npj computational materials

www.nature.com/npjcompumats

ARTICLE OPEN

Check for updates

# Machine-learning-based intelligent framework for discovering refractory high-entropy alloys with improved high-temperature yield strength

Stephen A. Giles[1]✉, Debasis Sengupta[1]✉, Scott R. Broderick[2] and Krishna Rajan[2]

Refractory high-entropy alloys (RHEAs) show significant elevated-temperature yield strengths and have potential to use as high-performance materials in gas turbine engines. Exploring the vast RHEA compositional space experimentally is challenging, and a small fraction of this space has been explored to date. This work demonstrates the development of a state-of-the-art machine learning framework coupled with optimization methods to intelligently explore the vast compositional space and drive the search in a direction that improves high-temperature yield strengths. Our yield strength model is shown to have a significantly improved predictive accuracy relative to the state-of-the-art approach, and also provides inherent uncertainty quantification through the use of repeated *k*-fold cross-validation. Upon developing and validating a robust yield strength prediction model, the coupled framework is used to discover RHEAs with superior high temperature yield strength. We have shown that RHEA compositions can be customized to have maximum yield strength at a specific temperature.

*npj Computational Materials* (2022)8:235 ; https://doi.org/10.1038/s41524-022-00926-0

## INTRODUCTION

High-entropy alloys (HEAs) are promising materials, which have garnered a tremendous amount of attention since their discovery in 2004[1–7]. Unlike traditional alloys, HEAs contain at least four elements in near-equal proportions, and the stability of the alloys are postulated to arise from higher configurational entropy[8]. Experimental studies have shown that refractory HEAs (RHEAs) possess superior high-temperature strength compared to super-alloys, making them an attractive class of alloys for further exploration[9–15] for potential use in high efficiency gas turbine engines. While HEAs provide tremendous opportunity due to the flexibility of the compositional space, they also pose stiff challenges to the material scientists tasked with exploring a design space with a huge number of possible compositions. Recently, Miracle et al. analyzed that using 75 elements that are not toxic, radioactive or noble gas, one can form 219 million 3–6 component base alloys[16]. If elemental composition is varied for each base alloy, the number of HEAs becomes more than 592 billion[16]. To date, only a tiny fraction of this composition space has been processed and characterized. One of the primary obstacles that impedes the accelerated development of HEA is the lack of generalized understanding of parameters that are responsible for dictating the mechanical and chemical behavior of these complex alloy systems. Although atomic and microstructural information can be obtained with modern high-resolution imaging techniques, these techniques are time-intensive and costly, thereby limiting the extent to which the vast composition space can be explored and characterized.

As a consequence, HEA research has mainly revolved around identifying rules for phase formation and atomic and micro-structural parameters potentially affecting mechanical properties and to develop criteria for classification of phases. Extensive work has been reported in the literature[17–19] to classify phases using parameters, such as the atomic size mismatch ($\delta$), the enthalpy of mixing ($\Delta H_{mix}$), the entropy of mixing, $\Delta S_{mix}$, and dimensionless quantities $\Omega$, $\Phi$, and $\varphi$. Ranges of these parameters have been proposed that can lead to different phases. However, developing guidelines for the improvement of mechanical properties is a relatively less studied area. Earlier reports in this area are primarily concentrated on experimental efforts concerning the modification of a base alloy and finding correlations between mechanical properties with parameters, such as lattice distortion[20–23], grain size[10,24,25], and phases[26–28]. While this approach can be beneficial when narrowly focused on a particular base alloy system, it has limited transferability to other systems. Recently, Maresca et al. have used an edge dislocation-based analytical model and generated ~10 million compositions with increasing theoretical yield strengths[29]. Using their analytical model, a large number of HEAs were discovered that would potentially result in yield strength improvement over the existing HEAs. Likewise, Rao et al. have used the Maresca et al. model recently as a guide for designing HEA compositions[30]. However, the use of ML models may offer superior accuracy for the prediction of mechanical, phase, and other physicochemical properties. In fact, a number of articles have recently appeared that have applied various ML methods to predict HEA phases[31–33]. In regards to other properties, Wen et al. developed an ML-based model to predict hardness in the $Al_xCo_yCr_zCu_uFe_vNi_w$ system, which was coupled with an experimental synthesis and optimization approach[34]. ML-based studies on the catalytic and thermal expansion properties of HEAs have also performed[33,35,36]. An ML-based framework informed by extant literature data is still a mostly untapped route for the development of HEAs with superior mechanical properties, particularly for RHEAs which possess the potential for high strength in high-temperature applications.

Currently, HEA development starts with processing a base alloy followed by generating a few HEAs with varying compositions. If the properties of the alloy compositions did not improve, the

[1]CFD Research Corporation, 6820 Moquin Dr. NW, Huntsville, AL 35806, US. [2]Department of Material Design and Innovation, University at Buffalo, 120 Bonner Hall, Buffalo, NY 14260, USA. ✉email: stephen.giles@cfd-research.com; debasis.sengupta@cfd-research.com

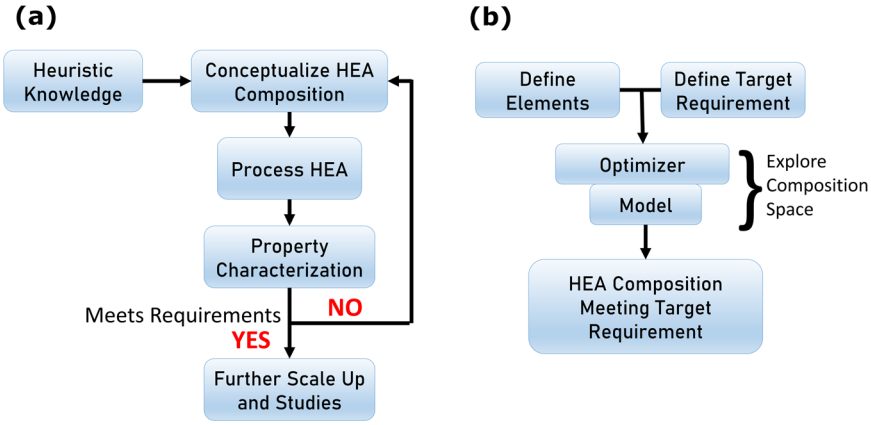


**Fig. 1 Alloy development strategies. a** Traditional HEA processing requires exploring the composition space experimentally to achieve a target property. This makes exploring a large HEA composition space and discovering HEAs difficult, time-consuming, and expensive; **b** the flowchart of the procedures which explore the composition space intelligently, using optimization coupled with a model to achieve a target requirement. The approach narrows down the search space for material scientists, potentially accelerating HEA discovery.

search for alloy compositions with improved properties continues (Fig. 1a). The work presented in this paper aims at substituting this "experiment-only" loop with intelligent ML-based models to screen HEAs and narrow down the search space providing material scientists only a few alloys potentially showing improved properties to process and characterize (Fig. 1b). This paper, as an example, focuses on yield strengths of refractory high-entropy alloys (RHEAs) and begins with developing a comprehensive forward ML model *via* identifying critical descriptors selected from a set of a large number of descriptors. The forward model was then coupled with a stochastic genetic algorithm[37] to discover RHEAs compositions with improved yield strengths. Valuable insights were gained with respect to identifying elements contributing to improving yield strength at room and high temperatures. The physical and thermodynamic descriptors were also analyzed to understand their roles in improved yield strength. Using the ML-based model, we discovered RHEA compositions with yield strengths customized for specific temperatures. Beginning with known experimental compositions of RHEAs, we have discovered several compositions with significant improvement in yield strength over the starting compositions. Although this work primarily focuses on the prediction of RHEA yield strength, the scope of future work based on this framework is much broader. Model development for other properties, such as ductility, creep behavior, and high-temperature oxidation resistance, will be undertaken to enable discovery of alloys satisfying multiple property requirements.

## RESULTS AND DISCUSSION

### Data collection and descriptor generation

The RHEA yield strength data were obtained from the publication by Couzinie et al.[38]. The data in the publication contained both compression and tension data. For the present work, only compression data was used as the tension data were insignificant in number. Since the article did not contain detailed processing conditions (e.g., annealing temperature, annealing time, etc.), we extracted the specific processing conditions from the original articles referenced by Couzinie et al.[38]. The revised dataset contained 280 temperature-dependent yield strength values for alloys. Of these 280 reported values, there were three measurements that underwent premature fracture, and thus, their yield strength measurements were more similar to a fracture strength. These measurements were kept in the dataset used for training and validation to account for the possibility of some alloys undergoing premature fracture. Also, as stated in the Methods section, the Maresca et al. model[39] was used to interpolate between experimental data to improve predictions at 200 °C and 400 °C. Following the inclusion of these data, the final dataset used for model training and validation had a total of 314 temperature-dependent yield strength values. Once the data were collected, a large number of composition-based descriptors were generated using the *matminer* library[40]. Calculation for several additional composition-based descriptors, such as lattice and modulus of distortion, were implemented in *matminer*. These descriptors are based on the nominal composition. It has been noted previously that interstitial elements (e.g., nitrogen and oxygen) can increase the yield strength by as much as 500 MPa at small concentrations[41], however, these variations in the interstitial content are expected to be at least partially absorbed within the dataset and the proclivity of nominal compositions to atomically bind with interstitial elements. Prior to training and validation of the ML models, the descriptor set was reduced by removing descriptors which had either undefined, low-variance, or linearly dependent values. Specific details of our approach are provided in the Methods section. A comprehensive list of the descriptors that were made available for the feature selection and ML training process is provided in the Supplementary Information (SI).

### Machine learning model development

Figure 2a shows the parity plot comparing the measured yield strength values to the predicted values following training and validation using the random forest model[42]. The six descriptors chosen by the sequential feature selection (SFS) method[43,44] are the test temperature, $\Omega$ (a dimensionless ratio of entropic and enthalpic contributions), atomic size mismatch ($\delta$), tantalum modulus distortion ($\delta G_{Ta}$), fractional composition of molybdenum ($x_{Mo}$), and a base strength, $\sigma_{0,\ min0.5}$, determined from the yield strength of the individual elements in the alloy. As stated in the Methods section, the dimensionless quantity $\Omega$ is defined as $\Omega = T_m \Delta S_{mix}/|\Delta H_{mix}|$. The model was found to have a cross-validation regression coefficient, $R^2$ (CV), of 89.5%. All data in the parity plot are colored by the temperature at which the yield strength measurements were made. As expected, the measurements performed above room temperature generally have lower yield strengths than measurements performed at or below room temperature. Our model is shown to provide very good quantitative agreement over the entire temperature range. Particularly noteworthy is that this model can predict the yield strength of both single phase *and* multi-phase alloys without explicit inclusion of phase information in the data. However, Yang and Zhang have described the ability of $\Omega$ and $\delta$ to be predictive

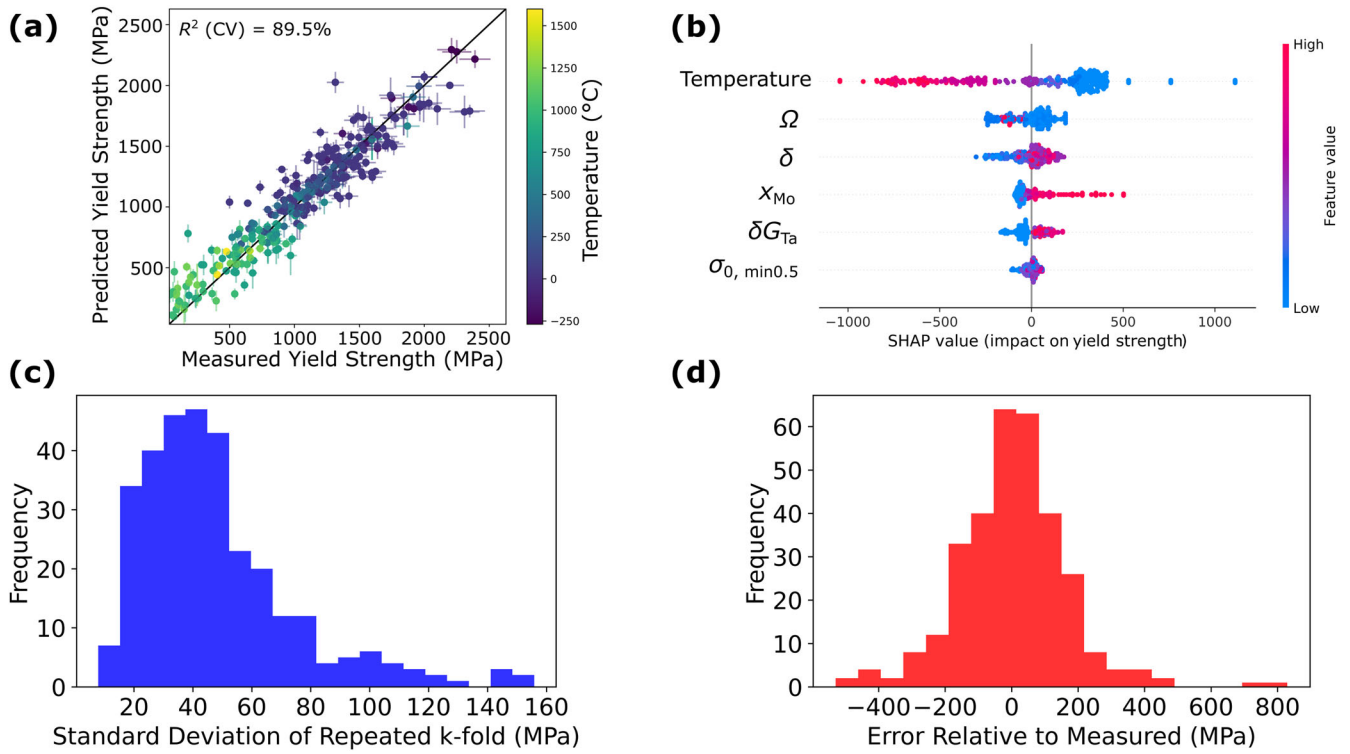


**Fig. 2 Model performance, explainability, and error analysis. a** Predicted vs. experimental yield strength for compression data using the model developed according to the procedures described in the Methods section. Note that we used a repeated k-fold method where 5-fold cross validation was performed 1000 times to compute the mean and standard deviation of each data point; **b** SHAP analysis is performed to identify descriptor importance. Features on the left are ordered according to their importance, with the most important feature (temperature) shown on top. The data points correspond to the individual alloy data points, where each have been colored according to the magnitude (high or low) of the feature in question. Positive SHAP values indicate that the yield strength is increased as a result of the feature value, whereas negative SHAP values indicate the yield strength is decreased due to the feature value. **c** Distribution of prediction standard deviations for each data point when 5-fold cross-validation is performed 1000 times. **d** Distribution of prediction errors relative to the measured values.

of the favorability of solid solution formation[19]. Therefore, the ability of our model to predict yield strength for both single and multi-phase RHEAs well is indicative of the importance of these descriptors. Models were also developed using LASSO regression[45], ridge regression[46], and gradient boosting regression[47]. These models are summarized in Supplementary Figs. 1–3 of the Supplementary Information. LASSO and ridge regression resulted in significantly lower accuracy due to their linear nature, whereas gradient boosting resulted in a $R^2$ (CV)comparable to that of random forest model shown in Fig. 2. The random forest and gradient boosting models are similar in that they are both ensemble models based on decision trees, and are capable of capturing complex, nonlinear relationships. Thus, it is not surprising that they give similar accuracies. The LASSO and ridge regressions are linear models with L1 and L2 regularization applied, respectively, and had lower accuracies due to missing inherent nonlinearity in the yield strength and descriptor relationships. Our present discussion is based on the results obtained with the random forest model.

To understand the effect and importance of each of the six selected descriptors on the predicted yield strength, we have applied the Shapley Additive Explanations (SHAP) technique[48,49] to physically interpret the forward model that was validated in Fig. 2a. SHAP analysis is a game-theoretic, local explanation method which computes the quantitative influence of model descriptors on the model output. Although never applied in materials science, SHAP analysis can provide some insights into the model over the entire data space which is otherwise difficult to obtain. As expected, the test temperature is shown via SHAP analysis to be the most important feature, with high temperature corresponding to a lower predicted yield strength and vice versa (Fig. 2b). As an additional study, we have examined the effect of replacing $T$ with the homologous temperature, $T/T_m$ (where $T_m$ is the melting point of the alloy using rule-of-mixture, see Method section for details), as Senkov et al. have suggested that this is the key variable dictating temperature-dependent yield strength[17]. We observe insignificant change in the model error and $R^2$ (CV), and the SHAP analysis showed virtually identical behavior of $T$ and $T/T_m$. The detailed results are included in SI (Figs. S4–S5). The other important conclusion of the SHAP analysis is that higher values of $\delta$, $\delta G_{Ta}$, and $x_{Mo}$ tend to positively contribute to the yield strength. The outlier red points for Ω result from smaller values of $|\Delta H_{mix}|$, thereby causing $\Omega$ to become very large (see Eq. (4)). The trend of $\Omega$ versus the SHAP impact on the model by $\Omega$ is shown in Supplementary Fig. 6. The trend reveals that the impact of $\Omega$ undergoes a such that for $\Omega < \sim 15$, $\Omega$ clearly has a positive impact on the yield strength, whereas for $\Omega > \sim 15$, $\Omega$ clearly has a negative impact on the yield strength. Therefore, this model implies the connection between the mechanical property and the parameters derived from atomic structure, and also provides a manner through which to directly link the two.

For the predicted values in Fig. 2a that were determined from performing 1000 repetitions of $k$-fold cross-validation (see the Method section for details), the standard deviation of the predicted value of each alloy was determined. The distribution of the standard deviations is shown in Fig. 2c. The distribution is log-normal due to all values being positive, by definition, and the average standard deviation was found to be 48 MPa. It should be pointed out that $k$-fold cross-validation is a well-established and effective validation method when the dataset size is small.

Although it is a common practice to validate model with one or two experimental measurements, this practice is inadequate for a statistically meaningful assessment of model accuracy. We additionally calculated the error in the predictions relative to the known experimental values. This error distribution is shown in Fig. 2d. In this case, the distribution is normal with a mean value near zero, and a characteristic mean *absolute* error (MAE) of 118 MPa. The corresponding percentage error distribution is shown in Supplementary Fig. 7. By contrast, the theoretical edge dislocation model proposed by Maresca and Curtin[39] was determined to have an MAE of 683 MPa for the same dataset, and consistently underpredicted the experimental yield strength. The corresponding parity plot and error distribution are provided in Fig. S8. The consistent underprediction of the yield strength by the edge dislocation model could be a result of many of the HEAs being screw-controlled. We note that Maresca et al. have also developed a complementary screw dislocation model[50]. However, the screw dislocation model involves a number of unknown material parameters that render it difficult to apply to HEA systems. Through repeated *k*-fold cross-validation, the ML model presented here is demonstrated to provide acceptable predictions for the whole range of RHEAs studied. Therefore, our model represents a significant step forward in the state-of-the-art as it not only has a better accuracy, but also concurs with some of the well-established physics.

### Experimental validation of machine learning model

We have also validated the developed machine learning model of yield strength with previously unseen experimental data. The validation of our model is shown in Fig. 3. The experimental data presented for CrMoNbV and NbTaTiV were taken from a recently published paper by Maresca et al.[41], who also presented yield strength predictions of an analytical model based on edge

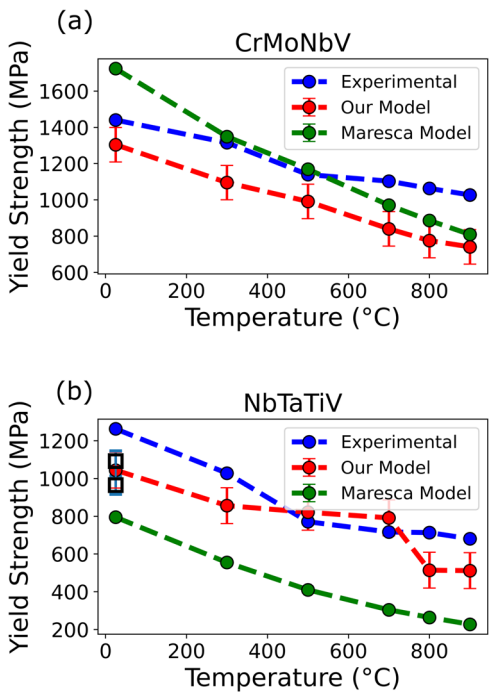


**Fig. 3 Validation of machine learning model with unseen experimental data, and comparison of the ML model to the edge dislocation model of Maresca et al. a** CrMoNbV, and **b** NbTaTiV. "Experimental" and "Maresca Model" data taken from Maresca et al[29]. For NbTaTiV, the black squares indicate the yield strength for the alloy reported by two other references[9,66]. A 95% confidence level error (±94 MPa) is shown for Our Model.

dislocations. Our model has a mean absolute error of 224 MPa for CrMoNbV and 147 MPa for NbTaTiV with respect to the experimental data. However, for NbTaTiV, other literature data has reported lower yield strength values at 25 °C which closely match our model's predictions. One potential explanation for the observed deviation from the experimental values is the presence of interstitials (e.g., nitrogen or oxygen) in the alloy. Relatively small concentrations of interstitials (e.g., ~2 at.%) have been shown to lead to changes in the yield strength as large as 500 MPa[51]. While our model has the advantage of being driven by experimental data, and thus is implicitly trained on data for alloys which likely contain interstitials and less susceptible to large errors from this source, there can still be significant deviation if the concentration of interstitials is relatively large in comparison to the data that has been used for model training. Overall, however, validation of our model demonstrates good quantitative accuracy and close agreement in the temperature dependence of yield strength.

### Alloy discovery: known base alloys as a starting point

Once the ML model is developed, various approaches can be followed in order to discover alloys with improved yield strength. The first approach, which experimentalists frequently follow, is to select a known alloy and improve its properties by varying the atomic fractions and/or adding elements. Our RHEA discovery approach mimics this procedure computationally and is schematically shown in Fig. 4a. The procedure varies the atomic composition as dictated by the genetic algorithm, and computes the yield strength using the forward model via computing the descriptors used in the model. Finally, convergence is achieved when the yield strength is maximized with respect to the elemental composition through the intelligent search of the compositional space towards the direction of improving yield strength. The details of the optimization process can be found in the Methods section.

As a demonstration case, we first selected $AlMo_{0.5}NbTa_{0.5}TiZr$ which already shows high yield strength at high-temperature. Therefore, it is logical to use this alloy as a starting point for the genetic algorithm optimization to examine whether it is possible to improve its yield strength further by manipulating the composition. To begin the optimization, we chose a temperature at which to maximize the yield strength. As an example, we are primarily concerned with discovering RHEA compositions with maximized yield strength at room temperature (25 °C) and at elevated temperature (1000 °C). The optimization progress was visualized by performing principal component analysis (Supplementary Figs. 10–11). Figure 4b compares the yield strength vs. temperature profiles of the base composition, the compositions that maximizes yield strength at 25 °C and 1000 °C. The base alloy and the 25 °C optimal alloy behave very similarly with respect to temperature. At 25 °C the optimized alloy improves upon the base alloy by 90 MPa, or 4%. The minor improvement in the yield strength indicates that the base alloy was already near optimal for yield strength at 25 °C. Furthermore, the base alloy and the 25 °C optimal alloy have a nearly identical temperature dependence. The 1000 °C optimal alloy exhibits a notably different temperature dependence, with the yield strength being approximately constant between 25–800 °C. The 1000 °C optimal alloy has a significantly reduced yield strength at 25 °C (1398 MPa), yet improves upon the base alloy yield strength by 13% at 1000 °C (848 MPa). The remarkable temperature insensitivity of the 1000 °C optimal alloy, which shows a statistically insignificant change in the yield strength between room temperature and 800 °C, is particularly worth noting. Experimentalists have frequently concerned themselves with searching for alloys that maintain high yield strength as they are exposed to increasingly high temperatures. The relative temperature insensitivity predicted here is similar to that observed by Senkov et al. for the $AlMo_{0.5}NbTa_{0.5}TiZr_{0.5}$ alloy[52]. In the case of the 1000 °C optimal

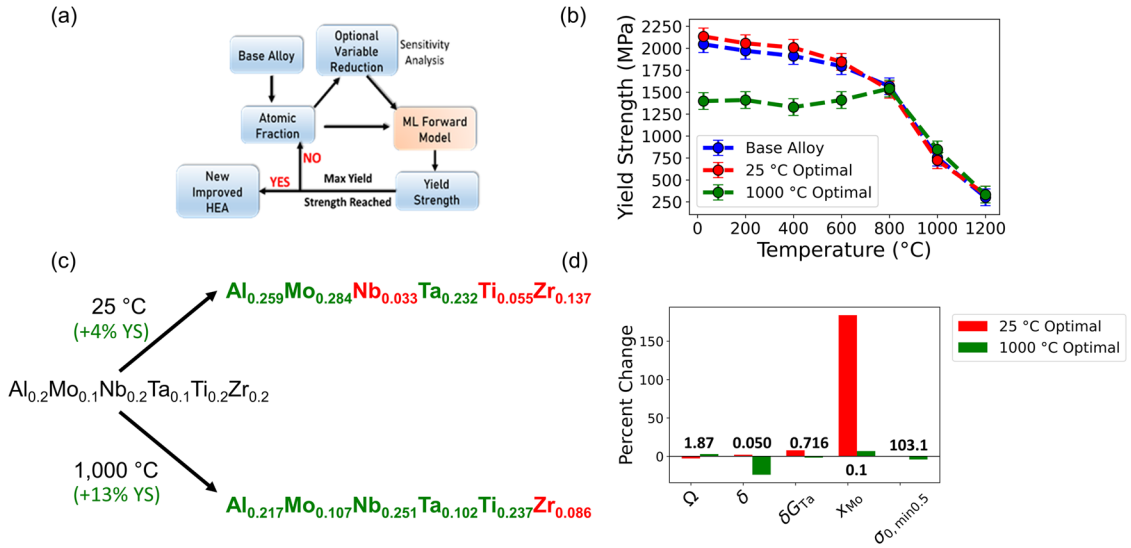


**Fig. 4 Optimization of AlMoNbTaTiZr for yield strength improvement. a** The flowchart outlines the steps for optimizing the HEA composition from a base HEA. The only input required for the model is a base HEA composition; **b** yield strength vs. temperature profile for the original base alloy and the two optimized alloys. **c** comparison of the element fractions for the original base alloy and the alloys optimized for 25 °C and for 1000 °C. Elements that were increased are shown in green text, while elements that were decreased are shown in red text; **d** percent change, relative to the base alloy, of each feature which serves as a direct input to the yield strength model. The numbers in bold black font are the values of each feature in the base alloy. A 95% confidence level error (±94 MPa) is shown for prediction.

alloy, given that the value of $\delta$ is the prevailing difference between the 1000 °C optimal alloy and the base alloy, our model suggests that lower values of δ could be correlated to a lessened temperature dependence of the yield strength. However, because there were three alloys in the dataset whose yield strength measurement was more similar to a fracture strength, we believe it is appropriate to caution readers that the temperature insensitivity observed here could be indicative of the 1000 °C optimal alloy being likely to prematurely fracture. In Fig. 4c the element fractions of the base alloy ($AlMo_{0.5}NbTa_{0.5}TiZr$) is compared to the final composition of the alloys optimized for 25 °C and 1000 °C yield strength. In the 25 °C alloy, the Mo fraction is shown to increase significantly relative to the base alloy. For the 1000 °C optimized alloy, on the other hand, the Mo fraction remains approximately unchanged while the Zr fraction decreases significantly. Simulation results thus indicate composition that improves room temperature strength does not necessarily improves high-temperature yield strength. The approach developed here therefore particularly useful for predicting RHEA compositions customized to have superior yield strength at a specific temperature of interest. Figure 4d shows the changes for each of the descriptors serving as direct inputs to the ML model at two different temperatures. Evident from observing the changes in the descriptor values relative to the base alloy is that the $x_{Mo}$ constitutes by far the largest change for the 25 °C optimal alloy. On the other hand, $\delta$, the atomic size mismatch, undergoes the largest change for the 1000 °C optimal alloy. That the optimal atomic size mismatch is significantly smaller for the 1000 °C optimal alloy indicates that increasing lattice distortion could have a deleterious impact on the yield strength, particularly at high temperature. Our results suggest that the impact of model descriptors on the yield strength is complex and temperature-dependent, thereby pointing towards a different physics which should be considered when designing alloys for a particular temperature range.

The aforementioned inverse optimization for both 25 °C and 1000 °C starting from $AlMo_{0.5}NbTa_{0.5}TiZr$ constitutes just one example of workflow that has been developed to improve upon a base alloy. The choice of the $AlMo_{0.5}NbTa_{0.5}TiZr$ RHEA to serve as the base alloy was informed by its already having an exceptional high-temperature yield strength. Other rationales could be used that could lead one to choose a different known alloy as a starting point for the optimization. For example, in addition to yield strength, ductility is a mechanical property of frequent concern. Yet, yield strength and ductility have a well-known tradeoff, where increasing ductility tends to lead to lower yield strength (Supplementary Fig. 12). Thus, it is logical to start with compositions that already possess a good room temperature ductility and manipulate them to maximize yield strength through optimization. While there is no certainty that the discovered compositions with improved yield strengths will retain their ductility, ML models can be developed to maximize both yield strength and ductility simultaneously. Such models are currently under development by us where one can find compositions that simultaneously improve multiple properties. In this paper, we have chosen two additional base alloys, $Mo_{0.3}NbTiV_{0.3}Zr$ and HfNbTaTiZr, which show high room temperature ductility. $Mo_{0.3}NbTiV_{0.3}Zr$ has a moderate yield strength at 25 °C (1312 MPa) and a ductility of 49.3%[53]. HfNbTaTiZr also has a high ductility (33.3%) but a very poor yield strength at 25 °C (929 MPa)[54].

The optimizations were performed on these two base compositions at 25 °C and 1000 °C, and the results of their temperature dependent yield strengths are shown in Fig. 5. Immediately apparent in Fig. 5 is that greater improvements in the yield strength are seen for these two cases relative to what was $AlMo_{0.5}NbTa_{0.5}TiZr$ case in Fig. 4d. This is a consequence of both of these base alloys being particularly sub-optimal with respect to yield strength as they were selected for their good *ductility* instead. It can be seen that significant improvements were achieved at both 25 °C and 1000 °C over the base alloy compositions; in particular for HfNbTaTiZr, 80% improvement (increase from 962 MPa to 1731 MPa) was achieved at 25 °C, and 36% improvement (252 MPa to 344 MPa) was obtained at 1000 °C by optimization of the elemental composition.

Inspecting the elemental compositions provided in Fig. 5 reveals some important trends with respect to which elements are modified to render an improved alloy. For the 25 °C optimal alloy in Fig. 5a, the concentrations of Nb, Ti, and Zr were decreased by approximately equivalent amounts relative to the base alloy, whereas the concentration of Mo and V were both increased significantly. By contrast, for the 1000 °C optimal alloy, the Ti fraction was increased and the V fraction was nearly

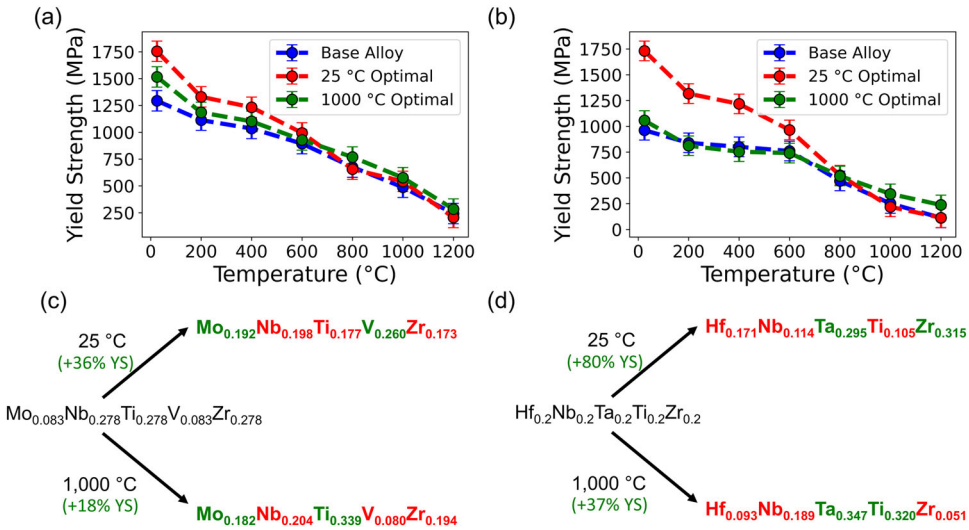


**Fig. 5 Composition optimization of known alloys to improve yield strength at 25 °C and 1000 °C. a** $Mo_{0.3}NbTiV_{0.3}Zr$ base alloy case, and the (**b**) HfNbTaTiZr base alloy case. The original base alloy is shown together with the optimized alloy for 25 °C and the optimized alloy for 1000 °C. The element compositions corresponding to each case are shown in panels (**c**) and (**d**). Element fractions that were increased during the optimization are shown in green text, while element fractions that were decreased are shown in red text. A 95% confidence level error (± 94 MPa) is shown for prediction.

unchanged. Similar to the alloy optimized for yield strength at 25 °C, however, the Mo concentration was also increased. That Mo generally tends to improved yield strength is consistent with the SHAP analysis discussed in Fig. 2b. Likewise, considering the HfNbTaTiZr base alloy case in Fig. 5b, Ti is also shown to be present in a high concentration for the 1000 °C optimal alloy and in low concentration for the 25 °C optimal alloy. Meanwhile, Ta and Zr in the HfNbTaTiZr base alloy exhibit a similar relationship with the yield strength that Mo and V exhibited in the $Mo_{0.3}NbTiV_{0.3}Zr$ base alloy. That increasing the Ta element fraction often corresponds to improved yield strength was also evident from the earlier SHAP analysis where the $\delta G_{Ta}$ was shown to have a positive impact on the yield strength.

### Virtual 10 element alloy discovery

So far, we have restricted the compositional variation within the elements of a base alloy without adding or exchanging any elements. A total of 10 elements (i.e., Al, Cr, Hf, Mo, Nb, Ta, Ti, V, W, and Zr) are present in the training dataset. Inclusion of additional elements was first performed via using the equimolar 10-element alloy as a starting point. Comparison of the temperature-dependent yield strength for the 10-element base alloy, the 25 °C optimal, and 1000 °C optimal is provided in Fig. 6a. The comparison reveals that both optimized alloys result in a significant improvement at 25 °C, and in fact, the two optimized alloys have very similar yield strengths at 25 °C. However, at 1000 °C, the 25 °C optimal alloy has a lower yield strength than the base alloy. The 1000 °C optimal alloy, however, maintains a high yield strength resulting in a 48% improvement upon the base alloy. The elemental compositions are provided in Fig. 6b. Common refractory elements (Hf, Nb, and Mo) are shown to be particularly beneficial when optimizing at 1000 °C, whereas lighter elements (Al and Ti) are more beneficial at 25 °C.

### Virtual base alloy search

To mimic typical HEA compositions, which often include only 4–6 principal elements, we also extended our approach to systematically predict the yield strength for every unique five-element RHEA from the set of 10 elements. Choosing 5 elements from this set of 10 elements yields 252 equiatomic alloy combinations, some of which have been reported in the literature. The forward ML model was used to predict the yield strength of all 252 equiatomic alloys, at both 25 °C and 1000 °C. The distribution of predicted yield strengths for 25 °C and 1000 °C are shown in Fig. 7a, b, respectively. At 25 °C the yield strengths are approximately normally distributed with a mean of 1457 MPa and a standard deviation of 175 MPa. The yield strength distribution at 1000 °C, however, is clearly bimodal. One mode is located at ~200 MPa, while the other mode is located at ~550 MPa. The bimodal behavior at 1000 °C is correlated with the presence of Mo. If Mo is present in the alloy, the alloy belongs to the class of materials with an average predicted yield strength of ~550 MPa at 1000 °C. In the absence of Mo, the alloy has a notably reduced high-temperature yield strength property. Existing theories of screw and edge dislocation mechanisms in HEAs by Maresca et al.[39,50] indicate that HEAs which predominantly follow the screw deformation mechanism typically do not retain their strength at high temperature, whereas edge-controlled deformation can retain mechanical strength at temperatures as high as 1900 K[39,41,50]. Furthermore, screw dislocations have been shown to increase in mobility relative to edge dislocations when the alloy transitions into ductile behavior (e.g., above a certain temperature)[55]. The ductile-to-brittle transition itself has been characterized by various criteria, including Pugh's criterion, which is the ratio of the shear modulus to the bulk modulus[56]. To this end, we have computed the mean value of Pugh's criterion for the Mo-containing RHEAs to be 0.368, whereas non-Mo-containing alloys had a mean Pugh's criterion of 0.295. Therefore, these two pieces of information, while not conclusive, do suggest Mo-containing RHEAs may have a greater likelihood of being edge-controlled, thus offering a theoretical justification for why a bimodal distribution is seen at high temperature but not at low temperature.

The equiatomic alloys, AlMoTaTiZr and AlHfMoTaTi, with the highest yield strengths at 25 °C and at 1000 °C, respectively, were identified and selected for further optimization. The two alloys are notably similar, with the only difference being the replacement of Zr with Hf. Both of these equiatomic alloys were not present in the compiled Couzinie et al. RHEA dataset[38] used for training and validation, nor, to the best of our knowledge, have they been reported in any of the RHEA literature. In Fig. 7c, d, the yield strengths of these two equiatomic alloys were further improved through optimization of their element fractions. Since both equiatomic base alloys had high yield strength, optimizing the

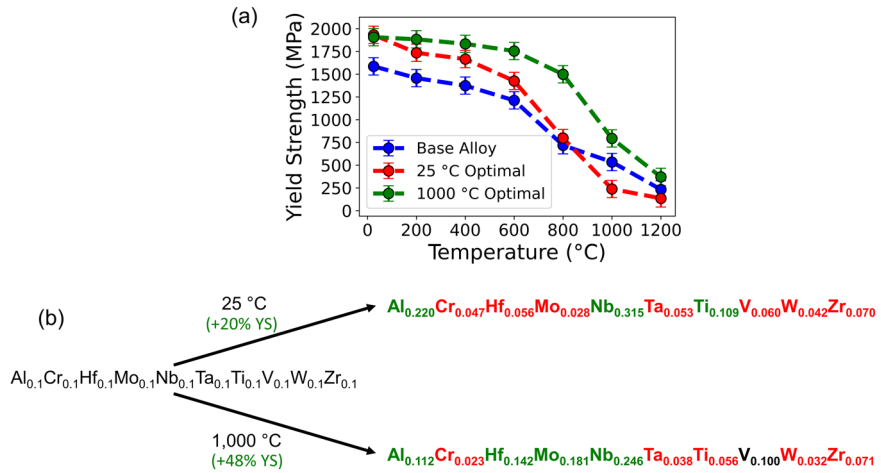


**Fig. 6 Composition optimization of a 10-element alloy system for 25 °C and 1000 °C. a** Temperature-dependent yield strength, and (**b**) corresponding changes in elemental composition starting from the equimolar base alloy and optimizing at 25 °C and 1000 °C. A 95% confidence level error (±94 MPa) is shown for prediction.

element fractions resulted in only slight improvements to the yield strength, with the $Al_{0.239}Mo_{0.123}Ta_{0.095}Ti_{0.342}Zr_{0.201}$ alloy resulting in an improvement of 7% at 25 °C, and the $Al_{0.151}Hf_{0.236}Mo_{0.137}$-$Ta_{0.131}Ti_{0.345}$ alloy resulting in an improvement of 2% at 1000 °C. Also, worth noting is that, assuming a rule-of-mixtures, the $Al_{0.239}Mo_{0.123}Ta_{0.095}Ti_{0.342}Zr_{0.201}$ alloy had a density of only 6.3 g cm$^{-3}$. According to the Couzinie et al. dataset[38], there is no HEA which is both stronger at 25 °C and less dense than the $Al_{0.239}Mo_{0.123}Ta_{0.095}Ti_{0.342}Zr_{0.201}$ RHEA. Furthermore, in Supplementary Fig. 13 we have compared the yield strengths of the AlMoTaTiZr and AlMoTaTiHf to the yield strength of all the reported equiatomic quinary HEAs in the literature. The 25 °C optimal equiatomic 5-element alloy, AlMoTaTiZr, is predicted to have the highest room-temperature yield strength (1992 MPa), while the 1000 °C optimal alloy has a yield strength nearly as high (805 MPa) as the highest reported alloy, HfMoTaTiZr, which had a yield strength of 855 MPa.

## Summary and future prospects

In this paper, we have demonstrated the concept of an intelligent computational framework based on machine learning and optimization to predict RHEA yield strength and discover RHEA compositions with theoretical improvement over the starting RHEA. The protocol developed here can be used by material scientists for quick screening of compositional space and identifying potential candidates with improved yield strength that merit processing and characterization. First, we have shown that repeated *k*-fold cross-validation coupled with feature selection is an effective approach to obtain a more statistically meaningful prediction of all data points in contrast to traditional used ML validation techniques. Using the robust ML-based yield strength prediction model with a clear understanding of the statistical errors, we have coupled this with a genetic algorithm to discover RHEA compositions with improved yield strengths. Given a baseline starting RHEA, the algorithm intelligently searches through the complex composition space to maximize yield strength. The concept was demonstrated for three different base alloys discussed in the RHEA literature, and optimal alloy compositions with improved yield strengths, as high as 80% were predicted for both 25 °C and 1000 °C. The alloys optimized for yield strength at 25 °C and at 1000 °C exhibited notable differences in composition and descriptors, underscoring that the mechanisms and criteria for maximizing strength at low temperature and high temperature can be quite different, and compositions maximizing yield strength at room temperature may not improve that for high temperatures. Finally, in a generalized approach, we predicted the low-temperature and high-temperature yield strength of 252 equiatomic RHEA chemistries. The top candidate for each temperature was further improved by tailoring the elemental composition using our generalized framework. Our ongoing work is extending this technique to predict other mechanical properties, such as hardness, ductility/plasticity, creep strength, and fatigue. The simultaneous optimization of multiple properties can also be incorporated into the framework. The ability to perform multi-property optimization will enable discovering HEAs with, for example, high ductility at room temperature and high strength at higher temperatures. Identifying HEAs which meet requirements for multiple properties experimentally is a grand challenge, and the work presented here creates a foundation for addressing this challenge.

## METHODS

### Descriptor calculations

Some noteworthy descriptors that bear discussion are ones that have been traditionally used for interpreting phases, such as the atomic size mismatch ($\delta$), the enthalpy of mixing ($\Delta H_{\text{mix}}$), the entropy of mixing ($\Delta S_{\text{mix}}$), $\Omega$, and $\Phi$. The equations defining these five descriptors are given as[17,18],

$$\delta = \sqrt{\sum_{i=1}^{n} x_i (1 - r_i/\bar{r})^2} \tag{1}$$

$$\Delta H_{\text{mix}} = \sum_{i=1, i\neq j}^{n} \Omega_{ij} x_i x_j \tag{2}$$

$$\Delta S_{\text{mix}} = -R \sum_{i=1, i\neq j}^{n} x_i \ln x_j \tag{3}$$

$$\Omega = \frac{T_m \Delta S_{\text{mix}}}{|\Delta H_{\text{mix}}|} \tag{4}$$

$$\Phi = \left| \frac{\Delta H_{\text{mix}} - T_m \Delta S_{\text{mix}}}{\Delta H_{\max} - T_{\max} \Delta S_{\max}} \right| \tag{5}$$

In Eqs. 1–5, $i$ and $j$ represent the $i^{\text{th}}$ and $j^{\text{th}}$ element, $r_i$ is the atomic radius of the $i^{\text{th}}$ element, $\bar{r}$ is the average atomic radius of the alloy, $\Omega_{ij} = 4\Delta H_{AB}^{\text{mix}}$ (where $\Delta H_{AB}^{\text{mix}}$ is the enthalpy of mixing of

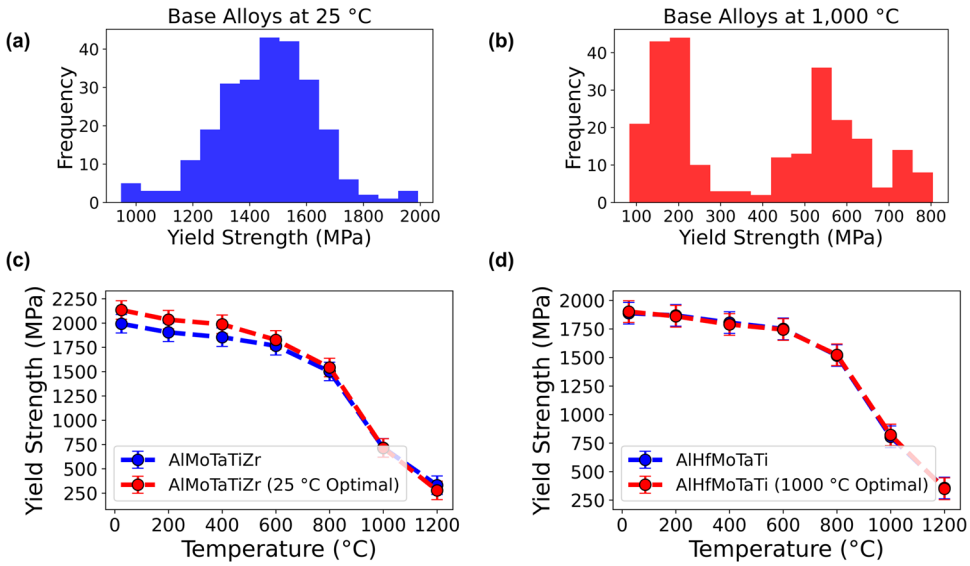


**Fig. 7 Yield strength predictions of 252 five-element equiatomic base alloys. a** Distribution of yield strength at 25 °C for the five-element equiatomic alloys. **b** Distribution of yields strength at 1000 °C. **c** AlMoTaTiZr and $Al_{0.239}Mo_{0.123}Ta_{0.095}Ti_{0.342}Zr_{0.201}$ (optimized for 25 °C). Yield strength improved by 7%. **d** AlHfMoTaTi and $Al_{0.151}Hf_{0.236}Mo_{0.137}Ta_{0.131}Ti_{0.345}$ (optimized for 1000 °C). Yield strength improved by 2%. A 95% confidence level error (±94 MPa) is shown for prediction.

binary alloys), $x$ is the element fraction, $T_m$ is melting temperature estimated from a rule-of-mixtures, $\Delta H_{\max}$ is the maximum enthalpy of mixing of a binary combination of the elements in the alloy, $T_{\max}$ is the maximum melting temperature of a single element in the alloy, and $\Delta S_{\max}$ is the maximum entropy of mixing of a binary alloy (i.e., an equiatomic binary alloy). We have investigated the accuracy of the rule-of-mixtures estimate for $T_m$ by comparing the estimates to those obtained from the liquidus line in calculated phase diagrams of 33 HEAs. The mean absolute error (MAE) in the rule-of-mixtures $T_m$ estimate is 148 K (Supplementary Fig. 14), which corresponds to an error of approximately 6%. This error has likely been absorbed within the random forest fitting of the data. While acceptability of this level of uncertainty in $T_m$ is certainly debatable, we argue that using the rule-of-mixtures is a rapid and effective method to estimate $T_m$ for yet-to-be-synthesized alloys.

Recently, lattice distortion and moduli of distortion are thought to be important parameters for determining phase formation and mechanical properties of HEAs[21,57]. Therefore, we have included them in our descriptor calculation. The lattice distortion around atom $i$ can be defined according to Senkov et al.[58] as,

$$\delta_{ai} = \frac{9}{8}\sum_j x_j \delta_{aij} \tag{6}$$

The 9 in the numerator is due to total number of atoms in the $i$-centered cluster in the BCC lattice, while the 8 in the denominator is due to the number of atoms around $i$ in the cluster (excluding $i$). The reduced atomic size difference, $\delta_{aij}$, is defined as,

$$\delta_{aij} = \frac{2(r_i - r_j)}{(r_i + r_j)} \tag{7}$$

Similarly, the modulus of distortion, $\delta_{Gi}$, is the defined as,

$$\delta_{Gi} = \frac{9}{8}\sum_j c_j \delta_{Gij} \tag{8}$$

$$\delta_{Gij} = 2(G_i - G_j)/(G_i + G_j) \tag{9}$$

where $G$ is the shear modulus. The atomic radii and moduli were collected from published values (AZO Materials (https://www.azom.com/))[17].

In addition to solid solution strengthening, grain boundary strengthening is another mechanism through which the mechanical properties of an alloy can be affected. The primary empirical equation governing the contribution of grain boundary strengthening to the observed mechanical property is the Hall-Petch relationship[59,60]. The Hall-Petch relationship is frequently expressed as,

$$\sigma_{HP} = \sigma_0 + Kd^{-\frac{1}{2}} \tag{10}$$

Where $\sigma_0$ is the base strength of the material, $K$ is the locking parameter, and $d$ is the grain size of the material. In alloys, the base strength is typically derived from the the individual element yield strengths using the rule-of-mixtures. The locking parameters of all elements of interest in this study have been tabulated in the literature[24]. Here, a rule-of-mixtures has been used to estimate the locking parameter of each alloy based on their elemental composition. The grain size, $d$, is closely related to the specific processing conditions affecting grain growth kinetics[61,62], and is typically not reported in the HEA literature. However, some HEA literature have performed detailed studies on the effect of ubiquitously reported processing conditions (e.g., annealing temperature, annealing time, etc.) on the grain size. Therefore, the specific processing conditions which we have collected from the original literature cited by Couzinie et al.[38] serve as a surrogate for a more detailed microstructural knowledge.

Finally, we have utilized the temperature-dependent yield strength model based on edge dislocation theory formulated by Maresca and Curtin[39] to augment the experimentally available HEA dataset at temperatures between 25 °C and 600 °C. The accuracy of ML approaches is always fundamentally limited by the availability of adequate data, and RHEA yield strengths at temperatures between 25 °C and 600 °C are typically not reported. Using the experimental yield strengths at 25 °C and 600 °C and the temperature-dependent yield strength model proposed by Maresca and Curtin[39], the yield strength for 200 °C and 400 °C were computed. More details and an example calculation of the yield strength at 200 °C and 400 °C based on experimentally known values at 25 °C and 600 °C are provided in the SI.

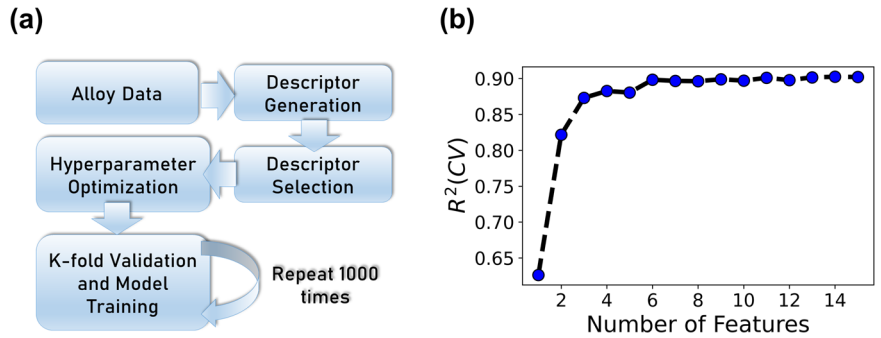


**Fig. 8 Model training and feature selection process. a** ML model development procedure followed in this work; **b** Changes in cross-validation coefficient as a function of number of descriptors. It shows there is no significant improvement beyond six (6) descriptors.

### Descriptor selection, model training and validation

Descriptor selection constitutes an important part of our work. Figure 8a outlines the procedure for training and validation of the ML model. We used the sequential forward selection (SFS) method[43,44] for selecting the best set of descriptors that describes the data within a given regression model. We have investigated four different regression models, namely: random forest[42], gradient boosting[47], LASSO[45], and ridge regression[46]. The procedure starts with first performing SFS coupled with a *k*-fold cross-validation technique where *k* is taken as five (5). In 5-fold cross-validation, the data were randomly divided into five sets, with four sets are used for training and the rest is for testing. The 5-fold cross-validation was repeated until all groups are used for validation. During this process, a set of best descriptors were identified with a specified number of maximum descriptors criteria. We have systematically varied the number of descriptors in the computed 5-fold validation to identify a critical number of descriptors beyond which there was negligible improvement of the cross-validation regression coefficient *R* (CV). The cross-validated $R^2$, $R^2$ (CV), remains constant beyond a six-descriptor model (Fig. 8b) which indicates that there is insignificant improvement in the predicted yield strength if more than six descriptors are chosen. We then performed optimization of the hyperparameters of the regression models using the six descriptors to identify parameters that maximize $R^2$ (CV). The following hyperparameters were used to train the random forest model: 100 estimators, a minimum of two samples to split an internal node, and a minimum of one sample required to be a leaf node. Once we have identified the descriptor set and optimized the hyperparameters, we performed a repeated *k*-fold validation. In the procedure, we performed the repeated 5-fold cross-validation step 1000 times. Each time, the members of each fold were chosen randomly which eliminated the biasness of a single 5-fold analysis. The repeated *k*-fold method provides a statistical variation of the predicted yield strength of *each* RHEA allowing us to compute the 95% confidence level of *each* yield strength data point. The *scikit-learn* library[63] was used to perform the model training and validation steps described here. We believe that our procedure provides a superior assessment of model's predictive ability compared to the commonly used single *k*-fold validation, or train-test splitting. At the end, we also performed 25 runs of SFS with six descriptors in order to examine variability of the selected features. The resulting feature selection probability distribution is provided in Supplementary Fig. 15. Replicates of the SFS process revealed that multiple composition-based descriptors had an approximately equal likelihood of being selected. Since multiple feature sets yielded nearly the same predictions, a representative feature set was chosen for the forward model and discovery process.

### Yield strength optimization

For solving the inverse problem, we used the differential evolution optimizer[37], a genetic algorithm, implemented within *scipy*[64] to design an atomic composition that maximizes yield strength at a specified temperature for a particular base alloy. Differential evolution is a stochastic population-based method that is frequently used for global optimization problems. At each pass through the population, the algorithm mutates each candidate solution by mixing with other candidate solutions to create a trial candidate. A central feature of the optimization problem is the definition of the *objective function*. For a minimization problem, the goal is to identify a solution that causes the objective function to be equal to zero, or minimizes the objective function to be as close to zero as possible. In this case, the goal is to maximize the yield strength. Therefore, the objective function should grow smaller as the yield strength grows larger. In the current investigation, we have found that defining the objective function as the reciprocal of the yield strength allows for alloys with increased yield strength to be found easily. The mutation constant was set between 0.5 and 1, with dithering employed. The recombination constant (i.e., cross-over probability) was 0.7. A linear constraint function was coupled with the optimization using Lampinen's approach[65] to ensure that the element fractions summed to 1. All individual element fractions that were present in a given alloy were bounded between 0.02 and 0.35, typical values that are representative of an HEA.

## DATA AVAILABILITY

A CSV file complete with all pertinent information (input features and yield strength data) has been made digitally available.

## CODE AVAILABILITY

Readers are requested to contact the authors.

Received: 17 May 2022; Accepted: 1 November 2022;
Published online: 12 November 2022

## REFERENCES

1. 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).
2. Chen, T. K., Shun, T. T., Yeh, J. W. & Wong, M. S. Nanostructured nitride films of multi-element high-entropy alloys by reactive DC sputtering. *Surf. Coat. Technol.* **188–189**, 193–200 (2004).
3. Hsu, C.-Y., Yeh, J.-W., Chen, S.-K. & Shun, T.-T. Wear resistance and high-temperature compression strength of Fcc $CuCoNiCrAl_{0.5}Fe$ alloy with boron addition. *Metall. Mater. Trans. A* **35**, 1465–1469 (2004).
4. Huang, P.-K., Yeh, J.-W., Shun, T.-T. & Chen, S.-K. Multi-principal-element alloys with improved oxidation and wear resistance for thermal spray coating. *Adv. Eng. Mater.* **6**, 74–78 (2004).

5. Yeh, J.-W. et al. Formation of simple crystal structures in Cu-Co-Ni-Cr-Al-Fe-Ti-V alloys with multiprincipal metallic elements. *Metall. Mater. Trans. A* **35**, 2533–2536 (2004).
6. Cantor, B., Chang, I. T. H., Knight, P. & Vincent, A. J. B. Microstructural development in equiatomic multicomponent alloys. *Mater. Sci. Eng. A* **375–377**, 213–218 (2004).
7. Ranganathan, S. Alloyed pleasures: multimetallic cocktails. *Curr. Sci.* **85**, 1404–1406 (2003).
8. Yeh, J.-W. In *High-Entropy Alloys* (eds. Gao, M. C., Yeh, J.-W., Liaw, P. K. & Zhang, Y.) 1–19 (Springer, 2016).
9. Senkov, O. N., Senkova, S. V., Miracle, D. B. & Woodward, C. Mechanical properties of low-density, refractory multi-principal element alloys of the Cr-Nb-Ti-V-Zr system. *Mater. Sci. Eng. A* **565**, 51–62 (2013).
10. Kang, B., Lee, J., Ryu, H. J. & Hong, S. H. Ultra-high strength WNbMoTaV high-entropy alloys with fine grain structure fabricated by powder metallurgical process. *Mater. Sci. Eng. A* **712**, 616–624 (2018).
11. Liu, Y. et al. Microstructure and mechanical properties of refractory $HfMo_{0.5}NbTiV_{0.5}$Sixhigh-entropy composites. *J. Alloy. Compd.* **694**, 869–876 (2017).
12. Maiti, S. & Steurer, W. Structural-disorder and its effect on mechanical properties in single-phase TaNbHfZr high-entropy alloy. *Acta Mater.* **106**, 87–97 (2016).
13. Senkov, O. N., Isheim, D., Seidman, D. N. & Pilchak, A. L. Development of a refractory high entropy superalloy. *Entropy* https://doi.org/10.3390/e18030102 (2016).
14. Zhang, M., Zhou, X. & Li, J. Microstructure and mechanical properties of a refractory CoCrMoNbTi high-entropy alloy. *J. Mater. Eng. Perform.* **26**, 3657–3665 (2017).
15. Zhang, Y., Yang, X. & Liaw, P. K. Alloy design and properties optimization of high-entropy alloys. *JOM* **64**, 830–838 (2012).
16. Miracle, D. B. High entropy alloys as a bold step forward in alloy development. *Nat. Commun.* **10**, 1–3 (2019).
17. Miracle, D. B. & Senkov, O. N. A critical review of high entropy alloys and related concepts. *Acta Mater.* **122**, 448–511 (2017).
18. Senkov, O. N., Miracle, D. B., Chaput, K. J. & Couzinie, J.-P. Development and exploration of refractory high entropy alloys—A review. *J. Mater. Res.* **33**, 3092–3128 (2018).
19. Yang, X. & Zhang, Y. Prediction of high-entropy stabilized solid-solution in multi-component alloys. *Mater. Chem. Phys.* **132**, 233–238 (2012).
20. Chen, H. et al. Contribution of lattice distortion to solid solution strengthening in a series of refractory high entropy alloys. *Metall. Mater. Trans. A Phys. Metall. Mater. Sci.* **49**, 772–781 (2018).
21. Coury, F. G., Clarke, K. D., Kiminami, C. S., Kaufman, M. J. & Clarke, A. J. High throughput discovery and design of strong multicomponent metallic solid solutions. *Sci. Rep.* **8**, 8600 (2018).
22. Yao, H. W. et al. Mechanical properties of refractory high-entropy alloys: Experiments and modeling. *J. Alloy. Compd.* **696**, 1139–1150 (2017).
23. Wang, Z., Fang, Q., Li, J., Liu, B. & Liu, Y. Effect of lattice distortion on solid solution strengthening of BCC high-entropy alloys. *J. Mater. Sci. Technol*. https://doi.org/10.1016/j.jmst.2017.07.013 (2018).
24. Cordero, Z. & Schuh, C. A. Six decades of the Hall—Petch effect—a survey of grain-size strengthening studies on pure metals. *Inter. Mater. Rev.* **61**, 495–512 (2016). *8*.
25. Asghari-Rad, P. et al. Effect of grain size on the tensile behavior of $V_{10}Cr_{15}Mn_5Fe_{35}Co_{10}Ni_{25}$ high entropy alloy. *Mater. Sci. Eng. A* **744**, 610–617 (2019).
26. Schuh, B. et al. Thermodynamic instability of a nanocrystalline, single-phase TiZrNbHfTa alloy and its impact on the mechanical properties. *Acta Mater.* **142**, 201–212 (2018).
27. Tang, Z., Zhang, S., Cai, R., Zhou, Q. & Wang, H. Designing high entropy alloys with dual fcc and bcc solid-solution phases: structures and mechanical properties. *Metall. Mater. Trans. A* **50**, 1888–1901 (2019).
28. He, J. Y. et al. A precipitation-hardened high-entropy alloy with outstanding tensile properties. *Acta Mater.* **102**, 187–196 (2016).
29. Maresca, F. et al. Edge dislocations can control yield strength in refractory body-centered-cubic high entropy alloys. *arXiv* https://doi.org/10.48550/arXiv.2008.11671 (2020).
30. Rao, Y., Baruffi, C., De Luca, A., Leinenbach, C. & Curtin, W. A. Theory-guided design of high-strength, high-melting point, ductile, low-density, single-phase BCC high entropy alloys. *Acta Mater.* **237**, 118132 (2022).
31. Lee, S. Y., Byeon, S., Kim, H. S., Jin, H. & Lee, S. Deep learning-based phase prediction of high-entropy alloys: Optimization, generation, and explanation. *Mater. Des.* **197**, 109260 (2021).
32. Huang, W., Martin, P. & Zhuang, H. L. Machine-learning phase prediction of high-entropy alloys. *Acta Mater.* **169**, 225–236 (2019).
33. Zhou, Z., Zhou, Y., He, Q., Ding, Z. & Li, F. Machine learning guided appraisal and exploration of phase design for high entropy alloys. *npj Comput. Mater.* **5**, 128 (2019).
34. Wen, C. et al. Machine learning assisted design of high entropy alloys with desired property. *Acta Mater.* **170**, 109–117 (2019).
35. Roy, D., Mandal, S. C. & Pathak, B. Machine learning assisted exploration of high entropy alloy-based catalysts for selective $CO_2$ reduction to methanol. *J. Phys. Chem. Lett.* **13**, 5991–6002 (2022).
36. Rao, Z. et al. Machine learning-enabled high-entropy alloy discovery. *arXiv* **2202**, 13753 (2022).
37. Storn, R. & Price, K. Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces. *J. Glob. Optim.* **11**, 341–359 (1997).
38. Couzinié, J. P., Senkov, O. N., Miracle, D. B. & Dirras, G. Comprehensive data compilation on the mechanical properties of refractory high-entropy alloys. *Data Br.* **21**, 1622–1641 (2018).
39. Maresca, F. & Curtin, W. A. Mechanistic origin of high strength in refractory BCC high entropy alloys up to 1900K. *Acta Mater.* **182**, 235–249 (2020).
40. Ward, L. et al. Matminer: an open source toolkit for materials data mining. *Comput. Mater. Sci.* **152**, 60–69 (2018).
41. Lee, C. et al. Strength can be controlled by edge dislocations in refractory high-entropy alloys. *Nat. Commun.* **12**, 6–13 (2021).
42. Breiman, L. Random forests. *Mach. Learn* **45**, 5–32 (2001).
43. Pudil, P., Novovičová, J. & Kittler, J. Floating search methods in feature selection. *Pattern Recognit. Lett.* **15**, 1119–1125 (1994).
44. Ferri, F. J., Pudil, P., Hatef, M. & Kittler, J. Comparative study of techniques for large-scale feature selection. *Pattern Recognit. Pract. IV* **16**, 403–413 (1994).
45. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B* **58**, 267–288 (1996).
46. Hoerl, A. E. & Kennard, R. W. Ridge regression: applications to nonorthogonal problems. *Technometrics* **12**, 69–82 (1970).
47. Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Ann. Stat.* **29**, 1189–1232 (2001).
48. Lundberg, S. M. & Lee, S. I. A unified approach to interpreting model predictions. *Neural Inf. Process. Syst.* https://doi.org/10.48550/arXiv.1705.07874 (2017).
49. Lundberg, S. M. et al. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2**, 56–67 (2020).
50. Maresca, F. & Curtin, W. A. Theory of screw dislocation strengthening in random BCC alloys from dilute to "High-Entropy" alloys. *Acta Mater.* **182**, 144–162 (2020).
51. Statham, C. D., Koss, D. A. & Christian, J. W. The thermally activated deformation of niobium-molybdenum and niobium-rhenium alloy single crystals. *Philos. Mag.* **26**, 1089–1103 (1972).
52. Senkov, O. N., Jensen, J. K., Pilchak, A. L., Miracle, D. B. & Fraser, H. L. Compositional variation effects on the microstructure and properties of a refractory high-entropy superalloy $AlMo_{0.5}NbTa_{0.5}TiZr$. *Mater. Des.* **139**, 498–511 (2018).
53. Wu, Y. D. et al. Phase composition and solid solution strengthening effect in TiZrNbMoV high-entropy alloys. *Mater. Des.* **83**, 651–660 (2015).
54. Senkov, O. N. et al. Microstructure and elevated temperature properties of a refractory TaNbHfZrTi alloy. *J. Mater. Sci.* **47**, 4062–4074 (2012).
55. Lu, Y., Zhang, Y.-H., Ma, E. & Han, W.-Z. Relative mobility of screw versus edge dislocations controls the ductile-to-brittle transition in metals. *Proc. Natl Acad. Sci. USA* **118**, e2110596118 (2021).
56. Senkov, O. N. & Miracle, D. B. Generalization of intrinsic ductile—to—brittle criteria by Pugh and Pettifor for materials with a cubic crystal structure. *Sci. Rep.* https://doi.org/10.1038/s41598-021-83953-z (2021).
57. Song, H. et al. Local lattice distortion in high-entropy alloys. *Phys. Rev. Mater.* **1**, 023404 (2017).
58. Senkov, O. N., Scott, J. M., Senkova, S. V., Miracle, D. B. & Woodward, C. F. Microstructure and room temperature properties of a high-entropy TaNbHfZrTi alloy. *J. Alloy. Compd.* **509**, 6043–6048 (2011).
59. Hall, E. O. The deformation and ageing of mild steel: III discussion of results. *Proc. Phys. Soc. Sect. B* **64**, 747 (1951).
60. Petch, N. J. The cleavage strength of polycrystals. *J. Iron Steel Inst.* **174**, 25–28 (1953).
61. Fazakas, E. et al. Experimental and theoretical study of $Ti_{20}Zr_{20}Hf_{20}Nb_{20}X_{20}$ (X = v or Cr) refractory high-entropy alloys. *Int. J. Refract. Met. Hard Mater.* **47**, 131–138 (2014).
62. Juan, C. C. et al. Simultaneously increasing the strength and ductility of a refractory high-entropy alloy via grain refining. *Mater. Lett.* **184**, 200–203 (2016).
63. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
64. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).

65. Lampinen, J. A constraint handling approach for the differential evolution algorithm. *Proc. 2002 Congr. Evol. Comput. CEC 2002* **2**, 1468–1473 (2002).
66. Yao, H. W. et al. NbTaV-(Ti,W) refractory high-entropy alloys: experiments and modeling. *Mater. Sci. Eng. A* **674**, 203–211 (2016).

## ACKNOWLEDGEMENTS

We would like to thank Prof. Peter Liaw for helpful discussions. We would also like to thank the reviewers for their valuable comments which helped revising the manuscript. This work was funded by the Office of Naval Research of the United States under the Small Business Technology Transfer program (contract #N68335-20-C-0402).

## AUTHOR CONTRIBUTIONS

S.A.G. performed literature search, developed code, performed computations, and wrote the manuscript; D.S performed literature search, developed code, performed computations, wrote the manuscript, and guided the research; S.R.B. and K.R. contributed significantly in the discussion, and reviewed the manuscript critically.

## 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/s41524-022-00926-0.

**Correspondence** and requests for materials should be addressed to Stephen A. Giles or Debasis Sengupta.