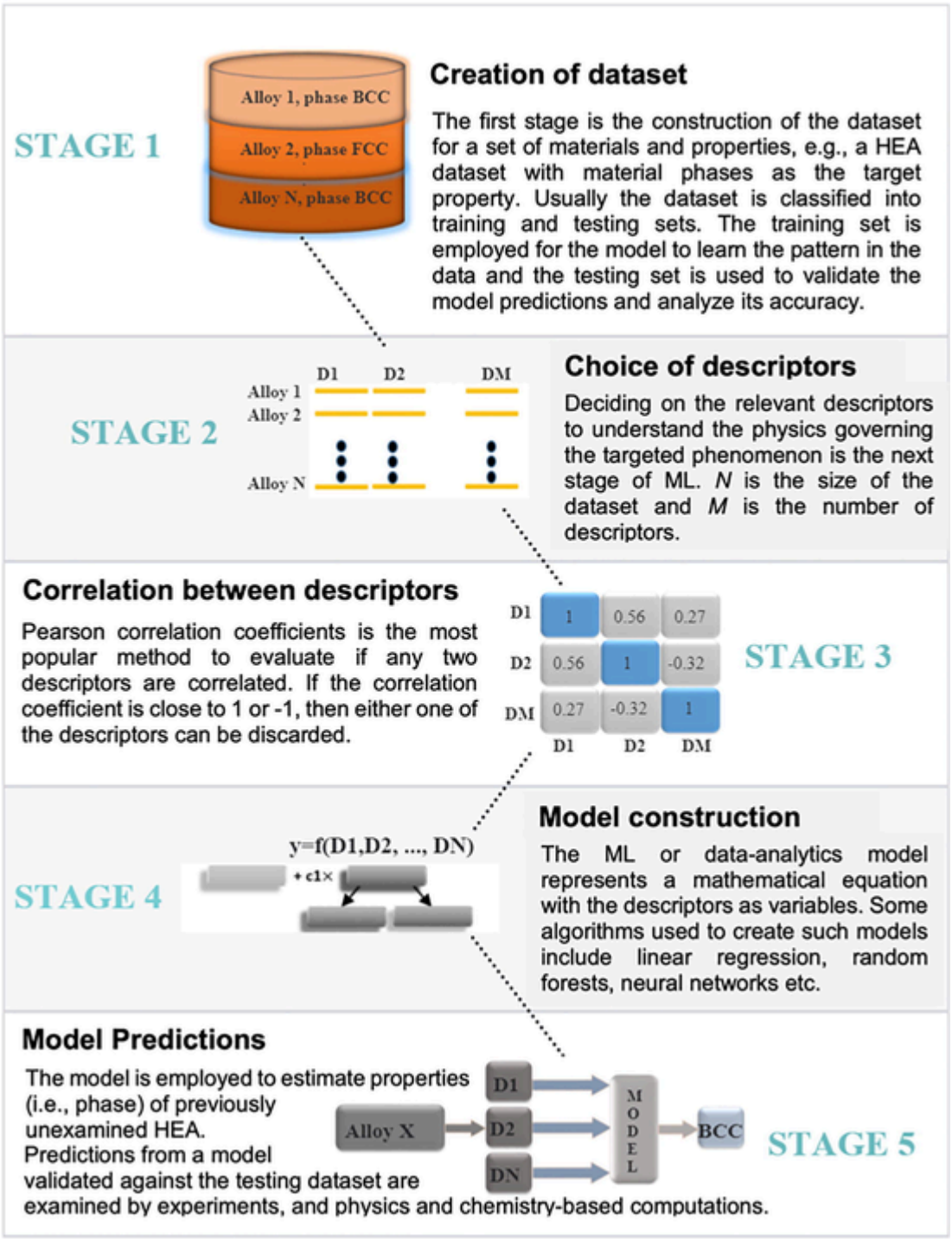

Fig. 1. The key stages associated with a ML and data-driven exploration of HEAs are presented. Stage 1: Collection of data from literature and from existing datasets. Stage 2: Formulation of the descriptors that can correlate with the target property. Stage 3: Down-selection of the feature descriptors using methods such as Pearson coefficients or genetic algorithm. Stage 4: Construction and validation of the ML model employing the testing set to obtain increasingly accurate predictions. Stage 5: Extrapolative predictions and physical scrutiny of the predictions with theory-based computations and experimental characterization.

(P), environmental descriptors (E) and/or chemical compositions (C). The type of descriptor to be used depends on the problem and target application. Early practices of ML in the realm of HEAs employed simple descriptors based on the Hume-Rothery rules and those related to the core principles of HEA, e.g., the mixing entropy. Fig. 3 illustrates the rationale behind considering these parameters as ML descriptors. The mixing entropy and the Hume-Rothery rule of solid solution solubility with the minimum atomic size difference criterion, were presumed to be the decisive factors for determining the phases in HEAs; a lower δ facilitates solid solution (SS) formation as in MoTaTi, while a higher δ promotes coexistence of multiple phases as in TiWZr. The fundamental need for ML lies in solving a multidimensional design problem and streamline experiments. As shown in Fig. 3 (c), the cuboidal region ($-22\text{kJ/mol} \leq \Delta H_m \leq 7\text{kJ/mol}$, $0 \leq \delta r \leq 8.5$ and $\frac{11\text{J}}{\text{Kmol}} \leq \Delta S_{mix} \leq 19.5\frac{\text{J}}{\text{Kmol}}$ [23]) depicts the parametric space for SS formation indicating the latter's dependence on atomic

Table 1
List of ML models available in literature with target property, features used, nature and size of dataset, and the performance metric. Most efforts have used experimental datasets but several reports have employed large volumes of data obtained from prior DFT calculations.

Sl. no.	Refs.	Dataset size and nature	Target application	Features used	Models used	Performance metric
1	Zhang et al. [38]	550, Experimental Phases	Phase prediction	1–26	1–6	Accuracy = 83.9%
2	Li et al. [30]	322, Experimental Phases	Phase prediction	2–5, 10, 27	6	Accuracy = 96.55%
3	Agarwal et al. [52]	-, Experimental Phases	Phase prediction	2–4, 27, 29	7	Accuracy = 84.21%
4	Dai et al. [39]	27500, First-principles	Atomic force field development by predicting interatomic force and energy	Encoded features by NN	8	Developed atomic force field gives fairly good estimation of elastic constants via MD.
5	Zhang et al. [37]	407, Experimental Phases	Phase prediction	2, 3, 8, 10, 30–34	6,9, 10	Accuracy = 97.87%
6	Wu et al. [31]	321, Experimental Phases	Designing eutectic HEAs	29	11	R = 0.995
7	Chang et al. [33]	91, Experimental Hardness	Designing HEAs with high hardness	29	8	R^2 = 0.94MAE = 36 HV. 5 new alloy compositions designed with optimum H.
8	Li et al. [66]	186 ($Cr_xCo_yNi_{1-x-y}$) Experimental UTS	Designing HEAs with high UTS	29	11	R = 98.75New compositions with high UTS were designed.
9	Islam et al. [35]	118, Experimental Phases	Phase prediction	2–4, 6	8	Accuracy = 80%
10	Huang et al. [67]	401, Experimental Phases	Phase prediction	2–4, 6, 27	6, 11, 12	Accuracy = 94.3%
11	Wen et al. [53]	155, Experimental Hardness	Designing HEAs with high hardness	2–4, 6–8, 11, 27, 29, 35–46	12–17	Accuracy ~ 80 %10 new compositions with high H were designed.
12	Liu et al. [40]	1400, First-principles	Effective Hamiltonian development for HEAs	47	8, 13	R^2 = 0.997RMSE = 0.43 meV
13	Kostiuchenko et al. [42]	200, First-principles	Atomic force field development	48, 49	12	Prediction error < 1 meV/atom
14	Pei et al. [47]	1252, Experimental Phases	Phase prediction	6, 10, 27, 50–58	18	Accuracy = 93%
15	Zhang et al. [60]	600, First-principles	Prediction of configurational energy of HEAs	47	19	RMSE ~ 0.6 meV
16	Roy et al. [36]	329, Experimental Phases. 87, Experimental E	Phase and E prediction	2–4, 6–8, 10, 59	20	Phase Prediction accuracy ~ 70%.MAE for E Prediction = 23.59 GPa
17	Zhou et al. [48]	601, Experimental Phases	Phase prediction	1–6, 10, 27, 51	6, 11, 21	Accuracy = 98.9%
18	Rickman et al. [16]	82, Experimental Hardness	Designing HEAs with high hardness	2–4, 8, 10, 60	22, 23	Correlation Factor of model = 0.812 New compositions with high H were designed.
19	Kim et al. [50]	6826, First-principles	B and G prediction	K Prediction-11, 26, 62-64G Prediction- 11, 26, 62, 65, 66	20	< 5% error for B and < 10% error for G predictions
20	Kauffmann et al. [49]	1798, First-principles	Phase prediction	27, 67–73	5	88% agreement with CALPHAD and 75% agreement with LTVC [68]
21	Tancret et al. [69]	322 Experimental Phases	Phase prediction	2, 4, 6, 8, 27, 37, 75–77	18	<0.5% error
22	Bhattacharya et al. [32]	114Experimental corrosion data	Rate constant of oxidation	29, 78–82	5, 12, 20	R^2 = 0.92
23	Yan et al. [34]	306 Experimental corrosion data	Corrosion rate prediciton	29, 83–92	3, 5, 6, 24–26	R^2 = 0.73
24	Arora et al. [70]	11,400 Stacking fault energy data from MD	Stacking fault energy prediction	93,94	25	RMSE = 0.57–2.76 mJ/m^2

Abbreviations for Table 1:
MAE: Mean Absolute Error
Hv: Vicker's Hardness
H: Hardness
UTS: Ultimate Tensile Strength
R: Regression Coefficient
RMSE: Root Mean Square Error
E: Young's Modulus
B: Bulk's Modulus
G: Shear Modulus
SFE: Stacking fault energy
MD: Molecular dynamics.

Table 2
Machine Learning Models as used in Table 1.

Sl. No.	Abbreviation	ML Model	Type of ML Task
1	Lda	Linear discriminant analysis	Classification
2	NBayes	Naïve Bayes classifier	Classification
3	Dtree	Decision Tree	Classification
4	Nnet	Neural Network classifier	Classification
5	RF	Random Forest	Classification
6	SVM	Support Vector Machine	Classification
7	ANFIS	Adaptive Neuro Fuzzy Interface System	Classification
8	NN	Neural Network	Regression
9	MLP	Multilayer Perceptron	Classification
10	GBDT	Gradient Boosting Decision Tree	Classification
11	ANN	Artificial Neural Network	Classification
12	KNN	K-nearest neighbours	Regression
13	LIN	Linear Regression	Regression
14	POLY	Polynomial Regression	Regression
15	SVR	Support Vector Regression	Regression
16	CART	Regression Tree Model	Regression
17	BPNN	Back Propagation Neural Network Model	Regression
18	GPC	Gaussian Process Classification	Classification
19	BR	Bayesian Regularized Regression	Regression
20	GB	Gradient Boost	Regression
21	CNN	Convolutional Neural Network	Classification
22	CCA	Canonical Correlation Analysis	Regression
23	GA	Genetic Algorithm	Regression
24	MLR	Multiple Linear Regression	Regression
25	RR	Ridge Regression	Regression
26	eXGB	eXtreme Gradient Boosting	Regression

and thermodynamic properties, thereby justifying their inclusion as ML descriptors.

While ΔH_m, ΔS_m, *VEC* and δr are typically considered as features in most ML efforts, additional descriptors like Ω and Φ have been shown to improve the performance of ML models for classifying phases or even to predict a physical or mechanical property for HEAs. A representative set of such descriptors employed in the literature are presented in Fig. 4. The relative accuracy of an atomic descriptor (r_d, denoted as δr in this work) in distinguishing BCC, FCC and NSP (not forming single phase solid solution) phases for a set of 322 as cast alloys is compared against a thermodynamic descriptor (ΔH denoted as ΔH_m in this work) in Fig. 4(a) [30]. Nevertheless, it was realized that these two parameters alone cannot classify the phases with sufficient accuracy. Hence, additional sophisticated thermodynamic (Ω) and atomic descriptors (VEC) were included for the phase classification problem. Such modifications rendered an improved classification of the BCC and FCC phases, although the NSP alloys remained distributed across the other two groups as shown in Fig. 4(b) [30]. A reasonably accurate categorization between FCC and BCC phases is obtained by using physical descriptors like density (ρ) and mean melting temperatures (T_m), as can be observed from Fig. 4 (c) [30]. The content map of Al-Co-Cr-Fe-Ni family of alloys using chemical composition to delineate the eutectic high entropy alloy (EHEA) formation region, reproduced in Fig. 4(b), is based on experimental data [31]. As observed, the eutectic composition is obtained when Al concentration lies between 15 and 20% and Cr content ranges from 0 to 25%. Likewise, the chemical composition of alloys has been used as a descriptor in several recent works [32,33] from which ML models can learn and predict a specific crystallographic phase or a physical property. Fig. 4(e,f) illustrate the use of ML (RF, GBDT and XGBoost techniques)by examining environmental data such as temperature and humidity, of a specific geographical location combined with the chemical composition descriptor for predicting corrosion rate of low-alloy steels SPA-H, SMA490 and SM490A using [34]. Detailed discussions for each type of these descriptors are provided in the following sections with citations to representative reports that have used those descriptors.

Models based on thermodynamic descriptors: A large number of ML efforts use thermodynamic parameters, denoted as T in Table 3, to predict the HEA crystallographic phases. Islam et al. [35] proposed the conversion of thermodynamic properties and Hume-Rothery rules into ML usable descriptors, and correlated the features to a predictive model for the phases of previously uncharacterized HEAs. Subsequent efforts that followed considered other material properties, such as Young's modulus, hardness, ultimate tensile strength (UTS) and stacking fault energy (SFE), using similar approaches. The enthalpy (ΔH_m) and entropy (ΔS_{mix}) of mixing together with the valence electron concentration(VEC) were found to have the highest relevance towards predicting phases [35], but another effort identified the mean melting point and electronegativity difference to be the most significant in determining phases [36]. In a related effort [37] aimed at estimating the phases formed in a HEA, the accuracy of model increased from 75 to 97% through use of an increased set of features by sub-categorization, e.g., ΔH_m as a sum of the mixing enthalpies for amorphous phase (H_{AM}), the intermetallic compound (H_{IM}), the solid-solution (H_{SS}) and the liquid phase (H_L). On the contrary, limiting the number of features from 70 to merely 4 contributed to an optimal prediction accuracy [38].

It is very important to understand and realize that the above model predictions are of an empirical nature because ML algorithm is simply a mathematical framework and the model output is predominantly rooted to the type of fitting function employed. The latter depends on the type of data, with no guidance from the physics of the underlying mechanisms.

Models based on atomic descriptors: Zhang et al. [38] employed 58 atomic scale descriptors, in addition to thermodynamic descriptors, to select the best model-descriptor combination. Genetic algorithm (GA) was employed to down select from a pool of 9 models and 70 descriptors. Eventually, the SVM with 4 down-selected features, *viz.,* the average atomic number, the difference in electronegativities, covalent radii and the boiling temperature emerged as the best model-descriptors combination with over 90% accuracy in predicting the crystal structure of HEAs.

Table 3
List of descriptors on HEAs employed in ML approaches. A, C, E, P and T denote atomic, chemical composition, environmental, physical and thermodynamic parameters.

Sl. No.	Description	Abbreviation	Formula	Type	Refs.
1	Mean Atomic Size	r	$\bar{r} = \sum_{i=1}^{n} c_i r_i$	T	
2	Atomic Size Mismatch	δr	$\delta r = \sqrt{\sum_{i=1}^{n} c_i \left(1 - \frac{r_i}{\bar{r}}\right)^2}$	A	[24]
3	Mixing Entropy	ΔS_m	$\Delta S_m = -R\sum_{i=1}^{n} c_i ln c_i$	T	[71]
4	Mixing Enthalpy	ΔH_m	$\Delta H_m = 4\sum_{i=1, i\neq j}^{n} c_i c_j H_{ij}^m$	T	[22]
5	Electronegativity Mismatch	χ	$\bar{\chi} = \sum_{i=1}^{n} c_i \chi_i$	A	[72]
6	Electronegativity Mismatch	δχ	$\delta\chi = \sqrt{\sum_{i=1}^{n} c_i \left(1 - \frac{\chi_i}{\bar{\chi}}\right)^2}$	A	
7	A geometrical parameter	λ	$\lambda = \frac{\Delta S_{mix}}{\delta^2}$	A	[27]
8	Parameter for predicting the solid-solution formation	Ω	$\Omega = \frac{T_m \Delta S_{mix}}{\lvert \Delta H_{mix} \rvert}$	T	[24]
9	Average number of Itinerant Electrons per Electrons	C_v	$C_v = \frac{\sum_{i=1}^{n} c_i e_{vi}}{\sum_{i=1}^{n} c_i Z_i}$	A	[38]
10	Melting Point	MT		P	
11	Cohesive Energy	CE		A	
12	Compression Modulus	CM		P	
13	First Ionization Energy	FIE		A	
14	Second Ionization Energy	SIE	$\overline{property} = \sum_{i=1}^{n} c_i property_i$	A	
15	Third Ionization Energy	TIE	$\delta(property) = \sqrt{\sum_{i=1}^{n} c_i \left(1 - \frac{property_i}{\overline{property}}\right)^2}$	A	
16	Work Function	WF		A	
17	Quantum Number	QN		A	
18	Column in the Periodic Table	C		A	
19	Relative Atomic Mass	RAM		A	
20	Atom Volume	VA		A	
21	Atomic Environment Number	AEN		A	
22	Chemical Potential	CP		A	
23	Effective Nuclear Charge	NCE		A	
24	Valence Electron Distance	DVE		A	
25	Core Electron Distance	DCE		A	
26	Density	D		P	
27	Valence Electron Concentration	VEC		A	
28	Parameter for predicting single-phase	φ	$\Phi = \frac{\Delta G_{SS}}{-\lvert \Delta G_{\max} \rvert}$	T	[28]
29	Composition of elements	Atomic %	–	C	[34]
30	Mixing Enthalpy of amorphous phase	H_{AM}	Calculated by Midema's theory [73] and Ouyang's model [74]	T	
31	Formation Enthalpy of intermetallic compound phase	H_{IM}		T	
32	Formation enthalpy of solid-solution phase	H_{SS}		T	
33	Mixing Enthalpy of liquid phase	H_L		T	
34	Elastic Energy of Alloy	H_E		P	
35	γ parameter	γ	$\gamma = \left(1 - \sqrt{\frac{(r+r_{\min})^2 - r^2}{(r+r_{\min})^2}}\right) / \left(1 - \sqrt{\frac{(r+r_{\max})^2 - r^2}{(r+r_{\max})^2}}\right)$	A	[53]
36	Mismatch of local electronegativity	D.χ	$D.\chi = \sum_{i=1}^{n}\sum_{j=1, i\neq j}^{n} C_i C_j \times \lvert \chi_i - \chi_j \rvert$	A	
37	Number of itinerant electrons	e/a	$\frac{e}{a} = \sum_{i=1}^{n} C_i \times (e/a)_i$	A	
38	Modulus mismatch in strengthening model	η	$\eta = \sum_{i=1}^{n} \frac{C_i \times \frac{G_i - G}{G_i + G}}{1 + 0.5 \times \left\lvert C_i \times \frac{2(G_i - G)}{G_i + G} \right\rvert}$	A	
39	Local size mismatch	D.r	$D.r = \sum_{i=1}^{n}\sum_{j=1, i\neq j}^{n} C_i C_j \times \lvert r_i - r_j \rvert$	A	
40	Energy term in strengthening model	A	$A = G \times \delta r \times (1+\mu)/(1-\mu)$	A	
41	Peierls-Nabarro factor	F	$F = 2G/(1-\mu)$	A	
42	Six square of work function	w	$w = \sum_{i=1}^{n} \left(C_i \times w_i\right)^6$	A	
43	Shear Modulus	G	$G = \sum_{i=1}^{n} C_i \times G_i$	P	
44	Difference of shear moduli	δG	$\delta G = \sqrt{\sum_{i=1}^{n} c_i \left(1 - \frac{G_i}{G}\right)^2}$	P	
45	Local modulus mismatch	D.G	$D.G = \sum_{i=1}^{n}\sum_{j=1, i\neq j}^{n} C_i C_j \times \lvert G_i - G_j \rvert$	P	
46	Lattice distortion energy	μ	$0.5 \times E \times \delta r$	A	
47	Short-range order parameters	SRO	$\alpha_m^{AB} = 1 - \frac{P_m^{A\|B}}{c_A}$	A	[41]
48	Number of nearest neighbors around an atom or the range of interaction	n	Check reference for formulation	A	[42]
49	Approximation rank controlling number of fitting parameters	r	Check reference for formulation	A	
50	Atomic Weight	AW		A	

Sl. No.	Description	Abbreviation	Formula	Type	Refs.
51	Bulk Modulus	B		P	
52	Covalent Radius	R_c	Analogous to 14 & 15	A	
53	Crust Abundance	CA		E	
54	Neutron Cross Section	NSC		A	
55	Fusion Heat	FH		P	
56	Boiling Point	BP		P	
57	Thermal Conductivity	K		P	
58	Vaporization Heat	V		P	
59	Lattice Constant	a		A	
60	Young's Modulus	E		P	
61	Holder mean	μ	$\mu_p(x) = \left(\left(\sum_{i=1}^{n} W_i \right)^{-1} \left(\sum_{i=1}^{n} W_i x_i^p \right) \right)^{1/p}$		[50]
62	Elemental group number	μ_4(g)	Formulation using 61	A	
63	Element atomic radius	μ_1(r)	Formulation using 61	A	
64	Holder mean Electronegativity	$\mu_4(\chi)$	Formulation using 61	A	
65	Element atomic radius	μ_2(r)	Formulation using 61	A	
66	Holder mean Electronegativity	$\mu_4(\chi)$	Formulation using 61	A	
67	BCC fraction at 1600 K	f_{bcc}	From CALPHAD	P	[49]
68	Single-phase start temperature of FCC	T_{fcc}	From CALPHAD	P	
69	Range of covalent radii in composition	$R_{covalent}$	$R_{covalent} = R_{max}$-R_{min}	A	
70	FCC fraction at 1200 K	F_{fcc}	From CALPHAD	P	
71	Fraction weighed mean of covalent radii	Fwm ($R_{covalent}$)	$fwm\left(R_{covalent}\right) = \sum_{i=1}^{n} c_i (R_{covalent})_i$	P	
72	Range of electronegativity in composition	R_χ	$R_\chi = \chi_{max} - \chi_{min}$	A	
73	Single-phase start temperature of BCC	T_{bcc}	From CALPHAD	P	
74	Gibbs energy	G_{energy}	$G_{energy} = H - T.S$	T	[69]
75	Ratio of melting temperature and critical spinodal decomposition temperature (T_{sc})	μ_t	$\mu_t = \frac{T_m}{T_{sc}}$	P	
76	Average Bulk modulus mismatch	K_m	Refer [75] for formulation	P	
77	Average interatomic spacing mismatch	S_m	Refer [75] for formulation	A	
78	Phase	–	α, near- α, α + β	P	[32]
79	Oxidation Temperatures	–	Temperature in Kelvin	E	[34]
80	Oxidation Time	–	Time in Hours	E	
81	Oxygen and water concentration	–	atm	E	
82	Mode of oxidation	–	Isothermal/cyclic oxidation	E	
83	Relative humidity	RH	15–79% range	E	
84	Duration of sunshine	SUNSHINE	Duration in hours	E	
85	Time of wetness	TOW	3700–5300 h	E	
86	Precipitation	PRECIPIT	1100–2300 mm	E	
87	Wind velocity	WIND_VEL	1.1–39.5 m/s	E	
88	Solar radiation	SOLAR	4100–6600 Mj/m^2	E	
89	Ultraviolet radiation	UV	180–370 Mj/m^2	E	
90	Chloride deposition rate	CHLORIDE	2–55 mg NaCl/m^2.d	E	
91	SO_2 deposition rate	SO_2	1.8–6.1 mg SO_2 /m^2.d	E	
92	Exposure period	TIME	1–10 years	E	
93	Species involved in binary bonds	–	Ni-Fe, Ni-Ni, Fe-Fe	A	[70]
94	Bond length	–	1 Å-7 Å	A	

With regards to forcefield development, Dai et al. [39] produced a deep learning potential for the $(Zr_{0.2}Hf_{0.2}Ti_{0.2}Nb_{0.2}Ta_{0.2})C$ HEA employing the encoding capability of deep neural networks to develop atomic descriptors that were free from human assigned features. In other words, instead of employing the stereotypical thermodynamic or atomic quantities, the local environments of atoms encoded by the neural network were mapped to descriptors. The developed potential was used to predict the elastic constants that agreed with existing reports within ~10% margin. Liu et al. [40] transformed Warren-Cowley short-range order parameters [41] into features that were then fitted to a neural network to obtain reasonably accurate estimates for the effective Hamiltonian of different HEAs, e.g.,. for NbMoTaW a root mean square error of 0.43 meV was noted. Likewise, Kostiuchenko et al. [42] used short range parameters for developing interatomic potentials but additionally considered lattice relaxation. They demonstrated with computational simulations that with a static lattice approach, the NbMoTaW alloy underwent a phase transition at 600 K but with the inclusion of lattice distortions, the HEA remained a solid-solution consistent with other research efforts [43–46]. Thus, inclusion of lattice relaxation successfully improved the accuracy of ML enabled potential based on short range order parameters. These efforts build the promise of the capability of atomic descriptors towards data-enabled predictions for HEAs.

Models based on physical descriptors: Unlike its predecessors, a recent work [47] adopted certain physical descriptors like bulk modulus, thermal conductivity, heat of boiling and fusion for predicting the crystallographic phases. Through a larger study [48], 4 ML models were examined on the same sets of data containing 1, 3, 4 and 13 features, with only the latter one including bulk modulus as a physical descriptor. As anticipated, the dataset with the greater number of features had the highest accuracy (97.8%) in predicting phases, and a low bulk modulus favored the formation of solid-solu-

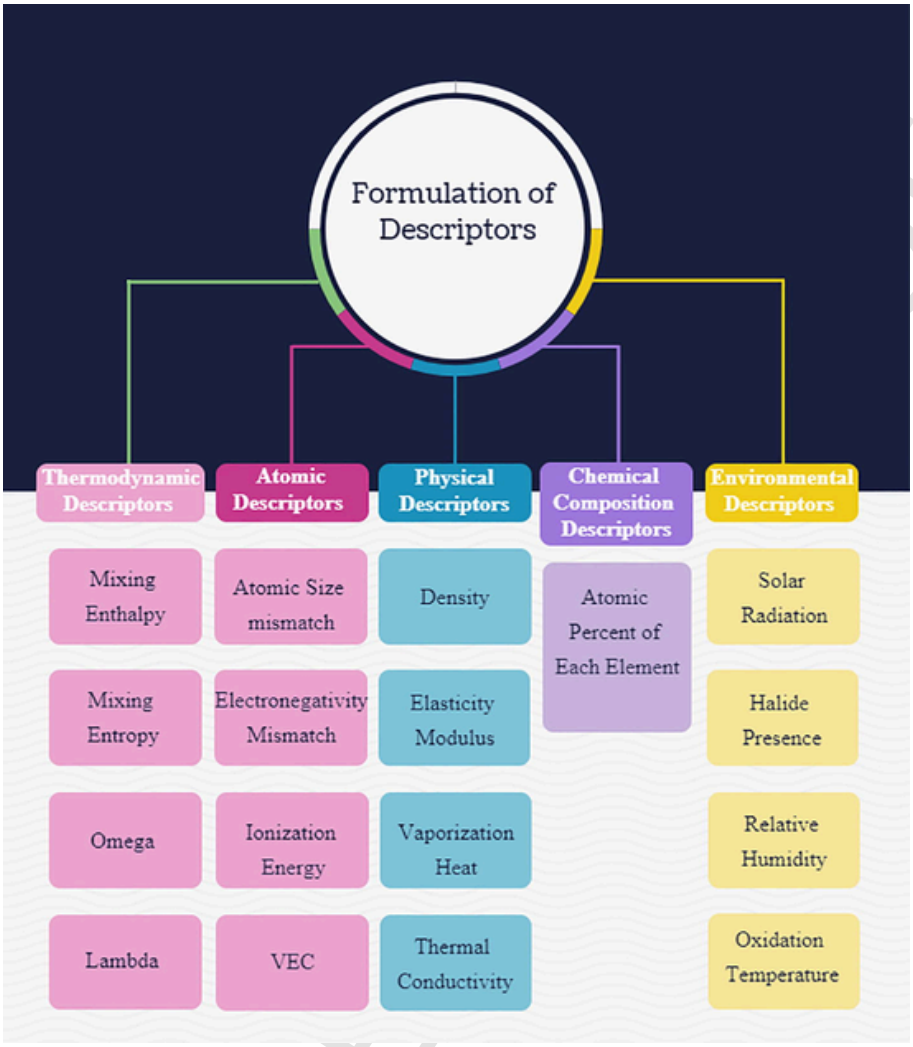

Fig. 2. A typical categorization of HEA descriptors is presented together with select examples. While thermodynamic descriptors represent information about the fundamental energy and transport relevant parameters, atomic descriptors embody the lattice and electronic properties of a HEA. Physical descriptors denote macroscopically measurable physical properties. For the problem on predicting corrosion behavior in HEAs, chemical composition (atomic fractions of constituent elements) and environmental factors (harshness of the surroundings) have been employed. These descriptors are further elaborated in Table 3.

tions. Rickman et al. [16] used Young's modulus asymmetry (representing stiffness) and mean melting temperature (signifying bond strengths) as features in a data-analytics model to propose compositions with high hardness. They employed GA to generate virtual alloy candidates that progressed from one generation to the next depending upon the magnitude of their "fitness" to obtain high hardness. The measure of hardness was performed using a canonical correlation analysis (CCA) trained on a dataset of 82 alloys. The approach resulted in the discovery of two alloys ($Co_{33}W_{07}Al_{33}Nb_{24}Cr_{03}$ and $Ti_{18}Ni_{24}Ta_{12}Cr_{22}Co_{24}$) with exceptional hardness (>1000 HV). The exceptional performance of this model can be attributed to the choice of features and the model, as well as the diversity of the dataset that was not restricted to one family of metals rather included a large palette of metals broadening the range of achievable values for the target property.

While data from experimental measurements have been the preferred choice for training the ML algorithm, some groups [49,50] have utilized data derived from density-functional theory (DFT) calculations (such as from the Materials Project database [51]) for enabling phase and mechanical property predictions. The advantage of using a DFT dataset is the availability of a large volume of fairly accurate data, and the relative ease of data generation without the time and resource requirements of physical experiments. Kim et al. [50] attempted to construct a model using mean of physical parameters (e.g. cohesive energy, density) in conjunction with thermodynamic descriptors to predict bulk and shear moduli. Results from their work, representing a tight coupling of DFT, ML and experiments, demonstrated a remarkably low error of ~5% in predicting the bulk modulus.

In brief, all strategies discussed above unanimously corroborate that correlating thermodynamic and physical parameters as features produces models with an improved predictive accuracy as compared to those using only a single category of descriptors.

Models based on chemical composition descriptors: To better understand the impact of chemical composition on a target property, researchers have attempted to incorporate the atomic percentages of the HEA elements as features. Agarwal et al. [52] performed phase prediction using two models, one that used elemental composition as descriptors and the other with thermodynamic descriptors. The testing accuracy was surprisingly higher for the composition-guided model (84.21%) relative to the other (80%). The higher accuracy could be attributed to the training dataset being dominated by the CoCrCuFeNi alloy family, and consequently the effect of elemental fractions on phase formation was distinctly identifiable. Wu et al.

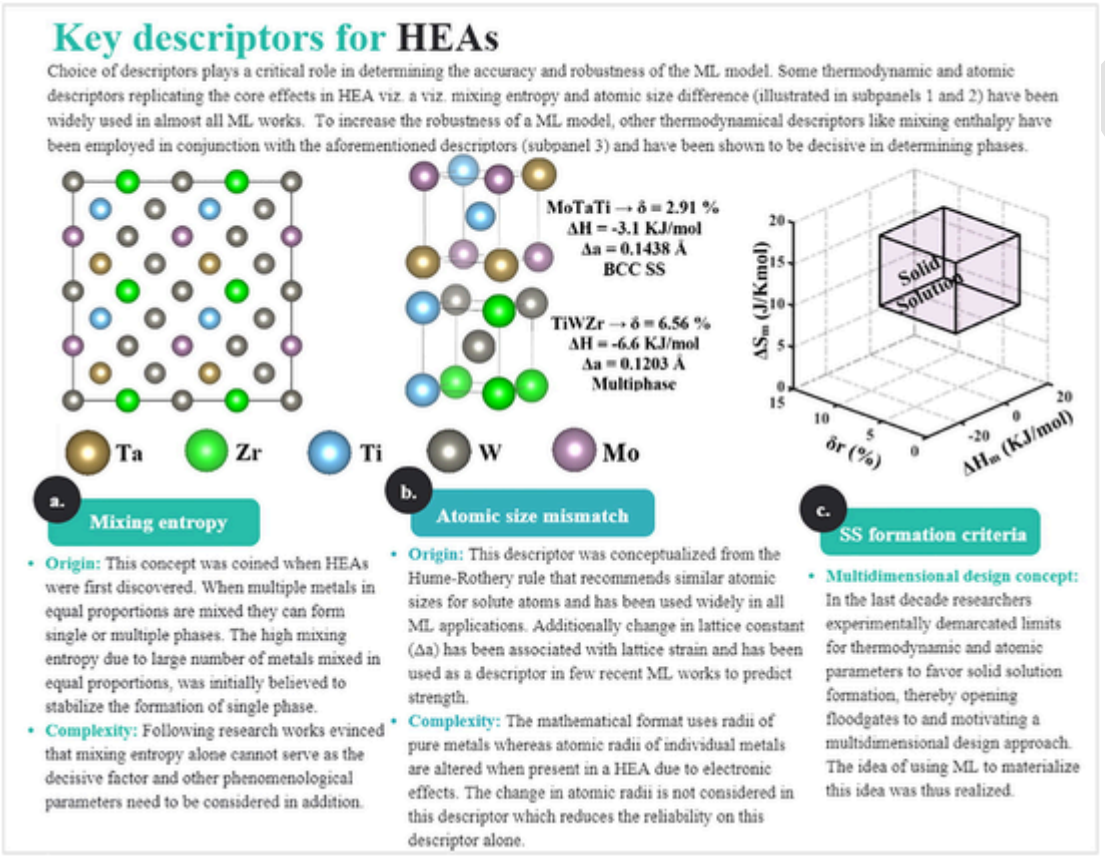

Fig. 3. Machine learning guided material phase classification using thermodynamic and atomic descriptors. Initial ML efforts to classify phases of HEAs used descriptors based on core effects such as (a) mixing entropy and (b) the Hume-Rothery rule for solid solutions employing minimum atomic size difference. These descriptors were initially considered to be decisive for determining the phases in HEAs. For instance, as illustrated in subpanel (b), a lower δ favors solid solution (SS) formation in MoTaTi and a higher δ results in the coexistence of multiple phases in TiWZr. The underlying motivation for ML driven classification of phases and prediction of properties arises from the need for multidimensional design of materials and processes for experiments. As presented in subpanel (c), the highlighted cuboidal region ($-22kJ/mol \leq \Delta H_m \leq 7kJ/mol$, $0 \leq \delta r \leq 8.5$ and $\frac{11J}{Kmol} \leq \Delta S_{mix} \leq 19.5\frac{J}{Kmol}$) [23] denotes the domain for SS formation and corroborates its dependence on the atomic and thermodynamic parameters, thereby justifying there use as descriptors in ML algorithms.

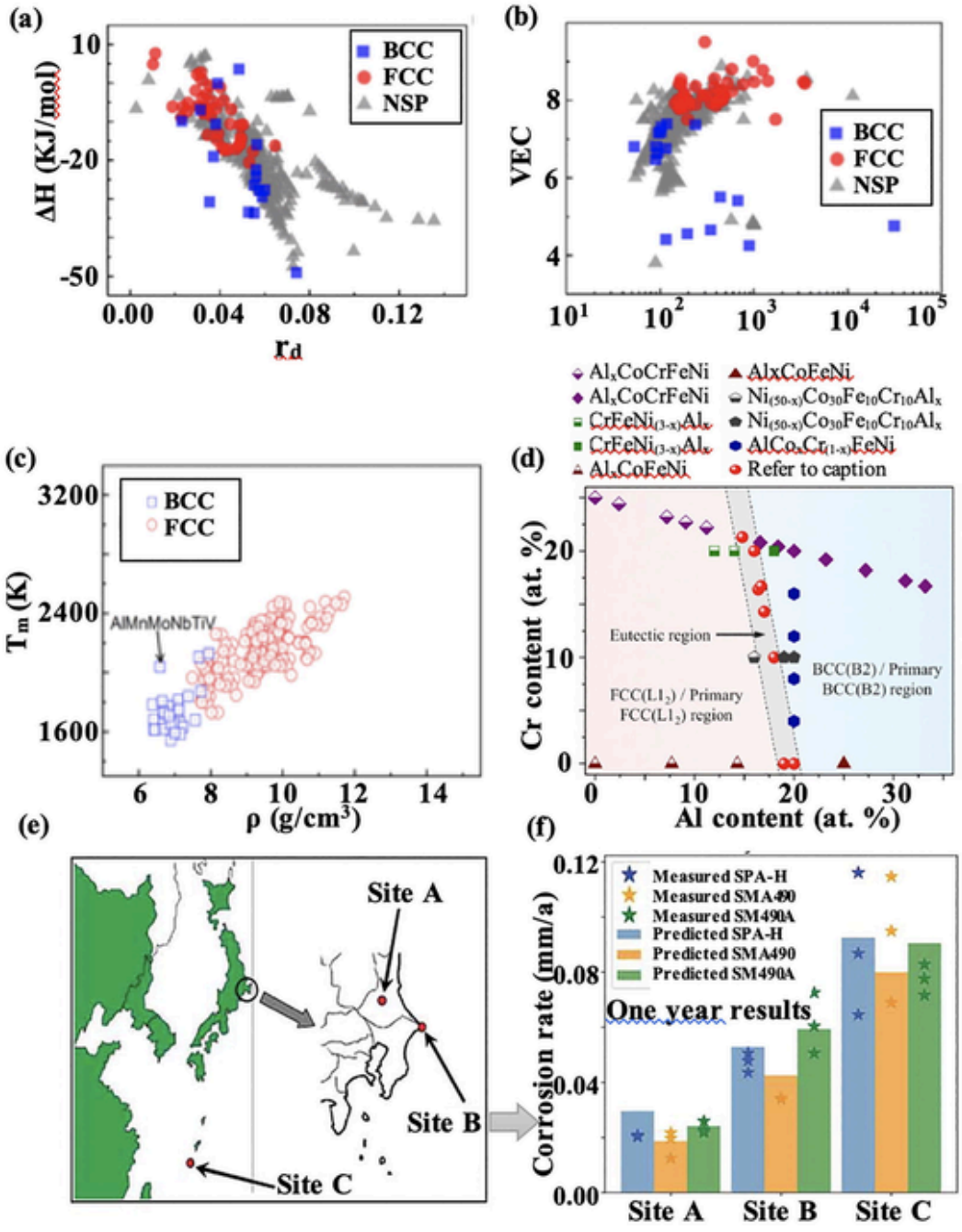

Fig. 4. A collated and representative illustration of ML from various categories of descriptors. (a) Statistical data of 322 as cast alloys comparing an atomic descriptor (r_d, denoted as δr in this paper) and a thermodynamic descriptor (ΔH denoted as ΔH_m in this paper) for their relative proficiency in accurately distinguishing between BCC, FCC and NSP (not forming single phase solid solution) crystal structures [30]. It was noted that these parameters alone cannot classify the phases with sufficient accuracy. Hence, as shown in (b), another pair of thermodynamic (Ω) and atomic (VEC) descriptors employed for the classification of phases revealed an improved predictability although the NSP alloys still overlapped with the other two lattice types [30]. (c) When using physical quantity descriptors like density (ρ) and mean melting temperatures (T_m), a definitive classification between FCC and BCC alloys is observed [30]. (d) Content map of Al-Co-Cr-Fe-Ni alloy compositions using the chemistry of the alloy to infer the eutectic high-entropy alloy (EHEA) formation based upon experimental data [31]. The eutectic composition identified by red circles is obtained when Cr content ranges from 0 to 25% and Al occupies between 15 and 20%. The chemical composition can be used as a descriptor from which models can learn to predict a specific phase formation or a physical property. (e) A ML model developed by learning on environmental data combined with the chemical composition for (f) prediction of corrosion resistance of low-alloy steels SPA-H, SMA490 and SM490A using RF, GBDT and XGBoost models (each denoted by a 'star') [34]. A good fit (with $R^2 = 0.94$) was obtained by training the GBDT model on the dataset. *Adapted with permission from ref.* [30] *((a), (b) and (c)), ref.* [31] *((d)) and ref.* [34] *((e) and (f)).*

[31] used a similar methodology to predict eutectic HEAs, employing chemical composition as the feature set to identify the most critical element in the HEA that affected the phase constitution, and to find the miscibility of other elements with this critical element. Chang et al. [33] trained an artificial neural network on a HEA dataset to design alloys with targeted high hardness. They coupled a simulated annealing (SA) algorithm with neural network and predicted a set of three HEAs composed of Al-Co-Cr-Cu-Fe-Ni elements with high hardness (>600 HV). Building upon the same technique, but additionally leveraging thermodynamic descriptors, Wen et al. [53] employed utility functions to minimize the number of screenings through the ML algorithm and filter out novel HEA composi-

tions with exceptional hardness, and in the process recommending 10 new compositions with high hardness (>800 HV). The plausible reasons for improvement could be the stronger correlation of thermodynamic parameters to the hardening mechanisms relative to the chemical composition descriptors.

Models based on environmental factors as descriptors: In the domain of corrosion/oxidation behavior of HEAs a number of studies have been conducted on the oxide formation [54,55] and reduction tendency [56] in HEAs. Nonetheless, only a few ML models have been developed and they utilize environmental factors like temperature, humidity, solar and UV radiations, chloride and SO_2 content in the electrolyte, to name a few. None of the thermodynamic or atomic descriptors have been considered relevant in such models. Bhattacharya et al. [32] constructed a model to predict the parabolic rate constant for the oxidation of Ti alloys at elevated temperatures, using a combination of environmental factors and chemical composition descriptors. They compiled experimental corrosion data and assigned the target property as the corrosion rate constant. Interestingly, the natural logarithm of the rate constant was set as the target property to reduce the skewness given the extensive variations in data. A notable precision was realized in the predictions using a gradient boosting model (coefficient of determination $R^2 = 0.92$). Along the similar lines, Yan et al. [34] constructed a model to simulate corrosion in a marine environment for low alloy steels. The environmental factors included additional features such as the presence of halide and SO_2, solar radiation and wind velocity. The results suggested that the chemical composition followed by chloride deposition rate and precipitation were the most significant features in determining the 3-year corrosion rate. Beyond that duration, the rust layer became sufficiently thick and environmental humidity becomes the most important factor in determining the corrosion rate.

In the review by Qiu et al. [57], the role of different alloying elements on the corrosion of HEAs was highlighted. An increasing fraction of Al increases the corrosion rate in depassivating environments (conditions that promote dissolution) whereas a protective film is formed by Al (e.g., in Ni-based corrosion resistant alloys) in passivating environments. The presence of Ti in an alloy, facilitates the formation of a TiO_2 protective film on the surface that impedes corrosion. A similar phenomenon is observed when Cr is added to an alloy; Cr_2O_3 protective layer is formed on the surface of the alloys which provides an enhanced corrosion resistance. In certain cases, Mo can conditionally provide protection from corrosion by forming a passive layer, but also Mo has been found to form Mo and Cr rich sigma (σ) phases that could remove Mo from the HEA matrix, making the alloy vulnerable to corrosion. Cu containing HEAs undergo a higher mass loss rate as Cu favors elemental segregation, and thus facilitates higher localized corrosion. Passivation of the alloy due to the formation of an oxide film may be hindered by the presence of halides, e.g., chloride ions. The widely accepted Point Defect Model (PDM) [58] explains the growth and disintegration of passive films on a metal surface via penetration of chloride ions at the atomic scale.

The pH of the environment plays a significant role in determining the corrosion rate. When an alloy is immersed in an aqueous solution, one or more metals ionize and form their precipitates. The reactions that occur are (i) $M \rightarrow M^{n+} + ne^{-1}$, which is the anodic or oxidation reaction and (ii) $nH^+ + ne^{-1} \rightarrow (n/2)H_2$ (gas) which is the cathodic or reduction reaction. To find the tendency of corrosion in the solution, it is important to know the potentials of the anodic ($E_{M^{++}/M}$) and cathodic reactions ($E_{H^{++}/\frac{1}{2}H_2}$). Then, the overall cell potential

$$\Delta E = \left(E_{H^{++}/\frac{1}{2}H_2} - E_{M^{++}/M}\right)$$

is

$$= \left(E^0_{H^{++}/\frac{1}{2}H_2} - E^0_{M^{++}/M}\right) - \frac{2.303RT}{nF}\left(2 \times pH + \log\left[M^{++}\right]\right)$$

[59]. This formulation suggests that the ΔE should be positive for ΔG to be negative so that the reaction is spontaneous in accord with the equation $\Delta G = -nFE$. Thus, under acidic conditions, a lower pH (acidic environment) will increase the ΔE and consequently increase the corrosion rate. Since pH of the test environment plays a critical role in determining the corrosion rate, it is a recommended descriptor for future ML efforts to predict corrosion behaviors.

Hierarchical Significance of Descriptors and Prospective ML Applications: A significant aspect of the ML framework is the numerical representation of descriptors necessary for constructing a quantitative prediction model. As mentioned before, formulating numerical descriptors requires a thorough knowledge of the domain to carefully assess the factors that may be related to the target property. Since the numerical descriptor serves to represent the real material, it may also be referred to as the signature of the material. Depending on the problem and the level of accuracy desired in the predictions, the signature can be formulated at varying levels of detail. If the goal is to understand the factors underlying a complex phenomenon such as phase formation and accuracy is less critical, then gross-level signatures may be defined. For instance, mixing entropy (ΔS_m) because it is a function of simply the concentrations of the elements in the alloy and not any inherent attributes of a specific alloy, i.e., two different alloys with same number of elements in identical proportions, will have equal ΔS_m even though the actual alloy compositions are different [4]. Other descriptors like environmental conditions (79–92 in Table 3) employed when predicting corrosion resistance, may also be considered as gross-level signatures because they represent the factors external to the alloy. On the other hand, if the objective is to determine material properties with reasonable level of accuracy, e.g., hardness, signatures with atomic scale information would be required. Descriptors like electronegativity and melting temperature that is reflective of the bond strength, are examples of atomic-level signatures [39]. In short, it is vital that material signatures describing all relevant attributes are incorporated within the dataset for accurate ML analysis. A pool of >94 descriptors currently available in the literature (Table 3) encompass signatures ranging from subatomic to macroscopic levels.

3. Critical challenges and strategies to overcome

First, since ML predictions are fundamentally statistical in nature, there is always an uncertainty associated with the predictions. Second, the predictions are highly unreliable if the testing data lies outside of the domain of the training data, i.e., extrapolative exploration. The extent to which a new alloy is located beyond of the domain of the training data can be quantified using uncertainty. Amongst the multiple techniques that exist for uncertainty quantification (UQ), e.g. Bayesian information criterion (BIC) [60] or the Gaussian process regression [61], have been implemented sparingly [62,63]. Alternate strategies include comparing the predictions of multiple closely related models (e.g., decision trees and random forests) with slightly different constructions [64]. We recommend that UQ should become a standard module of all data-enabled models as it would provide a transparent evaluation of the real accuracy and performance of a ML model.

A beneficial approach to exploit the capability of ML in exploring new design spaces for HEAs is by finding the "function maxima" or via "inverse mapping" [65], the objective being to filter materials that satisfy a target criteria. For ML as applied to HEAs, a forward pathway has been adopted to compute the target property by feeding the feature values of an unknown material, but determining the chemical composition from a preset target property value by reverse mapping is always a challenge. Some efforts [16,53,66] have

adopted the strategy of iteratively constructing compositions that evolve with successive generations simulating the process of reproduction and mutation. Though these efforts represent a major development, inverse mapping in a true sense still remains an elusive task due to the difficulty of identifying a maxima of a function in a high (>3) dimension space.

4. Summary and future directions

In this review, we present various class of descriptors for data-guided models used to explore HEAs. In particular, models based on thermodynamic descriptors have been widely used for phase identification using quantities such as enthalpy (ΔH_m) and entropy (ΔS_{mix}) of mixing. More recently, thermodynamic descriptors have been modified to capture finer phenomenological details via subcategorization, e.g., ΔH_m as a sum of the mixing enthalpies for amorphous (H_{AM}), the intermetallic compound (H_{IM}), the solid-solution (H_{SS}) and the liquid (H_L) phases. Models based on atomic descriptors that encode short range order parameters into numeric features have been used for developing interatomic forcefields. Physical quantities such as melting point and Young's modulus asymmetry have been adopted to predict properties like hardness, because these descriptors most closely resemble the interatomic bond strengths and have demonstrated reasonable predictive accuracy. The chemical composition and the environmental factors have been considered as descriptors for predicting oxidation resistance and corrosion rates in compositionally complex alloys. Their suitability for such investigations is attributed to corrosion being a surface phenomenon dependent on the environmental harshness and the elements present in the HEA. Some elements in the alloy tend to form protective layers that prevent further corrosion and hence the compositional descriptors play a significant role in such problems. The exponential rise in ML driven research for HEAs, especially over the last two years, is expected to continue. However, it is apparent that ML is most effective in the initial stages of such data-guided explorations, to get an estimate about the phases or properties, principally for those that are difficult to determine computationally, are challenging to measure experimentally or when the property is nondeterministic. Nevertheless, these approaches do have certain shortcomings that need to be overcome with further research.

First, ML techniques work best when the dataset volume is of the order of tens of thousands, but in the domain of HEAs, there are hardly ~600–700 experimental data points on phases, ~ 100 data points on experimental hardness measurements and similar for yield stress and Young's modulus. Conclusions derived on such limited training set can often be biased, which makes the need to use uncertainty quantification even more important. Second, most datasets are composed of data obtained at room temperature (RT). We note that one of the chief drivers for HEAs was its promise for good RT ductility and excellent high temperature strength, but most efforts thus far have focused only on RT properties like strength, hardness and phases. Third, another major drawback of the above discussed data-enabled methods is that the metadata associated with a physical phenomenon is generally unaccounted for. For example, ML cannot reveal the deformation mechanism, the failure mode, the creep mechanism or the fatigue properties. These properties still need to be investigated via experiments or high-performance computing (*ab initio* or molecular simulations) which are computationally expensive and time consuming but can produce reliable predictions.

It is important to remember that there are >500 billion compositions are possible with a 10% variation in the fraction of each element, but only around 100 new compositions have been examined. Several contemporary efforts also limit the investigations to slight variations of already discovered HEAs, in particular modifications of the Cantor alloy. Under such a scenario, there is a compelling need to expand the experimental dataset to the order of thousands, and include a wider palette of elements apart from those already explored, to improve the accuracy and build confidence in the results from fast ML techniques. Inverse mapping technique is the next potentially big step for ML explorations of HEAs. Data-centric efforts are anticipated to continue and expand for the foreseeable future, and their applications to HEAs will open floodgates to discovering physically relevant new candidate HEAs for a myriad of applications in sectors ranging from energy, transportation, defense and medicine.

CRediT authorship contribution statement

Ankit Roy: Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing - original draft, Writing - review & editing. **Ganesh Balasubramanian:** Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing - original draft, Writing - review & editing.

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.

Acknowledgements

The research was supported by the National Science Foundation (NSF) through the award # CMMI-1944040. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors' and do not necessarily reflect the views of the NSF.

References

[1] J.W. Yeh, S.K. Chen, S.J. Lin, J.Y. Gan, T.S. Chin, T.T. Shun, C.H. Tsau, S.Y. Chang, Nanostructured High-Entropy Alloys with Multiple Principal Elements: Novel Alloy Design Concepts and Outcomes, Adv. Eng. Mater. 6 (5) (2004) 299–303.
[2] P. Singh, A. Sharma, A.V. Smirnov, M.S. Diallo, P.K. Ray, G. Balasubramanian, D.D. Johnson, Design of high-strength refractory complex solid-solution alloys, npj Comput. Mater. 4 (1) (2018) 1–8.
[3] Y. Ikeda, B. Grabowski, F. Körmann, Ab initio phase stabilities and mechanical properties of multicomponent alloys: A comprehensive review for high entropy alloys and compositionally complex alloys, Mater. Charact. 147 (2019) 464–511.
[4] A. Roy, T. Sreeramagiri, B. Babuska, P.K.R. Krick, G. Balasubramanian, Lattice distortion as an estimator of solid solution strengthening in high-entropy alloys, Mater. Charact. (2021) 110877.
[5] L. Liu, J.B. Zhu, C. Zhang, J.C. Li, Q. Jiang, Microstructure and the properties of FeCoCuNiSnx high entropy alloys, Mater. Sci. Eng., A 548 (2012) 64–68.
[6] X.F. Wang, Y. Zhang, Y. Qiao, G.L. Chen, Novel microstructure and properties of multicomponent CoCrCuFeNiTix alloys, Intermetallics 15 (3) (2007) 357–362.
[7] B. Cantor, I.T.H. Chang, Knight, and A.J.B. Vincent,, Microstructural development in equiatomic multicomponent alloys, Mater. Sci. Eng., A 375–377 (2004) 213–218.
[8] O.N. Senkov, D.B. Miracle, K.J. Chaput, J.-Couzinie,, Development and exploration of refractory high entropy alloys—A review, J. Mater. Res. 33 (19) (2018) 3092–3128.
[9] D.B. Miracle, O.N. Senkov, A critical review of high entropy alloys and related concepts, Acta Mater. 122 (2017) 448–511.
[10] O.N. Senkov, G.B. Wilks, J.M. Scott, D.B. Miracle, Mechanical properties of Nb25Mo25Ta25W25 and V20Nb20Mo20Ta20W20 refractory high entropy alloys, Intermetallics 19 (5) (2011) 698–706.
[11] Roy A., M.J., and Balasubramanian G., Low energy atomic traps sluggardize the diffusion in compositionally complex refractory alloys. Intermetallics, 2021(in press).
[12] D. Miracle, High entropy alloys as a bold step forward in alloy development, Nat. Commun. 10 (1) (2019) 1–3.
[13] J. McCarthy, Artificial intelligence, logic and formalizing common sense, Philosophical logic and artificial intelligence, Springer, 1989, pp. 161–190.
[14] R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, C. Kim, Machine learning in materials informatics: recent applications and prospects, npj Comput. Mater. 3 (1) (2017) 1–13.
[15] S. Gorsse, M. Nguyen, O.N. Senkov, D.B. Miracle, Database on the mechanical properties of high entropy alloys and complex concentrated alloys, Data in brief 21 (2018) 2664–2678.

[16] J.M. Rickman, H.M. Chan, M. Harmer, J.A. Smeltzer, C.J. Marvel, A. Roy, G. Balasubramanian, Materials informatics for the screening of multi-principal elements and high-entropy alloys, Nat Commun 10 (1) (2019) 2618.
[17] Hume-Rothery, W., Atomic theory for students of metallurgy. 1952.
[18] M.C. Troparevsky, J.R. Morris, P.R.C. Kent, A.R. Lupini, G.M. Stocks, Criteria for Predicting the Formation of Single-Phase High-Entropy Alloys, Phys. Rev. X 5 (1) (2015) 011041.
[19] F. Wang, X., Y. Zhang, Y. Qiao, and G. L. Chen, Novel microstructure and properties of multicomponent CoCrCuFeNiTix alloys. Vol. 15. 2007. 357-362.
[20] K.B. Zhang, Z.Y. Fu, J.Y. Zhang, W.M. Wang, H. Wang, Y.C. Wang, Q.J. Zhang, Characterization of Nanocrystalline CoCrFeNiCuAl High-Entropy Alloy Powder Processed by Mechanical Alloying, Mater. Sci. Forum 620–622 (2009) 383–386.
[21] Y.J. Zhou, Y. Zhang, T.N. Kim, G.L. Chen, Microstructure characterizations and strengthening mechanism of multi-principal component AlCoCrFeNiTi0.5 solid solution alloy with excellent mechanical properties, Mater. Lett. 62 (17) (2008) 2673–2676.
[22] 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 (10) (2011) 103505.
[23] S. Guo, C.T. Liu, Phase stability in high entropy alloys: Formation of solid-solution phase or amorphous phase, Progress in Natural Science: Materials International 21 (6) (2011) 433–446.
[24] Yang, X. and Y. Zhang, Prediction of high-entropy stabilized solid-solution in multi-component alloys. Vol. 132. 2012. 233–238.
[25] A. Takeuchi, A. Inoue, Calculations of Amorphous-Forming Composition Range for Ternary Alloy Systems and Analyses of Stabilization of Amorphous Phase and Amorphous-Forming Ability, Mater. Trans. 42 (7) (2001) 1435–1444.
[26] Singh, S., N. Wanderka, B. Murty, U. Glatzel, and J. Banhart, Decomposition in multi component AlCoCrCuFeNi high entropy alloy. Vol. 59. 2011.
[27] 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.
[28] D.J.M. King, S.C. Middleburgh, A.G. McGregor, M.B. Cortie, Predicting the formation and stability of single phase high-entropy alloys, Acta Mater. 104 (2016) 172–179.
[29] Boer, F.d., R. Boom, W. Mattens, A. Miedema, and A.J.A. Niessen, North-Holland, Cohesion in metals: transition metal alloys, Vol. 1. 1988.
[30] Y. Li, W. Guo, Machine-learning model for predicting phase formations of high-entropy alloys, Physical Review Materials 3 (9) (2019) 095005.
[31] Q. Wu, Z. Wang, X. Hu, T. Zheng, Z. Yang, F. He, J. Li, J. Wang, Uncovering the eutectics design by machine learning in the Al–Co–Cr–Fe–Ni high entropy system, Acta Mater. 182 (2020) 278–286.
[32] S.K. Bhattacharya, R. Sahara, T. Narushima, Predicting the Parabolic Rate Constants of High-Temperature Oxidation of Ti Alloys Using Machine Learning, Oxid. Met. (2020) 1–14.
[33] Y.-J. Chang, C.-Y. Jui, W.-J. Lee, A.-C. Yeh, Prediction of the composition and hardness of high-entropy alloys by machine learning, JOM 71 (10) (2019) 3433–3442.
[34] L. Yan, Y. Diao, Z. Lang, K. Gao, Corrosion rate prediction and influencing factors evaluation of low-alloy steels in marine atmosphere using machine learning approach, Sci. Technol. Adv. Mater. (2020(just-accepted).).
[35] N. Islam, W. Huang, H.L. Zhuang, Machine learning for phase selection in multi-principal element alloys, Comput. Mater. Sci. 150 (2018) 230–235.
[36] A. Roy, T. Babuska, B. Krick, G. Balasubramanian, Machine learned feature identification for predicting phase and Young's modulus of low-, medium-and high-entropy alloys, Scr. Mater. 185 (2020) 152–158.
[37] L. Zhang, H. Chen, X. Tao, H. Cai, J. Liu, Y. Ouyang, Q. Peng, Y. Du, Machine learning reveals the importance of the formation enthalpy and atom-size difference in forming phases of high entropy alloys, Materials Design 108835 (2020).
[38] Y. Zhang, C. Wen, C. Wang, S. Antonov, D. Xue, Y. Bai, Y. Su, Phase prediction in high entropy alloys with a rational selection of materials descriptors and machine learning models, Acta Mater. 185 (2020) 528–539.
[39] F.-Z. Dai, B. Wen, Y. Sun, H. Xiang, Y. Zhou, Theoretical prediction on thermal and mechanical properties of high entropy (Zr0.2Hf0.2Ti0.2Nb0.2Ta0.2) C by deep learning potential, Journal of Materials Science Technology 43 (2020) 168–174.
[40] Liu, X., J. Zhang, M. Eisenbach, and Y. Wang, Machine learning modeling of high entropy alloy: the role of short-range order. arXiv preprint arXiv:1906.02889, 2019.
[41] J. Cowley, An approximate theory of order in alloys, Phys. Rev. 77 (5) (1950) 669.
[42] T. Kostiuchenko, F. Körmann, J. Neugebauer, A. Shapeev, Impact of lattice relaxations on phase transitions in a high-entropy alloy studied by machine-learning potentials, npj Comput. Mater. 5 (1) (2019) 1–7.
[43] F. Körmann, A.V. Ruban, M.H. Sluiter, Long-ranged interactions in bcc NbMoTaW high-entropy alloys, Materials Research Letters 5 (1) (2017) 35–40.
[44] F. Körmann, M.H. Sluiter, Interplay between lattice distortions, vibrations and phase stability in NbMoTaW high entropy alloys, Entropy 18 (8) (2016) 403.
[45] Y. Wang, M. Yan, Q. Zhu, W.Y. Wang, Y. Wu, X. Hui, R. Otis, S.-L. Shang, Z.-K. Liu, L.-Q. Chen, Computation of entropies and phase equilibria in refractory V-Nb-Mo-Ta-W high-entropy alloys, Acta Mater. 143 (2018) 88–101.
[46] P. Singh, A.V. Smirnov, D.D. Johnson, Ta-Nb-Mo-W refractory high-entropy alloys: anomalous ordering behavior and its intriguing electronic origin, Physical Review Materials 2 (5) (2018) 055004.
[47] Z. Pei, J. Yin, J.A. Hawk, D.E. Alman, M.C. Gao, Machine-learning informed prediction of high-entropy solid solution formation: Beyond the Hume-Rothery rules, npj Comput. Mater. 6 (1) (2020) 1–8.
[48] Z. Zhou, Y. Zhou, Q. He, Z. Ding, F. Li, Y. Yang, Machine learning guided appraisal and exploration of phase design for high entropy alloys, npj Comput. Mater. 5 (1) (2019) 1–9.
[49] K. Kaufmann, K.S. Vecchio, Searching for high entropy alloys: A machine learning approach, Acta Mater. (2020).
[50] G. Kim, H. Diao, C. Lee, A. Samaei, T. Phan, M. de Jong, K. An, D. Ma, P.K. Liaw, W. Chen, First-principles and machine learning predictions of elasticity in severely lattice-distorted high-entropy alloys with experimental validation, Acta Mater. 181 (2019) 124–138.
[51] A. Jain, S. Ong, G. Hautier, W. Chen, W.D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, Commentary: The Materials Project: A materials genome approach to accelerating materials innovation, APL Mater. 1 (1) (2013) 011002.
[52] A. Agarwal, A. Rao, Artificial Intelligence Predicts Body-Centered-Cubic and Face-Centered-Cubic Phases in High-Entropy Alloys, JOM 71 (10) (2019) 3424–3432.
[53] 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 Mater. 170 (2019) 109–117.
[54] B. Gorr, M. Azim, H.-J. Christ, T. Mueller, D. Schliephake, M. Heilmaier, Phase equilibria, microstructure, and high temperature oxidation resistance of novel refractory high-entropy alloys, J. Alloy. Compd. 624 (2015) 270–278.
[55] Liu, Y., Z. Chen, Y. Chen, J. Shi, Z. Wang, S. Wang, and F. Liu, Effect of Al content on high temperature oxidation resistance of AlxCoCrCuFeNi high entropy alloys (x= 0, 0.5, 1, 1.5, 2). Vacuum, 2019. 169: 108837.
[56] M. Gianelle, A. Kundu, K. Anderson, A. Roy, G. Balasubramanian, H.M. Chan, A novel ceramic derived processing route for Multi-Principal Element Alloys, Mater. Sci. Eng., A 793 (2020) 139892.
[57] Y. Qiu, S. Thomas, M.A. Gibson, H.L. Fraser, N. Birbilis, Corrosion of high entropy alloys, npj Mater. Degrad. 1 (1) (2017) 1–18.
[58] D.D. Macdonald, The point defect model for the passive state, J. Electrochem. Soc. 139 (12) (1992) 3434.
[59] Ahmad, Z., Principles of corrosion engineering and corrosion control. 2006: Elsevier.
[60] J. Zhang, X. Liu, S. Bi, J. Yin, G. Zhang, M. Eisenbach, Robust data-driven approach for predicting the configurational energy of high entropy alloys, Materials Design 185 (2020) 108247.
[61] Hastie, T., R. Tibshirani, and J. Friedman, The elements of statistical learning: data mining, inference, and prediction. 2009: Springer Science Business Media.
[62] D. Xue, P.V. Balachandran, J. Hogden, J. Theiler, D. Xue, T. Lookman, Accelerated search for materials with targeted properties by adaptive design, Nat. Commun. 7 (1) (2016) 1–9.
[63] T. Lookman, P.V. Balachandran, D. Xue, J. Hogden, J. Theiler, Statistical inference and adaptive design for materials discovery, Curr. Opin. Solid State Mater. Sci. 21 (3) (2017) 121–128.
[64] J. Behler, Representing potential energy surfaces by high-dimensional neural network potentials, J. Phys.: Condens. Matter 26 (18) (2014) 183001.
[65] S. Dudiy, A. Zunger, Searching for alloy configurations with target physical properties: impurity design via a genetic algorithm inverse band structure approach, Phys. Rev. Lett. 97 (4) (2006) 046401.
[66] J. Li, B. Xie, Q. Fang, B. Liu, Y. Liu, P.K. Liaw, High-throughput simulation combined machine learning search for optimum elemental composition in medium entropy alloy, Journal of Materials Science Technology (2020).
[67] W. Huang, Martin, and H.L. Zhuang,, Machine-learning phase prediction of high-entropy alloys, Acta Mater. 169 (2019) 225–236.
[68] Y. Lederer, C. Toher, K.S. Vecchio, S. Curtarolo, The search for high entropy alloys: a high-throughput ab-initio approach, Acta Mater. 159 (2018) 364–383.
[69] F. Tancret, I. Toda-Caraballo, E. Menou, P.E.J.R. Díaz-Del, Designing high entropy alloys employing thermodynamics and Gaussian process statistical analysis, Materials Design 115 (2017) 486–497.
[70] G. Arora, D.S. Aidhy, Machine Learning Enabled Prediction of Stacking Fault Energies in Concentrated Alloys, Metals 10 (8) (2020) 1072.
[71] A. Takeuchi, A. Inoue, Classification of bulk metallic glasses by atomic size difference, heat of mixing and period of constituent elements and its application to characterization of the main alloying element, Mater. Trans. 46 (12) (2005) 2817–2829.
[72] S. Fang, X. Xiao, L. Xia, W. Li, Y. Dong, Relationship between the widths of supercooled liquid regions and bond parameters of Mg-based bulk metallic glasses, J. Non-Cryst. Solids 321 (1–2) (2003) 120–125.
[73] F.R. De Boer, W. Mattens, R. Boom, A. Miedema, A. Niessen, Cohesion in metals. (1988).
[74] Y. Ouyang, X. Zhong, Y. Du, Y. Feng, Y. He, Enthalpies of formation for the Al–Cu–Ni–Zr quaternary alloys calculated via a combined approach of geometric model and Miedema theory, J. Alloy. Compd. 420 (1–2) (2006) 175–181.
[75] I. Toda-Caraballo, A. Rivera-Díaz-del-Castillo, criterion for the formation of high entropy alloys based on lattice distortion, Intermetallics 71 (2016) 76–87.