npj | Materials Degradation

www.nature.com/npjmatdeg

ARTICLE OPEN

Check for updates

Machine-learning-guided descriptor selection for predicting corrosion resistance in multi-principal element alloys

Ankit Roy[1,2,6], M. F. N. Taufique[2,6✉], Hrishabh Khakurel[3], Ram Devanathan[2], Duane D. Johnson[4,5] and Ganesh Balasubramanian[1]

More than $270 billion is spent on combatting corrosion annually in the USA alone. As such, we present a machine-learning (ML) approach to down select corrosion-resistant alloys. Our focus is on a non-traditional class of alloys called multi-principal element alloys (MPEAs). Given the vast search space due to the variety of compositions and descriptors to be considered, and based upon existing corrosion data for MPEAs, we demonstrate descriptor optimization to predict corrosion resistance of any given MPEA. Our ML model with descriptor optimization predicts the corrosion resistance of a given MPEA in the presence of an aqueous environment by down selecting two environmental descriptors (pH of the medium and halide concentration), one chemical composition descriptor (atomic % of element with minimum reduction potential), and two atomic descriptors (difference in lattice constant (Δa) and average reduction potential). Our findings show that, while it is possible to down select corrosion-resistant MPEAs by using ML from a large search space, a larger dataset and higher quality data are needed to accurately predict the corrosion rate of MPEAs. This study shows both the promise and the perils of ML when applied to a complex chemical phenomenon like corrosion of alloys.

npj Materials Degradation (2022)6:9 ; https://doi.org/10.1038/s41529-021-00208-y

INTRODUCTION

Corrosion of industrial machinery, bridges, engineering equipment, vehicles, and structures under different corrosive environments is a major concern for food processing and chemical industries, power plants, maritime and air transport. It is reported that in the United States 6.2% of the gross domestic product (GDP) is spent to replace engineering equipment in industries due to corrosion[1]. The development of corrosion-resistant materials and coatings capable of extending the lifespan of equipment and structures can yield substantial economic benefits. Birbilis et al. and Yeh et al. suggested that an emerging class of materials, high entropy alloys (HEAs) and multi-principal element alloys (MPEAs), are being considered as preferred candidates for corrosion-resistant applications compared to conventional alloys, such as stainless steels, Ni-based alloys, and Ti-based alloys[2,3]. MPEAs consist of 4 or more base elements that typically stabilize a single-phase (random) solid-solution crystallographic phase. Certain MPEA compositions have exhibited excellent hardness, exceptional wear properties, as well as high-temperature strength relative to traditional alloys[2,4–8].

In theory, MPEAs composed of 4 or more elements allow for extensive compositional freedom with perhaps greater property control. Due to the high entropy of mixing[2], which increases roughly as natural log N with number (N) of elements, the tendency to order or segregate is thus greatly reduced in MPEAs. Such alloys are then less likely to form ordered structures (unless a large enthalpy reduction occurs from a new elemental pair interaction, a pair probability that increases as N^2). Hence, MPEAs tend to form disordered solid solutions predominantly of face-centered cubic (FCC), body centered cubic (BCC), or hexagonal close-packed (HCP) structures[9–11], instead of more complex intermetallic phases. The locally disordered chemical environment within these solid solutions has been theorized to lead to the excellent corrosion properties of MPEAs[5]. For instance, Qiu et al.[12] reported that MPEAs containing Co, Cr, and Ni offer better corrosion resistance in the presence of NaCl solution compared to austenitic stainless steel.

Corrosion resistance of MPEAs depends on the elemental composition of the alloys and the corresponding corrosive environments[9,13]. The number of possible elemental compositions is much higher in MPEAs than that of traditional metallic alloys and the search can involve about a trillion combinations as we move away from the vertices of the multi-dimensional composition space toward the center[14]. Moreover, different corrosive environments, such as NaCl, HCl, and H_2SO_4 have different influences on MPEAs and it is extremely challenging to select the appropriate composition to optimize corrosion resistance for different environments by trial-and-error experiment or intuition[15]. Additionally, long-term atmospheric exposure testing for MPEAs is not commonly practiced in industries and hence such data are not readily available thereby limiting the applications for MPEAs in harsh aqueous environments[3]. Simulations and data informatics related to metallic alloys can be a potential tool to predict the corrosion resistance of MPEAs under different environments. Techniques like density functional theory (DFT)[16], molecular dynamics (MD), and thermodynamic modeling have been devoted to study phase stability, solidification, and crystallization kinetics of MPEAs[17–23]. These methods are computationally expensive and time-consuming even for a single composition and hence less readily usable on a large scale to narrow down the search space.

[1]Department of Mechanical Engineering and Mechanics, Lehigh University, Bethlehem, PA 18015, USA. [2]Pacific Northwest National Laboratory, Richland, WA 99354, USA. [3]Department of Mathematics, The University of Texas at Arlington, Arlington, TX 76019, USA. [4]Ames Laboratory, United States Department of Energy, Iowa State University, Ames, IA 50011, USA. [5]Department of Materials Science & Engineering, Iowa State University, Ames, IA 50011, USA. [6]These authors contributed equally: Ankit Roy, M. F. N. Taufique. ✉email: mohammadfn.taufique@pnnl.gov

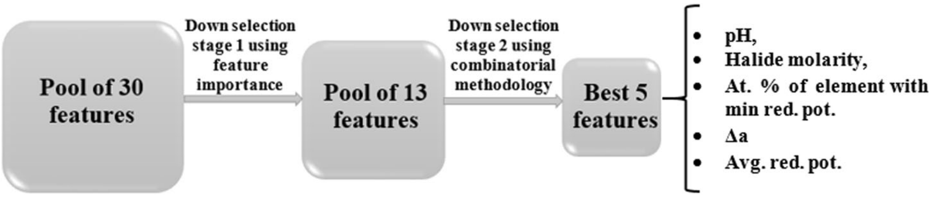

Pool of 30 features
Down selection stage 1 using feature importance
Pool of 13 features
Down selection stage 2 using combinatorial methodology
Best 5 features
pH,
Halide molarity,
At. % of element with min red. pot.
Δa
Avg. red. pot.

Fig. 1 Schematic diagram of the 2-stage feature down-selection process used in this work to select 5 features from a pool of 30. The 2-stage features down selection process adopted in this work to down select the features from a pool of 30 to a final 5.

Materials informatics offers a faster and less expensive path to develop new materials by identifying correlations and trends existing in the dataset, incorporating the underlying physics, predicting the properties of new materials, and guiding the next set of experiments[24]. Some examples of applications include predictions of crystal structures[25,26], physical properties[27], and corrosion rates[28]. The approach uses ML to construct a surrogate function (f) that relates the target property (Y) to the material features or descriptors (X). The depth up to which the descriptors capture the physical and chemical signatures of the materials combined with the mathematical accuracy of the models determine the robustness of the model. Therefore, for any given target property, the optimized combination of f and X needs to be selected. For a system with M models and N materials descriptors, there are $M \times (2^N - 1)$ potential combinations possible out of which only one would ideally have the desired accuracy. Material descriptors can be based on the composition, microstructure, processing conditions, and other thermodynamic parameters which necessitates the construction of a large pool of descriptors[29]. For instance, 10 descriptors and 5 models can produce 5115 different combinations out of which one will be selected based on minimum prediction error. Hence, the primary goal becomes choosing the best combination of model and descriptors to predict the target property of desired alloys.

Domain knowledge plays a crucial role in down selecting the descriptors[30,31]. For example, Rickman et al.[7] used Young's modulus asymmetry (representing stiffness) and mean melting temperature (proportional to bond strengths) as descriptors in conjunction with thermodynamic descriptors to generate compositions with high hardness. Kim et al.[27] used a DFT dataset (from the Materials Project database[32]) to construct a model using the average values of physical parameters (e.g., cohesive energy, density) in conjunction with thermodynamic parameters to predict the bulk and shear modulus. When the descriptor space becomes too large, principal component analysis can be used to reduce the descriptor space by using certain mathematical transformational laws but this operation may lack interpretability[33].

In the present study, we have utilized a gradient boost ML model coupled with a 2-stage feature down selection process to build a robust model for corrosion-rate prediction in MPEAs. This optimization process is based on the best performing combination of descriptors to predict the corrosion rate of MPEAs under different corrosive environments. The methodology and results will be explained further in the subsequent sections. We conclude by reasoning why certain features are the most impactful in determining corrosion rates and also carry out detailed data analysis to show the dependence of corrosion rate on those descriptors.

RESULTS AND DISCUSSION

According to the "no-free-lunch" theorem[34], there does not exist a perfect model for a given problem. There always exists room for improvement in accuracy by adopting different combinations of models and features. In the present work, we have chosen the best model for our problem as the Gradient Boosting Regressor (GBR) model. To choose the best features out of the pool of 30 features, we adopt a 2-stage down selection process as shown in Fig. 1. The first stage down selects top 13 out of the 30 features as obtained from the feature importance function. The second stage of down selection takes all possible combinations of 1, 2, 3, …, 13 features out of the 13 features from stage 1 and evaluates the mean squared error (MSE). The best number and combination of features are thus selected.

Features down selection stage 1

To reduce the complexity of the model, the number of descriptors can be limited to a certain value without compromising the performance of the model or sometimes even improving it. In fact, some reports claim that the model degrades the results when there are too many variables because the redundant feature variables interfere with the ML model[35]. Thus, it is essential to down select the most significant and relevant features to construct a robust model. Feature importance is determined by assigning a value to input features based on how relevant they are in predicting a target property. The feature importance was extracted from the model using the "feature_importances" function built into the Scikit-learn python module[36]. Gradient boosting models work by iteratively fitting several base models such as decision trees to data. The GBR fits one tree, then fits another tree to the residual error of the first tree. The GBR then fits a new tree to the residual error of the most recent tree and so on. Feature importance is computed as the mean and standard deviation of accumulation of the decrease in error with each progressing tree. Such an approach has been used by several other works in the past to eliminate redundant features from a larger feature pool[37,38]. The feature importance of all 30 features has been shown in Supplementary Fig. 1. With the objective of downsizing the feature pool to reduce the complexity of the model, the top 13 most important features out of the total 30 have been down selected. These top 13 descriptors with the relative importance ranging from 2 to 37% and corrosion-rate prediction by GBR model are provided in Fig. 2 and Table 1, respectively.

From feature importance, it is clear that the chemical environments during the corrosion reaction, such as pH of the reaction medium and presence of halide ion, have the maximum impact on the corrosion process. The pH of the environment plays a decisive role in the corrosion rate. To understand the effect of pH in greater detail, we take the example of a pure metal in an acidic environment. When a metal is submerged in a deaerated acid solution, the electrochemical reaction can be divided into two or more partial oxidation and reduction reactions as follows:

(i) $M \rightarrow M^{n+} + ne^{-1}$, which is the oxidation(anodic) reaction and
(ii) $nH^{+} + ne^{-1} \rightarrow (n/2)H_2$ (gas) which is the reduction (cathodic) reaction.

To find the propensity of corrosion in the given solution, the potentials of the anodic ($E_{M^{++}/M}$) and cathodic reaction ($E_{H^{++}/\frac{1}{2}H_2}$) should be known. The cell potential is formulated[39] as

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

where ΔE should be positive to make ΔG negative so that the reaction is spontaneous according to the equation $\Delta G = -nFE$. According to the above formulation under acidic conditions, a lower pH (acidic environment) will increase the ΔE and thereby enhance the corrosion rate. Therefore, the pH of the aqueous environment plays a crucial role in determining the corrosion rate.

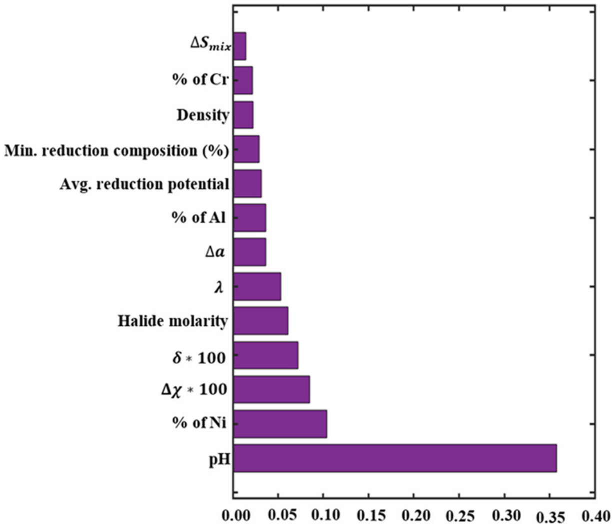


Fig. 2 The relative importance of the top 13 features to predict the corrosion rate. This down selected pool will be subsequently used for stage 2 feature down selection. The second stage may result in a reshuffling of the rankings of the top features.

Notably, this equation does not consider the formation of the passive film due to the presence of certain metals that form the protective oxide layer. Hence, this situation is considered by including the composition of metals like Cr, Al, Cu, Ti, Ni, Sn, and Mo[13] as separate features that will indirectly incorporate the effects of passive layer formation into the model. The addition of nickel to MPEAs has been investigated before. Ni was added to $Al_2CrFeCoCuTiNi_x$ (x = 0, 0.5, 1, 1.5, and 2) alloy system in 3.5 wt. % NaCl solution and 1 M NaOH solution. The composition with x = 1 was found to have the least corrosion current density in both salt and alkaline solution, implying that there is an optimum Ni content below and above which Ni has a detrimental impact on corrosion[40]. Ni has also been shown to be effective in resisting corrosion in nickel-aluminum bronze alloys, where increasing the concentration of Ni improves the corrosion resistance of the alloy[41].

Another important descriptor is the difference in electronegativity between the highest and lowest reduction potential of a specific alloy. A larger difference in reduction potential enhances the driving force for galvanic corrosion. Corrosion rate further increases with the presence of halide ion, such as chloride ion (Cl^{-1}) in the reaction medium. Cl^{-1} ions can penetrate the passive film on the alloy surface and hence increase the pitting corrosion rate. Chou et al. reported that pitting potential was proportional to the logarithm of the chloride concentrations at a constant temperature[42] and the potential will increase with increasing temperature and hence increase the pitting corrosion rate. It is

Table 1. Validation dataset with details of the corrosive environment, corresponding observed experimental corrosion rate[8,9] and predicted corrosion rate by considering the top 13 descriptors and the optimized descriptors.

	Alloy	Corrosion environment			Corrosion rate (mm/year)		
		pH	Temp (K)	Halide molarity (M)	Experiment	Predicted by top 13 descriptors	Predicted by descriptor optimization
1	$Al_2CoCrFeNi$	1.301	298	0.05	0.02965	0.05784	0.06498
2	$Al_2CrFeNiCoCuTi_{1.5}$	0.301	298	0	0.03558	0.04256	0.04256
3	$Fe_{68.59}Ni_{10.47}Co_{0.21}Mo_2Cr_{16.61}$	7	298	0.6	0.00115	0.06145	0.00299
4	$Al_2CrFeCoCuTiNi$	7	298	0.6	0.10533	0.09907	0.09908
5	$Fe_{24.85}Ni_{25.89}Co_{26}Mn_{0.51}Cr_{22.66}Al_{0.07}$	4	313	0	0.00803	0.00959	0.01394
6	$Fe_{46.86}Ni_{12.88}Co_{12.54}Mn_{11.72}$ $Cr_{15.8}Nb_{0.16}$	4	313	0	0.00959	0.07884	0.00339
7	FeCoNiCu	14	298	0	0.01561	0.60843	0.12714
8	$FeCoNiCuSn_{0.02}$	14	298	0	0.01177	0.33203	0.01977
9	$FeCoNiCuSn_{0.03}$	14	298	0	0.01043	0.336	0.01977
10	$FeCoNiCuSn_{0.04}$	14	298	0	0.01532	0.02754	0.01977
11	$Co_{1.5}CrFeNi_{1.5}Ti0.5Mo_{0.1}$	7	298	1	0.00121	0.00307	0.00307
12	$Co_{1.5}CrFeNi_{1.5}Ti0.5Mo_{0.5}$	7	298	1	0.00178	0.00414	0.00414
13	$Co_{1.5}CrFeNi_{1.5}Ti0.5Mo_{0.8}$	7	298	1	0.00355	0.04789	0.00707
14	$Co_{1.5}CrFeNi_{1.5}Ti0.5Mo_{0.1}$	14	298	0	0.00103	0.00233	0.00229
15	$Co_{1.5}CrFeNi_{1.5}Ti0.5Mo_{0.5}$	14	298	0	0.00143	0.00233	0.00229
16	$Co_{1.5}CrFeNi_{1.5}Ti0.5Mo_{0.8}$	14	298	0	0.00216	0.38237	0.38237
17	$Al_2CrFeCoCuTiNi_{0.5}$	7	298	0.6	0.25814	0.35711	0.35711
18	$Al_2CrFeCoCuTiNi_1$	7	298	0.6	0.10533	0.09907	0.09908
19	$Al_2CrFeCoCuTiNi_{1.5}$	7	298	0.6	0.52126	0.63757	0.33121
20	$Al_2CrFeCoCuTiNi_2$	7	298	0.6	0.54881	0.65705	0.33803
21	$FeCoNiCuSn_{0.02}$	7	298	0.6	0.0238	0.16663	0.02088
22	$FeCoNiCuSn_{0.03}$	7	298	0.6	0.03611	0.17072	0.02088
23	$FeCoNiCuSn_{0.04}$	7	298	0.6	1.08367	0.02688	0.02088
24	$FeCoNiCuSn_{0.05}$	7	298	0.6	0.0151	0.03016	0.03004
25	$FeCoNiCuSn_{0.07}$	7	298	0.6	0.02082	0.03007	0.03004

important to mention that corrosion rate in presence of Cl^{-1} is smaller for MPEAs compared to that for stainless steel. For instance, Chen et al. compared CuNiAlCoCrFeSi MPEA and compositionally similar 304SS with near-equal Cr content in deaerated 1 M NaCl solution at room temperature[43]. The MPEA was shown to be more noble and have a lower corrosion current density than 304SS in NaCl solution. Due to the presence of nearly amorphous structure, which had negligible or no grain boundaries and hence acted as passive film, the MPEA showed significant resistance to uniform corrosion. The passive region of the MPEA is smaller than that of 304SS and when the MPEA passive film is ruptured the alloy is more susceptible to nucleation and growth than 304SS.

We emphasize that the feature importance obtained from the ML model represents the ranking of the contribution of each feature to the mathematical expression of the target property when all the features are used. Hence, feature importance is basically the weight of each feature when that feature is used as a variable in the mathematical expression of corrosion rate formulated by the ML model. The top 13 features represent the features that have the highest weight. Since the formulation of this expression is based on statistics, it is only helpful in eliminating the redundant features. The order of the topmost important features may vary if the training dataset is changed. Hence, at the current stage with the limited dataset, we may not be able to conclusively determine the reasoning of the order of the top important features. But the approach may be employed to downsize the feature pool by eliminating the redundant features[35,37].

The choice of alloying elements is also an important descriptor. For instance, aluminum (Al) is used to synthesize lower density MPEAs because of light weight. In addition to that, Al helps to improve the mechanical strength of the MPEAs[44–50]. However, Al present in the Al based MPEAs is less noble (lower in the galvanic series) compared to the other alloying elements, such as Fe, Ni, Cr, Co, and Ti that are used to synthesize MPEAs[51]. Therefore, it is reasonable that in the presence of aqueous corrosive environments, Al atoms may be preferentially released from MPEA solid solution compared to other alloying elements and hence introduce dealloying. It is reported that depassivated Al-containing MPEAs have a higher corrosion rate and the rate increases with increasing Al content in the presence of a corrosion environment, which promotes dissolution. Interestingly, Al is beneficial when used as thin film for surface passivation as reported in the case of Ni-based corrosion-resistant alloys[52]. It is also reported that MPEAs containing Cr, Fe, and Ni showed good corrosion resistance in both NaCl and H_2SO_4 solution as compared to steel. Hsu et al. reported that due to the presence of Cr and Ni, CoCrFeNi MPEA showed better corrosion resistance as compared to 304LSS[53].

From Fig. 2, it is observed that thermodynamic parameters, such as mixing entropy (ΔS_{mix}), can be important for corrosion-rate prediction. The equation for ΔS_{mix} is given below:

$$\Delta S_{\mathrm{mix}} = -R\sum_{i=1}^{n}(C_i lnC_i) \qquad (2)$$

Where, C_i is the composition of element i in a specific MPEA, n is the total number of individual elements present on the MPEA and $R = 8.314\ J{\cdot}mol^{-1}{\cdot}K^{-1}$ is the molar gas constant, However, mixing entropy does not take into consideration the selection of elements; it only considers the number and fraction of elements. For example, CoCrFeNi and CoCuFeNi have the same mixing entropy, but the former exhibits better corrosion resistance due to the formation of the stable Cr_2O_3 film[54]. Hence, mixing entropy should not be used in isolation to make predictions about the corrosion rate.

Similarly, the difference in atomic radii (or δ) also appears as one of the top 13 most important features. δ is the representative of the strain in the lattice. Higher strain in the alloy lattice can introduce larger lattice distortion and promote the formation of secondary phases[55] which may be detrimental to the mechanical properties. But in a given set of single-phase alloys, the ones with a larger δ and hence larger strain are posited to have retarded diffusion thus making it difficult for Cl^{-} or O atoms to migrate inside the alloy matrix. Hence the alloy is protected from corrosion. Similar to δ, the Δa has an impact on lattice distortion and results in uniform lattice distortion[56]. Hence a higher difference in lattice constant might retard the diffusion and protect the alloy from corrosion.

As we are not able to draw concrete conclusions from our current list of top features, we adopt another ingenious technique to further downsize our feature list. This technique is explained in the following section and is used to build the final model for corrosion-rate prediction.

Features down selection stage 2

There is a tradeoff between trying to improve the accuracy of a ML model with a large number of descriptors and having a model that is easier to interpret with a smaller set of descriptors. Thus, it is necessary to track the model performance as a function of the number and combination of descriptors. To reduce the complexity of the model, we have performed descriptor set optimization by finding the optimum number and combination of descriptors that produce the minimum MSE.

In Fig. 3 (d), all combinations ($^{13}C_1$, $^{13}C_2$, $^{13}C_3$, …, $^{13}C_{13}$) of descriptors from the top 13 descriptors obtained in stage 1 down selection have been fitted to the GBR model to check dependency of performance on the number of descriptors. Such an approach has been successfully used in ref. [57] to increase the accuracy of the ML model from the case where a larger number of features is used. We plot the MSE of the GBR model as a function of the number and combination of descriptors shown in Fig. 4.

For a given number of descriptors (n), the best combination (lowest MSE out of $^{13}C_n$) is shown in Fig. 4 (solid-black frontier). The MSE does not show any trend with the number of descriptors and is minimum for a set of 5 descriptors, namely, pH and halide molarity of the medium, composition of element with minimum reduction potential, Δa, and average reduction potential.

At minimum marked in Fig. 4, the model makes the best predictions closest to the experimental values. The MSE at that point is a minimum (i.e., 0.055 mm/year) out of all possible cases. The predictions at this point (by optimization and by top 13 descriptors) and experimental corrosion rate are listed in Table 1. We can see in Table 1 that our ML model predicts the corrosion rate of different MPEAs in a variety of corrosive environments. However, the accuracy of the predictions increased when we applied optimized descriptors corresponding to the point of minimum error in Fig. 4. From this result, we can say that if the severity of the corrosion environment (pH of medium and halide molarity) is fixed then only two atomic descriptors (the Δa and average reduction potential), and one compositional descriptor (the composition of the element with minimum reduction potential) will be needed to predict the corrosion rate, saving considerable computational and experimental time for alloy down selection.

Furthermore, we have analyzed the effect of individual descriptors on corrosion rate to assign some physical significance to the descriptor. We have plotted in Fig. 5 (panel 1 and 2, respectively) the corrosion rate of 12 MPEA datasets from our training set versus Δa and average reduction potential. (Note: for each dataset, the corrosion environment is identical.) Details of the 12 datasets are presented in Supplementary Table 1. From Fig. 5 (panel 1), we see that only 6 datasets (marked with a star)

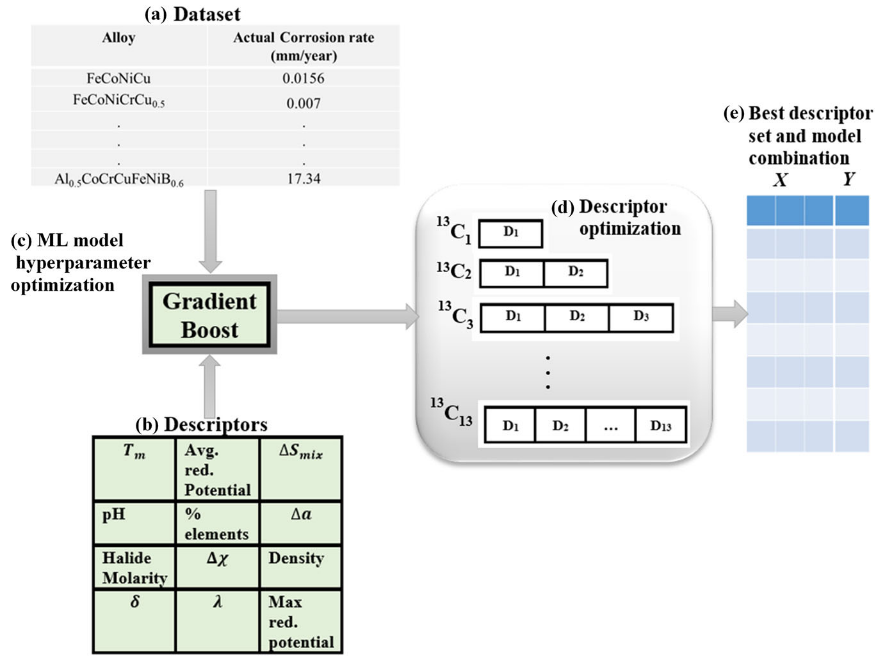


Fig. 3 Flowchart of the strategy used to obtain the best combination of ML model and descriptors for predicting the corrosion rates in MPEAs. **a** Composition and corrosion-rate dataset of 142 alloys was collected from the literature. **b** The descriptor space contained a total of 30 descriptors but down selected to 13 descriptors in stage 1 based on feature importance. **c** A ML model was chosen and hyperparameter optimization done to identify the best parameters with the lowest MSE. **d** Stage 2: an optimization technique down selected the best number and combination of descriptors from a total of 13 listed after stage 1. **e** The most accurate combination of the ML model and the descriptor set is then identified.

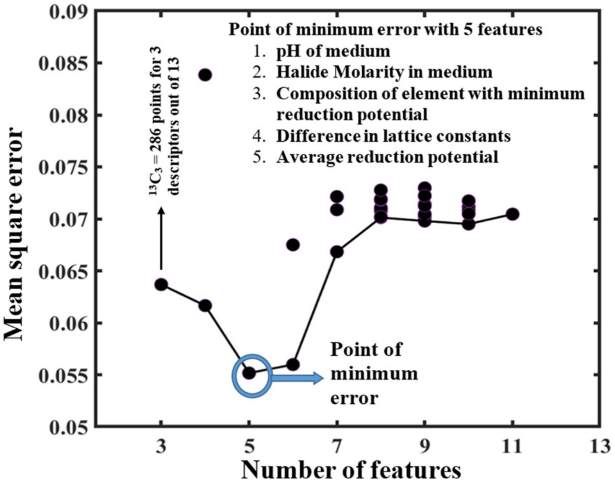


Fig. 4 MSE of the GBR model versus number of descriptors. Data are from numerous descriptor combinations from 1 to 13 for the absolute difference of predictions and experimental values, at least 19 out of 25 test data are within 2.5 times that of the experimental values. The lowest points represent the best regressor for a given number of descriptors. *y*-axis is limited for clarity to only those data that are <0.09. In actuality, there are ${}^{13}C_n$ values for any given number of descriptors (n), as shown for $n = 3$ a total of 286. MSE does not show a trend with the number of descriptors and is minimum for a set of 5 descriptors (listed in figure).

out of 12 alloy datasets show a trend where the corrosion rate tends to increase with increasing Δa. A larger Δa may introduce higher strain energy in the system and enhance the formation of secondary phases leading to a higher corrosion rate. However, this trend is not definitive for all the data, so a clear conclusion cannot be drawn, especially given the limited dataset.

Yet, from a material science point of view, only certain elements are reactive on surfaces of MPEAs, i.e., that segregate to a surface and form protective layers (oxide). Hence, within the datasets we find a trend of increasing corrosion rate with increasing Δa driven by the presence of a few specific elements. In Fig. 5 (a-l) (panel 1), the dataset 1, dataset 5 and dataset 7 show a trend in line with our hypothesis. As suspected, these datasets and their compositions (refer to Supplementary Table 2) reveal increasing Al in dataset 1 increases Δa and also increases the corrosion rate; this follows as Al may form a passive layer but the oxide film formed is porous[9] and fails to resist corrosion[13]. Similarly, in datasets 5 and 7, increasing Ti decreases the Δa and also the corrosion rate, which follows because Ti is known to form a protective TiO_2 film anticipated to improve the corrosion resistance of the alloy[58]. Therefore, we reiterate that it is difficult to draw a definitive conclusion regarding the role of Δa based on the limited dataset. Rather, it is the presence of certain protective reactive metals that decides corrosion resistance but directly affect a change in Δa of the MPEA.

From Fig. 5 (m-x) (panel 2), no clear trend is observed where lowering the average reduction potential has a negative impact on the corrosion rate of the MPEAs. Only 5 datasets (marked with a star) out of 12 show such a trend. Datasets 4, 5 and 7 show the opposite trend. Datasets 8 and 9 show a lot of scattering. The prediction of the ML model appears to be based on the synergistic effect of multiple descriptors, but the corrosion rate does not show a clear trend when considered individually with this selected descriptor.

To assess the improvement in model performance, we compare the predictions based on the top 13 descriptors and from the descriptor optimization (5 descriptors), as given in Fig. 6. We have observed 13 fairly accurate predictions (same order of magnitude as that of experiments) out of a validation set of 25 when we considered the top 13 descriptors. However, when we consider 5 optimized descriptors, namely, pH and halide molarity of the medium, composition of the element with minimum reduction

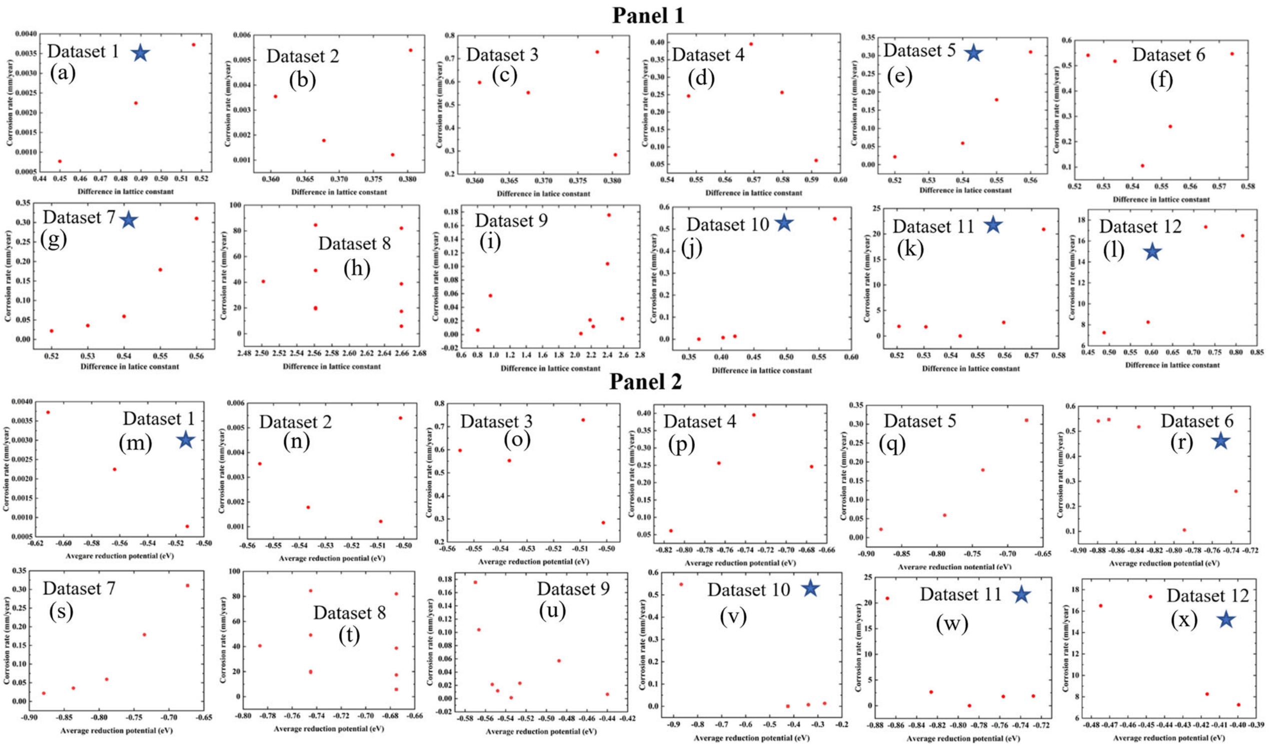


Fig. 5 Trends of corrosion rate with respect to 2 of the 5 down selected features. The corrosion rate of 12 MPEA datasets as a function of the difference in lattice constant and average reduction potential (eV) are presented in panel 1, (a–l) and panel 2, (m–x), respectively. The star mark in each panel indicates the datasets which are in agreement with our reasoning. Since the prediction of the ML model is based on the combined effect of the 5 down select features, the corrosion rate does not show a clear trend in some plots when plotted individually with the selected features. Details of the datasets are provided in Supplementary Table 1.

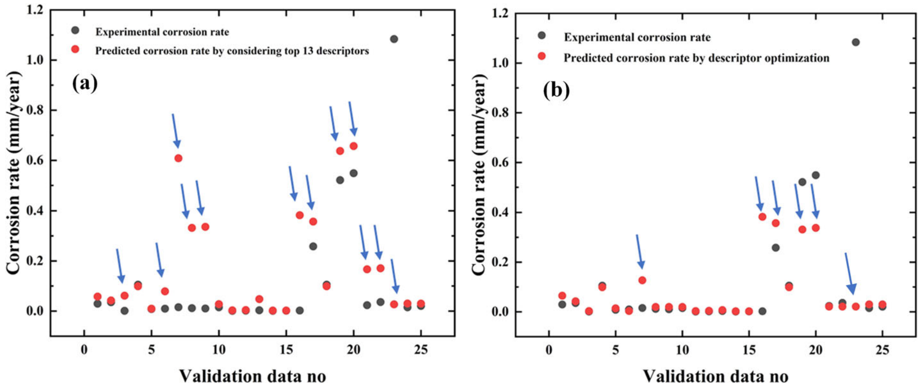


Fig. 6 Comparison between experimental and ML predicted corrosion rate. a Using top 13 features from stage 1 down selection, and **b** using 5 optimized features from stage 2 down selection. The bad prediction for each case is indicated by blue arrows.

potential, Δa, and average reduction potential, we achieve better predictions (19 out of 25 for the validation set were fairly accurate).

This study shows both the promise and perils of ML when applied to a complex phenomenon like corrosion of alloys. ML can help select descriptors and build a model that can reasonably predict trends in the corrosion rate of MPEAs in different environments to help narrow down the alloy search space. However, there remain large quantitative differences between the experimental corrosion rate and the rate predicted by this state-of-the-art ML model, which is not surprising given the orders of magnitude difference in corrosion rates in our training dataset. The data utilized in this study were collected from various sources in the literature. They have some inherent irreducible uncertainty and noise that makes it difficult to utilize this dataset to build models for a chemical space that is not part of the original training dataset. Sparse data and poor data quality are well-known limitations in materials science[59]. There is a need to collect high-quality corrosion data with metadata to understand the effect of descriptors, such as pH, halide ions, mixing entropy, and strain energy, on the corrosion behavior of MPEAs.

Our goal was to demonstrate the use and limitations of ML as a rapid-screening tool for corrosion resistance in the vast space of MPEAs. We developed a model that predicts corrosion rate (mm/year) within a relaxed threshold and significantly accelerates the identification of MPEAs with improved corrosion resistance. From our modeling exercise, we found that for most of the materials our predicted corrosion rates (mm/year) were within 2.5 times of experimental values. A larger dataset with higher quality data is needed to find more significant physical relationships between individual descriptors and the corrosion rate.

In summary, we have developed a domain-science-informed ML-based approach to predict the corrosion rate of MPEAs. We used the GBR model to predict the corrosion rate under different corrosive environments. We modified the predictive model by developing a count-based selection model to get better accuracy compared to conventional mean-squared error measurement. Our descriptor-optimization strategy predicts corrosion rates reasonably well by down selecting best descriptors—here five, namely, two environmental descriptors (pH of the medium and halide molarity), two atomic descriptors (Δa and average reduction potential), and one elemental descriptor (composition of the element with minimum reduction potential). Our main thrust was to examine the utility of ML models based on limited and noisy published data to predict corrosion rates for MPEAs. The model developed can search alloy space without requiring synthesis and testing of many alloys by trial-and-error in the laboratory. As such, it can dramatically reduce the time and cost involved in narrowing down the vast composition space. Our model offers a good prediction of the corrosion rate (19 out of 25 MPEAs in the validation dataset). However, it is challenging to rationalize the effect of individual descriptors in terms of materials science due to the limited training dataset of variable quality. The accuracy of the predictions could be further improved by increasing the training dataset size and data quality. Data sharing among researchers with standardized metadata is needed to realize this goal. The methods described in this work could also be applied to select degradation-resistant alloys based on the prediction of properties, such as Young's modulus, yield strength, and hardness, for a variety of engineering applications.

METHODS

Data collection

We have collected corrosion data of 142 MPEAs available in the literature from[9,13]. The results are from potentio-dynamic-polarization tests in acid, base, and salt solutions at room or higher temperatures. As the data includes corrosion current densities obtained from samples at the experimental scale, we assume that our model will predict localized corrosion at the point of active reactivity[60]. Here, we only screen single-phase disordered MPEAs, with the assumption that at the length scale of the corrosion process, the alloy surface will be corroded homogeneously at statistically distributed points on the surface with no segregation of a particular element in the microstructure. To quantify the extent of the corrosion, we converted the experimental current density to corrosion rate in mm/year. In a real application, the prediction of corrosion is made extremely challenging by the formation of a passivating film, the presence of impurities in the environment that can enhance or inhibit corrosion, and preferential dissolution of an element in the alloy[61]. Our aim is to offer a tool for the quick screening of MPEAs based on susceptibility to corrosion, just as a Pourbaix diagram[62] is used to assess the corrosion susceptibility of a metal. We show that despite the complexity of the corrosion process, variable quality of data drawn from different sources, and the sparseness of the dataset, actionable insights can be extracted about corrosion. The caveat is that this approach is based on uniform corrosion and may not adequately cover situations where the alloy corrodes incongruently. From an engineering perspective, it is anticipated that this machine learning aided corrosion-rate prediction can give a fair idea about the suitability of a set of alloys for a given application. The corrosion rate was obtained by converting the corrosion current density into corrosion rate (r_{corr}) by using the relation:

$$\text{Corrosion rate (mm/year)} = 3.27 \times 10^{-3} \times \frac{i_{corr}}{\rho} \times EW, \quad (3)$$

where ρ is the density in g/cc, i_{corr} is the corrosion current density (in μA/cm^2) and *EW* is the equivalent weight of the alloy given as:

$$EW = \left(\sum \frac{n_i f_i}{W_i}\right)^{-1} \quad (4)$$

where n_i is the valency of the *i*-th element of the alloy, f_i is the mass fraction of the *i*-th element, and W_i is the atomic weight of the *i*-th element in the alloy. To minimize the skew of 4–5 orders of magnitude in the target property i.e., corrosion rate (mm/year), the natural log of corrosion rates have been taken and set as target property. Out of the 142 MPEAs, we have included 117 in the training set, and the remaining 25 in the validation set. Once the model makes a prediction, the result is calculated as the antilog of the Ln (r_{corr}), which is in mm/year and can be directly compared with the test data r_{corr} values.

Materials descriptors space construction

One of the significant pillars of a ML algorithm is the set of descriptors used in constructing the algorithm. Since the degree of relevance of any given descriptor is unknown in the beginning, all potentially relevant descriptors even mildly affecting the target property should be listed while constructing the dataset. All descriptors based on thermodynamic, atomic, physical and chemical parameters should be considered. The thermodynamic parameters like mixing entropy (Ω)[63], a geometric parameter (λ) based on mixing entropy and atomic size mismatch[64], and mixing enthalpy control the process of MPEA phase formation. These parameters have been widely used for phase classification in MPEAs. Atomic parameters like electronegativity mismatch (χ) and atomic size mismatch (δr) have previously been used for classifying phases and predicting mechanical properties[7]. Physical parameters including melting point, compression modulus, and density have been used to predict stable phases[57] and hardness[7]. A rationale behind using melting point as a descriptor is that it represents the bond strength[65]. Additionally, previous work has also used chemical compositions to characterize the effect of addition of certain elements on specific target properties like oxidation behavior and corrosion rates[66]. The present work aims to predict the corrosion rate of general MPEAs (mm/year). For that purpose, we have included previously used parameters for corrosion-rate prediction and additional relevant parameters that have not been used to predict corrosion rate previously in the literature. A complete list of the 30 descriptors is given in Supplementary Table 2. The additional parameters include the maximum and minimum of the reduction potentials of all elements in the MPEA, and atomic percent of Cr, Al, Cu, Ti, Ni, Sn, and Mo. The elements Cr, Al, and Cu can form a protective oxide layer during corrosion and Mo triggers the Cr to form a protective layer[67], which makes them significant for consideration as descriptors in the ML algorithm. The presence of halide ion breaks the passive layer and reduces corrosion resistance due to halide attack[68]. Hence, concentration of halide in the corrosion test solution has been included as a descriptor. The severity of the environment characterized by the pH has been previously shown to have a significant influence on the pitting morphology[68]. The corrosion rate is higher in acidic or basic solutions relative to the rate in solutions with neutral pH[69].

We calculated the 30 descriptor values for all 142 alloys in our dataset and we set the r_{corr} as the target property. Since our descriptors vary in magnitude, we scaled down those descriptors that have a larger magnitude relative to other descriptors. More specifically, we scaled down ΔT_m by 100 and both Ω and T_m by 1000, and we scaled up Δχ and δ by 100. All descriptor values then ranged from 0 ~|10|. This approach is slightly different from regular normalization in the sense that only biasing is performed on the descriptors by multiplying with an appropriate bias value so scaled data are the same order of values for each descriptor. This minimizes the bias from one descriptor to another as seen by the ML model. In our dataset, some descriptors like Ω have values that range from 960 to 45800, with most values lying towards the 10^3 order of magnitude. In such a case a Min-Max normalization or a Z-score normalization would reduce majority of the values to the order of 10^{-2} for the descriptor Ω, while keep other descriptors in the order of 10^{-1}. Hence to avoid a wide variation in the values of the descriptors, such a scaling technique has been adopted. This ensures reducing the training time by starting the training process for each descriptor within the same timescale[70].

Machine-learning model

In our initial trials, we employed the least absolute shrinkage and selection operator (LASSO), random forest (RF), support vector machine (SVM) and the GBR to predict the corrosion rate. At this stage, we selected the algorithm with the maximum R^2 value and hence we present the R^2 values obtained from all models in Table 2. Ten-fold cross-validation was employed during the training. The GBR shows the highest training R^2 value of 0.93 and a testing R^2 value of 0.61 and is therefore selected as the final model for the prediction of corrosion rate.

GBR is an ensemble type algorithm. Ensemble methods are meta-algorithms that combine several base models to produce a better predictive model. To decrease variance, a bagging ensemble method can be used. To decrease bias, a boosting ensemble method can be used. Boosting was first introduced by Robert Schapire[71]. Gradient Boosting, Ada Boost and XGBoost all fall under boosting ensemble methods. A boosting method converts weak learners to strong ones. Usually, decision stumps are used as the base weak learners, but this is not always the case. Most boosting methods build models in stage-wise fashion and they generalize the model by optimizing an arbitrary differentiable loss function.

GBR is a machine-learning technique based on boosting method which can be used for both regression and classification[31,38]. The initial idea of gradient boosting was observed by Leo Breiman[72]. Friedman developed the explicit regression gradient boosting algorithm[73]. The algorithm for GBR is given below.

Table 2. The training and testing R^2 values obtained from the various models.

Model	Training R^2 value	Testing R^2 value
LASSO	0.21	0.11
RF	0.85	0.55
SVM	0.91	0.37
GBR	0.93	0.61

The GBR shows the highest score and hence is selected as the final model.

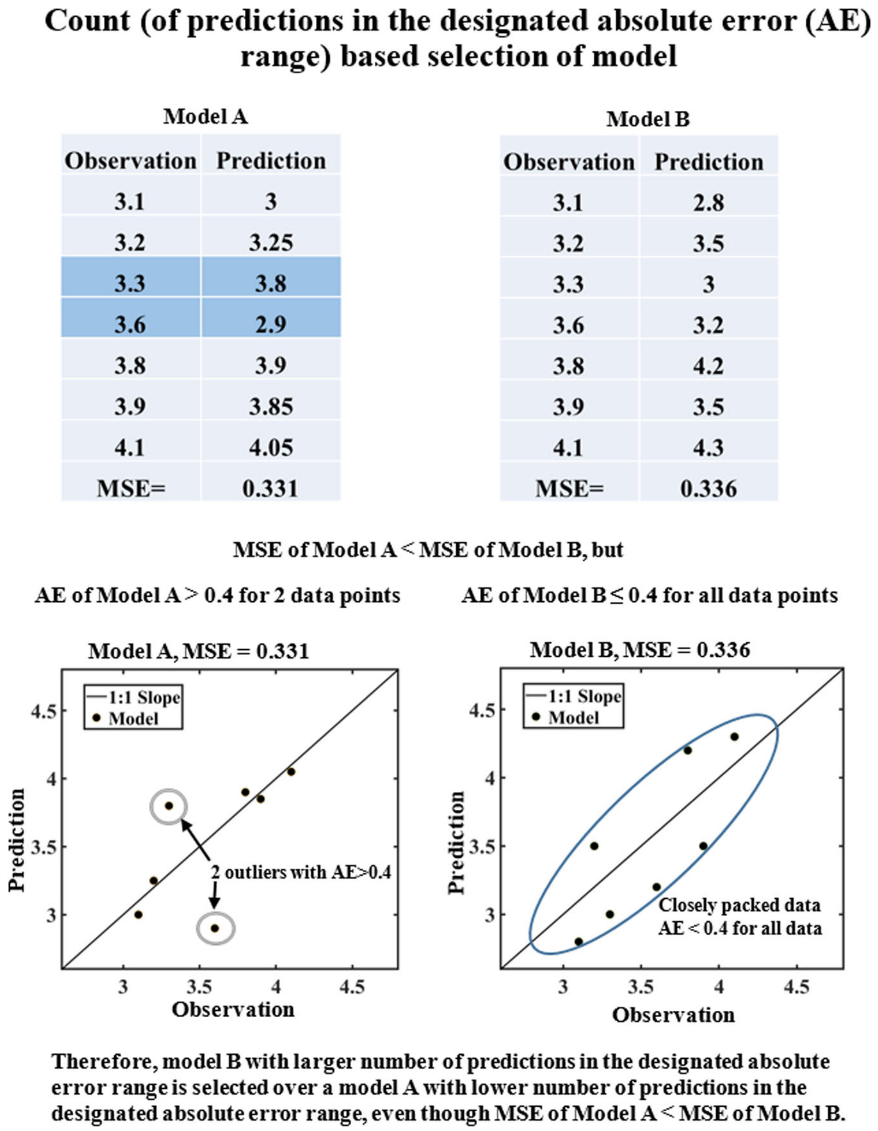
GBR Algorithm

```
Input: Training Data $\{(x_i, y_i)\}_{i=1}^{n}$,
  Differential Loss Function $L(y, f(x))$,
  Number of Iteration M

  1. Initialize $f_0(x) = argmin_\gamma \sum_{i=1}^{N} L(y_i, \gamma)$.
  2. For m = 1 to M:

 (a) For i = 1, 2, ……, N compute $r_{im} = -\left[\frac{\delta L(y_i, f(x_i))}{\delta f(x_i)}\right]_{f=f_{m-1}}$
 (b) Fit a regression tree to the targets $r_{im}$ giving terminal regions $R_{jm}$, j = 1, 2, …, $J_m$
 (c) For j = 1, 2 …, $J_m$ compute $\gamma_{jm} = argmin_\gamma \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma)$.
 (d) Update $f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$.

  3. Output $\hat{f}(x) = f_M(x)$
```



Count (of predictions in the designated absolute error (AE) range) based selection of model

Model A

Observation	Prediction
3.1	3
3.2	3.25
3.3	3.8
3.6	2.9
3.8	3.9
3.9	3.85
4.1	4.05
MSE=	0.331

Model B

Observation	Prediction
3.1	2.8
3.2	3.5
3.3	3
3.6	3.2
3.8	4.2
3.9	3.5
4.1	4.3
MSE=	0.336

MSE of Model A < MSE of Model B, but

AE of Model A > 0.4 for 2 data points

AE of Model B ≤ 0.4 for all data points

Therefore, model B with larger number of predictions in the designated absolute error range is selected over a model A with lower number of predictions in the designated absolute error range, even though MSE of Model A < MSE of Model B.

Fig. 7 A simplified illustration of the count (of predictions in the designated AE range) based selection of model: predictions of model A have smaller MSE than that of model B. However, the predictions of model B have a low AE ≤ 0.4 for all data as compared to model A that has AE > 0.4 for 2 data points. Hence, model B is more advantageous in making a greater number of good predictions than fewer more accurate predictions as done by model A. We, therefore, adopt this criterion for selecting the best model here. Data shown here for illustrative purpose only.

In this algorithm, we have used a training set $\{(x_1, y_1), \dots, (x_n, y_n)\}$ of known values of x and y. Our objective is to find an approximation $\hat{f}(x)$ to a function *f*(*x*) which minimizes the expected value of a loss function represented by $L(y, f(x))$. It does so by starting the model (step 1) with a constant function $f_0(x)$, given in the initialization step 1 above. In step 2 (a), the pseudo-residuals given by r_{im} are computed. In 2(b), a weak base learner was fitted to r_m, i.e., the model was trained using $\{(x, r_{im})\}_{i=1}^{n}$. Next, in step 2(c) the minimizer, γ_{jm} was found for $\sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma)$. In 2 (d), γ_{jm} obtained in 2(c), was used to update the model using expression $f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$, where $I(x \in R_{jm})$ is 1 if $x \in R_{jm}$ and 0 if $x \notin R_{jm}$. Step 2 is repeated M times to obtain the final model. For our study, we have used the GBR model through the sci-kit learn package[36].

Construction of regressor

Our problem demands the construction of a regressor to predict the Ln (r_{corr}) with the 30 descriptors. For this purpose, we have created a model using the GBR as it has previously been shown to be a robust algorithm for materials properties prediction[38]. To assess the accuracy of a model, we have devised a two-stage approach. To understand this, we present a sample dataset to explain the approach. Consider the dataset A and dataset B shown in Fig. 7. Predictions of model A has a smaller MSE than that of model B. But the predictions of model B have a low absolute error (AE) ≤ 0.4 for all data as compared to model A that has AE > 0.4 for 2 data points. The above difference is mathematically of lower magnitude but has an important consequence when the magnitudes of test data are of an extremely small order of magnitude such as those of corrosion rates reported in the literature (10^{-3} to 10^{-1} mm/year). Hence, we adopt this criterion of count (of predictions in the designated AE range) based selection of the model. This technique of model selection is advantageous in our case as it is more important to have a model that may have a higher MSE but can make fairly accurate predictions for all data points that are of nearly the same order of magnitude as that of the experimental dataset, rather than a model that has lower MSE but makes few very accurate predictions and a larger number of highly inaccurate predictions.

The model has been optimized for the best performance via hyperparameter optimization. The hyperparameters of the GBR model namely the learning rate and n-estimators, were varied from 0.005 to 0.20 and from 50 to 500, respectively, to find the point at which the minimum MSE is produced using the count-based model selection technique described above. It was found that a minimum MSE of 0.016 mm/year was obtained at a learning rate of 0.01 and 150 n-estimators.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

CODE AVAILABILITY

The codes that support the findings of this study are available from the corresponding author upon reasonable request.

Received: 30 April 2021; Accepted: 18 October 2021;
Published online: 31 January 2022

REFERENCES

1. Koch, G. in *Trends in Oil and Gas Corrosion Research and Technologies:Production and Transmission* 3–30 (Woodhead Publishing, 2017).
2. Yeh, J. W. et al. Nanostructured high-entropy alloys with multiple principal elements: novel alloy design concepts and outcomes. *Adv. Eng. Mater.* **6**, 299–303 (2004).
3. Birbilis, N. et al. A perspective on corrosion of multi-principal element alloys. *npj Mater. Degrad.* **5**, 1–8 (2021).
4. Tsai, M.-H. & Yeh, J.-W. High-entropy alloys: a critical review. *Mater. Res. Lett.* **2**, 107–123 (2014).
5. Roy, A. et al. Lattice distortion as an estimator of solid solution strengthening in high-entropy alloys. *Mater. Charact.* **172**, 110877 (2021).
6. Gianelle, M. et al. A novel ceramic derived processing route for Multi-Principal Element Alloys. *Mater. Sci. Eng. A* **793**, 139892 (2020).
7. Rickman, J. M. et al. Materials informatics for the screening of multi-principal elements and high-entropy alloys. *Nat. Commun.* **10**, 2618 (2019).
8. Roy, A., Munshi, J. & Balasubramanian, G. Low energy atomic traps sluggardize the diffusion in compositionally complex refractory alloys. *Intermetallics* **131**, 107106 (2021).
9. Shi, Y., Yang, B. & Liaw, P. K. Corrosion-resistant high-entropy alloys: a review. *Metals* **7**, 43 (2017).
10. Zhang, Y. et al. Microstructures and properties of high-entropy alloys. *Prog. Mater. Sci.* **61**, 1–93 (2014).
11. Gao, M. C. in *High-Entropy Alloys,* 369–398 (Springer International Publishing, 2016).
12. Qiu, Y. et al. Corrosion characteristics of high entropy alloys. *Mater. Sci. Technol.* **31**, 1235–1243 (2015).
13. Qiu, Y. et al. Corrosion of high entropy alloys. *npj Mater. Degrad.* **1**, 1–18 (2017).
14. Ye, Y. et al. High-entropy alloy: challenges and prospects. *Mater. Today* **19**, 349–362 (2016).
15. Miracle, D. B. & Senkov, O. N. A critical review of high entropy alloys and related concepts. *Acta Mater.* **122**, 448–511 (2017).
16. Taufique, M. et al. Impact of iodine antisite (I_{Pb}) defects on the electronic properties of the (110) $CH_3NH_3PbI_3$ surface. *J. Chem. Phys.* **149**, 164704 (2018).
17. Zhang, C. et al. Computational thermodynamics aided high-entropy alloy design. *JOM* **64**, 839–845 (2012).
18. Jiang, C. & Uberuaga, B. P. Efficient ab initio modeling of random multicomponent alloys. *Phys. Rev. Lett.* **116**, 105501 (2016).
19. Saal, J. E. et al. Equilibrium high entropy alloy phase stability from experiments and thermodynamic modeling. *Scr. Mater.* **146**, 5–8 (2018).
20. Lederer, Y. et al. The search for high entropy alloys: a high-throughput ab-initio approach. *Acta Mater.* **159**, 364–383 (2018).
21. Sanchez, J. M. et al. Phase prediction, microstructure and high hardness of novel light-weight high entropy alloys. *J. Mater. Res. Technol.* **8**, 795–803 (2019).
22. Tapia, A. J. S. F. et al. An approach for screening single phase high-entropy alloys using an in-house thermodynamic database. *Intermetallics* **101**, 56–63 (2018).
23. Senkov, O. et al. Accelerated exploration of multi-principal element alloys with solid solution phases. *Nat. Commun.* **6**, 1–10 (2015).
24. Butler, K. T. et al. Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).
25. Roy, A. et al. Machine learned feature identification for predicting phase and Young's modulus of low-, medium- and high-entropy alloys. *Scr. Mater.* **185**, 152–158 (2020).
26. Roy, A. & Balasubramanian, G. Predictive descriptors in machine learning and data-enabled explorations of high-entropy alloys. *Comput. Mater. Sci.* **193**, 110381 (2021).
27. Kim, G. et al. First-principles and machine learning predictions of elasticity in severely lattice-distorted high-entropy alloys with experimental validation. *Acta Mater.* **181**, 124–138 (2019).
28. Yan, L. et al. Corrosion rate prediction and influencing factors evaluation of low-alloy steels in marine atmosphere using machine learning approach. *Sci. Technol. Adv. Mater.* **21**, 359–370 (2020).
29. Ghiringhelli, L. M. et al. Big data of materials science: critical role of the descriptor. *Phys. Rev. Lett.* **114**, 105503 (2015).
30. Xue, D. et al. Material descriptors for morphotropic phase boundary curvature in lead-free piezoelectrics. *Appl. Phys. Lett.* **111**, 032907 (2017).
31. Khakurel, H. et al. Machine learning assisted prediction of the young's modulus of compositionally complex alloys. *Sci. Rep.* **11**, 17149 (2021).
32. Jain, A. et al. Commentary: The Materials Project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
33. Broderick, S. R. et al. Tracking chemical processing pathways in combinatorial polymer libraries via data mining. *J. Comb. Chem.* **12**, 270–277 (2010).
34. Wolpert, D. H. & Macready, W. G. No free lunch theorems for optimization. *IEEE Trans. Evol. Comput.* **1**, 67–82 (1997).
35. Zhang, L. et al. Machine learning reveals the importance of the formation enthalpy and atom-size difference in forming phases of high entropy alloys. *Mater. Des.* **193**, 108835 (2020).
36. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
37. Risal, S. et al. Improving phase prediction accuracy for high entropy alloys with Machine learning. *Comput. Mater. Sci.* **192**, 110389 (2021).
38. Roy, A. et al. Machine learned feature identification for predicting phase and Young's modulus of low-, medium-and high-entropy alloys. *Scr. Mater.* **185**, 152–158 (2020).
39. Ahmad, Z. *Principles of Corrosion Engineering and Corrosion Control* (Elsevier, 2006).
40. Qiu, X.-W. & Liu, C.-G. Microstructure and properties of Al2CrFeCoCuTiNix high-entropy alloys prepared by laser cladding. *J. Alloy. Compd.* **553**, 216–220 (2013).
41. Yang, F. et al. The role of nickel in mechanical performance and corrosion behaviour of nickel-aluminium bronze in 3.5 wt.% NaCl solution. *Corros. Sci.* **139**, 333–345 (2018).

42. Chou, Y. L. et al. Pitting corrosion of the high-entropy alloy Co1.5CrFeNi1.5-Ti0.5Mo0.1 in chloride-containing sulphate solutions. *Corros. Sci.* **52**, 3481–3491 (2010).
43. Chen, Y. et al. Microstructure and electrochemical properties of high entropy alloys—a comparison with type-304 stainless steel. *Corros. Sci.* **47**, 2257–2279 (2005).
44. Hsu, C.-Y. et al. Effect of aluminum content on microstructure and mechanical properties of Al x CoCrFeMo 0.5 Ni high-entropy alloys. *JOM* **65**, 1840–1847 (2013).
45. He, J. et al. Effects of Al addition on structural evolution and tensile properties of the FeCoNiCrMn high-entropy alloy system. *Acta Mater.* **62**, 105–113 (2014).
46. Wang, W.-R. et al. Effects of Al addition on the microstructure and mechanical property of AlxCoCrFeNi high-entropy alloys. *Intermetallics* **26**, 44–51 (2012).
47. Zhang, K. & Fu, Z. Effects of annealing treatment on phase composition and microstructure of CoCrFeNiTiAlx high-entropy alloys. *Intermetallics* **22**, 24–32 (2012).
48. Tang, Z. et al. Aluminum alloying effects on lattice types, microstructures, and mechanical behavior of high-entropy alloys systems. *JOM* **65**, 1848–1858 (2013).
49. Kao, Y.-F. et al. Microstructure and mechanical property of as-cast,-homogenized, and-deformed AlxCoCrFeNi (0≤ x≤ 2) high-entropy alloys. *J. Alloy. Compd.* **488**, 57–64 (2009).
50. Chou, H.-P. et al. Microstructure, thermophysical and electrical properties in AlxCoCrFeNi (0≤ x≤ 2) high-entropy alloys. *Mater. Sci. Eng. B.* **163**, 184–189 (2009).
51. Jones, D. A. *Principles and Prevention of Corrosion* (Prentice Hall, 1996).
52. Firouzdor, V. et al. Corrosion of a stainless steel and nickel-based alloys in high temperature supercritical carbon dioxide environment. *Corros. Sci.* **69**, 281–291 (2013).
53. Hsu, Y.-J., Chiang, W.-C. & Wu, J.-K. Corrosion behavior of FeCoNiCrCux high-entropy alloys in 3.5% sodium chloride solution. *Mater. Chem. Phys.* **92**, 112–117 (2005).
54. Muangtong, P. et al. The corrosion behaviour of CoCrFeNi-x (x= Cu, Al, Sn) high entropy alloy systems in chloride solution. *Corros. Sci.* **172**, 108740 (2020).
55. Guo, S. Phase selection rules for cast high entropy alloys: an overview. *Mater. Sci. Technol.* **31**, 1223–1230 (2015).
56. Wang, Z. et al. Atomic-size and lattice-distortion effects in newly developed high-entropy alloys with multiple principal elements. *Intermetallics* **64**, 63–69 (2015).
57. Zhang, Y. et al. Phase prediction in high entropy alloys with a rational selection of materials descriptors and machine learning models. *Acta Mater.* **185**, 528–539 (2020).
58. Mazzarolo, A. et al. Anodic growth of titanium oxide: electrochemical behaviour and morphological evolution. *Electrochim. Acta* **75**, 288–295 (2012).
59. Wenzlick, M. et al. Data assessment method to support the development of creep-resistant alloys. *Integr. Mater. Manuf. Innov.* **9**, 89–102 (2020).
60. Elsener, B. Corrosion rate of steel in concrete—measurements beyond the Tafel law. *Corros. Sci.* **47**, 3019–3033 (2005).
61. Qiu, Y. et al. Real-time dissolution of a compositionally complex alloy using inline ICP and correlation with XPS. *npj Mater. Degrad.* **4**, 1–6 (2020).
62. Virtanen, S. in *Encyclopedia of Electrochemical Power Sources* (Elsevier, 2009).
63. Yang, X. & Zhang, Y. Prediction of high-entropy stabilized solid-solution in multi-component alloys. *Mater. Chem. Phys.* **132**, 233–238 (2012).
64. Singh, A. K. et al. A geometrical parameter for the formation of disordered solid solutions in multi-component alloys. *Intermetallics* **53**, 112–119 (2014).
65. Rickman, J. M. Data analytics and parallel-coordinate materials property charts. *npj Comput. Mater.* **4**, 1–8 (2018).
66. Bhattacharya, S. K. et al. Predicting the parabolic rate constants of high-temperature oxidation of Ti alloys using machine learning. *Oxid. Met.* **94**, 1–14 (2020).
67. Hazza, M. & El-Dahshan, M. The effect of molybdenum on the corrosion behaviour of some steel alloys. *Desalination* **95**, 199–209 (1994).
68. Wang, Y. et al. Effect of pH and chloride on the micro-mechanism of pitting corrosion for high strength pipeline steel in aerated NaCl solutions. *Appl. Surf. Sci.* **349**, 746–756 (2015).
69. Ashley, G. & Burstein, G. Initial stages of the anodic oxidation of iron in chloride solutions. *Corrosion* **47**, 908–916 (1991).
70. Jayalakshmi, T. & Santhakumaran, A. Statistical normalization and back propagation for classification. *IJCTE* **3**, 1793–8201 (2011).
71. Schapire, R. E. The strength of weak learnability. *Mach. Learn.* **5**, 197–227 (1990).
72. Breiman, L. *Arcing the Edge* (Statistics Department, Univ. California, 1997).
73. Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Ann. Stat.* **29**, 189–1232 (2001).

ACKNOWLEDGEMENTS

This effort was principally supported by the U.S. Department of Energy's (DOE) Office of Energy Efficiency and Renewable Energy (EERE) under the Advanced Manufacturing Office through Ames Laboratory, which is operated for the U.S. DOE by Iowa State University under contract DE-AC02-07CH11358. H.K. was supported in part through the National Science Foundation (NSF) Mathematical Sciences Graduate Internship (MSGI) Program sponsored by the NSF Division of Mathematical Sciences. This program is administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy (DOE) and NSF. ORISE is managed for DOE by ORAU. This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of its employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

AUTHOR CONTRIBUTIONS

A.R. and M.F.N.T. contributed equally to this work as first authors. They performed the dataset construction, data analysis, and drafted the manuscript. H.K. calculated the features. G.B. and D.J. oversaw results, discussion, interpretation, and edited the manuscript. R.D. provided technical expertise on MPEA data, extracted data from literature, interpreted the results, and reviewed the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41529-021-00208-y.

Correspondence and requests for materials should be addressed to M. F. N. Taufique.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.