A systematic analysis of phase stability in refractory high entropy alloys utilizing linear and non-linear cluster expansion models

Chiraag Nataraj[a,*], Edgar Josué Landinez Borda[b], Axel van de Walle[a], Amit Samanta[b]

[a] *School of Engineering, Brown University, Providence, Rhode Island, 02912 USA*

[b] *Physics Division, Lawrence Livermore National Lab, Livermore, California, 94550 USA*

Abstract

The phase segregation behavior of three key refractory high entropy alloys (NbTiVZr, HfNbTaTiZr, and AlHfNbTaTiZr) is studied using first-principles calculations. Several linear and non-linear methods are utilized to generate surrogate models for three key refractory high entropy alloys (NbTiVZr, HfNbTaTiZr, and AlHfNbTaTiZr) via the cluster expansion formalism. The characteristics of each of the generated models is explored and the regression methods are compared. Finally, these surrogate models are utilized to generate Monte Carlo trajectories in order to explore the link between phase segregation and previously documented mechanical degradation in these materials.

Phase segregation and intermetallic phases documented in the experimental literature are reproduced in all three high entropy alloys. NbTiVZr forms vanadium and zirconium clusters at lower temperatures (250 K) which disperse into the single-phase matrix by 1000 K. HfNbTaTiZr forms HfZr, NbTa, and possibly TiZr intermetallic phases at lower temperatures (250 K). Unlike the other HEAs studied here, HfNbTaTiZr does not lose short-range ordering in the solid state until around 3500 K, which is above its melting temperature. AlHfNbTaTiZr forms NbTa and AlHfTiZr phases at lower temperatures (250 K), which are not observed at higher temperatures (1000 K).

Keywords: High-entropy alloys, Cluster expansion, Phase diagram, Phase stability, Multicomponent

*Corresponding author

Email addresses: chiraag_nataraj@brown.edu (Chiraag Nataraj), avdw@brown.edu (Axel van de Walle), samanta1@llnl.gov (Amit Samanta)

1. Introduction

Designing high entropy alloys (HEAs) and understanding their structure-property relationships are active topics of research [1, 15]. They are characterized by high numbers of elements, all at roughly equal atomic percentages. In particular, refractory high entropy alloys (such as those explored in this work — NbTiVZr, HfNbTaTiZr, and AlHfNbTaTiZr) are especially promising for a wide variety of reasons. Compared to some of the original HEAs, such as TaNbWMo and TaNbWMoV, these alloys are far less dense and more ductile while maintaining many of the superior high-temperature properties which make HEAs attractive candidates for use in the aerospace industry [23, 22, 24, 25].

Two important characteristics of any material are its mechanical properties, such as the stress-strain response and ductility, and its phase diagram. Understanding how the existence and stability of phases in a given material affect its mechanical properties is vital for engineering stronger, tougher materials. Analysis of phase stability in HEAs assumes particular importance because many physical properties of HEAs are affected by the presence of precipitates [13]. In addition, it is interesting to explore the length scales associated with short-range order (if any) present in the HEAs that do not form precipitates, as these length scales can impact the work strengthening or softening behavior [27]. Recent experiments suggest that the formation, the structure, and the composition of the precipitates are sensitive to the annealing temperature [33] and annealing time [32]. Thus, systematic analysis of the stable and metastable phases in the HEA systems using experiments is expensive and time consuming because the convergence towards thermodynamic equilibrium can be hindered by slow kinetics. Therefore, in this work we use atomistic simulations to analyze the stability of the solid solution phase in three important refractory HEAs.

In order to investigate how local chemistry impacts phase stability and mechanical properties from a computational standpoint, first principles (ab-initio density functional theory) calculations are most suitable. Since DFT scales quite poorly with respect to the system size, many scientists turn to training surrogate models in order to calculate thermodynamic quantities and properties of interest far more efficiently than is possible with brute-force DFT. The surrogate model of choice for crystalline alloys is the cluster expansion approach [4, 21], which expresses the energy of a structure in terms of pairwise, triplet, and higher-order interactions. The cluster expansion can then be used together with Monte Carlo (MC) methods to explore the impact of phase stability at a far lower computational cost than DFT+MC (calculating the energy of the candidate structure using DFT rather than a surrogate model) while simultaneously enabling the use of larger supercells to avoid finite-size artifacts in the predicted equilibrium structures. Therefore, Monte Carlo simulations using cluster expansion models are uniquely poised to tackle problems related to phase stability and analysis of compositions of (meta)stable phases because the different stable and metastable phases can be explored at very low computational cost. In addition, since these systems are being actively explored by experimentalists, this offers numerous

opportunities for mutual validation.

There are two fundamental questions that must be answered when constructing a cluster expansion. The first is the nature of the interactions considered, as modeled by the geometry of the clusters included and parametrized by the spatial extent of the cluster and the number of included atomic sites. The second is how much each cluster contributes to the total energy, as indicated by the so-called effective cluster interaction (ECI) coefficient of each cluster. This poses a challenging optimization problem, as each additional type of cluster adds computational time — both to the construction of the cluster expansion and to its use, e.g. in MC simulations. Thus, adding additional clusters when they do not meaningfully improve the accuracy of the cluster expansion should be avoided.

The chemical complexity of most HEAs presents significant challenges to the fitting of an accurate cluster expansion model, because the number of possible clusters (and thus the minimum number of structures required for a good fit) increases greatly as the number of elements is increased. Additionally, if the atomic sizes of the constituent elements vary drastically, significant distortion of the lattice complicates the cluster expansion model generation. In light of the unique nature of these challenges, it is instructive to consider several methods for fitting cluster expansion models for refractory high entropy alloys. In this work, various linear and non-linear algorithms are considered, including automated methods, a higher-order cluster augmentation method built on the automated method results, and alternative linear and non-linear solution methods. Further, the question of whether this type of surrogate model is effective in predicting stable or metastable phases is investigated. We also demonstrate that the resulting surrogate models can be used to probe the relationship between changes in phase stability and the degradation of mechanical properties.

In the next section, the methods used in this work are laid out, starting with the cluster expansion formalism, proceeding to the various linear and non-linear methods used, and ending with a description of the Monte Carlo method and short-range order parameters utilized in this work. The following section presents the results of each fitting method when applied to the various high entropy alloys mentioned earlier before exploring the nature of phase segregation in each of those alloys. Finally, some concluding thoughts on what this means for the use of these alloys in practice.

2. Methods

An overview of the cluster expansion formalism and a summary of the various linear and non-linear regression methods is given in the following subsections. Additionally, an overview is given of various error measures that are used to measure the quality of a given regression. Finally, a brief discussion of the Monte Carlo method utilized here and the definition of the short-range order parameters considered here conclude this section.

2.1. Cluster Expansion Formalism

The cluster expansion is a model which expands the energy of a crystal structure as a series of interactions between different clusters of atoms. Consider a crystal structure with n sites where each site i can be filled by one of M_i species. The configuration vector $\boldsymbol{\sigma}$ of this structure is a vector of length n with occupation variables σ_i indicating which type of atom sits at that lattice site and σ_i can range from 0 to $M_i - 1$. Even though the state of the system is represented by an ideal lattice, the energies associated with those states will account for the elastic contributions associated with relaxations away from ideal lattice sites. The generalized, multicomponent, multisublattice cluster expansion formulation is given in Eq. (1) [30].

$$E(\sigma) = \sum_{\alpha} m_{\alpha} J_{\alpha} \left\langle \prod_{i} \gamma_{\alpha_i', M_i}(\sigma_i) \right\rangle \tag{1}$$

- α is a cluster that is described by a vector of cluster variables α_i which each can take values from 0 to $M_i - 1$, reflecting either omission from the cluster (0) or various functional dependencies between the energy of a certain cluster and the occupation variable σ_i.
- The sum is taken over all symmetrically distinct clusters α, while the average of the occupation variables is taken over all clusters α' which are symmetrically equivalent to α.
- m_{α} is the multiplicity of a particular cluster and the term between $\langle\rangle$ is defined as the *correlation* of the structure with cluster α.
- J_{α} is the effective cluster interaction (ECI) coefficient.
- $\gamma_{\alpha_i', M_i}(\sigma_i)$ satisfies the properties defined in Eqs. (2) and (3), namely that γ for the null cluster and any combination of occupation variables is 1 and the γ functions are orthogonal.

$$\gamma_{0, M_i}(\sigma_i) = 1 \tag{2}$$

$$\frac{1}{M_i} \sum_{\sigma_i = 0}^{M_i - 1} \gamma_{\alpha_i, M_i}(\sigma_i)\, \gamma_{\beta_i, M_i}(\sigma_i) = \begin{cases} 1 & \text{if } \alpha_i = \beta_i \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

The functions γ_{α_i', M_i} can be any functions which satisfy the given properties. However, in this work, all calculations utilize the correlations output from ATAT, which uses the functions in Eq. (4).

$$\gamma_{\alpha_i, M_i}(\sigma_i) = \begin{cases} 1 & \text{if } \alpha_i = 0 \\ -\cos\left(2\pi \left\lceil \frac{\alpha_i}{2} \right\rceil \frac{\sigma_i}{M_i}\right) & \text{if } \alpha_i > 0 \text{ \& odd} \\ -\sin\left(2\pi \left\lceil \frac{\alpha_i}{2} \right\rceil \frac{\sigma_i}{M_i}\right) & \text{if } \alpha_i > 0 \text{ \& even} \end{cases} \tag{4}$$

2.2. Structure Generation and Cluster Expansion Fitting

Structures are generated using ATAT's `mmaps` code, which utilizes a variance reduction scheme to determine which structures to generate (possibly subject to a restricted concentration range for each species) [29, 30]. In the case of the high-entropy alloys studied here, no concentration limits are placed on any of the constituent elements. These structures are then relaxed with the Vienna Ab initio Simulation Package (VASP) [10, 11, 8, 9] and their energies are calculated. The cluster expansion model is fitted to formation energies computed relative to the bulk energies of the constituent elements. Since the cluster expansion formulation parametrizes the configuration dependence of the energy for a given lattice, structures which are excessively distorted must be excluded, as they would more appropriately belong to the cluster expansion for a different lattice. The change in the atomic structure following relaxation can be quantified by calculating the amount of distortion. If $\{\boldsymbol{t}_1, \boldsymbol{t}_2, \boldsymbol{t}_3\}$ are the lattice vectors of the original cell and $\{\boldsymbol{t}'_1, \boldsymbol{t}'_2, \boldsymbol{t}'_3\}$ are the lattice vectors of the relaxed cell, a measure of the amount of distortion is given by Eq. (6), where $\boldsymbol{T}$ contains the lattice vectors of the original cell as row vectors, $\boldsymbol{T}'$ contains the lattice vectors of the relaxed cell as row vectors, $|\boldsymbol{T}'|$ is the determinant of $\boldsymbol{T}'$, and sym calculates the symmetric part of the tensor: $\operatorname{sym} \boldsymbol{T} = \frac{1}{2}\left(\boldsymbol{T} + \boldsymbol{T}^\top\right)$.

$$\boldsymbol{\delta T} = \operatorname{sym}\left[\left(\frac{\boldsymbol{T}}{\sqrt[3]{|\boldsymbol{T}|}}\right)^{-1} \frac{\boldsymbol{T}'}{\sqrt[3]{|\boldsymbol{T}'|}}\right] - \boldsymbol{I} \tag{5}$$

$$\epsilon = \sqrt{\sum_{i,j} \delta T_{ij} \delta T_{ij}} \tag{6}$$

If $\boldsymbol{T}' = \boldsymbol{T}$, then $\boldsymbol{\delta T} = \boldsymbol{0}$ and $\epsilon = 0$ as expected. The exact threshold for what constitutes "too much" relaxation tends to be system-dependent and the exact values used for the systems at hand are listed in Section 3.

A wide variety of methods is used to obtain cluster expansion models for each given system. This section will simply provide an overview of these methods; more details can be found in the appendix (Sections A and B).

Three linear regression methods are explored in this work: ATAT's automated method, a higher-order cluster augmentation method, and a method utilizing the least angle regression. The **automated algorithm** used by ATAT's `mmaps` program first constructs a minimal (non-colinear) basis from the matrix of correlations before generating a least-squares fit. Additionally, it utilizes a physics-based algorithm to iterate hierarchically through the cluster choices by only including a cluster if all of its subclusters have been included and all clusters with the same number of points with a smaller diameter have been included [28], where the diameter of a cluster is the largest distance between two points in the cluster. The **higher-order cluster augmentation** method is built on

top of ATAT's automated method in order to test if the quality of the fit can be improved by deviating from `mmaps`' hierarchical rules. Starting from the automated fit, the next valid set of clusters is included and a two-step selection process is used to pick the most important clusters. Both of the previous methods use the Leave-One-Out Cross-Validation (LOOCV) score as the measure of error. The **least angle regression** method, proposed by Efron et al. in a seminal paper in 2004 [5], is an iterative and parsimonious alternative to the least squares-based fits described in the previous two methods. Unlike the previous methods, the coefficient of determination (the R^2 score) is used as the measure of error by setting aside a validation data set.

In addition to the aforementioned linear regression methods, several linear and non-linear regression methods are combined with Principal Component Analysis to explore non-linear features. Principal Component Analysis (PCA) is a linear transformation used to select a more compact set of features and decrease collinearity. The three regression methods examined here are ridge regression, kernel ridge regression, and Gaussian process regression. **Ridge regression** [6, 7, 18], also known as Tikhonov regularization, introduces a regularization matrix which often takes the form of a multiple of the identity matrix. The coefficient, the regularization parameter, functions as a constraint on the magnitude of the coefficients. The two other methods, **kernel ridge regression** and **Gaussian process regression** [20, 17], are kernel-based regression methods. The **kernel ridge regression** uses a Gaussian kernel and the hyperparameters are selected by optimizing the L_2 cost function using the Nelder-Mead method [16], while the coefficients are optimized by calculating the R^2 score using the three-way hold-out method (three data points are set aside as the validation data set instead of only one as in the LOOCV). By contrast, the **Gaussian process regression** assumes that the coefficients are normally distributed conditional on the data. Like in the previous case, a Gaussian kernel function is used and hyperparameters and coefficients are optimized as before.

2.3. Monte Carlo Simulations and Multicomponent Short-Range Order Parameters

Once the effective cluster interactions have been determined from one of the above methods, one can determine the thermodynamic equilibrium through canonical ensemble Monte Carlo simulations. The Metropolis Monte Carlo method, as implemented by the `memc2` utility from ATAT [30], is utilized here for generating trajectories of equilibrated structures. To this end, first, 5–6 starting structures are generated by randomly distributing elements throughout the structure. A Monte Carlo trajectory is launched from each of these structures and after equilibration, a set of 20 structures is saved. For the analysis of clustering and phase stability, 20 independent MC trajectories are launched and another set of 200–250 is saved per trajectory, leading to a total of 4000–5000 structures per starting structure per temperature.

In order to examine the phase stability of the high entropy alloys as a function of temperature, the same cluster expansion models are used to run Grand

Canonical Monte Carlo simulations. Due to the complexity of the alloys, the full chemical potential space is not examined in this work. Instead, the chemical potential of each species is varied individually while the other chemical potentials are kept fixed. Each chemical potential range is divided into 21 points (including both endpoints), which provides enough resolution to discern phase transitions.

Short-range order parameters are a class of descriptors which enables the quantification of the amount of order or disorder (in elemental distribution) in a crystalline material. The multicomponent generalization of the Warren-Cowley short-range order parameters [3], given in Eqs. (7) and (8), is utilized in this work. In these equations, P_{ij} is the probability of finding an atom of species i within a specified cut-off radius of species j, $\bar{c}_i$ is the average concentration of species i, and the second equation is utilized only if i is different from j. When the short-range order parameter, α_{ij}, is 0, $P_{ij} = \bar{c}_j$, meaning that the probability of finding an atom of species j is equal to the average concentration of species j in the system. In the case of clustering, the probability of finding species j in the neighborhood of species i is more than the overall concentration of species j: $P_{ij} > \bar{c}_j \implies \alpha_{ij} < 0$. On the other hand, when two species repel each other, $P_{ij} = 0 \Rightarrow \alpha_{ij} = 1$.

$$\alpha_{ii}(p) = \frac{P_{ii}(p) - \bar{c}_i}{1 - \bar{c}_i} \tag{7}$$

$$\alpha_{ij}(p) = 1 - \frac{P_{ij}(p)}{\bar{c}_j} \tag{8}$$

3. Fitting Cluster Expansion Models

3.1. Structure Generation and Selection

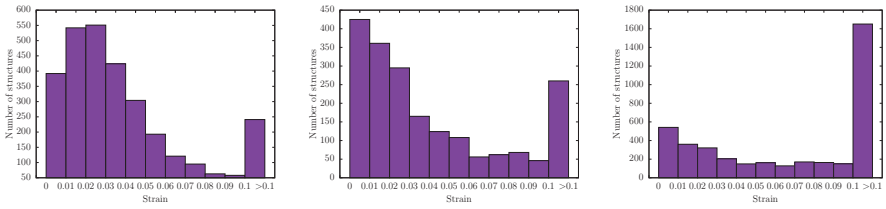
Table 1 details the number of structures generated for each alloy, while Fig. 1 shows the distribution of strain across the whole dataset. The atomic size mismatch introduces local atomic distortion in the HEAs. Since the cluster expansion formalism (Section 2.1) assumes a fixed lattice, highly distorted structures must be excluded from the set of structures used to fit the cluster expansion. Notably, the atomic size mismatch is more pronounced in the case of AlHfNbTaTiZr. Thus, a large number of structures have strains far greater than 0.1 and are thus excluded from all fitting procedures. For additional tables and graphs showing the strain distribution in each batch of structures as well as the distribution of the number of atoms in the structures, please see Section S.2.

3.2. Linear Fitting Methods

As mentioned earlier, all of the regressions here and in Section 3.3 fit per-atom formation energies calculated relative to the bulk energies of the constituent elements. Figure 2 shows the best fits obtained using the `mmaps` program, part of

HEA	Number of generated structures
NbTiVZr	2984
HfNbTaTiZr	1970
AlHfNbTaTiZr	4000

Table 1: The number of structures generated for each alloy. All structures are relaxed with the VASP parameters detailed in Section S.1 and then divided into bins based on their strain measures (see Eq. (6)).



(a) NbTiVZr (b) HfNbTaTiZr (c) AlHfNbTaTiZr

Figure 1: Distribution of strain (i.e. ϵ in Eq. (6)) over all generated structures for each HEA. The strain measure used is defined in Section 2.2. Because the cluster expansion formalism used here assumes a fixed lattice, highly distorted structures must be excluded from the set of structures used for fitting the cluster expansion. Compared to the NbTiVZr and HfNbTaTiZr systems, a far greater proportion of structures generated for AlHfNbTaTiZr have high strain and thus must be excluded from all fitting procedures.

the Alloy Theoretic Automated Toolkit (ATAT) [30]. It is noteworthy that none of the resulting cluster expansions include 4-body interactions; instead, large 2-body correlations seem to be used to describe those interactions. Additionally, the bulk of the clusters used seems to be 2-body correlations.

Starting from these cluster expansions, additional sizes of three-body clusters and, when possible, four-body clusters are added according to the higher-order cluster augmentation method described in Section 2.2. The cluster distributions of the resulting fits are shown in Fig. 3. Notably, attempting to add four-body clusters to the AlHfNbTaTiZr cluster expansion leads to an increase in the LOOCV and a worse fit. It is also interesting to note that a significant number of 4-body clusters with length 5.4 Å are included in the NbTiVZr case, while very few 4-body clusters are included at all in the HfNbTaTiZr case. In all of these materials, however, manually adding additional clusters (beyond those included by `mmaps`) improves the fit. The hierarchical rules used by `mmaps` enforce that clusters be included in groups of clusters of the same diameter and number of points. We find that allowing for some deviations from these rules can sometime results in a more predictive cluster expansion.

Moving from least-squares-based regressions to the least angle regression (LAR), expected trends in R^2 score over 3- and 4-body clusters are only obtained for NbTiVZr as LAR seems to generally require more structures for fitting than

some of the other methods. The change in R^2 score as a function of the size of validation data and the radii of clusters included is given in Table 1a in Section S.4. Looking at the table, it's clear that there is a saturation in the R^2 score as larger sizes of clusters are added, particularly 2-body clusters. The first size of 4-body clusters consistently doesn't seem to add much to the predictive power of the cluster expansion. However, the second size of 4-body clusters greatly increases the R^2 score in cases where the 2-body clusters are capped to a small size; this effect tapers off as the maximum size of 2-body clusters is increased. The distribution of included clusters at the point of R^2 saturation is given in Fig. 4a.

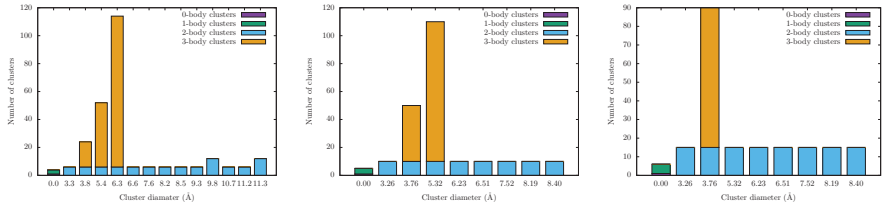
In the HfNbTaTiZr and AlHfNbTaTiZr cases, this same analysis is attempted (see Tables 1b and 1c in Section S.4). In these cases, the expected trend in R^2 score only occurs when excluding 4-body clusters altogether, which suggests that more data might be required; as the number of clusters increases, the amount of data necessary for a good fit also increases. For HfNbTaTiZr, the R^2 score saturates when 2-body clusters up to 8.5 Å and 3-body clusters up to 5.4 Å are included. In the AlHfNbTaTiZr case, the R^2 score saturates when 2-body clusters up to 9.3 Å and 3-body clusters up to 5.4 Å are included. The distribution of included clusters at the point of R^2 saturation for both materials is given in Figs. 4b and 4c.

Table 2 compares the best model obtained using each of the linear regression methods (automated fit, higher-order cluster augmentation, and least angle regression) for each high-entropy alloy. As is clear from this table, for the purposes of this comparison, 1485 structures are used to fit models for NbTiVZr, 810 structures are used to fit models for HfNbTaTiZr, and 2035 structures are used to fit models for AlHfNbTaTiZr. From these results, it is clear that the model obtained through least angle regression tends to have a higher LOOCV score than models obtained using the other fitting methods. `mmaps` systematically attempts all possible cluster choices as long as the total number of clusters is less than the number of input structures. In some cases, especially when the size of the input dataset is large, this procedure can be sped up by not attempting all available cluster choices and artificially limiting the maximum size of 3- or 4-body clusters, and the result can yield acceptable results while cutting the total computation time.

It is also interesting to investigate the general trends in the ECIs as a function of the regression method for each HEA. In the NbTiVZr case (see Figs. 2a, 3a and 4a in Section S.4), the manual cluster selection method seems to increase the spread of ECIs for the 3-body clusters, while both the automated method and LAR favor fits with smaller 3-body ECIs. As shown in Figs. 2b, 2c, 3b, 3c, 4b and 4c in Section S.4, the automated, manual, and LAR methods all seem to give similar ECI results for HfNbTaTiZr for 2- and 3-body clusters, while the LAR method selects a cluster expansion with a *far* wider range of ECIs for AlHfNbTaTiZr, as shown by the difference between the maximum and minimum ECI values for a given fit, than either the automated or manual methods.

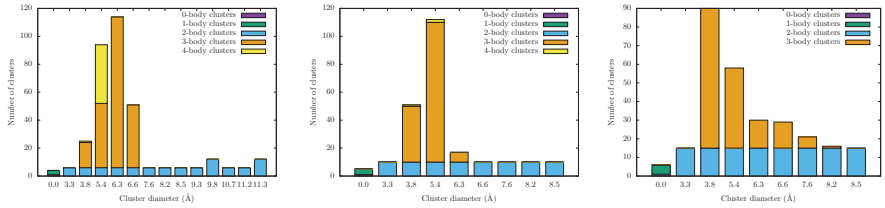
Alloy	Method	Strain cutoff	Number of structures	LOOCV	2b	3b	4b	$\lvert e_{\max}\rvert$
NbTiVZr	ATAT	0.03	1485	0.0098	102	172	0	0.0665
	Manual	0.03	1485	0.0094	102	190	0	0.0674
	LAR	0.03	1485	0.0109	56	139	17	0.0699
HfNbTaTiZr	ATAT	0.05	810	0.0085	80	140	0	0.0263
	Manual	0.05	810	0.0075	80	147	3	0.0248
	LAR	0.05	810	0.0100	79	139	0	0.0226
AlHfNbTaTiZr	ATAT	0.08	2035	0.0171	120	260	0	0.1060
	Manual	0.08	2035	0.0138	120	442	0	0.1001
	LAR	0.08	2035	—	195	735	0	0.0872

Table 2: Comparison of the best regressions obtained by the three different linear methods for each high-entropy alloy in question. "2b", "3b", and "4b" stand for the number of included 2-body, 3-body, and 4-body clusters, respectively. $\lvert e_{\max}\rvert$ is the maximum absolute error per atom for the regression in question. No LOOCV score is obtained for the LAR fit for AlHfNbTaTiZr due to numerical issues — the matrix in the denominator of Eq. (9) (see Appendix) becomes ill-conditioned. The LAR fit tends to include fewer clusters, which most likely contributes to a higher LOOCV score.



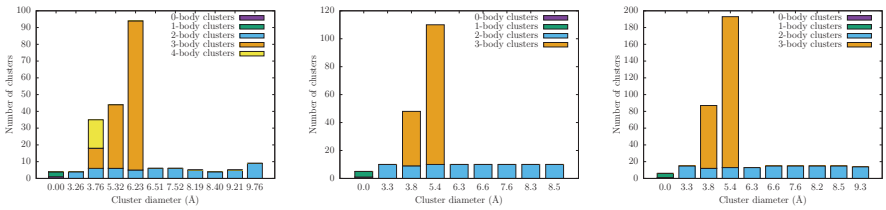
(a) NbTiVZr (b) HfNbTaTiZr (c) AlHfNbTaTiZr

Figure 2: Histograms of the different types of clusters included in the automated fit for each HEA as described in Section 2.2. These models include many longer-range 2-body interactions and some 3-body interactions. However, none of them include 4-body interactions. Additionally, very few 3-body clusters are included in the AlHfNbTaTiZr case, most likely due to a dearth of structures.



(a) NbTiVZr (b) HfNbTaTiZr (c) AlHfNbTaTiZr

Figure 3: Histograms of the different types of clusters present in the models obtained from the higher-order cluster augmentation method described in Section 2.2. In the AlHfNbTaTiZr case, adding 4-body clusters leads to an increase in the LOOCV score and a worse fit, which is not the case for NbTiVZr or HfNbTaTiZr. In all cases, the LOOCV score is improved by the inclusion of the additional clusters; quite notably, the LOOCV score of the AlHfNbTaTiZr fit is almost cut in half.



(a) NbTiVZr (b) HfNbTaTiZr (c) AlHfNbTaTiZr

Figure 4: Histograms of the different types of clusters included in the LAR fit for each HEA. No LOOCV score is obtained for the AlHfNbTaTiZr case due to numerical issues. In the HfNbTaTiZr and AlHfNbTaTiZr cases, the absence of 4-body clusters in the fits obtained by using the LAR method is likely due to the limited size of the training data set.

3.3. Nonlinear Fitting Methods

The nonlinear fitting methods explored in this work use the full set of AlHfNbTaTiZr structures with a strain cutoff of 0.08. This yields a dataset of 2035 structures. For all ridge- and kernel-based regression methods, the data set is randomly shuffled and the first 600 samples are utilized for training and validation—400 for training and 200 for validation—using the three-way holdout validation method, while the rest of the data is used as the test (unseen) dataset.

Given the large number of potential many-body correlations (1730) included in the original linear system, it is useful to reduce the dimensionality of the feature space to guarantee the applicability and quality of the method. As mentioned in Section 2.2, principal component analysis (PCA), a linear transformation which projects the features in the direction of largest variance, is used to reach this goal. PCA is useful because it adapts the fitting process to the distribution of the given pre-generated data.

Figure 5 shows the convergence of the coefficient of determination (R^2 score) for regressions fitting the total energy of each structure plotted against the number of features after applying PCA. In addition to plotting the curves for ridge regression, kernel ridge regression, and Gaussian process regression, it also shows two horizontal lines delineating the ideal R^2 score of 1.0 (implying that the variance in the data is fully explained by the model) and the R^2 score of the ridge regression fitted using all of the original 1730 many-body correlations. It is found that between 300 and 400 features (that is, between 17% and 23% of the original feature space) are selected as the principal components. It is also important to note that with only 400 features, the model is able to capture about 96% of the variance in the data and the differences between the three models are 1% or less after using only 300 features. The trend shows that selecting more than 300 features does not meaningfully alter the predictive power of the fit, which is to be expected as 400 samples are used to train the models and they are thus close to reaching the maximum of their learning capacity.

PCA projects the original features into a new subspace and selects the relevant

features in a robust and almost automatic way. In this way, it eliminates the collinearity between features and simplifies the fitting process. However, this projection makes a direct physical interpretation of these new orthogonal features difficult. The contribution of each of the new features can then be explored using linear ridge regression, kernel ridge regression, and Gaussian process regression.

Observing Fig. 6, which shows the predicted per-atom formation enthalpy plotted against the DFT-calculated per-atom formation enthalpy, it is clear that the models capture much of the variance in the data, with an R^2 score of over 0.99, using just 300 features. Table 3 shows different statistical indicators: the R^2 score, the Root Mean Square Error, and the Mean Absolute Error of the formation enthalpy per atom of each structure in the testing set. Table 4 shows other indicators of statistical error such as the mean error, maximum positive error, and maximum negative error, which again shows much similarity in predictive performance in statistical terms. The small differences in the statistical indicators in Tables 3 and 4 suggest that the additional degrees of freedom afforded by a non-linear model do not enhance the performance. This means that a linear combination of features is sufficient to capture the complexity of the energy landscape of these multicomponent alloys. However, following Occam's Razor, it is preferable to use the simplest model — in this case, the linear ridge regression with PCA.

It is hard to directly compare the methods described here with those of Section 3.2 due to the application of PCA. With the linear methods in Section 3.2, there is an easy and intuitive physical understanding that emerges from the preservation of the clusters as the features of interest. Once PCA is performed, that physical intuition is lost, although the automatic reduction in the number of coefficients is useful. As discussed above, introducing non-linearity into the model itself does not lead to significantly better results, so using any of the methods in Section 3.2 or this section should yield similar results.

	RR	KRR	GPR
R^2	0.9951	0.9949	0.9942
RMSE	0.0205	0.0209	0.0222
MAE	0.0118	0.0120	0.0130

Table 3: Comparison of error metrics for the various non-linear fitting methods: ridge regression (RR), kernel ridge regression (KRR), and Gaussian process regression (GPR). When the various regression methods are compared on all three metrics — the coefficient of determination score (see Section 2.2), the root-mean-square error (RMSE), and the mean absolute error (MAE) — it is clear that they all perform comparably. All errors are in units of $^{\mathrm{eV}}/_{\mathrm{atom}}$. Indeed, comparing the various plots of predicted energy versus actual energy (Fig. 6), the plots appear nearly identical, suggesting that the principal components selected through PCA largely account for the impact of structural variations on the energy.

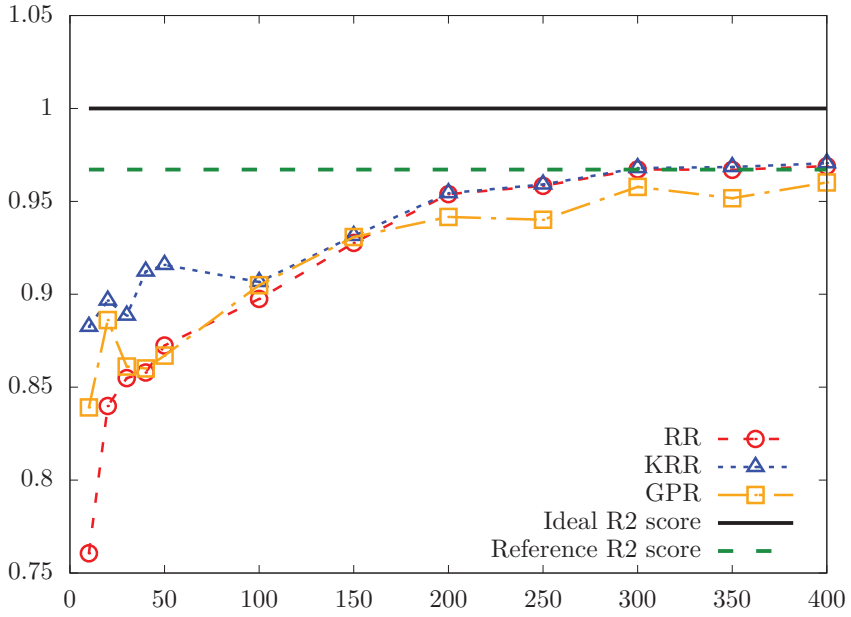
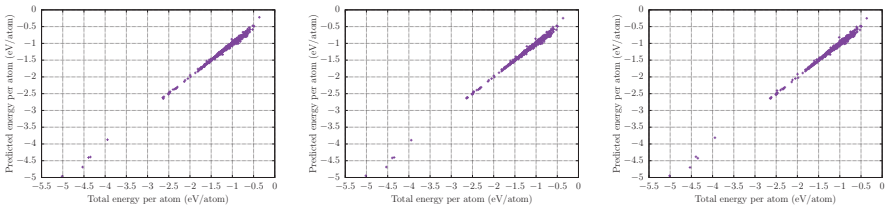


Figure 5: Convergence of the coefficient of determination (R^2 score) of the total energy of each structure against the number of new features obtained from the PCA projection. The regressions shown here are: linear ridge regression (RR), kernel ridge regression (KRR), and Gaussian process regression (GPR). The solid horizontal line denotes the ideal R^2 score of 1 (the model captures all of the variance in the data), while the thick dashed horizontal line is the R^2 score (0.9671) when fitting a model using the original 1730 features with linear ridge regression.



(a) Ridge regression with 300 principal components

(b) Kernel ridge regression with 300 principal components

(c) Gaussian process regression with 400 principal components

Figure 6: Plots of predicted versus actual energies for various linear and non-linear fitting methods. Unlike the linear methods mentioned earlier, several of these nonlinear methods involve a transformation into the principal component space, which prevents a similar analysis to Figs. 2 to 4. Thus, a plot of predicted energies versus actual energies, along with the R^2 score, is used as a measure of the quality of the fit.

	RR	KRR	GPR
Mean absolute error	0.0012	0.0016	0.0018
Max negative error	-0.1428	-0.1529	-0.1939
Max positive error	0.1557	0.1551	0.1674

Table 4: Comparison of additional error metrics for the various non-linear fitting methods: ridge regression (RR), kernel ridge regression (KRR), and Gaussian process regression (GPR). All errors are in units of $^{eV}/_{atom}$. As in Table 3, the fits all perform comparably, suggesting that the features in the PCA space contribute linearly to the energy calculation.

4. Evidence of Phase Separation

Analysis of phase stability assumes importance because many physical properties of HEAs are affected by the presence of precipitates. For example, all of the high entropy alloys examined here experience a drastic reduction in mechanical strength around 1000 K [22, 24, 25], shown in Fig. 7. Specifically, the stress-strain response is drastically different at 1073 K when compared to the behavior at 296 K–298 K. In addition, it is interesting to explore the length scales associated with short-range order, if any, present in the HEAs that do not form precipitates. For example, typical supercells used in DFT calculations contain 100–300 atoms. So, if the length scale associated with the decay in the SRO parameter (obtained from cluster expansion + Monte Carlo) is a few nanometers, then these results cannot easily be reproduced using such DFT supercells. On the other hand, if the SRO decays very quickly, then a random distribution of alloying elements can be used to model these complex alloys.

The cluster expansion models described in the previous sections are used to analyze phase stability and clustering in the three HEAs. LAR fits are used for NbTiVZr and AlHfNbTaTiZr, while a fit generated using the higher-order cluster augmentation method is used for HfNbTaTiZr. To study the phase stability of these alloys, Monte Carlo trajectories are obtained at various temperatures from 250 K to the melting point. Representative simulation cells are provided in Figs. 8 to 10 between 250 K and 1000 K for all high entropy alloy systems studied here. Visual evidence of phase segregation is corroborated by examining the multicomponent Warren-Cowley binary short-range order parameters described in Section 2.3 (see Section S.6 for the relevant figures). In this analysis, atoms in the first coordination shell are used to generate the SRO parameters.

4.1. Analysis of Phase Segregation in NbTiVZr

Figure 8 shows snapshots of representative equilibrium structures of NbTiVZr obtained from canonical Monte Carlo simulations. The clustering of vanadium atoms (shown in red) clearly shows that V atoms like to form clusters at low temperatures. In fact, at 250 K (Fig. 8a), almost all vanadium atoms in the material are present in the form of a cluster. The vanadium atoms gradually disperse into the rest of the material as the temperature is increased to 1000 K. This evolution is further corroborated by the trajectory of the V-V SRO parameter (see Fig. 11 in Section S.6), which begins close to -2 at 250 K and rises to ≈ -0.5

around 1000 K, where the segregation disappears. Concurrently, there is visual evidence in Fig. 8 of a first-order phase transition of a zirconium-rich phase between 500 K and 625 K. As in the vanadium-rich phase, almost all zirconium atoms in the material segregate into a separate (zirconium) cluster. However, unlike the phase transition of the vanadium-rich phase, in which the equilibrium concentrations of the precipitate and matrix change slowly with temperature, the zirconium-rich phase undergoes a sudden transition. This is reflected in the values of the Zr-Zr SRO parameter (see Fig. 11 in Section S.6), which decreases from near-zero at 250 K to ≈ -1 at 500 K. The SRO parameter then exhibits a kink and increases towards a value of 0 at slightly above 1000 K. Finally, the positive Nb-V SRO and negative Nb-Nb SRO at low temperatures suggests that Nb does not want to form clusters with V.

The results from grand canonical ensemble Monte Carlo simulations at 1000 K, 1500 K, 2000 K, and 2500 K are shown in Fig. 11. These graphs are ternary projections of the full tetrahedral phase diagram with Ti, V, and Zr at the corners. That is, in a regular Cartesian coordinate system, Ti is at the coordinate (0,0,0), V is at the coordinate (1,0,0), Zr is at the coordinate $(\frac{1}{2},\frac{\sqrt{3}}{2},0)$, and Nb is at the coordinate $(\frac{1}{2},\frac{\sqrt{3}}{2},\sqrt{\frac{2}{2}})$. Thus, the Nb "dimension" is flattened in these diagrams. In this set of simulations, a $1\,{}^{\mathrm{eV}}/_{\mathrm{atom}}$ interval is scanned in each direction around the chemical potential values that stabilize the equiatomic phase at 2500 K.

At 2500 K (Fig. 11d), the equiatomic phase is stable and there is a clear miscibility gap towards higher concentrations of Nb and Zr, as indicated by the lack of structures in that area of the phase diagram. This miscibility gap extends to all studied temperatures. At lower temperatures (Fig. 11a), there seem to be very quick transitions to extreme ends of the phase diagram even for minute changes in chemical potential, showing a marked contrast with diagrams at higher temperatures. For example, it can be seen in Fig. 11a that there are transitions from a Zr-rich phase to $Zr_{0.75}V_{0.25}$ to a V-rich phase along the right edge of the ternary projection. Given that the equiatomic phase occurs at the center of the ternary diagram (for all four ternary projections), it becomes clear that the equiatomic phase is stable at 2500 K and quite likely unstable at 1000 K.

Previous research [24] has shown that there is a large difference in the stress-strain behavior of this material between 298 K and 1073 K. Figure 7 reports the prior finding that NbTiVZr has very high ductility and yield strength at 298 K, but the yield strength drastically decreases at 1073 K. Given the change in the distribution of alloying elements in this HEA (i.e. the presence of V and Zr rich phases seen in Fig. 8), we conjecture that this degradation in mechanical properties is closely related to the changes observed as a function of temperature in the V- and Zr-rich phases. These phase changes, and the resulting changes in local chemistry, possibly play a role in the drastic reduction of strength of this material at higher temperatures.

In their work [24], Senkov et. al. found 3 phases — a Nb-rich untransformed

BCC matrix, a Zr-rich transformed BCC matrix, and V-rich precipitates. In experiments, they found Nb segregation, V segregation, and Zr segregation, and these segregations are more pronounced at 1273 K than at 298 K. In an experiment, phase transformation and chemical segregation is governed by several factors, including local distortion, local transformation strain, diffusion, configurational and vibrational entropy contributions, change in enthalpy, and free energy barriers. The cluster expansion model used here accounts for local distortions, configurational entropy, and changes in enthalpy. It does not take kinetics (diffusion, energy barriers, etc) into account, but that is not needed to detect a phase transition. Taking these qualifications into account, we believe that the evidence shown in Figs. 8 and 11 suggests that the V, Zr, and Nb clusters seen in this work could be early precursors of the phases observed by Senkov et al. Figures isolating the V and Zr atoms can be seen in Figs. 5 and 6 in Section S.5.

4.2. Analysis of Phase Segregation in HfNbTaTiZr

The phases in the case of HfNbTaTiZr are slightly harder to discern visually by looking at the snapshots of the simulation cells in Fig. 9, although the SROs (Fig. 12 in Section S.6) paint a clearer picture. Previous studies have observed a TaNb BCC phase and a HfZr HCP phase at low temperatures in this material [2, 31] (see Table 5). Given that the classic cluster expansion formalism cannot take multiple lattices into account, the HfZr phase cannot be seen here. However, the negative Hf-Zr SRO parameter is a manifestation of the presence of a local minimum on the energy surface and consequently points to the existence of this HfZr phase. The fact that the BCC cluster expansion constructed here could indicate the presence of an HCP phase is also plausible because the two lattices are closely related through the Burgers path. The BCC TaNb phase is also clearly seen through the negative Ta-Nb SRO parameter (≈ -0.5) and highly negative Nb-Nb SRO parameter (less than -1), and the similar magnitude of the Ti-Zr and Hf-Zr SRO parameters suggests the existence of a TiZr phase, although the exact lattice system of this phase cannot be discerned here. There is also a possibility that Hf, Ti, and Zr form the FCC HfTiZr phases noted at higher temperatures (see Table 5 and [31]). The increase in these SRO parameters at higher temperatures suggests that, in corroboration with previous research, the dominant phase at higher temperatures is the disordered single-phase "matrix". As in the case of NbTiVZr, there is a correlation between these phase transformations and the change in stress-strain behavior of this material as seen in Fig. 7. The changes in local chemistry induced as a result of the disordering of the HfZr, TaNb, and TiZr phases at higher temperatures most likely play a role in the drastic reduction of the strength of this material.

The results from grand canonical Monte Carlo simulations at 1000 K and 4000 K are shown in Fig. 12. It is relevant to note here that 4000 K is higher than the melting temperature in this material and is not physically accessible. However, using such temperatures allows us to better observe the limiting high-temperature behavior of the material. In this set of simulations, a $1\,^{\mathrm{eV}}/_{\mathrm{atom}}$

interval is scanned in each direction around the chemical potential values that stabilize the equiatomic phase at 4000 K. In Figs. 12a to 12c, it's clear that there is a transition from $Hf_{0.5}Ti_{0.5}$ to $Hf_{0.25}Ti_{0.2}Zr_{0.55}$ as the chemical potential of Hf is varied. Further, at 1000 K, Nb does not appear in any of the scanned chemical potential region (Fig. 12a), while Fig. 12d shows significant fractions of Nb at different concentrations (from ≈ 0.1 to ≈ 0.6, depending on the part of the chemical potential space) at 4000 K. In addition, at 1000 K, there is a transition from $Hf_{0.5}Ti_{0.5}$ to $Hf_{0.45}Ta_{0.53}Ti_{0.02}$ as the chemical potential of Ta is varied, a transition from $Hf_{0.5}Ta_{0.5}$ to $Hf_{0.5}Ti_{0.5}$ as the chemical potential of Ti is varied, and a transition from $Hf_{0.5}Ti_{0.5}$ to $Hf_{0.25}Ti_{0.2}Zr_{0.55}$ as the chemical potential of Zr is varied (Figs. 12a to 12c). Comparing Figs. 12a to 12c with Figs. 12d to 12f, the lack of jumps in concentration around the equiatomic phase at 4000 K shows that the equiatomic HfNbTaTiZr is stable at 4000 K and unstable at 1000 K. Indeed, at least in this chemical potential range, the equiatomic phase does not appear at 1000 K, although further verification is needed. Calculations at intermediate temperatures are provided in the supplementary materials.

Interestingly, the behavior of the SRO parameters in this system contrast sharply with those of both NbTiVZr and AlHfNbTaTiZr. Many of the SRO parameters remain significantly different from zero even at the melting point, indicating short-range ordering. This concords with the HfTiZr phases noted at higher temperatures [31]. This suggests that, contrary to NbTiVZr and AlHfNbTaTiZr, the mechanical properties of this material may depend more strongly on other factors than on the local chemistry and state of disorder.

4.3. Analysis of Phase Segregation in AlHfNbTaTiZr

At low temperatures, AlHfNbTaTiZr (Fig. 10) develops a highly ordered phase and undergoes a phase transformation between 500 K and 600 K. Ta and Nb segregate to one side of the crystal, while Al, Hf, Ti, and Zr segregate to the other (this can clearly be seen visually when isolating groups of species as in Figs. 7 and 8 in Section S.5).[1] This is also apparent in the SRO parameters (see Fig. 13 in Section S.6), where the Al-Zr, Hf-Zr, Al-Ti, and Nb-Ta parameters are less than -1, reflecting a segregation tendency. Interestingly, unlike HfNbTaTiZr, the Zr-Ti, Hf-Ti and Zr-Hf SROs do not show any tendency to segregate.

Senkov et al. [25] reported that the Al containing HEA with a composition of $Al_{0.4}Hf_{0.6}NbTaTiZr$ contains a HCP phase (with a volume fraction of $\approx 13\%$ at 973 K) that is rich in Al and Zr. Zr and Hf lie in the same column in the periodic table (iso-electronic). In addition, Zr, Ti and Hf are HCP metals even though their atomic radii differ slightly. If the Al-Zr clustering observed by Senkov et al. corresponds to the Al-Zr clusters in our simulations, then we conjecture based on the SRO profiles that the HCP phase in $Al_{0.4}Hf_{0.6}NbTaTiZr$ is also rich in Hf and Ti.

[1] It should be noted here that MC simulations that do not include vibrations could overestimate the transition temperature.

Additionally, the kink in the Al-Hf, Al-Ti, Al-Zr, Hf-Zr, Hf-Hf, Hf-Nb, Nb-Nb, Nb-Ta, Nb-Zr, Ta-Ta, Ta-Ti, Ti-Ti, and Ti-Zr SRO parameters between 500 K and 600 K (see Fig. 13 in Section S.6) suggests a phase transition which is, indeed, reflected visually in Fig. 10. Above those temperatures, the segregation largely disappears and the only phase present is the disordered single-phase "matrix". As with the other alloys, this segregation at low temperature and the phase transition described here may play a crucial role in explaining the decline in strength seen in Fig. 7.

The segregation of Ta and Nb observed in our simulations (Figs. 7, 8 and 10) and also quantitatively shown in the SRO plots (Fig. 13 in Section S.6) is in line with the observations made by Lin et al. [12]. Lin et al. conjectured that Nb and Ta have higher melting points than the other alloying elements and hence form Ta and Nb rich dendrites as the melt is cooled. Our results suggest that apart from their high melting points, there is also an underlying energetic driving force, in the solid state, for the observed segregation.

The results from grand canonical ensemble Monte Carlo simulations at 1000 K and 2500 K are shown in Fig. 13. In these simulations, a $1\,^{\mathrm{eV}}/_{\mathrm{atom}}$ interval is scanned in each direction around the chemical potential values that stabilize the equiatomic phase at 2500 K. There is a phase transition from a roughly equiatomic alloy phase to a Ta-rich phase at 1000 K as the chemical potential of Ta is changed, which can be seen in the jumps in the "Ta" curve in Figs. 13a to 13c. Similarly, still at 1000 K, there is a jump from a Nb-Ta rich phase to a Al-Hf-Nb-Ti-Zr rich phase as the chemical potential of Al is changed and a jump in the opposite direction as the chemical potential of Nb is changed. In contrast, the target phase is stable at 2500 K, as can be seen by the smooth curves in Figs. 13d to 13f. Comparing Figs. 13a to 13c and Figs. 13d to 13f, it is clear that the equiatomic phase of interest is stable at a broader range of compositions at 2500 K than at 1000 K due to the lack of jumps in composition around that phase.

Senkov et al. in 2014 [25] found that at temperatures below $\approx$ 1300 K, the aluminum containing HEA, $Al_{0.4}Hf_{0.6}NbTaTiZr$, contains three phases: a Ta and Nb rich BCC phase, a Ti and Zr rich and Nb, Ta depleted BCC phase and a Al and Hf rich HCP phase. While the true energy landscape of this alloy can be very complicated due to the differences in chemistry and atomic radii, the kinetics of such a system can also be very complicated. Our results correspond to metastable or stable phases, but cannot account for strain or multiple lattice systems. Nevertheless, the fact that the Nb-Ta SRO is less than 0 suggests some sort of BCC or HCP precipitate can be present in this alloy.

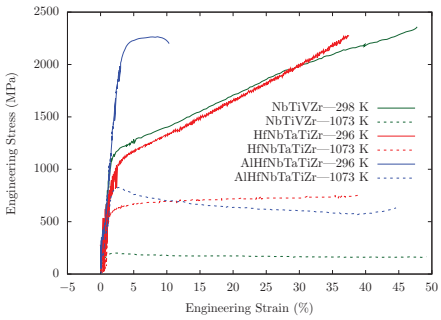
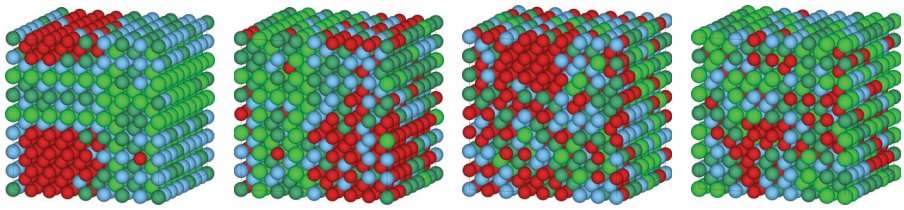
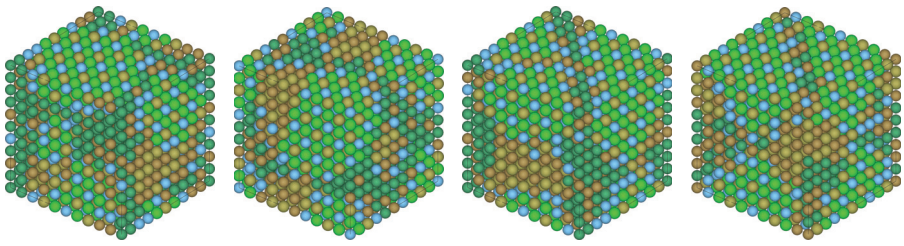


Figure 7: Stress-Strain curves showing a drastic decrease in mechanical strength at 1073 K for all HEAs studied here. WebPlotDigitizer [19] is utilized to extract the stress-strain curves plotted in Fig. 7 from the plots in [22, 24, 25].



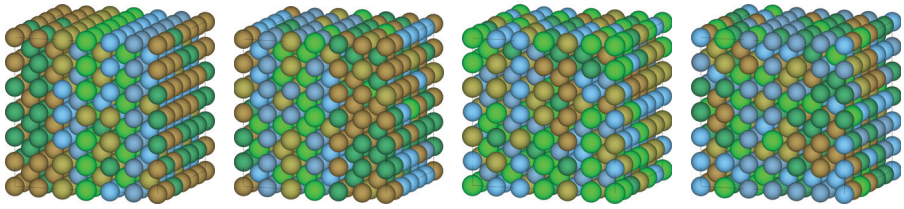
(a) Representative simulation cell at 250 K (b) Representative simulation cell at 500 K (c) Representative simulation cell at 750 K (d) Representative simulation cell at 1000 K

Figure 8: Representative NbTiVZr simulation cells at several temperatures between 250 K and 1000 K. At 250 K, vanadium atoms form a cluster that gradually decreases in size with an increase in temperature. Key: Nb, Ti, V, Zr.



(a) Representative simulation cell at 200 K (b) Representative simulation cell at 400 K (c) Representative simulation cell at 800 K (d) Representative simulation cell at 1000 K

Figure 9: Representative HfNbTaTiZr simulation cells at several temperatures between 200 K 1000 K. At lower temperatures, several intermetallic phases (HfZr, TaNb, and TiZr) form. These slowly disperse into the matrix as the temperature is increased. However, other intermetallic phases form at higher temperatures (FCC HfTiZr). Key: Hf, Nb, Ta, Ti, Zr.

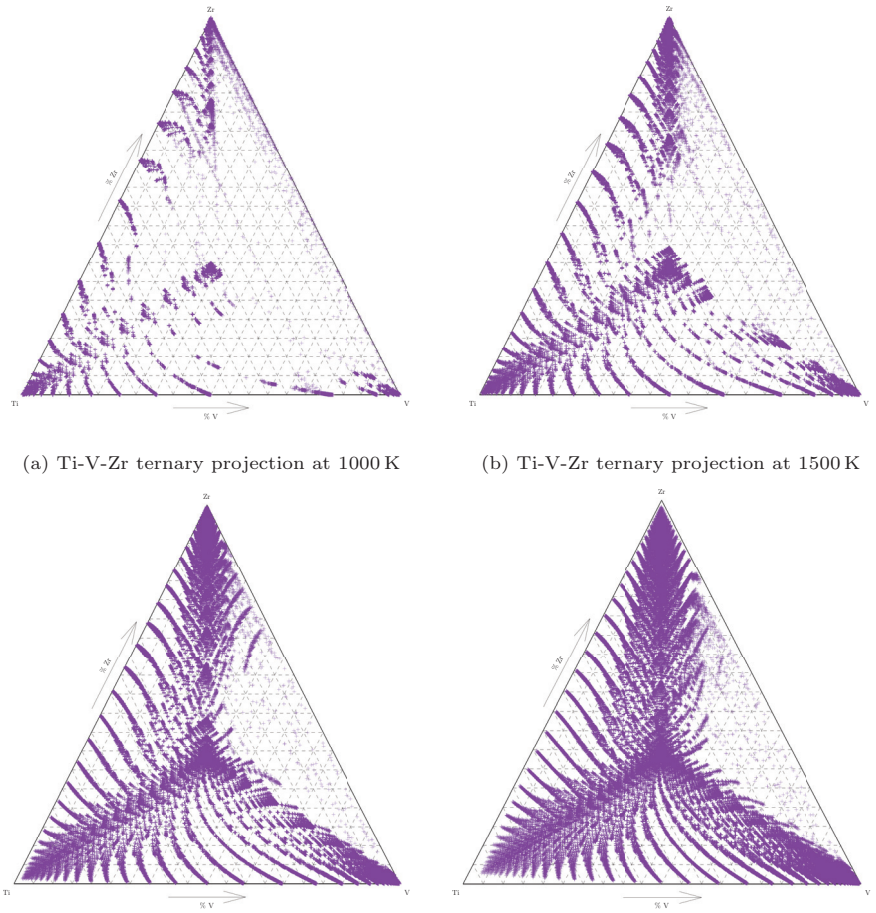


(a) Representative simulation cell at 250 K (b) Representative simulation cell at 500 K (c) Representative simulation cell at 750 K (d) Representative simulation cell at 1000 K

Figure 10: Representative AlHfNbTaTiZr simulation cells at several temperatures between 250 K and 1000 K. At 250 K, there is clear segregation of Al, Hf, Ti, and Zr from Nb and Ta, forming two distinct phases. The alloy slowly becomes more disordered as the temperature is increased. Key: Al, Hf, Nb, Ta, Ti, Zr.

Temperature (°C)	Phases
550	TaNb BCC + minor matrix BCC + HfZr HCP
700	TaNb BCC + HfZr HCP + Matrix BCC
	TaNb: (5.39%, 34.56%, 40.76%, 14.98%, 4.30%)
	HfZr: (36.21%, 0.69%, 4.69%, 15.65%, 42.77%)
	Matrix: (20.97%, 17.23%, 20.82%, 21.10%, 19.89%)
900	Minor TaNb BCC + matrix BCC
1000	HfZr HCP + Matrix BCC
	HfZr: (37.37%, 2.96%, 3.81%, 10.38%, 45.48%)
	Matrix: (20.06%, 19.09%, 19.78%, 21.82%, 19.23%)
1450	Two HfTiZr FCC + Matrix BCC
	HfTiZr FCC1: (36.53%, 0.94%, 0.49%, 23.31%, 38.74%)
	HfTiZr FCC2: (39.28%, 0.78%, 0.19%, 14.17%, 45.58%)
As-homogenized	Matrix BCC

Table 5: Phase composition of HfNbTaTiZr at various temperatures. Most relevant to this work is the existence of a BCC TaNb phase and a HCP HfZr phase at lower temperatures and the two FCC HfTiZr phases at higher temperatures. Extracted from [2, 31].



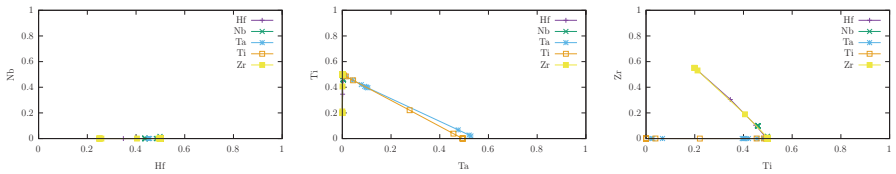
(a) Ti-V-Zr ternary projection at 1000 K

(b) Ti-V-Zr ternary projection at 1500 K

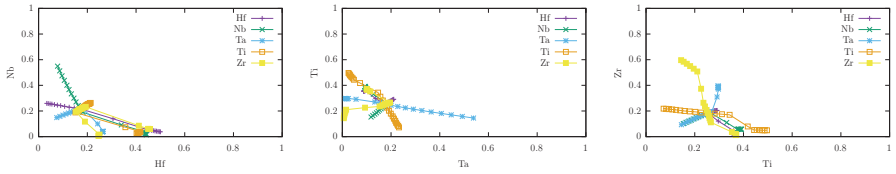
(c) Ti-V-Zr ternary projection at 2000 K

(d) Ti-V-Zr ternary projection at 2500 K

Figure 11: Results from Grand Canonical Monte Carlo simulations at 1000 K, 1500 K, 2000 K, and 2500 K for NbTiVZr. These diagrams are ternary projections of the full tetrahedral phase diagram and the Nb "dimension" is flattened. It is apparent from these simulations that there is a miscibility gap at increased concentrations of V and Zr that persists at all of these temperatures. Further, at lower temperatures (Fig. 11a), there are quick phase transitions to the extreme ends of the phase diagram over small changes in chemical potential. For example, at the right edge of the triangle in Fig. 11a, there is a transition from a Zr-rich phase to $Zr_{0.75}V_{0.25}$ to a V-rich phase. The equiatomic phase occurs at the center of the phase diagram, and from these ternary projections, it is clear that the equiatomic phase is stable at 2500 K and likely unstable at 1000 K.

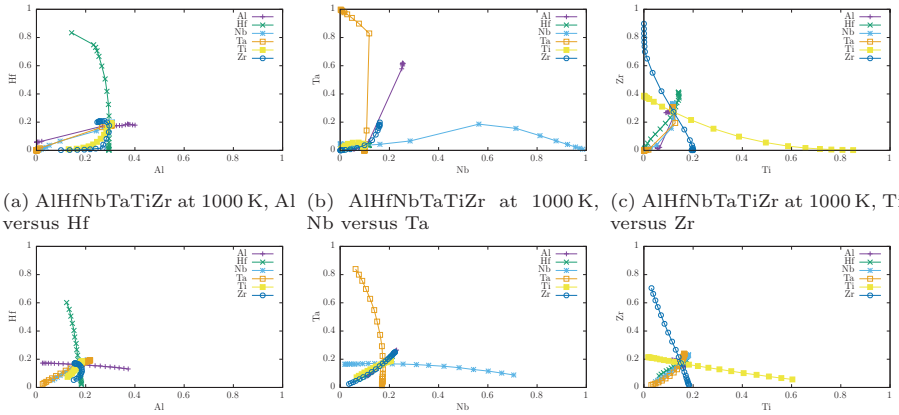


(a) HfNbTaTiZr at 1000 K, Hf versus Nb (b) HfNbTaTiZr at 1000 K, Ta versus Ti (c) HfNbTaTiZr at 1000 K, Ti versus Zr



(d) HfNbTaTiZr at 4000 K, Hf versus Nb (e) HfNbTaTiZr at 4000 K, Ta versus Ti (f) HfNbTaTiZr at 4000 K, Ti versus Zr

Figure 12: Results from Grand Canonical Monte Carlo simulations at 1000 K (Figs. 12a to 12c) and 4000 K (Figs. 12d to 12f) for HfNbTaTiZr. Each curve tracks how concentrations of the various elements vary as the chemical potential of that species is varied. Jumps in concentrations signify a phase transition. For example, in Fig. 12a, it is apparent that there is no significant fraction of Nb in the scanned chemical potential region (although this is not the case at higher temperatures). Further, in Figs. 12a to 12c, it's clear that there is a transition from $Hf_{0.5}Ti_{0.5}$ to $Hf_{0.25}Ti_{0.2}Zr_{0.55}$ as the chemical potential of Hf is varied. There is also a phase transition from $Hf_{0.5}Ti_{0.5}$ to $Hf_{0.45}Ta_{0.53}Ti_{0.02}$ as the chemical potential of Ta is varied and a transition from $Hf_{0.5}Ti_{0.5}$ to $Hf_{0.5}Ta_{0.5}$ as the chemical potential of Ti is varied. Finally, there is a phase transition from $Hf_{0.5}Ti_{0.5}$ to $Hf_{0.25}Ti_{0.2}Zr_{0.55}$ as the chemical potential of Zr is varied. At 4000 K, the equiatomic phase is stable, as shown in Figs. 12d to 12f — all composition curves are smooth, indicating a lack of phase transitions at this temperature. Further calculations at intermediate temperatures are provided in the supplementary materials.



(a) AlHfNbTaTiZr at 1000 K, Al versus Hf

(b) AlHfNbTaTiZr at 1000 K, Nb versus Ta

(c) AlHfNbTaTiZr at 1000 K, Ti versus Zr

(d) AlHfNbTaTiZr at 2500 K, Al versus Hf

(e) AlHfNbTaTiZr at 2500 K, Nb versus Ta

(f) AlHfNbTaTiZr at 2500 K, Ti versus Zr

Figure 13: Results from Grand Canonical Monte Carlo simulations at 1000 K (Figs. 13a to 13c) and 2500 K (Figs. 13d to 13f) for AlHfNbTaTiZr. Each curve tracks how concentrations of the various elements vary as the chemical potential of that species is varied. Jumps in concentrations signify a phase transition. At 1000 K, there is a sudden jump from a phase rich in all 6 elements (slightly favouring Al and Zr) to a Ta-rich phase as the chemical potential of Ta is changed. There is similarly a jump from a Nb-Ta rich phase to a Al-Hf-Nb-Ti-Zr rich phase as the chemical potential of Al is changed and a jump in the opposite direction as the chemical potential of Nb is changed. At 2500 K, the equiatomic phase is stable, as shown in Figs. 13d to 13f — all composition curves are smooth, indicating a lack of phase transitions at this temperature. Further calculations at intermediate temperatures are provided in the supplementary materials.

5. Conclusion

Several linear and non-linear cluster expansion models are used to study phase stability and elemental segregation effects in high entropy alloys. However, it is shown here that generating a cluster expansion (CE) model for high entropy alloys is a challenging task; due to the increase in the number of alloying elements, the number of possible clusters (i.e. many-body correlations) increases exponentially. Correspondingly, the number of structures required to generate the CE model also increases. Even generating a suitable database becomes tricky — while the alloying elements in a HEA are present in nearly equal atomic proportions, the database required to generate the CE model must include a sufficiently diverse set of structures to capture possible phase segregation and precipitation effects. Such a diverse database can often contain structures that undergo large distortions and hence have to be removed from the database. An additional challenge is that, apart from local lattice distortions, the lattice structure itself may change during relaxation; due to the fixed-lattice nature of the cluster expansion formalism used here, these structures must be discarded as well. These issues are most clearly evident for AlHfNbTaTiZr.

It is clear from these results that 2-body and 3-body interactions are necessary to generate a predictive model. As shown in Table 2, the number of 3-body interactions utilized in the fit is nearly twice the number of 2-body clusters used. This stands in sharp contrast to simpler binary and ternary alloys, where 2-body clusters are the dominant type of cluster necessary (e.g. the cluster expansions generated for Ni_3Al with dopants by Sun et al. in [26]). Additionally, as demonstrated by the higher-order cluster augmentation method (see Section 3.2), adding 4-body clusters and larger 3-body clusters can often improve the predictive power of the fit. Further, alternatives like the least angle regression can be utilized in parallel with least-squares-based methods such as ATAT's automated method and the higher-order cluster augmentation method in cases where the need for a larger dataset makes the latter methods computationally demanding.

The cluster expansion formalism for alloys is a widely used method with a solid mathematical foundation, and the features it uses are intuitive, physically-motivated clusters. The fact that the cluster expansion has been successful, especially in the field of alloy theory, shows that these intuition-based features are very good at capturing the complexity of the underlying energy landscape. When these features are transformed into a different space using Principal Component Analysis, a set of orthogonal features is obtained at the cost of losing the physical intuition underlying the original features. Interestingly, the number of orthogonal features required to generate a cluster expansion model with similar predictive power (i.e. R2 score > 0.99) drastically decreases compared to the number of intuition-based, physically-motivated clusters shown in Table 2. Moreover, as in the case of the original cluster expansion formalism, these orthogonal features also contribute linearly towards the energy. This is evident from the fact that there is no statistical difference between the ridge regression,

kernel ridge regression, and Gaussian process regression results (Section 3.3). Therefore, a linear regression-based fitting process is sufficient to capture the complexities of the energy surface.

Cluster expansion models for each alloys are used to reproduce various intermetallic phases observed in experiments utilizing Monte Carlo methods and a connection is drawn between these phases and the sudden drop in material strength around 1000 K in all of these high entropy alloys (Section 4). NbTiVZr forms vanadium and zirconium clusters at lower temperatures (250 K) which disperse into the single-phase matrix by 1000 K. This is seen both through examining the structures visually and looking at the numerical values of the short-range order parameters and is most likely related to the intermetallic phases and precipitates seen in Senkov et al.'s work [24].

HfNbTaTiZr forms NbTa and HfZr intermetallic phases at 250 K, in addition to a possible TiZr intermetallic phase. The HfZr intermetallic phase may correspond to the HCP HfZr intermetallic phase seen in [31]. Additionally, the possible TiZr intermetallic phase seen here may actually correspond to the HfTiZr FCC phase noted at higher temperatures in previous work [31]. Furthermore, in contrast with NbTiVZr and AlHfNbTaTiZr, HfNbTaTiZr does not lose short-range ordering until well above the melting temperature (≈ 3500 K).

Finally, in AlHfNbTaTiZr, Nb-Ta and Al-Hf-Ti-Zr phase segregate at lower temperatures, while the solid solution phase emerges at higher temperatures. The Al-Hf-Ti-Zr phase seen here may be related to the experimentally-observed Al-Zr phase reported in the literature, though further verification is needed. The phase transitions described above most likely play a role in the sudden decline in mechanical strength seen at higher temperatures in all of these alloys, especially in the cases of NbTiVZr and AlHfNbTaTiZr.

6. Acknowledgment

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. The authors also acknowledge support from the National Science Foundation through grant DMR-2001411. Computing support for this work came from the Lawrence Livermore National Laboratory (LLNL) institutional computing facility.

References

[1] Jian Chen, Xueyang Zhou, Weili Wang, Bing Liu, Yukun Lv, Wei Yang, Dapeng Xu, and Yong Liu. A review on fundamental of high entropy alloys with promising high–temperature properties. *Journal of Alloys and Compounds*, 760:15–30, 2018. ISSN 0925-8388. doi: https://doi.org/10.1016/j.jallcom.2018.05.067.

[2] S.Y. Chen, Y. Tong, K.-K. Tseng, J.-W. Yeh, J.D. Poplawsky, J.G. Wen, M.C. Gao, G. Kim, W. Chen, Y. Ren, R. Feng, W.D. Li, and P.K. Liaw. Phase transformations of HfNbTaTiZr high-entropy alloy at intermediate temperatures. *Scripta Materialia*, 158:50–56, 2019. ISSN 1359-6462. doi: https://doi.org/10.1016/j.scriptamat.2018.08.032.

[3] D. de Fontaine. The number of independent pair-correlation functions in multicomponent systems. *Journal of Applied Crystallography*, 4(1):15–19, Feb 1971. doi: 10.1107/S0021889871006174.

[4] F. Ducastelle. *Order and Phase Stability in Alloys*. Cohesion and structure. North-Holland, 1991. ISBN 9780444869739. URL `https://books.google.com/books?id=2ZNTAAAAMAAJ`.

[5] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *Ann. Statist.*, 32(2):407–499, 04 2004. doi: 10.1214/009053604000000067.

[6] Manus Foster. An Application of the Wiener-Kolmogorov Smoothing Theory to Matrix Inversion. *Journal of the Society for Industrial and Applied Mathematics*, 9(3):387–392, 1961. doi: 10.1137/0109031.

[7] Arthur E. Hoerl and Robert W. Kennard. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1):55–67, 1970. doi: 10.1080/00401706.1970.10488634.

[8] G. Kresse and J. Furthmüller. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B*, 54: 11169–11186, 10 1996. doi: 10.1103/PhysRevB.54.11169.

[9] G. Kresse and J. Furthmüller. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Computational Materials Science*, 6(1):15–50, 1996. ISSN 0927-0256. doi: 10.1016/0927-0256(96)00008-0.

[10] G. Kresse and J. Hafner. Ab initio molecular dynamics for liquid metals. *Phys. Rev. B*, 47:558–561, 01 1993. doi: 10.1103/PhysRevB.47.558.

[11] G. Kresse and J. Hafner. Ab initio molecular-dynamics simulation of the liquid-metal–amorphous-semiconductor transition in germanium. *Phys. Rev. B*, 49:14251–14269, 05 1994. doi: 10.1103/PhysRevB.49.14251.

[12] Chun-Ming Lin, Chien-Chang Juan, Chia-Hsiu Chang, Che-Wei Tsai, and Jien-Wei Yeh. Effect of Al addition on mechanical properties and microstructure of refractory AlxHfNbTaTiZr alloys. *Journal of Alloys and Compounds*, 624:100–107, 2015. ISSN 0925-8388. doi: https://doi.org/10.1016/j.jallcom.2014.11.064.

[13] W.H. Liu, T. Yang, and C.T. Liu. Precipitation hardening in CoCrFeNi-based high entropy alloys. *Materials Chemistry and Physics*, 210:2–11, 2018. ISSN 0254-0584. doi: https://doi.org/10.1016/j.matchemphys.2017.07.037. URL http://www.sciencedirect.com/science/article/pii/S0254058417305461. High-Entropy Materials.

[14] C. L. Mallows. Some Comments on CP. *Technometrics*, 15(4):661–675, 1973. ISSN 00401706. URL http://www.jstor.org/stable/1267380.

[15] D.B. Miracle and O.N. Senkov. A critical review of high entropy alloys and related concepts. *Acta Materialia*, 122:448 – 511, 2017. ISSN 1359-6454. doi: https://doi.org/10.1016/j.actamat.2016.08.081.

[16] J. A. Nelder and R. Mead. A Simplex Method for Function Minimization. *The Computer Journal*, 7(4):308–313, 01 1965. ISSN 0010-4620. doi: 10.1093/comjnl/7.4.308. URL https://doi.org/10.1093/comjnl/7.4.308.

[17] Ricardo Olea. *Geostatistics for Engineers and Earth Scientists*. Springer US, 1999. doi: 10.1007/978-1-4615-5001-3.

[18] David L. Phillips. A Technique for the Numerical Solution of Certain Integral Equations of the First Kind. *J. ACM*, 9(1):84–97, January 1962. ISSN 0004-5411. doi: 10.1145/321105.321114.

[19] Ankit Rohatgi. *WebPlotDigitizer*. San Francisco, California, USA, 4.2 edition, 2019. URL https://automeris.io/WebPlotDigitizer.

[20] Matthias Rupp. Machine learning for quantum mechanics in a nutshell. *International Journal of Quantum Chemistry*, 115(16):1058–1073, 2015. doi: 10.1002/qua.24954. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/qua.24954.

[21] J.M. Sanchez, F. Ducastelle, and D. Gratias. Generalized cluster description of multicomponent systems. *Physica A: Statistical Mechanics and its Applications*, 128(1):334–350, 1984. ISSN 0378-4371. doi: https://doi.org/10.1016/0378-4371(84)90096-7. URL https://www.sciencedirect.com/science/article/pii/0378437184900967.

[22] O. N. Senkov, J. M. Scott, S. V. Senkova, F. Meisenkothen, D. B. Miracle, and C. F. Woodward. Microstructure and elevated temperature properties of a refractory TaNbHfZrTi alloy. *Journal of Materials Science*, 47:4062–4074, 05 2012. doi: 10.1007/s10853-012-6260-2.

[23] O.N. Senkov, J.M. Scott, S.V. Senkova, D.B. Miracle, and C.F. Woodward. Microstructure and room temperature properties of a high-entropy TaNbHfZrTi alloy. *Journal of Alloys and Compounds*, 509(20):6043–6048, 2011. ISSN 0925-8388. doi: https://doi.org/10.1016/j.jallcom.2011.02.171.

[24] O.N. Senkov, S.V. Senkova, D.B. Miracle, and C. Woodward. Mechanical properties of low-density, refractory multi-principal element alloys of the Cr–Nb–Ti–V–Zr system. *Materials Science and Engineering: A*, 565:51–62, 2013. ISSN 0921-5093. doi: https://doi.org/10.1016/j.msea.2012.12.018.

[25] O.N. Senkov, S.V. Senkova, and C. Woodward. Effect of aluminum on the microstructure and properties of two refractory high-entropy alloys. *Acta Materialia*, 68:214–228, 2014. ISSN 1359-6454. doi: https://doi.org/10.1016/j.actamat.2014.01.029.

[26] Ruoshi Sun, Christopher Woodward, and Axel van de Walle. First-principles study on $Ni_3Al(111)$ antiphase boundary with Ti and Hf impurities. *Phys. Rev. B*, 95:214121, Jun 2017. doi: 10.1103/PhysRevB.95.214121. URL `https://link.aps.org/doi/10.1103/PhysRevB.95.214121`.

[27] A. van de Walle and M. Asta. First-principles investigation of perfect and diffuse antiphase boundaries in HCP-based Ti-Al alloys. *Metallurgical and Materials Transactions A*, 33:735–741, 2002. doi: 10.1007/s11661-002-0139-9. URL `https://doi.org/10.1007/s11661-002-0139-9`.

[28] A. van de Walle and G. Ceder. Automating first-principles phase diagram calculations. *Journal of Phase Equilibria*, 23(4):348–359, 2002. doi: 10.1361/105497102770331596.

[29] A. van de Walle, M. Asta, and G. Ceder. The alloy theoretic automated toolkit: A user guide. *Calphad*, 26(4):539–553, 2002. ISSN 0364-5916. doi: https://doi.org/10.1016/S0364-5916(02)80006-2.

[30] Axel van de Walle. Multicomponent multisublattice alloys, nonconfigurational entropy and other additions to the Alloy Theoretic Automated Toolkit. *Calphad*, 33(2):266–278, 2009. ISSN 0364-5916. doi: https://doi.org/10.1016/j.calphad.2008.12.005. Tools for Computational Thermodynamics.

[31] Cheng Yang, Kenta Aoyagi, Huakang Bian, and Akihiko Chiba. Microstructure evolution and mechanical property of a precipitation-strengthened refractory high-entropy alloy HfNbTaTiZr. *Materials Letters*, 254:46–49, 2019. ISSN 0167-577X. doi: https://doi.org/10.1016/j.matlet.2019.07.027.

[32] Rui Zhou, Yong Liu, Bin Liu, Jia Li, and Qihong Fang. Precipitation behavior of selective laser melted FeCoCrNiC0.05 high entropy alloy. *Intermetallics*, 106:20–25, 2019. ISSN 0966-9795. doi: https://doi.org/10.1016/j.intermet.2018.12.001. URL `http://www.sciencedirect.com/science/article/pii/S0966979518306770`.

[33] Y.X. Zhuang, H.D. Xue, Z.Y. Chen, Z.Y. Hu, and J.C. He. Effect of annealing treatment on microstructures and mechanical properties of FeCoNiCuAl high entropy alloys. *Materials Science and Engineering: A*,

572:30–35, 2013. ISSN 0921-5093. doi: https://doi.org/10.1016/j.msea.2013.01.081. URL http://www.sciencedirect.com/science/article/pii/S0921509313001871.

Appendices

Appendix A Linear Regression Methods

Three linear regression methods are explored in this work: ATAT's automated method, a higher-order cluster augmentation method, and a method utilizing the least angle regression. An overview of all three methods is given below. In all of the discussions below, $\boldsymbol{A}$ is the matrix containing the correlations of n structures with m clusters, $\boldsymbol{y}$ is the vector of n DFT-calculated energies, and $\hat{\boldsymbol{y}}$ is the vector of n energies predicted by the model. Indexing notation convention denotes that e.g. y_i is the energy of structure i, $1 \leq i \leq n$.

The **automated algorithm** utilized by ATAT's `mmaps` program first constructs a minimal (non-colinear) basis from the matrix of correlations before generating a least-squares fit. Additionally, it utilizes a physics-based algorithm to iterate hierarchically through the cluster choices by only including a cluster if all of its subclusters have been included and all clusters with the same number of points with a smaller diameter have been included [28], where the diameter of a cluster is the largest distance between two points in the cluster. Model selection is done using the leave-one-out cross-validation score, which is a way of measuring the predictive power of a fit. In the general case, it is calculated as follows:

1. Remove one datapoint from the system (y_i)
2. Find the best fit for the $n-1$ datapoints still in the system
3. Predict the value of the omitted datapoint ($\hat{y}_i$)
4. Repeat for each datapoint in the system

Finding the squared norm of the difference between the two vectors $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$ and dividing by n yields the LOOCV score. In the specific case of a least-squares fit, Eq. (9) provides a more efficient method of calculating the LOOCV score which only requires calculating the full least-squares regression (the division here is element-wise) [28]:

$$\text{CV score} = \frac{1}{\sqrt{n}} \left|\left| \frac{\hat{\boldsymbol{y}} - \boldsymbol{y}}{\boldsymbol{1} - \text{diag}\left(\boldsymbol{A}(\boldsymbol{A}^\top \boldsymbol{A})^{-1} \boldsymbol{A}^\top\right)} \right|\right| \tag{9}$$

The `mmaps` code uses the LOOCV score as a metric to assess the quality of the fit and prevent over-fitting.

The **higher-order cluster augmentation** method is built on top of the automated method described above in order to test if the quality of the fit can be improved by deviating from `mmaps`' hierarchical rules. The method proceeds as follows:

1. Start with the results from the method outlined above. This fit will include n_2 2-body clusters with maximum cluster diameter d_2, n_3 3-body clusters with maximum cluster diameter d_3, and n_4 4-body clusters with maximum cluster diameter d_4. These fulfill the condition that $d_2 \geq d_3 \geq d_4$.
2. Increase d_3 such that it includes one more size of 3-body clusters (this could include different *types* of clusters, but they all have the same diameter).
3. For $n = 1, 2, \ldots, n_{d_3} - 1$, pick the first n clusters out of the clusters just included and generate a fit using the method above. Out of these, pick the fit with the lowest LOOCV score. This determines the number of clusters to use from the clusters just included (call this number n^*).
4. Now, pick $\min\left(\binom{n_{d_3}}{n^*}, C\right)$ different combinations of n^* clusters from the clusters just included. Select the one with the lowest LOOCV score and remove all of the other clusters of this size from future fits.
5. Repeat this procedure until $d_3 = d_2$. Then, repeat the procedure with d_4 instead of d_3 until $d_4 = d_3 = d_2$.

In this work, $C = 500$. As is shown in Section 3, manually augmenting the automated results obtained from the `mmaps` code with additional clusters, particularly larger 3-body clusters and some 4-body clusters, often improves the fit up to a point. The key issue encountered with this method is that as d_3 and d_4 are increased, the number of clusters grows exponentially. Since the LOOCV is a statistical estimator of the true predictive power, increasing the number of trial cluster choices increases the risk of finding a lower LOOCV purely by chance, even if the true predictive power of the corresponding cluster expansion is not truly better.

The **least angle regression** method, proposed by Efron et al. in a seminal paper in 2004 [5], is an alternative to the least squares-based fits described in the previous two methods. Unlike the least squares fit, the least angle regression method is iterative and introduces sparsity into the model — that is, it has the ability to set some of the coefficients in the solution vector to 0. Unlike other, similarly parsimonious methods, the iterative approach eliminates the need for a regularization parameter. The method starts by setting the coefficient vector (the vector of ECIs) equal to zero. Then, the correlations $\hat{\boldsymbol{c}}$ are calculated from $\hat{\boldsymbol{c}} = \boldsymbol{A}^\top (\boldsymbol{y} - \hat{\boldsymbol{y}})$, where $\boldsymbol{A}$ is the correlation matrix from the cluster expansion formalism, $\hat{\boldsymbol{y}}$ is the vector of predicted energies given by $\boldsymbol{A}\boldsymbol{x}$ (where $\boldsymbol{x}$ is the vector of ECIs), and $\boldsymbol{y}$ is the vector of actual energies being fitted. The set of clusters with the maximum absolute correlation is then defined as the "active set" and the ECIs are updated such that in the next iteration of the algorithm, one more cluster enters the active set. This procedure is repeated, including one more cluster in the active set at each iteration, until all of the clusters have been included in the active set (which corresponds to the ordinary least-squares fit). Hence, the least angle regression method obtains a trajectory of fits ranging from no clusters included in the active set (initial state) to all clusters included in the active set (the full least-squares fit).

Before fitting with LAR, the full dataset is split into a training dataset and a validation dataset. During model selection (described below), all calculations happen on the training dataset. To measure the predictive power of the dataset, the selected model is then used to predict the energies of the structures in the validation dataset.

Following the procedure used by Efron et al. in [5], Mallows' C_p criterion [14] is utilized here to select the best fit from the trajectory of fits returned by LAR. For the least angle regression, the equation for the C_p criterion reduces to Eq. (10) [5], where $\hat{\boldsymbol{x}}_{\text{OLS}}$ is the least-squares solution and k is the number of non-zero elements of $\hat{\boldsymbol{x}}$. As noted above, the C_p criterion is calculated based on the predicted energies of the structures in the training dataset.

$$C_p\left(\hat{\boldsymbol{x}}, k\right) = n\left(\frac{\left|\left|\boldsymbol{y} - \hat{\boldsymbol{y}}\left(\hat{\boldsymbol{x}}\right)\right|\right|^2}{\left|\left|\boldsymbol{y} - \hat{\boldsymbol{y}}\left(\hat{\boldsymbol{x}}_{\text{OLS}}\right)\right|\right|^2} - 1\right) + 2k \tag{10}$$

The $2k$ term in Eq. (10) avoids overfitting by increasing the score when the number of non-zero ECIs is increased.

Once the model has been selected, the coefficient of determination (Eq. (11)), or R^2 score for short, is calculated based on the predicted energies of the structures in the validation dataset to assess the predictive capability of the fit. In Eq. (11), $\boldsymbol{y}_v$ is the vector of energies from DFT for the validation dataset, $\hat{\boldsymbol{y}}_v$ is the vector of energies predicted by the cluster expansion for the validation set, $\bar{y}_v$ is the mean of $\boldsymbol{y}_v$, and $\mathbf{1}$ is a vector of 1s with the same length as $\boldsymbol{y}_v$.

$$R_v^2 = 1 - \frac{\left|\left|\boldsymbol{y}_v - \hat{\boldsymbol{y}}_v\right|\right|^2}{\left|\left|\boldsymbol{y}_v - \bar{y}_v\mathbf{1}\right|\right|^2} \tag{11}$$

Appendix B Non-Linear Regression Methods

In addition to the aforementioned linear regression methods, several linear and non-linear regression methods are combined with Principal Component Analysis to explore non-linear features. The three methods examined here are ridge regression, kernel ridge regression, and Gaussian process regression.

Before utilizing the aforementioned regression methods, Principal Component Analysis (PCA) is used to select a more compact set of features and decrease collinearity. The basic idea of PCA is that the highest variance directions provide the most information, so the directions with smallest variance can be neglected. If $\boldsymbol{A}$ is the correlation matrix (of size $n \times m$) containing the correlations of n structures with m clusters, the goal is to find a representation $\hat{\boldsymbol{A}}$ of size $n \times k$, with $k \ll m$ orthogonal features. This is done by means of an orthogonal linear transformation, obtained with the diagonalization of the covariance matrix $\boldsymbol{\Sigma}$:

$$\boldsymbol{\Sigma} = \frac{1}{n}\boldsymbol{A}^\top \boldsymbol{A} \tag{12}$$

A matrix $\boldsymbol{U}$ is then formed by arranging the eigenvalues of the covariance matrix $\boldsymbol{\Sigma}$ such that the corresponding eigenvalues are in descending order. Since the magnitude of the eigenvalue corresponds to the magnitude of the variance, a larger eigenvalue corresponds to a larger variance of the corresponding eigenvector. Then, to reduce the number of features, the first k column vectors are selected from this matrix $\boldsymbol{U}$ and the original correlations are projected into this space:

$$\hat{\boldsymbol{A}} = \boldsymbol{A}\boldsymbol{U} \tag{13}$$

How many principal components are chosen depends on how much variance those components capture. To analyze this, each eigenvalue is normalized by the sum of the eigenvalues and the k eigenvectors are selected such that the sum of the corresponding normalized eigenvalues is greater than 0.9. These k features are then used in the regression methods described below.

Ridge regression [6, 7, 18], also known as Tikhonov regularization, is a method of regularizing a least-squares problem by introducing a regularization matrix $\boldsymbol{\Gamma}$ and solving the modified minimization problem $\min_{\boldsymbol{x}} ||\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}||_2^2 + ||\boldsymbol{\Gamma}\boldsymbol{x}||_2^2$. Often, the matrix $\boldsymbol{\Gamma}$ takes the form of a multiple, λ, of the identity matrix. The regularization parameter λ functions as a constraint on the magnitude of the coefficients, $\boldsymbol{x}$, which can be calculated as:

$$\boldsymbol{x} = \left(\boldsymbol{A}^\top \boldsymbol{A} + \lambda \boldsymbol{I}\right)^{-1} \boldsymbol{A}^\top \boldsymbol{y} \tag{14}$$

The two other methods, **kernel ridge regression** and **Gaussian process regression** [20, 17], are kernel-based regression methods. In the kernel ridge regression, the energy of a structure with feature vector $\hat{\boldsymbol{p}}$ is given by $\hat{y}\left(\hat{\boldsymbol{p}}\right) = \sum_{i=1}^{N} \alpha_i \phi\left(\hat{\boldsymbol{p}}, \boldsymbol{p}_i\right)$, where $\boldsymbol{p}_1$, $\boldsymbol{p}_2$, $\cdots$, $\boldsymbol{p}_N$ are feature vectors such that $\boldsymbol{p}_i$ corresponds to the correlations of structure i projected onto the principal component space. The kernel $\phi\left(\boldsymbol{p}_i, \boldsymbol{p}_j\right) = K_{ij}$ measures the similarity between two structures i and j with feature vectors $\boldsymbol{p}_i$ and $\boldsymbol{p}_j$. The coefficients α_i are given by $\boldsymbol{\alpha} = \text{argmin}_{\boldsymbol{\alpha}} \left|\left|\hat{\boldsymbol{y}} - \boldsymbol{y}\right|\right|^2 + \lambda \left|\left|\hat{\boldsymbol{y}}\right|\right|^2$. Here, $\hat{\boldsymbol{y}} = \boldsymbol{K}\boldsymbol{\alpha}$ is the vector containing the values of the function at all the training points, $\boldsymbol{\alpha}$ is a vector that contains all the coefficients, λ is a regularization parameter and $||\hat{\boldsymbol{y}}||$ is the norm of the function in the feature space. This cost function is equivalent to $\boldsymbol{\alpha} = \text{argmin}_{\boldsymbol{\alpha}} \left|\left|\boldsymbol{K}\boldsymbol{\alpha} - \boldsymbol{y}\right|\right|^2 + \lambda \boldsymbol{y}^\top \boldsymbol{K}\boldsymbol{y} \Rightarrow \boldsymbol{\alpha} = \left(\boldsymbol{K} + \lambda \boldsymbol{I}\right)^{-1}\boldsymbol{y}$. In our analysis, a Gaussian kernel is used: $K_{ij} = K_{ji} = \phi\left(\boldsymbol{p}_i, \boldsymbol{p}_j\right) = \gamma^2 \exp\left(-\frac{||\boldsymbol{p}_i - \boldsymbol{p}_j||^2}{2\sigma^2}\right)$, where the hyper-parameters γ and σ determine the smoothness of the fit and the length that scales the distance between two feature vectors, respectively. The parameters γ, σ, and λ are selected by optimizing the L_2 cost function using the

Nelder-Mead method [16], while the coefficients α are optimized by calculating the R^2 score using the three-way hold-out method (three data points are set aside as the validation data set instead of only one as in the LOOCV).

The Gaussian process regression is a non-parametric Bayesian analysis technique that assumes that the coefficients, $\{\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots\}$, of the expansion $\hat{y}\left(\hat{\boldsymbol{p}}\right) = \sum_{i=1}^{N} \alpha_i \phi\left(\hat{\boldsymbol{p}}, \boldsymbol{p}_i\right)$ are normally distributed conditional on the data. As a result, the predictions are likewise normally distributed, and the exact distribution and its error is determined by the covariance, which in turn is determined by the kernel function used. Thus, as in case of the kernel ridge regression, the coefficients that minimize the posterior square loss function are given by $\boldsymbol{\alpha} = \left(\boldsymbol{K} + \lambda \boldsymbol{I}\right)^{-1} \boldsymbol{y}$, but the hyper-parameter λ is interpreted as the Gaussian prior amplitude rather than a regularization coefficient. A Gaussian kernel is used in our analysis and the parameters γ, σ, and λ are selected by optimizing the log-likelihood cost function using the Nelder-Mead method [16], while the coefficients α are optimized by calculating the R^2 score using the three-way hold-out method (three data points are set aside as the validation data set instead of only one as in the LOOCV).